Solution to Problem 2

(A)

(A.1)

$$\begin{aligned}
\textbf{States:} \quad & s = (x_{1:k}, y_k, b_k) \\
\textbf{Actions:} \quad & a = \{\text{keep } y_k, \text{change to } \bar{y}_k\} \\
\textbf{Time-horizon:} \quad & N \\
\textbf{Reward function:} \quad & -\,||\tilde{\theta}_N \;-\; \theta_0||_1 \\
\textbf{Transition probabilities:} \quad & \text{see (2)} \\
\textbf{Constraint:} \quad & \text{number of times the action} \\
& \text{“change to } \bar{y}_k\text{”is taken} \leq b
\end{aligned} \tag{1}$$

In more details:

- *States*: The states $\mathcal{S}$ of the MDP are tuples containing: *i)* $x_{1:k}$ – an $M \times 1$ vector with the number of times each observation has been seen until time k; *ii)* $b_k \in \mathbb{N}_0$ – the current budget left to use at time k; *iii)* y_k – the observation received at time k. Terminal states are ones where all the observations have been received, that is, where $\sum_{l=1}^{M} x_l = N$.
- *Actions*: The possible actions are to keep the last observation y_k or change it to a certain value $\tilde{y}_k \in \mathcal{Y}$. The number of actions is $\text{card}(A) = M$.
- *Reward function*: The reward is zero in all states except in the terminal states, where it is inversely proportional to the error of the estimate computed after the teacher's alterations.
- *Transition probabilities*:

$$\begin{aligned}
&\mathbb{P}\{s' = (x', b, y_{k+1}) \mid s = (x, b, y_k), a = \text{“keep } y_k\text{”}\} = p(y_{k+1}), \\
&\qquad \text{where } [x']_{y_{k+1}} = [x]_{y_{k+1}} + 1 \\
&\mathbb{P}\{s' = (x', b-1, y_{k+1}) \mid s = (x, b, y_k), a = \text{“}\tilde{y}_k\text{”}\} = p(y_{k+1}), \\
&\qquad \text{where } [x']_{y_{k+1}} = [x]_{y_{k+1}} + 1, [x']_{y_k} = [x]_{y_k} - 1, \text{and} \\
&\qquad [x']_{\tilde{y}_k} = [x]_{\tilde{y}_k} + 1, \\
&\text{and } \mathbb{P}\{\text{others}\} = 0.
\end{aligned} \tag{2}$$

If the action is “keep y_k”, the next state depends, with probability $p(y_{k+1})$, on the next received observation y_{k+1}. The value of the next state is obtained by simply replacing the last value of the previous state y_k by the new observation received, and adding one to that entry of the vector x, $[x']_{y_{k+1}} = [x]_{y_{k+1}} + 1$. If the action is “change to $\tilde{y}_k$”, the next state will have the same probability as in the previous case, where one is added to $[x]_{y_{k+1}}$. However, it will now have a one subtracted from the previous observation in $[x]_{y_k}$ and a one added in $[x]_{\tilde{y}_k}$ (since y_k was altered to $\tilde{y}_k$), as well as a budget of $b_{k+1} = b_k - 1$.

Note that the chosen formulation of the states and actions satisfies the Markovian property.

(A.2) The changes are added in blue in the solution of (A.1). The constraint is enforced by attributing an infinitely negative reward to transitions to states where the budget would be $b_{k+1} < 0$.

(A.3) It scales with N since the state only saves the proportion. With M not so well but it doesn't have an impact as big.

(A.4) *i)* To delay spending its budget as much as possible. Only changing an observation whenever it is seen more than the corresponding proportion. *ii)* to change all the observations for a sequence that is the closest possible to the true parameter θ.

(A.5) The problem would become deterministic.

$$\begin{aligned}
\textbf{States:}\quad & s = (x_{1:N}, b) \\
\textbf{Actions:}\quad & a = a_{1:N} \\
\textbf{Time-horizon:}\quad & N = 1 \\
\textbf{Reward function:}\quad & -\,||\tilde{\theta}_N \;-\; \theta_0||_1 \\
\textbf{Transition probabilities:}\quad & \mathbb{P}\{s' = (a_{1:N}, b) \mid s = (x_{1:N}, b), a = a_{1:N}\} = 1 \\
& \mathbb{P}\{\text{others}\} = 0 \\
\textbf{Constraint:}\quad & |a_{1:N} - x_{1:N}|_1 < b
\end{aligned} \tag{3}$$

The time horizon would be $N = 1$. The state would not need the last observation y, it would be $s = (x, b)$. The actions would be which observations to change and to which value, for example represented by an $M \times 1$ vector a with the corrected number of times each observation has been seen. The budget constraint would be that $|a_{1:N} - x_{1:N}|_1 < b$.

(A.6) The reward function for an adversarial teacher would have the opposite sign – it would be larger the larger the difference between the student's estimate and the true value, e.g. $||\tilde{\theta}_N \;-\; \theta_0||_1$.

Solution to Problem 3

(i) We model the problem as an infinite discounted MDP with discount factor q. We then choose to define

- The state space is defined as $\mathcal{S} = \mathcal{P} \cup \{X\}$. When a sate $s \in \mathcal{P}$, it means that we are at a state where we found a pair of skis to buy at a price s and we haven't bought yet a pair. When a state $s = X$, it means that we are at a state where we have already bought a pair of skis.
- The action space is defined as $\mathcal{A}_s = \{B, R\}$ for $s \in \mathcal{P}$ and $\mathcal{A}_s = \emptyset$ for $s = X$. The action B stands for buy and the action R stands for rent.
- The transition probabilities

$$\begin{aligned}
\forall p, p' \in \mathcal{P}, \quad \mathbb{P}(p'|p, R) = f(p') \quad \text{and} \quad \mathbb{P}(X|p, R) = 0 \\
\forall p, p' \in \mathcal{P}, \quad \mathbb{P}(X|p, B) = 1 \quad \text{and} \quad \mathbb{P}(p'|p, B) = 0 \\
\forall p \in \mathcal{P}, \quad \mathbb{P}(X|X) = 1 \quad \text{and} \quad \mathbb{P}(p|X) = 0
\end{aligned}$$

- The rewards are defined for all $p \in \mathcal{P}$

$$\begin{aligned}
r(p, B) &= -p \\
r(p, R) &= -c \\
r(X) &= 0
\end{aligned}$$

- The objective is

$$\max_{\pi} \quad \mathbb{E}\left[\sum_{t=0}^{\infty} q^t r(s_t, a_t)\right]$$

(ii) We start by writing Bellman's equation. When $s = X$, we have

$$\begin{aligned}
V^\star(X) &= r(X) + qV^\star(X) \\
&= 0 + qV^\star(X)
\end{aligned}$$

the above implies that $V^\star(X) = 0$. Now when $s = p \in \mathcal{P}$ we have

$$\begin{aligned}
\forall p \in \mathcal{P}, \qquad V^\star(p) &= \max\left\{r(p, R) + q\sum_{p' \in \mathcal{P}} f(p')V^\star(p'), r(p, R) + qV^\star(X)\right\} \\
&= \max\left\{-c + q\sum_{p \in \mathcal{P}} f(p)V^\star(p), -p\right\}
\end{aligned}$$

Now, observe that existence of an optimal policy and its corresponding value function $V^\star$ is guaranteed in our case, and that $-c+q\sum_{p' \in \mathcal{P}} f(p')V^\star(p')$

is constant independent of p. Thus, by setting $p_0 = c - q\sum_{p\in\mathcal{P}} f(p)V^\star(p)$, we note that the optimal policy is to buy a pair of skis at price p if and only if $p < p_0$.

(iii) We establish that $p_0 > c$. Let us note that $V^\pi(p) < 0$ for all $p \in \mathcal{P}$. In particular, $V^\star(p) < 0$. Thus, we clearly see that $p_0 = c - q\sum_{p\in\mathcal{P}} f(p)V^\star(p) > c$. So if we find a pair of skis at a price $p \le c < p_0$, then we should definitely buy.

(iv) (MDP reformulation) We reformulate our MDP as follows:

- State space is $\mathcal{S} = \mathcal{P} \cup \mathcal{W}$.
- Action spaces are defined as follows: $\mathcal{A}_s = \{B, R\}$ if $s \in \mathcal{P}$ and $\mathcal{A}_s = \{S, K\}$ if $s \in \mathcal{W}$. S stands for sell and K stands for keep.
- Transition probabilities: we have for all $p, p' \in \mathcal{P}, w, w' \in \mathcal{W}$,

$$
\begin{aligned}
\mathbb{P}(p'|p, R) &= f(p') \\
\mathbb{P}(w|p, B) &= 1 \\
\mathbb{P}(w'|w, K) &= g(w') \\
\mathbb{P}(p'|w, S) &= 1
\end{aligned}
$$

- Reward functions: for all $p \in \mathcal{P}, w \in \mathcal{W}$

$$
\begin{aligned}
r(p, R) &= -c \\
r(p, B) &= -p \\
r(w, K) &= 0 \\
r(w, S) &= w - c
\end{aligned}
$$

- We keep the same infinite horizon discounted objective.

(MDP solution) Bellman's equations give

$$
\begin{aligned}
V^\star(p) &= \max\left\{ r(p, R) + q\sum_{p'\in\mathcal{P}} f(p')V^\star(p'), r(p, B) + q\sum_{w'\in\mathcal{W}} g(w')V^\star(w') \right\} \\
&= \max\left\{ -c + q\sum_{p'\in\mathcal{P}} f(p')V^\star(p'), -p + q\sum_{w'\in\mathcal{W}} g(w')V^\star(w') \right\}
\end{aligned}
$$

and

$$
\begin{aligned}
V^\star(w) &= \max\left\{ r(w, K) + q\sum_{w'\in\mathcal{W}} g(w')V^\star(w'), r(w, S) + q\sum_{p'\in\mathcal{P}} f(p')V^\star(p') \right\} \\
&= \max\left\{ q\sum_{w'\in\mathcal{W}} g(w')V^\star(w'), w - c + q\sum_{p'\in\mathcal{P}} f(p')V^\star(p') \right\}
\end{aligned}
$$

We note that existence of an optimal policy and its corresponding optimal value $V^\star$ is guaranteed. Thus, by setting,

$$\alpha = c + q \sum_{w' \in \mathcal{W}} g(w') V^\star(w') - q \sum_{p' \in \mathcal{P}} f(p') V^\star(p')$$

we note that we buy a pair of skis at price p iff $p < \alpha$ and sell a pair of skis at price w iff $w \geq \alpha$. we observe indeed that $\alpha = p_1 = w_1$

Solution to Problem 4

(A)

(A.1) The algorithm is on-policy because it is optimizing the same policy used to perform exploration.

(A.2) The evaluation step is performed in the following line

$$Q(S_t, A_t) \leftarrow \text{ average}(\mathit{Returns}(S_t, A_t)).$$

The improvement step is performed in the following line

$$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a).$$

(B)

(B.1) Given a tuple (s, a, r, s', a'), the Q-function update rule for SARSA is the following:

$$Q(s,a) = Q(s,a) + \alpha \left[r + \gamma Q(s', a') - Q(s,a)\right]$$

From the trajectory we get the following (s, a, r, s', a') tuples:

$$\{(1,2,0.3,3,2),(3,2,0.1,4,1),(4,1,-0.7,1,2),(1,2,0.3,3,3),\\ (3,3,-0.1,2,2),(2,2,0.3,4,3),(4,3,1,4,1)\}.$$

We perform three updates of the Q-value function:

$$\begin{aligned}
Q(1,2) &= Q(1,2) + 0.5\left[0.3 + 0.5 \cdot Q(3,2) - Q(1,2)\right] \\
&= 0 + 0.5\left[0.3 + 0.5 \cdot 0 - 0\right] = 0.15 \\
Q(3,2) &= Q(3,2) + 0.5\left[0.1 + 0.5 \cdot Q(4,1) - Q(3,2)\right] \\
&= 0 + 0.5\left[0.1 + 0.5 \cdot 0 - 0\right] = 0.05 \\
Q(4,1) &= Q(4,1) + 0.5\left[-0.7 + 0.5 \cdot Q(1,2) - Q(4,1)\right] \\
&= 0 + 0.5\left[-0.7 + 0.5 \cdot 0.15 - 0\right] = -0.313 \\
Q(1,2) &= Q(1,2) + 0.5\left[0.3 + 0.5 \cdot Q(3,3) - Q(1,2)\right] \\
&= 0.15 + 0.5\left[0.3 + 0.5 \cdot 0 - 0.15\right] = 0.225 \\
Q(3,3) &= Q(3,3) + 0.5\left[-0.1 + 0.5 \cdot Q(2,2) - Q(3,3)\right] \\
&= 0 + 0.5\left[-0.1 + 0.5 \cdot 0 - 0\right] = -0.05 \\
Q(2,2) &= Q(2,2) + 0.5\left[0.3 + 0.5 \cdot Q(4,3) - Q(2,2)\right] \\
&= 0 + 0.5\left[0.3 + 0.5 \cdot 0 - 0\right] = 0.15 \\
Q(4,3) &= Q(4,3) + 0.5\left[1 + 0.5 \cdot Q(4,1) - Q(4,3)\right] \\
&= 0 + 0.5\left[1 + 0.5 \cdot -0.313 - 0\right] = 0.422
\end{aligned}$$

So, the new Q-value function can be represented as follows

$$Q(s,a) = \begin{array}{c c} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 0 & 0.225 & 0 \\ 0 & 0.15 & 0 \\ 0 & 0.05 & -0.05 \\ -0.313 & 0 & 0.422 \end{bmatrix} \end{array}.$$

The greedy policy is therefore

$$\pi = \begin{array}{c c} & a \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 2 \\ 2 \\ 2 \\ 3 \end{bmatrix} \end{array}.$$

(B.2) We would expect SARSA to converge to a higher value. Since SARSA optimizes the value of the exploration policy, the agent would waste less time falling into the black tiles during exploration. On the contrary, the Q-Learning agent would be more reckless and would fall into the black tiles more often when exploring, since it does not care about the performance under the exploration policy.

(C)

(C.1) The only way for the agent to end up on a black tile is to pick the RIGHT (R) action three consecutive times. Let's define B as the set of black tiles. We can write

$$\begin{aligned} \mathbb{P}(s_4 \in B) &= \pi(a_1 = \text{ R } | s_1)\pi(a_2 = \text{ R } | s_2)\pi(a_3 = \text{ R } | s_3) && (4) \\ &= (0.2)^3 = \frac{8}{1000}. && (5) \end{aligned}$$

The regret is computed as

$$\begin{aligned} \text{Regret}(\pi) &= \mathbb{E}_{\pi^\star} \sum_{t=1}^{3} r_t - \mathbb{E}_{\pi} \sum_{t=1}^{3} r_t \\ &= -3 - (-3 \cdot \mathbb{P}(s_4 \notin B) - 102 \cdot \mathbb{P}(s_4 \in B)) \\ &= -3 - (-3(1 - \mathbb{P}(s_4 \in B)) - 102 \cdot \mathbb{P}(s_4 \in B)) \\ &= -3 - (-3 \cdot \frac{992}{1000} - 102 \cdot \frac{8}{1000}) \\ &= \frac{99}{125} \end{aligned}$$

(D)

(D.1) We can expand $G_t - Q(S_t, A_t)$ as follows

$$\begin{aligned}
G_t - Q(S_t, A_t) &= R_{t+1} + \gamma G_{t+1} - Q(S_t, A_t) \\
&= R_{t+1} + \gamma G_{t+1} - Q(S_t, A_t) + \gamma Q(S_{t+1}, A_{t+1}) - \gamma Q(S_{t+1}, A_{t+1}) \\
&= \delta_t + \gamma(G_{t+1} - Q(S_{t+1}, A_{t+1})) \\
&= \delta_t + \gamma(R_{t+2} + \gamma G_{t+2} - Q(S_{t+1}, A_{t+1})) \\
&= \delta_t + \gamma(R_{t+2} + \gamma G_{t+2} - Q(S_{t+1}, A_{t+1}) + \gamma Q(S_{t+2}, A_{t+2}) - \gamma Q(S_{t+2}, A_{t+2})) \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 (G_{t+2} - Q(S_{t+2}, A_{t+2})) \\
&= \dots \\
&= \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k
\end{aligned}$$

Solution to Problem 5

(A.1) The eligibility vector is

$$x(s,a) - \frac{\nabla_\theta \sum_b e^{\theta^\top x(s,b)}}{\sum_b e^{\theta^\top x(s,b)}} = x(s,a) - \sum_b x(s,b)\pi(b|s)$$

Therefore in $s = (1,0)$ and $a = 1$ we get

$$\begin{bmatrix}1\\0\\1\end{bmatrix} - \left(\begin{bmatrix}1\\0\\0\end{bmatrix}\pi(0|(1,0)) + \begin{bmatrix}1\\0\\1\end{bmatrix}\pi(1|(1,0)) + \begin{bmatrix}1\\0\\2\end{bmatrix}\pi(2|(1,0))\right)$$

We have $\pi(0|(1,0)) = \frac{e}{3e} = 1/3$. Since $\theta_3 = 0$, we obtain that $\pi(0|(1,0)) = \pi(1|(1,0)) = \pi(2|(1,0))$. Therefore the result is given by

$$\begin{bmatrix}1\\0\\1\end{bmatrix} - \frac{1}{3}\begin{bmatrix}3\\0\\3\end{bmatrix} = 0$$

(A.2) Let $f(s)$ be the density of $\mathcal{N}(0,I)$. Then, the expectation is given by

$$\mathbb{E}_{s\sim\rho,a\sim\pi_\theta(\cdot|s)}[r(s,a)] = \int \sum_a \pi_\theta(a|s)\|s\|^2 f(s)ds = \int \|s\|^2 f(s)ds$$

where the last equality follows from the fact that the reward does not depend on the action. Write $\|s\|^2 = s_1^2 + s_2^2 + s_1 s_2$, where (s_1, s_2) are the two components of s. Since the covariance matrix is an identity, the two variables are independent and thus $\mathbb{E}[s_1 s_2] = 0$. From which follows that

$$\mathbb{E}_{s\sim\rho,a\sim\pi_\theta(\cdot|s)}[r(s,a)] = \int (s_1^2 + s_2^2) f(s)ds = \mathrm{Var}(s_1) + \mathrm{Var}(s_2) = 2.$$

(A.3) The policy gradient is given by

$$\nabla_\theta V^{\pi_\theta} = \mathbb{E}_{s\sim\rho,a\sim\pi_\theta(\cdot|s)}[\nabla_\theta \ln \pi_\theta(a|s) Q^{\pi_\theta}(s,a)]$$

From exercise $(A.1)$ we obtain

$$\nabla_\theta V^{\pi_\theta} = \mathbb{E}_{s\sim\rho,a\sim\pi_\theta(\cdot|s)}\left[\left(x(s,a) - \sum_b x(s,b)\pi_\theta(b|s)\right)\ln(1 + \theta^\top x(s,a))\right]$$

We approximate the gradient using the sequence of observations provided: $((0,0),0), ((1,0),0), ((0,1),1), ((2,1),1)$.
Let $g(s,a) = (x(s,a) - \sum_b x(s,b)\pi_\theta(b|s))\ln(1 + \theta^\top x(s,a))$. Then we can approximate the expectation using the empirical average:

$$\nabla_\theta V^{\pi_\theta} \approx \frac{1}{4}\left(g((0,0),0) + g((1,0),0) + g((0,1),1) + g((2,1),1)\right)$$

Since $\theta = (0,0,1)$ we get

$$g(s,a) = \left(\begin{bmatrix} 0 \\ 0 \\ a \end{bmatrix} - \sum_b \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix} \frac{e^b}{1+e+e^2} \right) \ln(1+a) = \begin{bmatrix} 0 \\ 0 \\ a - \frac{e+e^2}{1+e+e^2} \end{bmatrix} \ln(1+a)$$

We find $g((0,0),0) = g((1,0),0) = 0$, and

$$g((0,1),1) = g((2,1),1) = \begin{bmatrix} 0 \\ 0 \\ 1 - \frac{e+e^2}{1+e+e^2} \end{bmatrix} \ln(2)$$

Therefore the approximate gradient is

$$\nabla_\theta V^{\pi_\theta} \approx \frac{\ln(2)}{2} \begin{bmatrix} 0 \\ 0 \\ 1 - \frac{e+e^2}{1+e+e^2} \end{bmatrix}$$

(B.1) First, note that for $k = 2$ we obtain

$$\begin{aligned} \hat{A}_t^{(2)} &= r_t + \lambda r_{t+1} + \lambda^2 V_{\theta_t}(s_{t+2}) - V_{\theta_t}(s_t), \\ &= r_t + \lambda r_{t+1} + \lambda^2 V_{\theta_t}(s_{t+2}) - V_{\theta_t}(s_t) \pm \lambda V_{\theta_t}(s_{t+1}), \\ &= r_t + \lambda V_{\theta_t}(s_{t+1}) - V_{\theta_t}(s_t) + \lambda r_{t+1} + \lambda^2 V_{\theta_t}(s_{t+2}) - \lambda V_\theta(s_{t+1}), \\ &= \delta_t + \lambda \delta_{t+1} \end{aligned}$$

The conclusion follows by an induction argument.

(B.2) The conclusion follows from the following sequence of equations

$$\begin{aligned} (1-\alpha)\sum_{n=1}^{\infty} \alpha^{n-1} \hat{A}_t^{(n)} &= (1-\alpha)(\hat{A}_t^{(1)} + \alpha \hat{A}_t^{(2)} + \alpha^2 \hat{A}_t^{(3)} + \cdots), \\ &= (1-\alpha)(\delta_t + \alpha(\delta_t + \lambda\delta_{t+1}) + \alpha^2(\delta_t + \lambda\delta_{t+1} + \lambda^2\delta_{t+2}) + \cdots), \\ &= (1-\alpha)(\delta_t \sum_{n\geq 0} \alpha^n + \alpha\lambda\delta_{t+1} \sum_{n\geq 0} \alpha^n + (\alpha\lambda)^2 \delta_{t+2} \sum_{n\geq 0} \alpha^n + \cdots), \\ &= (1-\alpha)(\delta_t \frac{1}{1-\alpha} + \alpha\lambda\delta_{t+1} \frac{1}{1-\alpha} + (\alpha\lambda)^2 \delta_{t+2} \frac{1}{1-\alpha} + \cdots), \\ &= \delta_t + \alpha\lambda\delta_{t+1} + (\alpha\lambda)^2 \delta_{t+2} + \cdots, \\ &= \sum_{n=0}^{\infty} (\alpha\lambda)^n \delta_{n+1} \end{aligned}$$