Reinforcement Learning – Exam, Second Session

Nicolas Gast Panayotis Mertikopoulos

April 2022 (academic year 2021-2022)

Please send up your exam completed by *email* before Wednesday, April 13, 11pm CET. The exam will be followed by a oral examination On Thursday (10min per person) that will focus mostly on this exam.

For each answer, please justify carefully your answer (try to be **concise** and **precise**).

Exercise 1 Red-jack (20 points)

We consider a modified version of the Blackjack (different from the one of the exam!). You draw cards one by one from an infinite deck. The deck contains cards 2 to 10 and "A". Each card is equally likely to be drawn each time. The value of a card is the number shown on the card, except for the "A" that can be valued 1 or 11.

At each turn you may "hit" or "stay". If you "hit", you draw a new card (and receive a reward +1). If you "stay", then you sum the values of your card, and you choose if each "A" is valued 1 or 11. If the sum tht you obtain is 15, you earn 0. If it is between 16 and 21, you earn 10. If it is (strictly) lower than 15 or (strictly) larger than 21, your reward is -10. Note that when the sum is larger than 21 (when counting the A as 1), the action "hit" is not available. For instance, the cards A, 4, 5 give you "+3" (for the three cards) plus "+10" at the end (because you will choose A=11 and $11+4+5=20 \in \{15 \dots 21\}$) whereas A, 6, 6 give you 3 points for the three cards plus "-10" at the end (because $A+6+6$ equals 13 or 23).

1. (5 points) We propose to model the problem as a Markov decision process. Please explicit the state space that you choose, the reward vector, and the transition probabilities.
2. (9 points) Compute the value function and the optimal policy (**you may use a computer**, in which case you might be asked to show us the code during the oral exam). Explain (in words) what is the optimal policy. Explain also the algorithm that you use to compute the value function and the optimal policy.
3. (6 points) We now suppose that you do not know the transition probabilities. You are using Q-learning.
   At some point in time, suppose that you start with en empty hand and draws 5 cards: 5, 6, 7, A, 5 (with actions hit, hit, hit, hit, hit, stay).
   a) What is the corresponding sequence of states?
   b) Recall the Q-learning update.
   c) What are the corresponding Q-values update, assuming that initially all values are to 0. The learning rate is $\alpha = 0.1$ and the discount factor is $\gamma = 0.5$.

Exercise 2 EXP3 with explicit exploration (20 points)

This problem considers a variant of the standard EXP3 algorithm that we studied in class. The setup is that of an adversarial multi-armed bandit: at each stage $t = 1, 2, \ldots$, the learner selects an action α_t from some finite set $\mathcal{A} = \{1, \ldots, n\}$. Simultaneously, nature (the player's "adversary") selects a payoff vector $v_t = (v_{1,t}, \ldots, v_{n,t}) \in [0,1]^n$, and the learner gets as reward the component $v_{\alpha_t,t}$ of v_t. As we discussed in class, if the learner chooses the action $\alpha_t \in \mathcal{A}$ of the t-th stage based on a probability distribution (mixed strategy) $x_t \in \Delta(\mathcal{A})$, the agent's regret relative to a fixed strategy $p \in \Delta(\mathcal{A})$ defined as

$$\mathrm{Reg}_p(T) = \sum_{t=1}^{T} \langle v_t, p - x_t \rangle \tag{1}$$

Throughout this problem, we assume that the only payoff information available to the learner is the payoff $u_t = v_{\alpha_t,t}$ that they got at time t.

1. *(2 points)* Recall the definition of the importance weighted estimator

$$\hat{v}_{\alpha,t} = \frac{\mathbb{1}\{\alpha = \alpha_t\}}{x_{\alpha,t}} v_{\alpha_t,t} \tag{IWE}$$

   Discuss whether the learner has sufficient information to implement (IWE) at each stage of the process and show that $\mathbb{E}[\hat{v}_{\alpha,t}] = v_\alpha$ for all $\alpha \in \mathcal{A}$.

2. *(4 points)* Write down the EXP3 algorithm for the above setting and state (without proof) the guarantee of EXP3 for the learner's mean regret $\mathbb{E}[\mathrm{Reg}_p(T)]$.

In the sequel, we will consider a variant of the EXP3 algorithm, known as *EXP3 with explicit exploration* (EXP3.expl). The algorithm is based on the following recursion:

$$\begin{aligned} &\text{Mixed strategy:} && x_t = (1-\varepsilon)\,\Lambda(y_t) + \varepsilon\,\mathrm{unif} \\ &\text{Action selection:} && \alpha_t \sim x_t \\ &\text{Update step:} && y_{t+1} = y_t + \gamma \hat{v}_t \end{aligned} \tag{EXP3.X}$$

where:

- $y_t \in \mathbb{R}^{\mathcal{A}}$ is a sequence of vectors initialized at $y_1 = 0$ (and subsequently updated as above).
- The "logit map" $\Lambda \colon \mathbb{R}^n \to \Delta(\mathcal{A})$ is defined as

$$\Lambda(y) = \frac{(e^{y_1}, \ldots, e^{y_n})}{e^{y_1} + \cdots + e^{y_n}} \tag{2}$$

- $\mathrm{unif} = (1/n, \ldots, 1/n)$ denotes the uniform distribution on $\mathcal{A}$.
- $\varepsilon, \gamma > 0$ are parameters.

3. *(1 point)* For which choice of parameters do we recover EXP3 from EXP3.X?

4. *(1 point)* Show that, under EXP3.X, the estimator (IWE) enjoys the bound

$$\max_{\alpha\in\mathcal{A}}|\hat{v}_\alpha| \le \frac{n}{\varepsilon} \qquad \text{for all } \alpha = 1,\dots,n \tag{3}$$

In the rest of this problem we will focus on deriving the regret guarantees of EXP3.X. For this, as in the case of EXP3, we will consider the potential function

$$F(p,y) = \sum_{\alpha\in\mathcal{A}} p_\alpha \log p_\alpha + \log \sum_{\alpha\in\mathcal{A}} \exp(y_\alpha) - \langle y, p\rangle \tag{4}$$

[As we discussed in class, you may take for granted that $F(p,y) \ge 0$ for all $p \in \Delta(\mathcal{A})$ and all $y \in \mathbb{R}^n$.]

5. *(1 point)* Prove that, for all $y \in \mathbb{R}^n$, we have $\nabla_y F(p,y) = \Lambda(y) - p$, that is:

$$\frac{\partial F}{\partial y_\alpha} = \Lambda_\alpha(y) - p_\alpha = \frac{e^{y_\alpha}}{e^{y_1} + \cdots + e^{y_n}} - p_\alpha \qquad \text{for all } \alpha = 1,\dots,n. \tag{5}$$

6. *(1 point)* Show that

$$\frac{\partial^2 F}{\partial y_\alpha \partial y_\beta} = \delta_{\alpha\beta}\,\Lambda_\alpha(y) - \Lambda_\alpha(y)\,\Lambda_\beta(y) \qquad \text{for all } y \in \mathbb{R}^n \tag{6}$$

where $\delta_{\alpha\beta}$ is equal to 1 if $\alpha = \beta$ and 0 otherwise.

7. *(2 points)* Show that

$$\sum_{\alpha,\beta=1}^{n} \frac{\partial^2 F}{\partial y_\alpha \partial y_\beta} w_\alpha w_\beta \le \max_{\alpha\in\mathcal{A}} w_\alpha^2 \qquad \text{for all } y, w \in \mathbb{R}^n \tag{7}$$

8. *(2 points)* Combining the above with Taylor's theorem (or otherwise) conclude that

$$F(p, y+w) \le F(p,y) + \langle \Lambda(y), y - w\rangle + \frac{1}{2}\max_{\alpha\in\mathcal{A}} w_\alpha^2 \tag{8}$$

To proceed with the regret analysis of EXP3.X, let $F_t = F(p, y_t)$.

9. *(2 points)* Combining Questions 4 and 8, show that

$$F_{t+1} \le F_t + \frac{\gamma}{1-\varepsilon}\langle \hat{v}_t, x_t - p\rangle + \frac{\varepsilon\gamma}{1-\varepsilon}\langle \hat{v}_t, p - \text{unif}\rangle + \frac{\gamma^2 n^2}{2\varepsilon^2} \tag{9}$$

and conclude that

$$\frac{\gamma}{1-\varepsilon}\,\mathbb{E}[\text{Reg}_p(T)] \le F_1 + \frac{\varepsilon\gamma}{1-\varepsilon}\sum_{t=1}^{T}\langle v_t, p - \text{unif}\rangle + \frac{\gamma^2 n^2}{2\varepsilon^2} T \tag{10}$$

10. *(2 points)* Show that $|\langle v_t, p - \mathrm{unif}\rangle| \leq 2$ and conclude that

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq \frac{F_1}{\gamma} + 2\varepsilon T + \frac{n^2}{2}\frac{\gamma}{\varepsilon^2}T \tag{11}$$

11. *(2 points)* What is the best dependence of the bound (11) on T if you use parameters of the form $\gamma \propto 1/T^p$ and $\varepsilon \propto 1/T^q$ for $p, q \geq 0$? How does this bound compare to the bound of EXP3?

Notation. In the above, $\langle u, v\rangle = \sum_{i=1}^n u_i v_i$ denotes the ordinary pairing between $u, v \in \mathbb{R}^n$.