2023-24 M-R. Amini, P. Gaillard and N. Gast

Mathematical Fundations of Machine Learning

Final Examination

The duration of the exam is 2 hours. A single two-sided sheet of handwritten notes (with any content) is allowed. Answers can be written in French or English. This exam is made of 3 independent parts.

Part 1. Appetizers

1. Let $(x_1, y_1), \dots, (x_T, y_T)$ be a sequence of i.i.d. random variables following a distribution ν on a bounded subset of $\mathbb{R}^d \times \mathbb{R}$ for some $d \geq 1$. We consider a convex compact decision set $\Theta \subseteq \mathbb{R}^d$ and the loss $\ell_t(\theta) = \big|\langle \theta, x_t\rangle - y_t\big|$.
   (a) Give the definition of the (adversarial) regret R_T of an algorithm that chooses $\theta_t \in \Theta$ at each round t.
   (b) Provide the pseudo-code of Online Gradient Descent and the order in T of its regret bound.
   (c) Denoting by $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \theta_t$ the average iterate, show that
   $$\mathcal{R}(\bar{\theta}_t) - \inf_{\theta \in \Theta} \mathcal{R}(\theta) \leq \frac{\mathbb{E}[R_T]}{T} \quad \text{where} \quad \mathcal{R}(\theta) = \mathbb{E}\big[|\langle \theta, X\rangle - Y|\big], \quad (X, Y) \sim \nu\,.$$

2. In stochastic bandits, what are the drawbacks of the Explore-Then-Commit algorithm compared to UCB?

3. We are using Q-learning for a MDP with 4 states and 3 actions, and at some point our algorithm has the following Q-table:

State Action	s_1	s_2	s_3	s_4
a_1	1	2	3	1
a_2	2	1	1	2
a_3	3	0	4	3

   (a) If we use ε-greedy with $\varepsilon = 0.3$, what is the probability of choosing the action a_1 if we are in state s_2?
   (b) We are using Q-learning with a learning rate $\alpha = 0.1$ and a discount factor $\gamma = 0.5$. Suppose that we are in the state s_3, that we choose the action a_1 and observe a instantaneous reward $R = 2$ and next state s_1. Which values of the Q-table will the algorithm update and what are their values?

4. We consider infinite-horizon MDP with a discount factor γ.
   (a) What range of values can take the discount factor γ?
   (b) What is the difference between setting γ very small or setting γ very large?
   (c) Give an example of a MDP whose optimal policy depends on γ.

Part 2. Continuous Exponential Weights

Setting and notation. Let $\Theta \subseteq \mathbb{R}^d$ be a compact convex decision space and $\eta > 0$. A function $f : \Theta \mapsto \mathbb{R}$ is said η-exp-concave if $x \mapsto e^{-\eta f(x)}$ is concave over Θ. We consider the following setting. At each round $t \geq 1$, the learner chooses $\theta_t \in \Theta$, then the environment chooses a continuous η-exp-concave loss $\ell_t : \Theta \to [0,1]$ and reveals it to the learner.

5. Determine the set of η (if any) for which the the squared loss $f(x) = \|x - x^*\|_2^2$ for $\|x^*\|_2 \leq B$ are η-exp-concave.

6. We consider the continuous exponentially weighted average forecaster (EWA) that predicts

$$\theta_t = \frac{\int_{\theta\in\Theta} \theta w_{t-1}(\theta) d\mu(\theta)}{\int_{\theta\in\Theta} w_{t-1}(\theta) d\mu(\theta)}\,, \qquad \text{where} \qquad w_{t-1}(\theta) = \exp\Big(- \eta \sum_{s=1}^{t-1} \ell_s(\theta)\Big)\,,$$

and where $d\mu$ is the Lebesgue measure.

(a) Show that

$$W_T \overset{\text{def}}{=} \int_\Theta w_T(\theta) d\mu(\theta) \leq \mu(\Theta) \exp\left(-\eta \sum_{t=1}^T \ell_t(\theta_t)\right)\,.$$

(b) Let $\varepsilon \in (0,1)$ and $\theta^* \in \arg\min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta)$. Define $\Theta_\varepsilon \overset{\text{def}}{=} \big\{(1-\varepsilon)\theta^* + \varepsilon\theta,\ \theta \in \Theta\big\}$. Show that for all $t \geq 1$ and all $\theta \in \Theta_\varepsilon$, we have

$$\exp\big(- \eta \ell_t(\theta)\big) \geq (1-\varepsilon) \exp\big(- \eta \ell_t(\theta^*)\big)\,.$$

(c) Using that $\mu(\Theta_\varepsilon) \geq \varepsilon^d \mu(\Theta)$ (no proof needed), show that

$$W_T \geq \mu(\Theta) \varepsilon^d (1-\varepsilon)^T \exp\Big(- \eta \sum_{t=1}^T \ell_t(\theta^*)\Big)\,.$$

(d) Conclude that the regret is upper-bounded as

$$\mathrm{R}_T(\theta^*) \overset{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t) - \sum_{t=1}^T \ell_t(\theta^*) \leq \frac{1 + d\log(T+1)}{\eta}\,.$$

Part 3. MDP and optimal stopping

Setting and notation. Consider an unbiased d face dice with faces numbered from 1 to d. You can throw the dice as many times as you want and ach throw is independent of the previous ones. You gain money as follows:

- After any throw, you can stop and earn the sum of the values that you obtained.
- If, after a throw, you obtain a value that you have already observed, then the game stops and you earn nothing.

For instance, if your sequence of throws is "1, 3", you can stop and earn 4 or throw again. If you throw again: if you obtain a "2", you can stop and earn 6 or throw again, but if you obtain "3", you have to stop and earn nothing.

We suppose that you want to maximize your expected gain.

7. (a) Formulate the problem as a MDP (*i.e.*, explain what is the state space, the rewards, the action space, and the transition probabilities).

   (b) Compute the optimal policy for $d = 3$. What is its expected gain?

   (c) For how large d could your computer solve this problem exactly (you are not asked for an exact value but an estimation)? What could you do for larger d?

8. We now consider a variant of the game when the game is lost only when your new throw is exactly the same as the one just before (for instance, if your sequence of throw is $2, 1, 2$ you can throw again but if your sequence is $2, 1, 1$, you stop and gain nothing).

   (a) Formulate the problem as a MDP. What is the difference with the previous question?

   (b) Compute the optimal strategy for $d = 3$. What is its expected gain?

   (c) For how large d could your computer solve this problem exactly?