CS/ECE/STAT-861: Theoretical Foundations of Machine Learning

University of Wisconsin–Madison, Fall 2025 *Instructor: Kirthevasan Kandasamy*

Homework 4. Due 11/08/2025, 11.59 pm

Instructions:

1. Homework is due on Canvas by 11:59 pm on the due date. Please plan to submit well before the deadline. Refer to the course website for policies on late submission.
2. Homework must be typeset using appropriate software, such as LaTeX. Handwritten and scanned submissions will **not** be accepted.
3. Your solutions will be evaluated on correctness, clarity, and conciseness.
4. Unless otherwise specified, you may use any result we have already proved in class. Clearly state which result you are using.
5. Solutions to some of the problems may be found in the recommended textbook or other resources. Unless stated otherwise, you should attempt the problems on your own. You may **not** search the internet for solutions or use LLM-based tools.
6. If you use any external references, please cite them in your submission.
7. **Collaboration:** You may collaborate in groups of size up to 3 to solve problems indicated by a star ($\star$). You may not collaborate on other problems. If you collaborate, please indicate your collaborators at the beginning of the problem. Even if you collaborate, *you must write the solution in your own words*.

1 Explore-then-commit for K–armed bandits *

In this question, we will upper and lower bound the regret for the explore-then-commit algorithm, described below, on a K–armed stochastic bandit problem.

Algorithm 1 Explore-then-Commit

Given: the K arms $[K]$, time horizon T, number of exploration rounds m $(< T/K)$
Pull each arm m times in the first mK rounds.
Set $A = \text{argmax}_{i\in[K]} \frac{1}{m}\sum_{t=1}^{mK} \mathbb{1}(A_t = i)X_t$.
Pull arm A for the remaining $T - mK$ rounds.

Let $\nu = \{\nu_i\}_{i\in[K]}$ be a σ sub-Gaussian K-armed bandit model, i.e each ν_i is a σ sub-Gaussian distribution. Let $\mu_i = \mathbb{E}_{X\sim\nu_i}[X]$ denote the mean of arm i, $\mu^\star = \max_{j\in[K]} \mu_j$ be the highest mean, and let $\Delta_i = \mu^\star - \mu_i$ be the gap between the optimal arm and the i^{th} arm. Assume, without loss of generality, that $\mu_i \in [0, 1]$ for all i. Let $R_T(m, \nu)$ denote the regret when we execute the above algorithm on ν with m exploration rounds,

$$R_T(m, \nu) = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right].$$

1. **[5 pts]** *(Gap-dependent bound)* Show that there exists global constants C_1, C_2 such that

$$R_T(m, \nu) \le m \sum_{i;\Delta_i>0} \Delta_i + C_1(T - mK) \sum_{\Delta_i>0} \Delta_i \exp\left(\frac{-m\Delta_i^2}{C_2\sigma^2}\right).$$

2. **[5 pts]** *(Gap-independent bound)* Let $\mathcal{P}$ denote the class of all σ sub-Gaussian bandits whose means are bounded between 0 and 1. Show that for a suitable choice of m, say m' (possibly dependent on T and K), that we have

$$\sup_{\nu\in\mathcal{P}} R_T(m', \nu) \in \tilde{\mathcal{O}}(K^{1/3}T^{2/3}).$$

3. **[12 pts]** *(Lower bound)* Show that the result in part 2 cannot be improved (say via a tighter upper bound analysis) for the explore-then-commit algorithm. That is, show

$$\inf_{m\in\mathbb{N}} \sup_{\nu\in\mathcal{P}} R_T(m, \nu) \in \Omega(K^{1/3}T^{2/3}).$$

   Hint: One approach is to adopt a similar technique to the proof of the general lower bound for K-armed bandits, but adapt it to the structure of the explore-then-commit algorithm. Your alternatives will need to depend on the specific choice of m to get a tight lower bound. To do so, you should carefully consider the failure cases if m is picked to be too large or too small.

2 Do losses need to be non-negative for Hedge/EXP3?

Suppose we are in a sequential decision-making environment with adversarial losses that are known to be in $[-1, 0]^K$. The losses are bounded, but are negative. If we have full information feedback, we will use the Hedge algorithm, and if we have bandit feedback, we will use EXP3. In this question, we will explore if it is necessary to shift the losses so that they are non-negative before applying either of these algorithms.

1. **[2 pts]** *(Shifting losses in Hedge)* Intuitively, explain if you expect the Hedge algorithm to still work if the losses are negative, but are bounded in $[-1, 0]$.

2. **[2 pts]** Justify your answer in part 1 by showing where the proof we did in class breaks down without the non-negativity assumption, or by arguing that the proof will still carry through.

3. **[2 pts]** *(Shifting losses in EXP3)* Intuitively, explain if you expect EXP3 to still work if the losses are negative, but are bounded in $[-1, 0]$. You may reason using the exploration-exploitation tradeoff.

4. **[2 pts]** Justify your answer in part 3 by showing where the proof we did in class breaks down without the non-negativity assumption, or by arguing that the proof will still carry through.

3 Lower bounds for learning from expert advice

1. **[12 pts]** *(Finite time lower bound)* Recall the experts problem, where the regret for a policy π and a sequence of loss vectors $\ell = (\ell_1, \ldots, \ell_T) \in [0, 1]^{K \times T}$ as follows:

$$R_T(\pi, \ell) = \mathbb{E}_\pi \left[\sum_{t=1}^{T} \ell(A_t) \right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell(a).$$

Here, the expectation $\mathbb{E}_\pi$ is with respect to the randomness of the policy. You may assume an oblivious adversary. Show that the minimax regret satisfies,

$$\inf_{\pi} \sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \in \Omega\left(\sqrt{T}\right).$$

Hint: You may use a similar argument to the lower bound proof for adversarial bandits we did in class. If so, it is not necessary to repeat all steps. You can summarize the key ideas and point out the main differences.

2. *(An asymptotic lower bound)* Recall that the Hedge algorithm achieves $\mathcal{O}(\sqrt{T \log(K)})$ regret. While the result in part 1 is tight in the $\sqrt{T}$ term, it has no dependence on the number of experts K. In this question, you will show the following asymptotic lower bound on the regret for the experts problem which captures this dependence (but in a weaker asymptotic sense).

$$\inf_{\pi} \lim_{T \to \infty} \lim_{K \to \infty} \sup_{\ell_1, \ldots, \ell_T} \frac{R_T(\pi, \ell)}{\sqrt{T \log(K)}} \geq C.$$

Here, C is a universal constant.

   (a) **[6 pts]** Let π be given. Fix the values of T and K. Let P be the uniform distribution over $\{0, 1\}^{K \times T}$ for the loss vectors $\ell = (\ell_1, \ldots, \ell_T)$. Show that

$$\sup_{\ell_1, \ldots, \ell_T} R_T(\pi, \ell) \geq \frac{1}{2} \mathbb{E}_{\ell \sim P} \left[\max_{a \in [K]} \sum_{t=1}^{T} (1 - 2\ell_t(a)) \right].$$

   (b) **[2 pts]** Prove the asymptotic lower bound shown above.

   Hint: You may use the following well-known result about independent Rademacher random vectors $\sigma_1, \ldots, \sigma_T$, where $\sigma_t \in \{-1, 1\}^K$.

$$\lim_{T \to \infty} \lim_{K \to \infty} \mathbb{E}_{\sigma_1, \ldots, \sigma_T} \frac{\left[\max_{i \in [K]} \sum_{t=1}^{T} \sigma_t(i) \right]}{\sqrt{T \log(K)}} = \sqrt{2}.$$

4 Online linear optimization with FTPL

Consider the following one dimensional online linear optimization problem. An oblivious adversary chooses losses $\ell = \{\ell_t\}_{t \in [T]} \in [-1, 1]^T$ ahead of time. On round t, a learner chooses $\omega_t \in [-1, 1]$ and incurs loss $\omega_t \ell_t$. At the end of

the round, she observes ℓ_t. We wish to design a policy π to minimize the regret $R_T(\pi, \ell)$ for all $\ell \in [-1, 1]^T$, where

$$R_T(\pi, \ell) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t \omega_t\right] - \min_{\omega \in [-1,1]} \sum_{t=1}^{T} \ell_t \omega.$$

[12 pts] Consider the following instantiation of FTPL for this problem. Here, $\mathrm{Unif}\,(a, b)$ denotes a uniform distribution in the interval (a, b). Specify a choice of η and bound the regret for this policy.

Algorithm 2 FTPL for linear losses

Given: time horizon T, parameter η.
Sample $\ell_0 \sim \mathrm{Unif}\,(-1/\eta, 1/\eta)$.
for $t = 1, \ldots, T$ **do**
$\omega_t \leftarrow \operatorname{argmin}_{\omega \in [-1,1]} \sum_{s=0}^{t-1} \omega \ell_s$.
end for

5 Online learning in two-player zero sum games

In class, we saw that the standard definition of regret—comparing an algorithms performance to that of the best fixed action in hindsight—is not always meaningful with adaptive adversaries. Nevertheless, in this question, you will examine a specific setting in game theory where this notion of regret leads to useful outcomes.

In a two-player zero sum game, there are two players and a game matrix $Q \in [-R, R]^{m \times n}$. Player 1 can choose among m possible actions, and player 2 can choose between n possible actions. If player 1 chooses action i and player two chooses j, then player 2 pays player 1 an amount $Q(i, j)$ (if $Q(i, j) < 0$, then player 1 pays player 2). Instead of choosing actions (also called pure strategies), players can choose probability distributions over these actions (also called mixed strategies, hence forth simply strategies); an action will then be sampled and played. If player 1 and 2 choose $x \in \Delta([n]), y \in \Delta([m])$ respectively, then the expected payoff for player 1 is $x^\top Q y$.

A fundamental question in studying such games is how a player should choose their strategy, given that a player's optimal choice depends on the opponent's strategy. For example, if player 1 knows player 2 will play y', their optimal response is $x' = \arg\max_{x \in \Delta([m])} x^\top Q y'$. However, if player 2 anticipates that player 1 will play x', she will adjust to $y'' = \arg\min_{y \in \Delta([n])} x'^\top Q y$. In this problem, we will study sa explore two common solution concepts in game theory, safety strategies and Nash equilibria (which happen to coincide in two-player zero-sum games).

Safety strategies. The first solution concept we will look at are safety strategies. Suppose player 1 has to announce her strategy ahead of time (and stick to it), and player 2 gets to choose after observing player 1. If player 1 announces x', she knows player 2 will choose $y'' = \operatorname{argmin}_{y \in \Delta([n])} x'^\top Q y$. Hence, player 1's *safety strategy* is to choose $\widetilde{x}$ to maximize her worst case payoff, i.e $\widetilde{x} = \operatorname{argmax}_{x \in \Delta([n])} \min_{y \in \Delta([n])} x^\top Q y$. Similarly, player 2's safety strategy is $\widetilde{y} = \operatorname{argmin}_{y \in \Delta([n])} \max_{x \in \Delta([n])} x^\top Q y$. Intuitively, we would expect that the player who has to announce their strategy first will be at a disadvantage. This can be seen via the following simple argment for any continuous function f defined on a compact space:

$$\max_x \min_y f(x, y) \leq \min_y \max_x f(x, y). \tag{1}$$

When $f(x, y) = x^\top Q y$, above, the LHS is the amount player 2 will play player 1 if player 1 announces her straegy first, while the RHS is the amount if player 2 announces first.

Minimax theorem. The minimax theorem states that both these quantities are in fact equal in a two-player zero sum game. Therefore, there is no disadvantage to announcing your strategy first. We have,

$$\max_{x \in \Delta([m])} \min_{y \in \Delta([n])} x^\top Q y = \min_{y \in \Delta([n])} \max_{x \in \Delta([m])} x^\top Q y.$$

1. *(Proof of minimax theorem)* In part 1 of this problem, we will prove the minimax theorem using online learning. Consider the following scenario where the game is repeated over T rounds in the following manner. On round t, players 1 and 2 *simultaneously* choose (mixed) strategies x_t and y_t respectively. The expected payoff for players 1 and 2 is $x_t^\top Q y_t$ and $-x_t^\top Q y_t$. At the end of the round, the strategies x_t, y_t are revealed to each other. Players 1 and 2 will choose their strategy based on their knowledge of Q and the history of previous actions via policies $\pi^{(1)}$ and $\pi^{(2)}$ respectively. We can define the regret $R_T^{(1)}$ of player 1 and the regret $R_T^{(2)}$ of player 2 in this environment as follows,

$$R_T^{(1)}(\pi^{(1)}) = \max_{\mathbf{x}\in\Delta[m]} \sum_{t=1}^{T} \mathbf{x}^\top Q y_t \; - \; \sum_{t=1}^{T} x_t^\top Q y_t, \qquad R_T^{(2)}(\pi^{(2)}) = \sum_{t=1}^{T} x_t^\top Q y_t \; - \; \min_{\mathbf{y}\in\Delta[n]} \sum_{t=1}^{T} x_t^\top Q \mathbf{y}.$$

   (a) **[8 pts]** *(Minimax theorem via no-regret learning)* Show that if there exists policies $\pi^{(1)}$ and $\pi^{(2)}$ such that both players achieve sublinear regret simultaneously, i.e $R_T^{(1)}(\pi^{(1)}) \in o(T)$ and $R_T^{(1)}(\pi^{(2)}) \in o(T)$, then

$$\min_{y\in\Delta[n]} \max_{x\in\Delta[m]} x^\top Q y \leq \max_{x\in\Delta[m]} \min_{y\in\Delta[n]} x^\top Q y.$$

   (b) **[1 pts]** *(Oblivious or adaptive)* From the perspective of player 1 (or player 2), should the environment be viewed as an oblivious or adaptive adversary?

   (c) **[4 pts]** *(Designing policies)* Design policies $\pi^{(1)}, \pi^{(2)}$ for players 1 and 2 respectively, so that we simultaneously achieve sublinear regret $R_T^{(1)}, R_T^{(2)}$. When combined with (1), this implies the minimax theorem.

   Hint: Can you frame the problem facing each individual player as a problem we have studied in class?

Nash equilibrium. The second solution concept we will look at are Nash equilibria. A pair of strategies $(x_\star, y_\star)$ is said to be a Nash equilibrium if player 1 (player 2) has no incentive to deviate from $x_\star$ ($y_\star$), provided that player 2 (player 1) is playing $y_\star$ ($x_\star$). That is,

$$x_\star^\top Q y_\star \geq x^\top Q y_\star \quad \forall\, x \in \Delta([m]), \qquad \text{and} \qquad x_\star^\top Q y_\star \leq x_\star^\top Q y \quad \forall\, y \in \Delta([n]).$$

Recall that any pair of safety strategies is a Nash equilibrium, and conversely any Nash equilibrium is a safety strategy in a two-player zero sum game. We can also define an ϵ–approximate Nash equilibrium (ϵ–ANE) $(\overline{x}, \overline{y})$ as follows,

$$x^\top Q \overline{y} - \epsilon \leq \overline{x}^\top Q \overline{y} \leq \overline{x}^\top Q y + \epsilon, \qquad \forall\, x \in \Delta([m]),\, y \in \Delta([n]).$$

2. *(Computing Nash equilibria)* In part 2, we will design an algorithm to compute an ϵ–ANE.

   (a) **[4 pts]** *(Approximate Nash equilibrium via online learning)* Consider the two-player game in part 1. Denote $\overline{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ and $\overline{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$. Show that after T rounds, we arrive at an ϵ_T–approximate Nash equilibrium. You should express ϵ_T in terms of $R_T^{(1)}(\pi^{(1)})$ and $R_T^{(2)}(\pi^{(2)})$.

   (b) **[4 pts]** *(Finding an approximate NE)* Let ϵ be given. Based on your solution to part 1 and part 2a, outline an algorithm to compute an ϵ–APNE. Specify, how many iterations you will need, in terms of ϵ.