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

University of Wisconsin–Madison, Fall 2024 *Instructor: Kirthevasan Kandasamy*

Homework 4. Due 11/09/2024, 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; handwritten and scanned submissions will **not** be accepted. If you typeset your homework using LaTeX, you will receive 5 percent extra credit.
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 try the problems on your own instead of searching for answers. If you used any external references, please cite them in your submission.
6. **Collaboration:** You may collaborate in groups of size up to 3 on this homework. If you collaborate, please indicate your collaborators at the beginning of your homework. In any case, you must write the solution in your own words.

1 Optimistic algorithms for structured bandits

1. *(Linear bandits)* Recall the stochastic linear bandit problem, where, on each round a learner chooses an action A_t and observes $X_t = \theta_\star^\top A_t + \epsilon_t$. In class, we studied the LinUCB algorithm which chooses action $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \mathrm{UCB}_{t-1}(a)$ on round t, where $\mathrm{UCB}_{t-1}(a) = \widehat{\theta}_{t-1}^\top a + \beta_t \|a\|_{V_{t-1}^{-1}}$. Here,

$$\widehat{\theta}_t \triangleq \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\lambda \|\theta\|_2^2 + \sum_{s=1}^{t} (X_s - \theta^\top A_s)^2 \right) \qquad \text{and,} \quad V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top,$$

We showed that UCB_{t-1} is an upper confidence bound for $\theta_\star^\top a$ for all a and all t, for appropriately chosen β_t.

   (a) **[2 pts]** *(Expression for $\widehat{\theta}_t$)* Show that, $\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s$.

   (b) *(Deriving LinUCB via the optimism principle)* Recall the "optimism under uncertainty principle", which states that we should pretend that the environment is as nice as statistically possible, given the data, and then behave myopically. A common application of this idea for linear bandits is as follows (see for instance, LS Chapter 19): construct a confidence region C_t for $\theta_\star$, and then choose the action which *optimistically* maximizes the expected reward in C_t. That is,

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\theta \in C_{t-1}} \theta^\top a.$$

   In our proof of LinUCB we showed that the following C_t traps $\theta_\star$ with probability at least $1 - 1/T^2$,

$$C_t = \left\{ \theta \in \mathbb{R}^d;\ \|\theta - \widehat{\theta}_t\|_{V_t} \le \beta_{t+1} \right\}.$$

   Here, $\widehat{\theta}_t$ and β_t are as given above. You are encouraged to go through the proof we did in class and identify precisely where we showed this.

   [6 pts] Show that the above choice of C_t yields precisely the same action selection rule as LinUCB.

2. *(Reading exercise, GP bandits)* Let $\mathcal{A}$ be an action space and let $f : \mathcal{A} \to \mathbb{R}$ be an (unknown) function defined on $\mathcal{A}$. Consider a bandit algorithm π which achieves pseudo-regret

$$R_T(\pi) = T f(a_\star) - \sum_{t=1}^{T} f(A_t).$$

   Here $a_\star \in \operatorname{argmax}_{a \in \mathcal{A}} f(a)$. Recall the following recipe for designing optimistic algorithms for structured bandits, and bounding the pseudo-regret.

   (a) **Algorithm:** on each round choose $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \mathrm{UCB}_{t-1}(a)$, where UCB_t is an upper confidence bound for f based on data collected up to round t.

   (b) Let G be the good event in which the confidence intervals trap the true function f. Use martingale concentration to upper bound $\mathbb{P}(G^c)$.

   (c) Under G, bound the instantaneous pseudo-regret by $f(a_\star) - f(A_t) \le 2 \times e_{A_t, t-1}$, where $e_{A_t,t-1}$ is the width of the confidence interval of A_t at round $t-1$.

   (d) Bound the summation $\sum_{t=1}^{T} e_{A_t,t-1}$.

(**correction:** There were some minor typos which have been fixed. Thanks to Travis and Zhexuan for pointing them out. The confidence interval for selecting A_t should be C_{t-1} (not C_t) and the probability of trapping the true parameter should be $1 - 1/T^2$ (not $1 - 1/T$). –KK)

[7 pts] Read the paper *Gaussian process optimization in the bandit setting* by Srinivas, Krause, Kakade, Seeger, 2010, available at arxiv.org/pdf/0912.3995. In particular, focus on the frequentist result in Theorem 3, when f lies in an RKHS. Identify the specific equation, section, lemma, or theorem in the paper that corresponds to each of the four steps outlined above.

N.B. You do not need to understand what an RKHS is to answer this question. My recommendation would be to read sections 2 and 3 to understand the setting and then focus on Theorem 3. You can then go over the proof of Theorem 3 in Appendix A and B. While Appendix A is written in the Bayesian setting, many results (including those relevant to this question) are also applicable in the frequentist setting.

2 Lower bounds for learning from expert advice

1. **[8 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 $\sigma_t \in \{-1, 1\}^K$.

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

3 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 1 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

4 Online learning in two-player zero sum games

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) **[10 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) **[5 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.