CSC 696H Homework 2

Chicheng Zhang

October 2024

- This homework is due on Oct 31, 5pm.
- Your solutions to these problems will be graded based on both correctness and clarity. Your arguments should be clear: there should be no room for interpretation about what you are writing. Otherwise, I will assume that they are wrong, and grade accordingly.
- If you feel hard to make progress on any of the questions, you can post your questions on Piazza. Try posing your questions to be as general as possible, so that it can promote discussion among the class.
- You are encouraged to discuss the homework questions with your classmates, but the discussions should only be at a high level, and you should write your solutions in your own words. For every question you have had discussions on, please mention explicitly whom you have discussed with; otherwise it may be counted as academic integrity violation.
- Feel free to use existing theorems from the course notes / the textbook.

Problem 1 (16pts)

Linear bandits and elliptical potential. Suppose we are in the stochastic linear bandit setting. The sequence of (context, action)'s are represented by feature vectors $\{\phi_t\}_{t=1}^T$ such that

$$\phi_t = \begin{cases} (1,2) & t \text{ is odd} \\ (3,1) & t \text{ is even} \end{cases}$$

and we see rewards $\{r_t\}_{t=1}^T$ such that all $r_t = 0$.

- Recall the definition of $V_t(\lambda) = \sum_{s=1}^t \phi_s \phi_s^\top + \lambda I$ is the (scaled) regularized data covariance matrix up to round t; Compute $V_6(1), \hat{\theta}_7(1)$ and $V_{12}(1), \hat{\theta}_{13}(1)$. (You may find it useful to use matrix computation tools such as numpy or matlab to calculate these.)
- Define the confidence set at time step t for θ^*, $\Theta_t := \left\{\theta : \|\theta - \hat{\theta}^t(1)\|_{V_{t-1}(1)} \le 2\right\}$. (Note that the norm bound here is changed from $\beta_t(1)$ in the original lecture to 2 for simplicity.) In one plot, draw the graphs of Θ_7 and Θ_{13} and label them. You can use any plotting software you like; I recommend `https://www.desmos.com/calculator`.
- For $x = (1,1)$ and $y = (1,-1)$, calculate $[\min_{\theta\in\Theta_{13}} \langle\theta, x\rangle, \max_{\theta\in\Theta_{13}} \langle\theta, x\rangle]$ and $[\min_{\theta\in\Theta_{13}} \langle\theta, y\rangle, \max_{\theta\in\Theta_{13}} \langle\theta, y\rangle]$ respectively. These are confidence intervals for $\langle\theta^*, x\rangle$, $\langle\theta^*, y\rangle$ respectively. Which confidence interval has a higher uncertainty? Does this match your intuition?

  (Hint: we have an alternative expression for $\max_{\theta\in\Theta_{13}} \langle\theta, x\rangle$ in class that is simpler to calculate.)
- calculate the cumulative elliptical potential $\sum_{s=1}^t \|\phi_s\|^2_{V_s(1)^{-1}}$ for $t = 1, \ldots, 100$ and plot it as a function of t. Is your result consistent with what the elliptical potential lemma predicts? Why?

Problem 2 (8pts)

Prove the claim we omitted in the class: for the LinUCB algorithm, reg_t, the instantaneous regret at time step t, is upper bounded:

$$\text{reg}_t \leq 2b_t(a_t),$$

where $b_t(a) = \beta_t(1)\|\phi(x_t, a)\|_{V_{t-1}(1)^{-1}}$ is the exploration bonus of action a at time step t. Provide your justification in all steps. Why is this a useful claim? (Hint: (1) We proved a similar result in the MAB lecture. (2) You can refer to my lecture notes if you like.)

Problem 3 (8pts)

Reducing multi-armed bandits to contextual linear bandits. In our lecture, we saw that any A-armed bandit instance $(f^*(1), \ldots, f^*(A))$ can be viewed as a $d = A$-dimensional linear bandit instance with context $x_t = z_0$ (a dummy context), and feature map

$$\phi(x, a) = \begin{bmatrix} 0 \\ \cdots \\ 1 \\ \cdots \\ 0 \end{bmatrix} \leftarrow a\text{th entry}$$

- How should we define the reward predictor θ^* in the corresponding linear bandit instance?
- Suppose we run the LinUCB algorithm (with $\lambda = 0$) to solve this linear bandit instance (and thus solving the original MAB problem). Write down an analytical expression of the linear bandit confidence set Θ_t in terms the arm pull counts $\{N_{t-1}(a)\}_{a=1}^A$ and sample mean $\left\{\hat{f}_t(a)\right\}_{a=1}^A$. Your answer should not involve other quantities other than constants. For simplicity you can assume that $N_{t-1}(a)$'s are all nonzero.
- Continuing the question above, write down an analytical expression of a_t^{LinUCB}, the action chosen by LinUCB in terms of $\{N_{t-1}(a)\}_{a=1}^A$ and $\left\{\hat{f}_t(a)\right\}_{a=1}^A$. How does it compare with the action chosen by the UCB algorithm?

Problem 4 (10pts)

Consider the following family of action selection rule (call it UCB(λ)) for multi-armed bandits:

> At timestep t:
>
> choose action $a_t = \text{argmax}_{a \in \mathcal{A}} I_t(a)$, where for every action a, define $I_t(a) = \hat{f}_t(a) + \lambda b_t(a)$.

Of course, when $\lambda = 1$, this is the UCB algorithm we know and love. Evaluate UCB(λ) with

$$\lambda = 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 0, -1$$

in an multi-armed bandit environment you like, and report pseudo regret $\sum_{t=1}^T f^*(a^*) - f^*(a_t)$ as a function of T. Which algorithm works best? Is your finding consistent with the lectures? Did UCB(0) or UCB(-1) work well, if not, why?

(To obtain robust evaluation, do not forget to run each algorithm 5-10 times and plot your learning curves with error bars, see e.g. this link. Without error bars, it is impossible to assess whether the superiority of an algorithm is just by random chance.)

Problem 5 (2pts)

- How much time did it take you to complete this homework?
- What paper are you planning to present?