Homework 1

6501 Reinforcement Learning (Spring 2025)

Deadline: 11:59pm, February 7, 2025

You may type or handwrite your solution. If handwritten, take photos and compile them into a PDF file before submitting it on Gradescope.

1 ϵ-greedy for contextual bandits

In this problem, we will derive the regret bound of ϵ-greedy in contextual bandits with a regression oracle (Page 40 here). Consider the algorithm below.

Algorithm 1 ϵ-Greedy

Parameter: $\epsilon \in [0,1]$, A (number of actions)
Given: A regression oracle
for $t = 1, 2, \ldots, T$ **do**
  Receive x_t, and obtain $\hat{R}_t$ from the regression oracle.
  Define

$$\pi_t(a) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{A} & \text{if } a = \operatorname{argmax}_{a'} \hat{R}_t(x_t, a') \\ \frac{\epsilon}{A} & \text{otherwise} \end{cases}$$

  Sample $a_t \sim \pi_t$, and receive $r_t = R(x_t, a_t) + w_t$, where w_t is a zero-mean noise.

Define $a_t^\star = \operatorname{argmax}_a R(x_t, a)$. Assume that $R(x,a) \in [0,1]$ and $\hat{R}_t(x,a) \in [0,1]$ for any x, a, t.

(a) (5%) Show the following inequality. Note that the left-hand side is the expected regret at round t, and the right-hand side is ϵ plus the estimation error of the regression oracle.

$$R(x_t, a_t^\star) - \mathbb{E}_{a \sim \pi_t}\left[R(x_t, a)\right] \le \epsilon + \underbrace{R(x_t, a_t^\star) - \hat{R}_t(x_t, a_t^\star)}_{\text{estimation error on } a_t^\star} + \underbrace{\mathbb{E}_{a \sim \pi_t}\left[\hat{R}_t(x_t, a) - R(x_t, a)\right]}_{\text{estimation error on } a_t}.$$

Proof.

$$\begin{aligned}
&\text{Right hand side} - \text{Left hand side} \\
&= \epsilon + \mathbb{E}_{a \sim \pi_t}[\hat{R}_t(x_t, a)] - \hat{R}_t(x_t, a_t^\star) \\
&= \epsilon + (1 - \epsilon) \max_{a'} \hat{R}_t(x_t, a') + \frac{\epsilon}{A} \sum_{a=1}^{A} \hat{R}_t(x_t, a) - \hat{R}_t(x_t, a_t^\star) && \text{(by the definition of } p_t\text{)} \\
&\ge \epsilon + \frac{\epsilon}{A} \sum_{a=1}^{A} \hat{R}_t(x_t, a) - \epsilon \hat{R}_t(x_t, a_t^\star) && (\max_{a'} \hat{R}_t(x_t, a') \ge \hat{R}_t(x_t, a_t^\star)) \\
&\ge 0. && (0 \le \hat{R}_t(x,a) \le 1)
\end{aligned}$$

□

(b) (5%) Show that the two estimation error terms in (a) can be bounded as

$$\mathbb{E}_{a\sim\pi_t}\left[\hat{R}_t(x_t,a)-R(x_t,a)\right] \le \sqrt{\mathbb{E}_{a\sim\pi_t}\left[\left(\hat{R}_t(x_t,a)-R(x_t,a)\right)^2\right]},$$

$$R(x_t,a_t^\star)-\hat{R}_t(x_t,a_t^\star) \le \sqrt{\frac{1}{\pi_t(a_t^\star)}\mathbb{E}_{a\sim\pi_t}\left[\left(\hat{R}_t(x_t,a)-R(x_t,a)\right)^2\right]},$$

respectively.

Proof. First inequality is by the well-known fact that $\mathbb{E}[X] \le \sqrt{\mathbb{E}[X^2]}$ for any random variable X. Second inequality:

$$\begin{aligned} R(x_t,a_t^\star)-\hat{R}_t(x_t,a_t^\star) &\le \sqrt{\left(R(x_t,a_t^\star)-\hat{R}_t(x_t,a_t^\star)\right)^2} \\ &= \sqrt{\frac{1}{\pi_t(a^\star)}\pi_t(a^\star)\left(R(x_t,a_t^\star)-\hat{R}_t(x_t,a_t^\star)\right)^2} \\ &\le \sqrt{\frac{1}{\pi_t(a^\star)}\sum_{a\in\mathcal{A}}\pi_t(a)\left(R(x_t,a)-\hat{R}_t(x_t,a)\right)^2}. \end{aligned}$$

□

(c) (5%) Combining (a) and (b), show that the one-step expected regret at round t can be upper bounded as

$$R(x_t,a_t^\star)-\mathbb{E}_{a\sim p_t}\left[R(x_t,a)\right] \le \epsilon + 2\sqrt{\frac{A}{\epsilon}\cdot\mathbb{E}_{a\sim\pi_t}\left[\left(\hat{R}_t(x_t,a)-R(x_t,a)\right)^2\right]}.$$

(Hint: note that $\pi_t(a_t^\star) \ge \frac{\epsilon}{A}$)

Proof. Combining the bounds in (a) and (b), and upper bound $1 \le \frac{1}{\pi_t(a^\star)}$ we get

$$R(x_t,a_t^\star)-\mathbb{E}_{a\sim\pi_t}\left[R(x_t,a)\right] \le \epsilon + 2\sqrt{\frac{1}{\pi_t(a^\star)}\mathbb{E}_{a\sim\pi_t}\left[\left(R(x_t,a)-\hat{R}_t(x_t,a)\right)^2\right]}.$$

Then use the hint. □

(d) (5%) Show that expected total regret $\mathbb{E}\left[\sum_{t=1}^T\left(R(x_t,a_t^\star)-R(x_t,a_t)\right)\right]$ can be upper bounded by

$$\epsilon T + 2\sqrt{\frac{AT\mathbb{E}\left[\text{Err}\right]}{\epsilon}},$$

where Err is the total regression error defined as

$$\text{Err} = \sum_{t=1}^T \mathbb{E}_{a\sim\pi_t}\left[\left(\hat{R}_t(x_t,a)-R(x_t,a)\right)^2\right].$$

Proof. For simplicity, denote $\text{reg}_t = R(x_t,a_t^\star)-\mathbb{E}_{a\sim\pi_t}\left[R(x_t,a)\right]$ and $e_t = \mathbb{E}_{a\sim\pi_t}\left[\left(\hat{R}_t(x_t,a)-R(x_t,a)\right)^2\right]$. From (c) we have $\text{reg}_t \le \epsilon + 2\sqrt{\frac{A}{\epsilon}e_t}$. Summing this over t and use Cauchy-Schwarz inequality we get

$$\sum_{t=1}^T \text{reg}_t \le \epsilon T + 2\sum_{t=1}^T\sqrt{\frac{A}{\epsilon}e_t}$$

$$\leq \epsilon T + 2\sqrt{\frac{AT}{\epsilon}\sum_{t=1}^{T} e_t}$$
$$= \epsilon T + 2\sqrt{\frac{AT}{\epsilon}\text{Err}}.$$

Finally, taking expectation on two sides, and using the fact $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$ finishes the proof.

□

2 $O(\log T)$ regret bound for UCB

In the class, we showed that the regret incurred in UCB is bounded by $\tilde{O}(\sqrt{AT})$. In this problem, we will show that the same algorithm in fact ensures a more favorable $O(A \log T)$ regret when the mean of the best arm has a constant gap with the mean of other arms. We first define the following quantities: Let $R(a) \in [0, 1]$ be the true mean of the reward of arm a. Define $R^\star \triangleq \max_{a\in[A]} R(a)$ and $\Delta(a) = R^\star - R(a)$.

We consider the UCB algorithm described in Algorithm 2 (the same as presented in the class). Assume that the number of arms A is less or equal to the number of rounds T, and assume w_t is zero-mean and 1-sub-Gaussian.

Algorithm 2 UCB for multi-armed bandits

Input: A (number of arms), T (total number of rounds), δ (failure probability).
for $t = 1, \ldots, A$ **do**
  Draw $a_t = t$ and observe $r_t = R(a_t) + w_t$.
for $t = A + 1, \ldots, T$ **do**
  Define

$$N_t(a) = \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}, \quad \hat{R}_t(a) = \frac{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} r_s}{N_t(a)}, \quad \textit{conf}_t(a) = \sqrt{\frac{2\log(2/\delta)}{N_t(a)}}, \quad \tilde{R}_t(a) = \hat{R}_t(a) + \textit{conf}_t(a).$$

  Draw $a_t = \text{argmax}_a \tilde{R}_t(a)$ and observe $r_t = R(a_t) + w_t$.

Recall that the regret is defined as $\text{Regret} = TR^\star - \sum_{t=1}^{T} R(a_t)$.

(a) (4%) Use Theorem 1 to show that with probability $1 - AT\delta$, $|\hat{R}_t(a) - R(a)| \leq \textit{conf}_t(a)$ for all time t and arm a.

Proof. By the definition of $\hat{R}_t(a)$,

$$\left|\hat{R}_t(a) - R(a)\right| = \left|\frac{1}{N_t(a)}\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} r_s - R(a)\right|, \tag{1}$$

where the expectation of r_s in the summation is equal to $R(a)$ because $a_s = a$. For each possible a and $N_t(a)$, by Theorem 1, Eq. (1) is upper bounded by $\sqrt{\frac{2\log(2/\delta)}{N_t(a)}} = \textit{conf}_t(a)$ with probability $1 - \delta$. Using a union bound over all possible a and $N_t(a)$, we have Eq. (1) for all a and t with probability at least $1 - AT\delta$. □

(b) (4%) Assume that the inequality in (a) holds for all t and a. Prove that for any sub-optimal arm a (i.e., $\Delta(a) > 0$), if $N_t(a) > \frac{8\log(2/\delta)}{\Delta(a)^2}$, then $\textit{conf}_t(a) < \frac{\Delta(a)}{2}$ and $\tilde{R}_t(a) < R(a) + \Delta(a)$.
(Hint: For the second inequality, you have to use the inequality in (a))

Proof. When $N_t(a) > \frac{8\log(2/\delta)}{\Delta(a)^2}$, we have

$$\textit{conf}_t(a) = \sqrt{\frac{2\log(2/\delta)}{N_t(a)}} < \sqrt{\frac{2\log(2/\delta)}{\frac{8\log(2/\delta)}{\Delta(a)^2}}} = \frac{\Delta(a)}{2}. \tag{2}$$

Therefore,

$$\begin{aligned}
\tilde{R}_t(a) &= \hat{R}_t(a) + \mathit{conf}_t(a) && \text{(by the definition of } \tilde{R}_t(a)\text{)} \\
&\le R(a) + 2\mathit{conf}_t(a) && \text{(by the assumption that the inquality in (a) holds)} \\
&< R(a) + \Delta(a) && \text{(by Eq. (2))}
\end{aligned}$$

□

(c) (4%) Assume that the inequality in (a) holds for all t and a. Prove that for any optimal arm $a^\star$ (i.e., $\Delta(a^\star) = 0$), $\tilde{R}_t(a^\star) \ge R(a^\star)$ for all t.

Proof. for any optimal arm $a^\star$, we have

$$\begin{aligned}
\tilde{R}_t(a^\star) &= \hat{R}_t(a^\star) + \mathit{conf}_t(a^\star) && \text{(by the definition of } \tilde{R}_t(a^\star)\text{)} \\
&\ge R(a^\star) - \mathit{conf}_t(a^\star) + \mathit{conf}_t(a^\star) && \text{(by the assumption that the inequality in (a) holds)} \\
&= R(a^\star).
\end{aligned}$$

□

(d) (4%) Assume that the inequality in (a) holds for all t and a. Prove that for any sub-optimal arm a (i.e., $\Delta(a) > 0$), if $N_t(a) > \frac{8\log(2/\delta)}{\Delta(a)^2}$, then arm a will NOT be drawn at round t.
(Hint: Use (b) and (c) to show $\tilde{R}_t(a) < \tilde{R}_t(a^\star)$)

Proof. Combining (b) and (c), we get $\tilde{R}_t(a) < R(a) + \Delta = R^\star \le \tilde{R}_t(a^\star)$. Since the algorithm chooses the arm that maximizes $\tilde{R}_t(\cdot)$, this suboptimal arm a will not be chosen. □

(e) (4%) Assume that the inequality in (a) holds for all t and a. Argue that any sub-optimal arm a will not be drawn more than $\frac{8\log(2/\delta)}{\Delta(a)^2}+1$ times. Then show that the regret is upper bounded by $\sum_{a:R(a)\neq R^\star} \min\left\{T\Delta(a),\ \frac{8\log(2/\delta)}{\Delta(a)} + \Delta(a)\right\}$ with probability at least $1 - AT\delta$.

Proof. When the inequality in (a) holds (which happens with probability at least $1 - AT\delta$), by (b), we know that when arm a has been drawn for $\lceil \frac{8\log(2/\delta)}{\Delta(a)^2} \rceil$ many times, it will not been drawn again. This implies that

$$\sum_{t=1}^{T} \mathbb{I}\{a_t = a\} \le \min\left\{ \left\lceil \frac{8\log(2/\delta)}{\Delta(a)^2} \right\rceil, T \right\}. \tag{3}$$

In this case, the regret can be upper bounded as

$$\begin{aligned}
\text{Regret} &= \sum_{t=1}^{T} (R^\star - R(a_t)) \\
&= \sum_{a:\ R(a)\neq R^\star} \sum_{t=1}^{T} \mathbb{I}\{a_t = a\}(R^\star - R(a)) \\
&\le \sum_{a:\ R(a)\neq R^\star} \min\left\{ \left\lceil \frac{8\log(2/\delta)}{\Delta(a)^2} \right\rceil, T \right\} \times \Delta(a) && \text{(by Eq. (3) and the definition of } \Delta(a)\text{)} \\
&\le \sum_{a:\ R(a)\neq R^\star} \min\left\{ \Delta(a) + \frac{8\log(2/\delta)}{\Delta(a)}, T\Delta(a) \right\}.
\end{aligned}$$

□

Finally, we can choose $\delta = \frac{1}{AT^2}$, so that the expected regret is upper bounded by

$$(1 - AT\delta) \sum_{a:\ R(a) \neq R^\star} \min\left\{\Delta(a) + \frac{8\log(2/\delta)}{\Delta(a)}, T\Delta(a)\right\} + \underbrace{AT\delta}_{\text{failure probability}} \times \underbrace{T}_{\text{regret in the failure case}}$$
$$\leq \sum_{a:\ R(a) \neq R^\star} \min\left\{\Delta(a) + \frac{8\log(2AT^2)}{\Delta(a)}, T\Delta(a)\right\} + 1.$$

3 Survey

(a) (3%) How much time did you use to complete each of the problems in this homework? Do you have any suggestion for the course? (e.g., the pace of the lecture, the length of the homework)

Appendix

Theorem 1 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \ldots, X_N$ be i.i.d. σ-sub-Gaussian random variables with mean μ. Then with probability at least $1 - \delta$,*

$$\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \leq \sigma\sqrt{\frac{2\log(2/\delta)}{N}}.$$