

# 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 $\epsilon$ -greedy for contextual bandits

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

---

**Algorithm 1**  $\epsilon$ -Greedy

---

**Parameter:**  $\epsilon \in [0, 1]$ ,  $A$  (number of actions)

**Given:** A regression oracle

**for**  $t = 1, 2, \dots, T$  **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  $r_t = R(x_t, a_t) + w_t$ .

---

Define  $a_t^* = \operatorname{argmax}_a R(x_t, a)$ . 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  $\epsilon$  plus the estimation error of the regression oracle.

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

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

$$\begin{aligned} \mathbb{E}_{a \sim \pi_t} [\hat{R}_t(x_t, a) - R(x_t, a)] &\leq \sqrt{\mathbb{E}_{a \sim \pi_t} \left[ \left( \hat{R}_t(x_t, a) - R(x_t, a) \right)^2 \right]}, \\ R(x_t, a_t^*) - \hat{R}_t(x_t, a_t^*) &\leq \sqrt{\frac{1}{\pi_t(a_t^*)} \mathbb{E}_{a \sim \pi_t} \left[ \left( \hat{R}_t(x_t, a) - R(x_t, a) \right)^2 \right]}, \end{aligned}$$

respectively.

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

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

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

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

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

where Err is the total regression error defined as

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

## 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  $R^* \triangleq \max_{a \in [A]} R(a)$  and  $\Delta(a) = R^* - R(a)$ .

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),  $\delta$  (failure probability).

**for**  $t = 1, \dots, A$  **do**

    Draw  $a_t = t$  and observe  $r_t = R(a_t) + w_t$ .

**for**  $t = A + 1, \dots, T$  **do**

    Define

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

    Draw  $a_t = \arg\max_a \tilde{R}_t(a)$  and observe  $r_t = R(a_t) + w_t$ .

---

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

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

(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  $\text{conf}_t(a) < \frac{\Delta(a)}{2}$  and  $\tilde{R}_t(a) < R(a) + \Delta(a)$ .

(Hint: For the second inequality, you have to use the inequality in (a))

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

(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  $\tilde{R}_t(a) < \tilde{R}_t(a^*)$ )

(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  $\sum_{a: R(a) \neq R^*} \min \left\{ T\Delta(a), \frac{8 \log(2/\delta)}{\Delta(a)} + \Delta(a) \right\}$  with probability at least  $1 - AT\delta$ .

### 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  $X_1, X_2, \dots, X_N$  be i.i.d.  $\sigma$ -sub-Gaussian random variables with mean  $\mu$ . Then with probability at least  $1 - \delta$ ,*

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