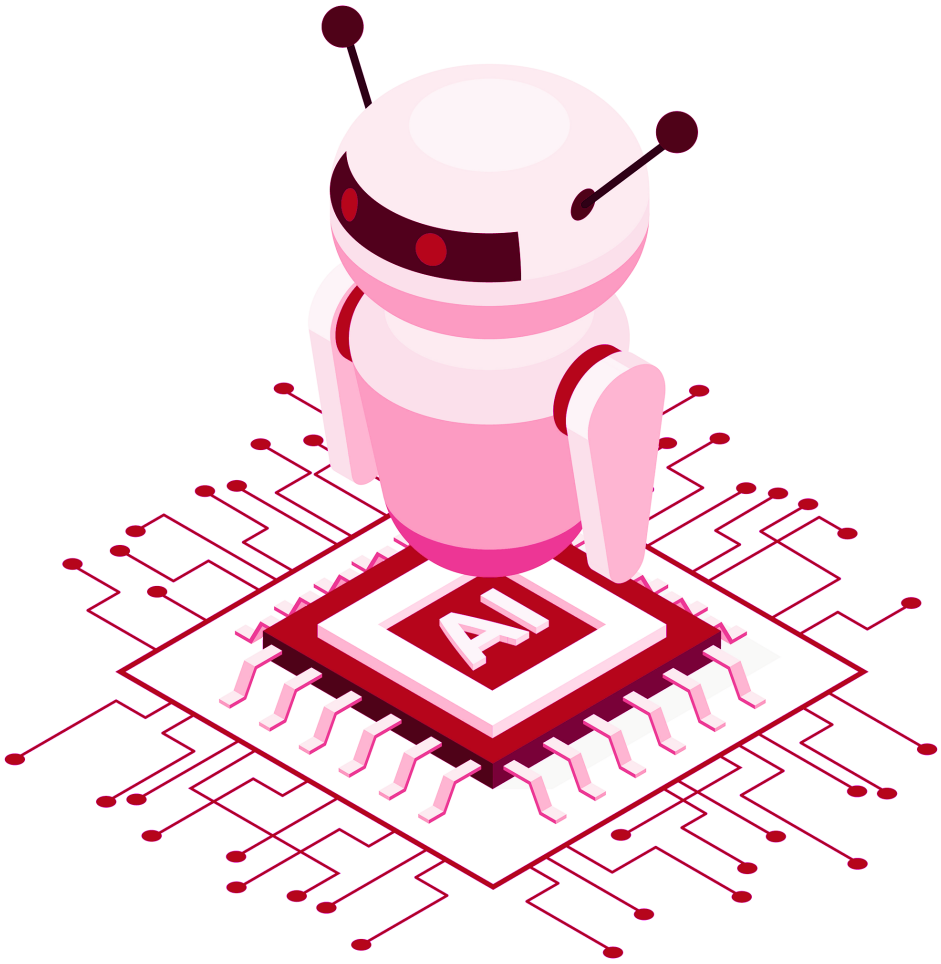

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution forHomework [9]

[Exploration Methods]

By:

[Full Name]

[Student Number]

RIML

Deep RL Course

Spring 2025

Contents

1 Light-tailed Distributions[25-points]

1.1 Hoeffding's Inequality[10-points]

1.1.1 a)[6-points]

We first write X as a convex combination of its bounds. Define

$$\theta = \frac{X-a}{b-a},$$

so that $X = (1-\theta)a + \theta b$ and $\theta \in [0,1]$. From $\mathbb{E}[X] = 0$ we get

$$0 = \mathbb{E}[X] = \mathbb{E}[(1-\theta)]\, a + \mathbb{E}[\theta]\, b,$$

so

$$\mathbb{E}[\theta] = \frac{-a}{b-a}, \qquad \mathbb{E}[1-\theta] = \frac{b}{b-a}.$$

Next, by convexity of the exponential function, for each possible x we have:

$$e^{sX} = e^{s[(1-\theta)a+\theta b]} \leq (1-\theta)\, e^{sa} + \theta\, e^{sb}.$$

Taking expectation on both sides gives

$$\mathbb{E}[e^{sX}] \;\leq\; \mathbb{E}[\,(1-\theta)\,]\, e^{sa} \;+\; \mathbb{E}[\,\theta\,]\, e^{sb} \;=\; \frac{b}{b-a}e^{sa} \;-\; \frac{a}{b-a}e^{sb}.$$

Denote this upper bound by

$$L \;=\; \frac{be^{sa} - ae^{sb}}{b-a}.$$

We now show $L \leq \exp(s^2(b-a)^2/8)$.

To simplify L, set

$$p = \frac{-a}{b-a} \in [0,1], \quad 1-p = \frac{b}{b-a}, \quad u = s(b-a).$$

Then $a = -p(b-a)$, $b = (1-p)(b-a)$. Substitute into L:

$$L = \frac{(1-p)(b-a)\, e^{s(-p(b-a))} \;-\; \big(-p(b-a)\big)\, e^{s((1-p)(b-a))}}{b-a} = (1-p)e^{-pu} + pe^{(1-p)u} =: \phi(u).$$

Define $f(u) = \ln \phi(u)$. Note first $\phi(0) = (1-p) + p = 1$, so $f(0) = 0$. Compute derivatives:

$$\phi'(u) = (1-p)(-p)e^{-pu} + p(1-p)e^{(1-p)u} = p(1-p)\big(e^{(1-p)u} - e^{-pu}\big),$$

so

$$f'(u) = \frac{\phi'(u)}{\phi(u)}, \quad f'(0) = \frac{p(1-p)(1-1)}{\phi(0)} = 0.$$

For the second derivative,

$$f''(u) = \frac{\phi''(u)\,\phi(u) - (\phi'(u))^2}{\phi(u)^2}.$$

Think of $f''(u)$ as variance a two-point distribution with weights $(1-p)$ and p on values $-p$ and $1-p$; the length of that interval is 1, so its variance is at most $1/4$. Therefore

$$f''(u) \le \tfrac{1}{4} \quad \text{for all } u.$$

Now we write Taylor's theorem around $u = 0$: for some ξ between 0 and u,

$$f(u) = f(0) + f'(0)\,u + \tfrac{1}{2} f''(\xi)\,u^2 \le 0 + 0 + \tfrac{1}{2} \cdot \tfrac{1}{4}\,u^2 = \frac{u^2}{8}.$$

Therefore, $\phi(u) = e^{f(u)} \le e^{u^2/8}$. Since $u = s(b-a)$, we have

$$L = \phi(u) \le \exp\big(s^2(b-a)^2/8\big).$$

Finally, going back to the bound on the moment generating function,

$$\mathbb{E}[e^{sX}] \le L \le \exp\big(s^2(b-a)^2/8\big).$$

1.1.2 b)[4-points]

Define $\mu_i = \mathbb{E}[Z_i]$ and $S = \sum_{i=1}^n (Z_i - \mu_i)$. We want to bound $\Pr\big(\frac{1}{n}S \ge t\big)$.

For any $s > 0$. By Markov inequality we have:

$$\Pr(S \ge nt) = \Pr\big(e^{sS} \ge e^{snt}\big) \;\le\; \frac{\mathbb{E}[e^{sS}]}{e^{snt}}.$$

Because the Z_is are independent, we can bring the $\mathbb{E}$ inside:

$$\mathbb{E}[e^{sS}] = \mathbb{E}\Big[\prod_{i=1}^n e^{s(Z_i-\mu_i)}\Big] = \prod_{i=1}^n \mathbb{E}[\,e^{s(Z_i-\mu_i)}\,].$$

For each i, $X_i = Z_i - \mu_i$ has mean zero and lies in an interval of length $b-a$. By part (a), for each i,

$$\mathbb{E}[e^{s(Z_i-\mu_i)}] \;\le\; \exp\big(s^2(b-a)^2/8\big).$$

So for the whole S we have:

$$\mathbb{E}[e^{sS}] \;\le\; \exp\big(n\,s^2(b-a)^2/8\big) \Rightarrow \Pr(S \ge nt) \;\le\; \exp\big(ns^2(b-a)^2/8 - snt\big).$$

We choose s to minimize the exponent. View $f(s) = ns^2(b-a)^2/8 - snt$. Differentiate: $f'(s) = n(b-a)^2 s/4 - nt$:

$$f'(s) = 0 \Rightarrow s = \frac{4t}{(b-a)^2}.$$

Now put it in the equation:

$$\begin{aligned} ns^2(b-a)^2/8 - snt &= n \cdot \frac{16t^2}{(b-a)^4} \cdot \frac{(b-a)^2}{8} - n \cdot \frac{4t}{(b-a)^2} \cdot t \\ &= n \cdot \frac{16t^2}{8(b-a)^2} - n \cdot \frac{4t^2}{(b-a)^2} \\ &= \frac{2nt^2}{(b-a)^2} - \frac{4nt^2}{(b-a)^2} \\ &= -\frac{2nt^2}{(b-a)^2}. \end{aligned}$$

So the expected result is proven on one end:

$$\Pr\big(\tfrac{1}{n}\sum_{i=1}^{n}(Z_i - \mu_i) \geq t\big) = \Pr(S \geq nt) \;\leq\; \exp\Big(-\frac{2nt^2}{(b-a)^2}\Big).$$

For the lower tail, $\Pr(\frac{1}{n}S \leq -t)$, we apply the same idea to $-S$. We have $-S = \sum_{i=1}^{n}(\mu_i - Z_i)$, and each $\mu_i - Z_i$ has mean zero and lies still is in an interval of length $b-a$. Repeating the above with $s > 0$ on $\Pr(-S \geq nt)$ gives the same bound:

$$\Pr\big(S \leq -nt\big) = \Pr(-S \geq nt) \;\leq\; \exp\Big(-\frac{2nt^2}{(b-a)^2}\Big).$$

1.2 Sub-Gaussian[15-points]

1.2.1 a-1)[2-points]

When X is sub-Gaussian with parameter σ, we have:

$$\mathbb{E}\big[e^{s(X-\mathbb{E}[X])}\big] \leq \exp\big(s^2\sigma^2/2\big) \quad \forall s \in \mathbb{R}.$$

For any $t > 0$, because of Markov's inequality and the property of sub-gaussianness we have:

$$\Pr(X - \mathbb{E}[X] \geq t) = \Pr\big(e^{s(X-\mathbb{E}[X])} \geq e^{st}\big) \leq \frac{\mathbb{E}[e^{s(X-\mathbb{E}[X])}]}{e^{st}} \leq \exp\big(s^2\sigma^2/2 - st\big).$$

Minimize the exponent in s: set $s = t/\sigma^2$, which yields exponent

$$-\frac{t^2}{2\sigma^2}.$$

So we have:

$$\Pr(X - \mathbb{E}[X] \geq t) \leq \exp\big(-t^2/(2\sigma^2)\big).$$

1.2.2 a-2)[2-points]

Using the exact same way of proof for $-X$ gives

$$\Pr(X - \mathbb{E}[X] \leq -t) \leq \exp\big(-t^2/(2\sigma^2)\big).$$

1.2.3 a-3)[2-points]

By applying the union inequality on union of two ways of the abs we have:

$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq \Pr(X - \mathbb{E}[X] \geq t) + \Pr(X - \mathbb{E}[X] \leq -t) \leq 2\exp\big(-t^2/(2\sigma^2)\big).$$

1.2.4 b)[3-points]

We have $Y_i = X_i - \mu_i$ so $\mathbb{E}[Y_i] = 0$ and each Y_i is sub gaussian with σ_i. Define $S_n = \sum_{i=1}^n Y_i$.

For any $\lambda \in \mathbb{R}$:

$$\mathbb{E}\left[e^{\lambda S_n}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda Y_i}\right] \leq \prod_{i=1}^n \exp\left(\frac{\lambda^2\sigma_i^2}{2}\right) = \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2\right)$$

So we proved S_n is sub-Gaussian with parameter $\sigma = \sqrt{\sum_{i=1}^n \sigma_i^2}$.

As we proved in the last question, with $\mathbb{E}[S_n] = 0$:

$$\Pr\left(|S_n| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) = 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$$

1.2.5 c)[4-points]

Let $Y_i = X_i - \mu$ so $\mathbb{E}[Y_i] = 0$ and each Y_i is σ-sub-Gaussian. Define $S_n = \sum_{i=1}^n Y_i$.

Similar to the last section, S_n is sub-Gaussian as you will see:

$$\mathbb{E}\left[e^{\lambda S_n}\right] = \mathbb{E}\left[e^{\lambda\sum_{i=1}^n (X_i-\mu)}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i-\mu)}\right] \leq \prod_{i=1}^n e^{\lambda^2\sigma^2/2} = \exp\left(\frac{\lambda^2}{2}\cdot n\sigma^2\right)$$

So S_n is sub-Gaussian with parameter $\sigma\sqrt{n}$.

$$\Pr\left(\frac{S_n}{n} \geq \epsilon\right) = \Pr\left(S_n \geq n\epsilon\right) \leq \exp\left(-\frac{(n\epsilon)^2}{2\cdot(\sigma\sqrt{n})^2}\right) = \exp\left(-\frac{n^2\epsilon^2}{2n\sigma^2}\right) = \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$$

Now for the next inequality. Set $\epsilon = \sqrt{\frac{2\sigma^2\ln(1/\delta)}{n}}$. Then:

$$\Pr\left(\frac{S_n}{n} \geq \epsilon\right) \leq \exp\left(-\frac{n}{2\sigma^2}\cdot\frac{2\sigma^2\ln(1/\delta)}{n}\right) = \delta$$

Therefore:

$$\Pr\left(\frac{S_n}{n} < \epsilon\right) \geq 1 - \delta$$

2 UCB[75-points]

2.1 The Upper Confidence Bound Algorithm[40-points]

2.1.1 a)[2-points]

In the multi-armed bandit problem, exactly one arm is selected at each time step t. The total regret R_n is defined as:

$$R_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} r_t\right]$$

where $\mu^* = \max_i \mu_i$ is the optimal mean reward, and r_t is the reward obtained at time t.

Let A_t be the arm selected at time t. The cumulative reward can be individually summed by arms:

$$\sum_{t=1}^{n} r_t = \sum_{i=1}^{K} \sum_{t:A_t=i} r_t = \sum_{i=1}^{K} \sum_{j=1}^{T_i(n)} X_{i,j}$$

where $T_i(n)$ is the number of times arm i is pulled by time n, and $X_{i,j}$ is the reward from the j-th pull of arm i.

Taking expectations:

$$\mathbb{E}\left[\sum_{t=1}^{n} r_t\right] = \sum_{i=1}^{K} \mathbb{E}\left[\sum_{j=1}^{T_i(n)} X_{i,j}\right]$$

For each arm i, by the tower property of expectation:

$$\mathbb{E}\left[\sum_{j=1}^{T_i(n)} X_{i,j}\right] = \sum_{j=1}^{T_i(n)} \mathbb{E}[X_{i,j}] = \mathbb{E}\left[T_i(n)\mu_i\right] = \mu_i \mathbb{E}[T_i(n)]$$

since $\mathbb{E}[X_{i,j}] = \mu_i$ (rewards are independent of pull counts).

Using the fact that $n = \sum_{i=1}^{K} T_i(n)$:

$$R_n = \sum_{i=1}^{K} \mu^* \mathbb{E}[T_i(n)] - \sum_{i=1}^{K} \mu_i \mathbb{E}[T_i(n)] = \sum_{i=1}^{K} (\mu^* - \mu_i)\mathbb{E}[T_i(n)] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$$

2.1.2 b)[4-points]

When δ is fixed as a constant (e.g., $\delta = 0.1$), a bad event B can cause linear regret. Define:

$$B = \{\exists t > K : \mathrm{UCB}_1(t, \delta) \leq \mu_1\}$$

This event means the optimal arm (arm 1) appears suboptimal at some time $t > K$. If B occurs, there might be some other arm i which has $\mathrm{UCB}_i(t,\delta) > \mathrm{UCB}_1(t,\delta)$ and get pulled after time K until the end, which is $O(n)$ times, and the regret grows linearly: $R_n \geq c\Delta_i n$ for some $c > 0$

The probability of B is bounded by:

$$\begin{aligned}\Pr(B) &\leq \Pr\left(\exists s \in \{1,\dots,n\} : \hat{\mu}_{1,s} + \sqrt{\frac{2\ln(1/\delta)}{s}} \leq \mu_1\right) \\ &\leq \sum_{s=1}^{n} \Pr\left(\hat{\mu}_{1,s} - \mu_1 \leq -\sqrt{\frac{2\ln(1/\delta)}{s}}\right) \\ &\leq \sum_{s=1}^{n} \delta = n\delta\end{aligned}$$

since rewards are 1-sub-Gaussian, giving the tail bound $\Pr(\hat{\mu}_{1,s} - \mu_1 \leq -a) \leq e^{-sa^2/2} \leq \delta$ for $a = \sqrt{2\ln(1/\delta)/s}$.

To avoid linear regret, choose $\delta = \delta_n \to 0$ such that $n\delta_n \to 0$. The optimal choice is $\delta = 1/n^2$:

$$\Pr(B) \leq n \cdot (1/n^2) = 1/n \to 0$$

This ensures the bad event occurs with almost zero probability in the limit, allowing sub-linear regret almost surely.

2.1.3 c)[4-points]

Suppose G_i holds but $T_i(n) > u_i$. Let τ be the time of the (u_i+1)-th pull of arm i. Then $T_i(\tau-1) = u_i$, so

$$\mathrm{UCB}_i(\tau-1,\delta) = \hat{\mu}_{i,u_i} + \sqrt{\tfrac{2\ln(1/\delta)}{u_i}} < \mu_1 \quad \text{(by the second part of } G_i\text{)},$$

while

$$\mathrm{UCB}_1(\tau-1,\delta) \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta) > \mu_1 \quad \text{(by the first part of } G_i\text{)}.$$

Therefore we have: $\mathrm{UCB}_i(\tau-1,\delta) < \mathrm{UCB}_1(\tau-1,\delta)$, which is contradicting that arm i is chosen at time τ.

2.1.4 d)[4-points]

Write

$$T_i(n) = T_i(n)\,\mathbf{1}_{G_i} + T_i(n)\,\mathbf{1}_{G_i^c} \leq u_i + n\,\mathbf{1}_{G_i^c},$$

since on event of G_i we are sure $T_i(n) \leq u_i$ for other is and always $T_i(n) \leq n$ either way. So taking expectation gives the inequality.

2.1.5 e)[6-points]

Since each reward is 1-sub-Gaussian, the sum of u_i of them is $\sqrt{u_i}$-sub-Gaussian, so the average $\hat{\mu}_{i,u_i}$ is $\frac{1}{\sqrt{u_i}}$-sub-Gaussian.

The condition gives $\sqrt{\frac{2\ln(1/\delta)}{u_i}} \leq (1-c)\Delta_i$. If the event holds, then

$$\hat{\mu}_{i,u_i} \geq \mu_1 - \sqrt{\tfrac{2\ln(1/\delta)}{u_i}} \geq (\mu_i + \Delta_i) - (1-c)\Delta_i = \mu_i + c\Delta_i.$$

Now if $\hat{\mu}_{i,u_i} + \sqrt{\frac{2\ln(1/\delta)}{u_i}} \geq \mu_1$, then $\hat{\mu}_{i,u_i} - \mu_i \geq \mu_1 - \mu_i - \sqrt{\frac{2\ln(1/\delta)}{u_i}} \geq c\Delta_i$.

Since $\hat{\mu}_{i,u_i} - \mu_i$ is $1/\sqrt{u_i}$-sub-Gaussian, the tail bound gives $\Pr(\hat{\mu}_{i,u_i} - \mu_i \geq c\Delta_i) \leq \exp(-u_i c^2 \Delta_i^2/2)$.

2.1.6 f)[4-points]

We have:

$$G_i^c = \left\{\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{i,u_i} + \sqrt{\frac{2\ln(1/\delta)}{u_i}} \geq \mu_1\right\}.$$

Union bound gives:

$$\Pr(G_i^c) \leq \Pr\left(\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right) + \Pr\left(\hat{\mu}_{i,u_i} + \sqrt{\frac{2\ln(1/\delta)}{u_i}} \geq \mu_1\right).$$

Part (e) bounds the second term by $\exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$.

First term is equivalent to $\exists t \in [n]$ such that $\mu_1 \geq \mathrm{UCB}_1(t,\delta)$. At any time t with $s = T_1(t)$ pulls of arm 1:

$$\mathrm{UCB}_1(t,\delta) = \hat{\mu}_{1,s} + \sqrt{\frac{2\ln(1/\delta)}{s}}.$$

The event $\mu_1 \geq \mathrm{UCB}_1(t,\delta)$ implies:

$$\hat{\mu}_{1,s} \leq \mu_1 - \sqrt{\frac{2\ln(1/\delta)}{s}}.$$

Since it is sub gaussian:

$$\Pr\left(\hat{\mu}_{1,s} \leq \mu_1 - \sqrt{\frac{2\ln(1/\delta)}{s}}\right) \leq \delta.$$

Union bound over $s = 1, \ldots, n$:

$$\Pr\left(\exists t \in [n] : \mu_1 \geq \mathrm{UCB}_1(t,\delta)\right) \leq \sum_{s=1}^{n} \delta = n\delta.$$

Therefore we have, $\Pr(G_i^c) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$.

2.1.7 g)[6-points]

Set $\delta = 1/n^2$. From parts (d) and (f):

$$\mathbb{E}[T_i(n)] \leq u_i + n\left[n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right] = u_i + n^2 \cdot (1/n^2) + n\exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right) = u_i + 1 + n\exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right).$$

Choose $c = 1/2$ and $u_i = \frac{16\ln n}{\Delta_i^2}$. So we have:

$$\Delta_i - \sqrt{\frac{2\ln(1/\delta)}{u_i}} = \Delta_i - \sqrt{\frac{2\cdot 2\ln n}{16\ln n/\Delta_i^2}} = \Delta_i - \sqrt{\frac{\Delta_i^2}{4}} = \Delta_i - \frac{\Delta_i}{2} = \frac{\Delta_i}{2} = c\Delta_i.$$

Now we bound the exponential term:

$$n\exp\left(-\frac{u_i(1/2)^2\Delta_i^2}{2}\right) = n\exp\left(-\frac{u_i\Delta_i^2}{8}\right) = n\exp\left(-\frac{16\ln n\cdot\Delta_i^2}{8\Delta_i^2}\right) = n\exp(-2\ln n) = n\cdot n^{-2} = n^{-1} \leq 1.$$

so:

$$\mathbb{E}[T_i(n)] \leq \frac{16\ln n}{\Delta_i^2} + 1 + 1 = \frac{16\ln n}{\Delta_i^2} + 2.$$

Since u_i may not be integer, use $\lceil u_i \rceil \leq u_i + 1$:

$$\mathbb{E}[T_i(n)] \leq \left(\frac{16\ln n}{\Delta_i^2} + 1\right) + 1 + 1 = 3 + \frac{16\ln n}{\Delta_i^2}.$$

2.1.8 h)[5-points]

From part (a):

$$R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)].$$

For optimal arm i^*, $\Delta_{i^*} = 0$, so it makes 0. For suboptimal arms ($\Delta_i > 0$), part (g) gives $\mathbb{E}[T_i(n)] \leq 3 + \frac{16 \ln n}{\Delta_i^2}$.

so:

$$R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)] \leq \sum_{i:\Delta_i>0} \Delta_i \left(3 + \tfrac{16 \ln n}{\Delta_i^2}\right) = 3 \sum_{i:\Delta_i>0} \Delta_i + 16 \ln n \sum_{i:\Delta_i>0} \tfrac{1}{\Delta_i}.$$

Since $\sum_{i:\Delta_i>0} \Delta_i = \sum_{i=1}^{K} \Delta_i$ (as $\Delta_{i^*} = 0$):

$$R_n \leq 3 \sum_{i=1}^{K} \Delta_i + \sum_{i:\Delta_i>0} \frac{16 \ln n}{\Delta_i}.$$

2.1.9 i)[5-points]

Choose $\Delta = \sqrt{\dfrac{16\,K \ln n}{n}}$. Split arms into $\{i : \Delta_i < \Delta\}$ and $\{i : \Delta_i \geq \Delta\}$. Then

$$R_n = \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[T_i(n)] = \sum_{\Delta_i<\Delta} \Delta_i \, \mathbb{E}[T_i(n)] + \sum_{\Delta_i\geq\Delta} \Delta_i \, \mathbb{E}[T_i(n)].$$

For $\Delta_i < \Delta$, since $T_i(n) \leq n$ and $\Delta_i < \Delta$,

$$\sum_{\Delta_i<\Delta} \Delta_i \, \mathbb{E}[T_i(n)] \leq n\,\Delta.$$

For $\Delta_i \geq \Delta$, use part(g): $\mathbb{E}[T_i(n)] \leq 3 + \frac{16 \ln n}{\Delta_i^2}$. So

$$\sum_{\Delta_i\geq\Delta} \Delta_i \, \mathbb{E}[T_i(n)] \leq \sum_{\Delta_i\geq\Delta} \left(3\,\Delta_i + \frac{16 \ln n}{\Delta_i}\right) \leq 3 \sum_{i=1}^{K} \Delta_i \; + \; \frac{16\,K \ln n}{\Delta}.$$

Putting these together,

$$R_n \leq n\,\Delta + \frac{16\,K \ln n}{\Delta} + 3 \sum_{i=1}^{K} \Delta_i.$$

Now put the $\Delta = \sqrt{16\,K \ln n / n}$ in to get

$$n\,\Delta + \frac{16\,K \ln n}{\Delta} = 2\sqrt{n \cdot 16\,K \ln n} = 8\sqrt{nK \ln n}.$$

Therefore

$$R_n \leq 8\sqrt{nK \ln n} + 3 \sum_{i=1}^{K} \Delta_i.$$

2.2 Power of 2 version of UCB Algorithm$^*(Bonus)[35 - points]$

3 Online Learning[50-points]

3.1 Randomized Weighted Majority Algorithm[35-points]

3.1.1 a)[5-points]

Let $S_t = \sum_i w_i(t)$. The weight update is:

$$S_{t+1} = \sum_{\text{correct } i} w_i(t) + \sum_{\text{wrong } i} w_i(t)(1-\varepsilon) = S_t - \varepsilon \sum_{\text{wrong } i} w_i(t).$$

The algorithm's mistake probability at round t is $\Pr(\text{mistake}_t) = \mathbb{E}\left[\sum_i \frac{w_i(t)}{S_t} \mathbf{1}_{\{i \text{ wrong}\}}\right]$. So:

$$\mathbb{E}\left[\sum_{\text{wrong } i} w_i(t)\right] = \mathbb{E}\left[S_t \cdot \Pr(\text{mistake}_t \mid \mathcal{F}_t)\right] = \mathbb{E}[S_t]\Pr(\text{mistake}_t).$$

so:

$$\mathbb{E}[S_{t+1}] = \mathbb{E}[S_t] - \varepsilon\mathbb{E}[S_t]\Pr(\text{mistake}_t) = \mathbb{E}[S_t]\left(1 - \varepsilon\Pr(\text{mistake}_t)\right).$$

3.1.2 b)[8-points]

From (a) we know $\mathbb{E}[S_{t+1}] = \mathbb{E}[S_t](1 - \varepsilon\Pr(\text{mistake}_t))$. Using $1 - x \le e^{-x}$ for $x \in [0,1]$:

$$\mathbb{E}[S_{t+1}] \le \mathbb{E}[S_t] e^{-\varepsilon \Pr(\text{mistake}_t)}.$$

When we open the inequality from $t = 1$ to T (with $S_1 = N$) we have:

$$\mathbb{E}[S_{T+1}] \le N \exp\left(-\varepsilon \sum_{t=1}^{T} \Pr(\text{mistake}_t)\right).$$

3.1.3 c)[15-points]

For any expert i, $w_i(T+1) = (1-\varepsilon)^{M_i}$ and $w_i(T+1) \le S_{T+1}$. By (b) we know:

$$(1-\varepsilon)^{M_i} \le \mathbb{E}[S_{T+1}] \le N e^{-\varepsilon\mathbb{E}[M]}.$$

Taking logs and using the fact that $\ln(1-\varepsilon) \ge -\varepsilon - \varepsilon^2$ (for $\varepsilon \in [0, 0.5]$):

$$M_i \ln(1-\varepsilon) \le \ln N - \varepsilon\mathbb{E}[M] \implies M_i(-\varepsilon - \varepsilon^2) \le \ln N - \varepsilon\mathbb{E}[M].$$

So finally we have $\varepsilon\mathbb{E}[M] \le M_i(\varepsilon + \varepsilon^2) + \ln N$, so:

$$\mathbb{E}[M] \le (1+\varepsilon)M_i + \frac{\ln N}{\varepsilon}.$$

3.1.4 d)[7-points]

From (c), for the best expert i^*:

$$\mathbb{E}[M] \leq (1+\varepsilon)\min_i M_i + \frac{\ln N}{\varepsilon} \leq \min_i M_i + \varepsilon T + \frac{\ln N}{\varepsilon}.$$

Set $\varepsilon = \sqrt{\frac{\ln N}{T}}$. Then:

$$\varepsilon T + \frac{\ln N}{\varepsilon} = \sqrt{T \ln N} + \sqrt{T \ln N} = 2\sqrt{T \ln N}.$$

This is a good regret bound. It's $O(\sqrt{T \log N})$, which is sublinear in T and logarithmic in N. So average regret $\to 0$ as $T \to \infty$. This is near-optimal for adversarial settings.

3.2 Hedge Algorithm$^*(Bonus)[15-points]$

3.2.1 a)[6-points]

Update: $w_i(t+1) = w_i(t)e^{-\varepsilon \ell_{t,i}}$. So:

$$S_{t+1} = \sum_i w_i(t) e^{-\varepsilon \ell_{t,i}}.$$

Using $e^{-x} \leq 1 - x + x^2$ for $|x| \leq 1$ (since $\ell_{t,i} \in [-1,1]$ and $\varepsilon \leq 1$):

$$S_{t+1} \leq \sum_i w_i(t)\left(1 - \varepsilon \ell_{t,i} + \varepsilon^2 \ell_{t,i}^2\right) = S_t - \varepsilon S_t \sum_i p_t(i)\ell_{t,i} + \varepsilon^2 S_t \sum_i p_t(i)\ell_{t,i}^2.$$

3.2.2 b)[7-points]

From (a), $S_{t+1} \leq S_t \exp\left(-\varepsilon \sum_i p_t(i)\ell_{t,i} + \varepsilon^2 \sum_i p_t(i)\ell_{t,i}^2\right)$. Opening the inequality:

$$S_{T+1} \leq S_1 \exp\left(-\varepsilon \sum_{t=1}^{T}\sum_i p_t(i)\ell_{t,i} + \varepsilon^2 \sum_{t=1}^{T}\sum_i p_t(i)\ell_{t,i}^2\right).$$

Since $S_1 = N$ and $S_{T+1} \geq w_i(T+1) = \exp\left(-\varepsilon \sum_{t=1}^{T} \ell_{t,i}\right)$ for any i:

$$\exp\left(-\varepsilon \sum_{t=1}^{T} \ell_{t,i}\right) \leq N \exp\left(-\varepsilon \sum_{t=1}^{T}\sum_i p_t(i)\ell_{t,i} + \varepsilon^2 \sum_{t=1}^{T}\sum_i p_t(i)\ell_{t,i}^2\right).$$

So this derives the bound by taking logarithms.

Hedge generalizes RWM to $[-1,1]$ losses. For 0-1 losses, $\ell_t(i)^2 = \ell_t(i)$, so the bound becomes $\frac{\ln N}{\varepsilon} + \varepsilon \mathbb{E}[M]$, matching RWM's form.

3.2.3 c)[2-points]

From (b) and $\ell_{t,i}^2 \leq 1$:

$$\text{Regret} \leq \frac{\ln N}{\varepsilon} + \varepsilon \sum_{t=1}^{T} 1 = \frac{\ln N}{\varepsilon} + \varepsilon T.$$

Set $\varepsilon = \sqrt{\frac{\ln N}{T}}$. Then:

$$\frac{\ln N}{\varepsilon} + \varepsilon T = \sqrt{T \ln N} + \sqrt{T \ln N} = 2\sqrt{T \ln N}.$$
