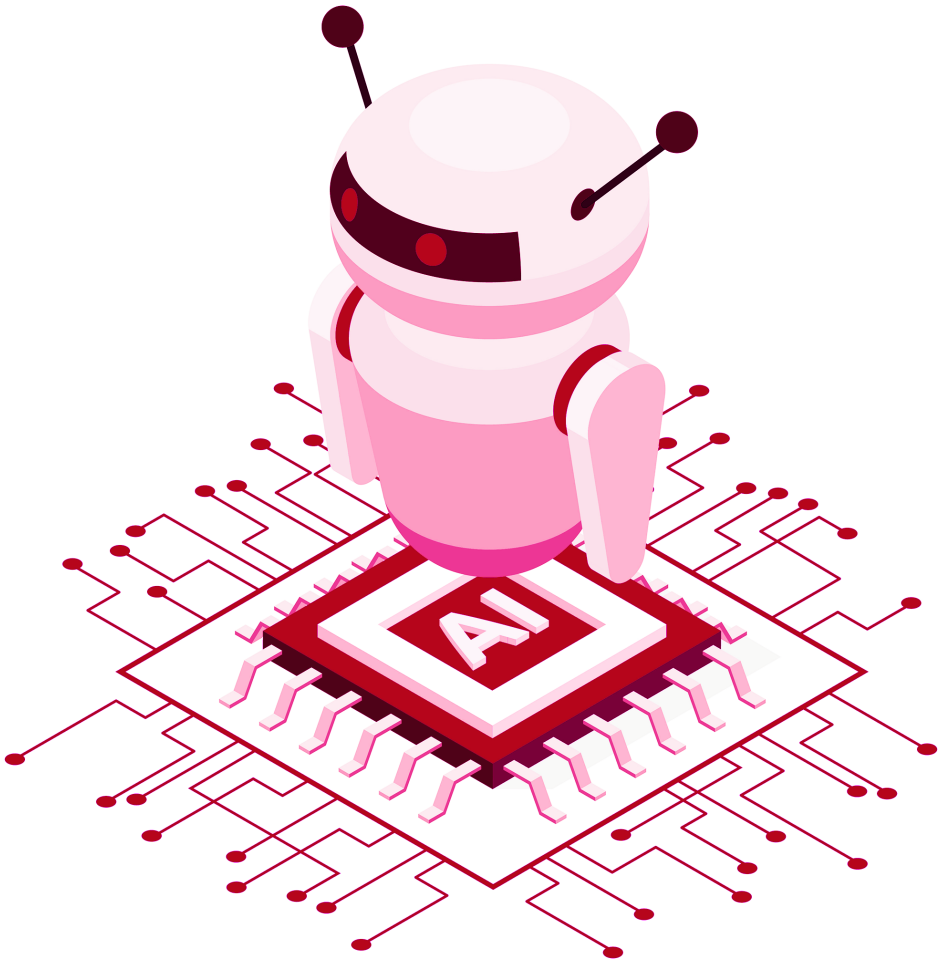

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution forHomework [9]

Exploration Methods

By:

Danial Parnian

401110307

RIML

Spring 2025

Contents

1 Light-tailed Distributions[25-points] **1**

1.1 Hoeffding's Inequality[10-points] 1

1.1.1 a)[6-points] 1

1.1.2 b)[4-points] 2

1.2 Sub-Gaussian[15-points] 4

1.2.1 a-1)[2-points] 4

1.2.2 a-2)[2-points] 5

1.2.3 a-3)[2-points] 5

1.2.4 b)[3-points] 5

1.2.5 c)[4-points] 6

2 UCB[75-points] **8**

2.1 The Upper Confidence Bound Algorithm[40-points] 8

2.1.1 a)[2-points] 8

2.1.2 b)[4-points] 9

2.1.3 c)[4-points] 9

2.1.4 d)[4-points] 10

2.1.5 e)[6-points] 10

2.1.6 f)[4-points] 11

2.1.7 g)[6-points] 12

2.1.8 h)[5-points] 13

2.1.9 i)[5-points] 14

2.2 Power of 2 version of UCB Algorithm$^*(Bonus)[35-points]$ 16

3 Online Learning[50-points] **18**

3.1 Randomized Weighted Majority Algorithm[35-points] 18

3.1.1 a)[5-points] 18

3.1.2 b)[8-points] 18

3.1.3 c)[15-points] 19

3.1.4 d)[7-points] 20

Declaration: I used ChatGPT to enhance my writing and correct grammatical errors.

1 Light-tailed Distributions[25-points]

1.1 Hoeffding's Inequality[10-points]

1.1.1 a)[6-points]

Solution. I'll prove it step by step.

Step 1: The function $f(x) = e^{sx}$ is convex for any real s. Any point x in $[a, b]$ can be written as a convex combination of its endpoints:

$$x = \frac{b-x}{b-a}a + \frac{x-a}{b-a}b$$

By Jensen's inequality, we can write:

$$e^{sX} \leq \frac{b-X}{b-a}e^{sa} + \frac{X-a}{b-a}e^{sb}$$

Step 2: Take the expectation of both sides of the inequality:

$$\begin{aligned}\mathbb{E}[e^{sX}] &\leq \mathbb{E}\left[\frac{b-X}{b-a}e^{sa} + \frac{X-a}{b-a}e^{sb}\right] \\ &= \frac{b-\mathbb{E}[X]}{b-a}e^{sa} + \frac{\mathbb{E}[X]-a}{b-a}e^{sb}\end{aligned}$$

Step 3: Given that $\mathbb{E}[X] = 0$, the inequality simplifies to:

$$\mathbb{E}[e^{sX}] \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}$$

Step 4: Let $p = \frac{-a}{b-a}$. Since $a \leq \mathbb{E}[X] \leq b$, it follows that $a \leq 0 \leq b$, which ensures $p \in [0, 1]$. Also we conclude that $a = -p(b-a)$ and $b = (1-p)(b-a)$. Substituting these into the inequality:

$$\mathbb{E}[e^{sX}] \leq (1-p)e^{-sp(b-a)} + pe^{s(1-p)(b-a)}$$

Let $u = s(b-a)$. We define a log-generating function $L(u)$:

$$L(u) = \log\left((1-p)e^{-pu} + pe^{(1-p)u}\right) = -pu + \log\left(1-p+pe^{u}\right)$$

Our goal is to show that $L(u) \leq \frac{u^2}{8}$.

Step 5: We expand $L(u)$ around $u = 0$ using Taylor Expansion. First, we compute its value and the first two derivatives at $u = 0$:

- $L(0) = -p(0) + \log(1 - p + pe^0) = \log(1) = 0$.
- $L'(u) = -p + \frac{pe^u}{1-p+pe^u}$. At $u = 0$, $L'(0) = -p + \frac{p}{1-p+p} = 0$.
- $L''(u) = \frac{p(1-p)e^u}{(1-p+pe^u)^2}$.

To bound the second derivative, one can show that $L''(u) \leq \frac{1}{4}$ for all u. This can be proven by letting $z = \sqrt{\frac{p}{1-p}}e^{u/2}$, which transforms $L''(u)$ to $\frac{1}{(z+1/z)^2}$, and since $(z + 1/z)^2 \geq 4$, the bound holds.

By Taylor's theorem, there exists some ξ between 0 and u such that:

$$L(u) = L(0) + uL'(0) + \frac{u^2}{2}L''(\xi)$$

Substituting the values:

$$L(u) = 0 + u(0) + \frac{u^2}{2}L''(\xi) \leq \frac{u^2}{2} \cdot \frac{1}{4} = \frac{u^2}{8}$$

Step 6: Conclusion Substitute back $u = s(b - a)$:

$$\log(\mathbb{E}[e^{sX}]) \leq L(s(b-a)) \leq \frac{(s(b-a))^2}{8} = \frac{s^2(b-a)^2}{8}$$

Exponentiate both sides to get the desired result:

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2(b-a)^2}{8}}$$

□

1.1.2 b)[4-points]

Solution. First we'll prove following lemmas:

Lemma 1 (Markov's Inequality)**.** *For any non-negative random variable X and any constant $a > 0$:*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof. From the definition of expectation for a non-negative continuous random variable:

$$\mathbb{E}[X] = \int_0^\infty xf(x)dx = \int_0^a xf(x)dx + \int_a^\infty xf(x)dx$$

Since the first integral is non-negative, $\mathbb{E}[X] \geq \int_a^\infty xf(x)dx$. Inside this integral, $x \geq a$, so we can write:

$$\mathbb{E}[X] \geq \int_a^\infty af(x)dx = a\int_a^\infty f(x)dx = a \cdot \mathbb{P}(X \geq a)$$

Rearranging gives the result. □

Lemma 2 (The Chernoff Bound). *For any random variable Y, any constant t, and any $s > 0$:*

$$\mathbb{P}(Y \geq t) \leq e^{-st}\mathbb{E}[e^{sY}]$$

Proof. The event $\{Y \geq t\}$ is identical to the event $\{sY \geq st\}$ since $s > 0$. Because the exponential function is monotonically increasing, this is also identical to the event $\{e^{sY} \geq e^{st}\}$.

$$\mathbb{P}(Y \geq t) = \mathbb{P}(e^{sY} \geq e^{st})$$

Let's define a new random variable $X = e^{sY}$. Since $e^{sY} > 0$, X is a non-negative random variable. We can apply Markov's inequality to X with the constant $a = e^{st}$:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Substituting the expressions for X and a:

$$\mathbb{P}(e^{sY} \geq e^{st}) \leq \frac{\mathbb{E}[e^{sY}]}{e^{st}}$$

This directly leads to the Chernoff bound:

$$\mathbb{P}(Y \geq t) \leq e^{-st}\mathbb{E}[e^{sY}]$$

□

Now Let $X_i = Z_i - \mathbb{E}[Z_i]$. The variables X_i are independent with $\mathbb{E}[X_i] = 0$. The sum is $S_n = \sum_{i=1}^{n} X_i$.

1. Proof of the Upper Tail We want to bound $\mathbb{P}\left(\frac{S_n}{n} \geq t\right) = \mathbb{P}(S_n \geq nt)$. Using the Chernoff bound (lemma 2) for any $s > 0$:

$$\mathbb{P}(S_n \geq nt) \leq e^{-snt}\mathbb{E}[e^{sS_n}]$$

Because the X_i's are independent, we'll have:

$$\mathbb{E}[e^{sS_n}] = \mathbb{E}\left[e^{s\sum X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{sX_i}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{sX_i}]$$

Each X_i is a random variable with mean 0. Since $Z_i \in [a, b]$, X_i is bounded in an interval of length $b - a$. By Hoeffding's Lemma (part a), we have:

$$\mathbb{E}[e^{sX_i}] \leq e^{s^2(b-a)^2/8}$$

Substituting this into the product:

$$\prod_{i=1}^{n} \mathbb{E}[e^{sX_i}] \leq \prod_{i=1}^{n} e^{s^2(b-a)^2/8} = e^{ns^2(b-a)^2/8}$$

Plugging this back into our first bound:

$$\mathbb{P}(S_n \geq nt) \leq e^{-snt} \cdot e^{ns^2(b-a)^2/8} = \exp\left(-snt + \frac{ns^2(b-a)^2}{8}\right)$$

This bound holds for any $s > 0$. To find the tightest bound, we minimize the exponent by choosing an optimal s. Let $g(s) = -snt + \frac{ns^2(b-a)^2}{8}$.

$$\frac{dg}{ds} = -nt + \frac{2ns(b-a)^2}{8} = 0 \implies s = \frac{4t}{(b-a)^2}$$

Since $t \geq 0$, our choice of s is valid. Substituting this optimal s back into the exponent:

$$-\left(\frac{4t}{(b-a)^2}\right)nt + \frac{n(b-a)^2}{8}\left(\frac{4t}{(b-a)^2}\right)^2 = \frac{-4nt^2}{(b-a)^2} + \frac{2nt^2}{(b-a)^2} = \frac{-2nt^2}{(b-a)^2}$$

This gives the final bound for the upper tail:

$$\mathbb{P}\left(\frac{S_n}{n} \geq t\right) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

2. Proof of the Lower Tail To prove the lower tail, we can apply the same argument to the variables $-X_i$.

$$\mathbb{P}\left(\frac{S_n}{n} \leq -t\right) = \mathbb{P}(-S_n \geq nt) = \mathbb{P}\left(\sum_{i=1}^{n}(-X_i) \geq nt\right)$$

Let $Y_i = -X_i$. The variables Y_i are independent, have mean 0, and are also bounded in an interval of length $b-a$. Therefore, the entire proof for the upper tail applies directly to the sum $\sum Y_i$. This immediately gives the result:

$$\mathbb{P}\left(\frac{S_n}{n} \leq -t\right) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

This completes the proof. □

1.2 Sub-Gaussian[15-points]

1.2.1 a-1)[2-points]

Solution. We want to prove $\Pr[X - \mu > t] \leq e^{-t^2/2\sigma^2}$. For any $\lambda > 0$, we apply the Chernoff bound:

$$\Pr[X - \mu \geq t] = \Pr[e^{\lambda(X-\mu)} \geq e^{\lambda t}] \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}$$

Using the sub-Gaussian property:

$$\Pr[X - \mu \geq t] \leq \frac{e^{\lambda^2\sigma^2/2}}{e^{\lambda t}} = \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda t\right)$$

This bound holds for any $\lambda > 0$. To find the tightest bound, we minimize the exponent $g(\lambda) = \frac{\lambda^2\sigma^2}{2} - \lambda t$ with respect to λ:

$$\frac{dg}{d\lambda} = \lambda\sigma^2 - t = 0 \implies \lambda = \frac{t}{\sigma^2}$$

This λ is positive since $t > 0$. Substituting it back into the exponent:

$$\frac{1}{2}\left(\frac{t}{\sigma^2}\right)^2\sigma^2 - \left(\frac{t}{\sigma^2}\right)t = \frac{t^2}{2\sigma^2} - \frac{t^2}{\sigma^2} = -\frac{t^2}{2\sigma^2}$$

Thus, $\Pr[X - \mu \geq t] \leq e^{-t^2/2\sigma^2}$. Since $\Pr[X > \mu + t] \leq \Pr[X \geq \mu + t]$, the inequality holds. □

1.2.2 a-2)[2-points]

Solution. We want to prove $\Pr[X < \mu - t]$, which is equivalent to $\Pr[X - \mu < -t]$. We use the Chernoff bound again. For any $\lambda' > 0$, we can write:

$$\Pr[-(X-\mu) \geq t] = \Pr[e^{-\lambda'(X-\mu)} \geq e^{\lambda' t}] \leq \frac{\mathbb{E}[e^{-\lambda'(X-\mu)}]}{e^{\lambda' t}}$$

Let $\lambda = -\lambda'$. Since the sub-Gaussian definition holds for all $\lambda \in \mathbb{R}$, we have:

$$\Pr[X - \mu \leq -t] \leq e^{\lambda t}\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda t}e^{\lambda^2\sigma^2/2} = \exp\left(\frac{\lambda^2\sigma^2}{2} + \lambda t\right)$$

To minimize the exponent for $\lambda < 0$, we find the optimal $\lambda = -t/\sigma^2$. Substituting this in yields result:

$$\Pr[X < \mu - t] \leq e^{-t^2/2\sigma^2}$$

□

1.2.3 a-3)[2-points]

Solution. This result follows directly from the first two parts using the union bound. The event $|X - \mu| \geq t$ is the union of two disjoint events: $\{X - \mu \geq t\}$ and $\{X - \mu \leq -t\}$.

$$\begin{aligned}\Pr[|X-\mu| \geq t] &= \Pr[(X-\mu \geq t) \cup (X-\mu \leq -t)] \\ &= \Pr[X \geq \mu + t] + \Pr[X \leq \mu - t]\end{aligned}$$

Using the bounds from parts (1) and (2):

$$\Pr[|X-\mu| \geq t] \leq e^{-t^2/2\sigma^2} + e^{-t^2/2\sigma^2} = 2e^{-t^2/2\sigma^2}$$

This completes the proof. □

1.2.4 b)[3-points]

Solution. First we'll prove following lemma.

Lemma 3. *Let $X_1, \ldots, X_n$ be independent random variables, where each X_i is sub-Gaussian with parameter σ_i^2 and mean μ_i. Their sum, $S_n = \sum_{i=1}^n X_i$, is a sub-Gaussian random variable with parameter $\sigma_S^2 = \sum_{i=1}^n \sigma_i^2$.*

Proof. To prove the lemma, we must show that S_n satisfies the sub-Gaussian condition with respect to its mean, $\mu_S = \mathbb{E}[S_n] = \sum_{i=1}^n \mu_i$. That is, we must show that for any $\lambda \in \mathbb{R}$:

$$\mathbb{E}[e^{\lambda(S_n - \mu_S)}] \leq e^{\lambda^2\sigma_S^2/2}$$

We start from the left-hand side:

$$\begin{aligned}\mathbb{E}[e^{\lambda(S_n-\mu_S)}] &= \mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right)}\right] \\ &= \mathbb{E}\left[e^{\sum_{i=1}^n \lambda(X_i-\mu_i)}\right] \\ &= \mathbb{E}\left[\prod_{i=1}^n e^{\lambda(X_i-\mu_i)}\right]\end{aligned}$$

Due to the independence of the $\{X_i\}$ variables, we have:

$$\mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda(X_i-\mu_i)}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda(X_i-\mu_i)}]$$

By the definition of a sub-Gaussian variable, we have $\mathbb{E}[e^{\lambda(X_i-\mu_i)}] \leq e^{\lambda^2\sigma_i^2/2}$ for each i. Substituting this into the product:

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

By defining $\sigma_S^2 = \sum_{i=1}^{n}\sigma_i^2$, we arrive at the desired condition:

$$\mathbb{E}[e^{\lambda(S_n-\mu_S)}] \leq e^{\lambda^2\sigma_S^2/2}$$

This completes the proof of the lemma. □

Now lets solve the problem step by step:

Step 1: Let $Y_i = X_i - \mu_i$. Each Y_i is an independent, zero-mean random variable that is sub-Gaussian with parameter σ_i^2. Let $S = \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}(X_i - \mu_i)$. Our goal is to bound $\Pr[|S| \geq t]$.

Step 2: According to the lemma 3, the sum S is a sub-Gaussian random variable. Its mean is $\mathbb{E}[S] = \sum_{i=1}^{n}\mathbb{E}[Y_i] = 0$. The sub-Gaussian parameter is $\sigma_S^2 = \sum_{i=1}^{n}\sigma_i^2$.

Step 3: In part a-3, we proved that for any sub-Gaussian variable X with mean μ and parameter σ^2, the following holds:

$$\Pr[|X - \mu| \geq t] \leq 2e^{-t^2/2\sigma^2}$$

We substitute S for X, its mean $\mathbb{E}[S] = 0$ for μ, and its parameter $\sigma_S^2 = \sum_{i=1}^{n}\sigma_i^2$ for σ^2:

$$\Pr[|S - 0| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sigma_S^2}\right)$$

Step 4: Conclusion This yields the final result:

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

□

1.2.5 c)[4-points]

Solution. Note that replacing t with $n\epsilon$ in the previous part, leads to something similar but with an extra factor of 2 in the RHS. To reach the desired result, we do every step again:

1. Proof of the First Inequality We use the Chernoff bound method. Let $Y_i = X_i - \mathbb{E}[X]$. The variables $\{Y_i\}$ are i.i.d. with $\mathbb{E}[Y_i] = 0$ and are σ-sub-Gaussian. We want to bound $\mathbb{P}(\frac{1}{n}\sum Y_i \geq \epsilon) = \mathbb{P}(\sum Y_i \geq n\epsilon)$.

For any $\lambda > 0$, the Chernoff bound is:

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq n\epsilon\right) \leq e^{-\lambda n\epsilon}\mathbb{E}\left[e^{\lambda \sum Y_i}\right]$$

By independence and the sub-Gaussian property of each Y_i:

$$\mathbb{E}\left[e^{\lambda \sum Y_i}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda Y_i}] \leq \prod_{i=1}^{n} e^{\lambda^2\sigma^2/2} = e^{n\lambda^2\sigma^2/2}$$

Substituting this back into the bound:

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq n\epsilon\right) \leq e^{-\lambda n\epsilon} e^{n\lambda^2\sigma^2/2} = \exp\left(-\lambda n\epsilon + \frac{n\lambda^2\sigma^2}{2}\right)$$

To get the tightest bound, we minimize the exponent by choosing the optimal λ. Setting the derivative to zero gives $\lambda = \epsilon/\sigma^2$. Plugging this value back:

$$-n\epsilon\left(\frac{\epsilon}{\sigma^2}\right) + \frac{n\sigma^2}{2}\left(\frac{\epsilon}{\sigma^2}\right)^2 = -\frac{n\epsilon^2}{2\sigma^2}$$

This gives the final inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$$

2. Proof of the Second Inequality This statement is derived by inverting the inequality from the first part. Let $\bar{X}_n = \frac{1}{n}\sum X_i$. From part 1, we know:

$$\mathbb{P}(\bar{X}_n - \mu \geq \epsilon) \leq e^{-n\epsilon^2/(2\sigma^2)}$$

We want the probability of this event to be at most δ. So we set the bound equal to δ and solve for the deviation ϵ.

$$e^{-n\epsilon^2/(2\sigma^2)} = \delta$$

Taking the logarithm:

$$-\frac{n\epsilon^2}{2\sigma^2} = \ln(\delta) = -\ln(1/\delta)$$

$$\epsilon^2 = \frac{2\sigma^2\ln(1/\delta)}{n} \implies \epsilon = \sqrt{\frac{2\sigma^2\ln(1/\delta)}{n}}$$

So, for this specific value of ϵ, we have shown:

$$\mathbb{P}\left(\bar{X}_n - \mu \geq \sqrt{\frac{2\sigma^2\ln(1/\delta)}{n}}\right) \leq \delta$$

The event we are interested in is the complement of the one above. The probability of a complement event A^c is $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. This leads us to the final result:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu < \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}}\right) \geq 1 - \delta$$

□

2 UCB[75-points]

2.1 The Upper Confidence Bound Algorithm[40-points]

2.1.1 a)[2-points]

Solution. We'll rewrite the standard definition of total regret by decomposing it across the arms. First lets rename some definitions. Let $\mu_a = \mathbb{E}[r|\text{arm } a \text{ is pulled}]$ be the true mean reward of arm a (previously defined as R_i). Let $\mu^* = \max_{a \in \mathcal{A}} \mu_a$ be the mean reward of the optimal arm (previously defined as R_{max}). The reward gap for arm a is $\Delta_a = \mu^* - \mu_a$. By definition, $\Delta_a \geq 0$ for all arms, and the gap for the optimal arm is 0. Also note that by the definition of $T_a(n)$ we have $n = \sum_{a \in \mathcal{A}} T_a(n)$.

The total regret, R_n, is the expected difference between the total reward obtained by an optimal policy (always pulling the best arm) and the total reward obtained by the policy being evaluated.

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

We can rewrite both terms in this expression using $T_a(n)$.

Since $n = \sum_{a \in \mathcal{A}} T_a(n)$, and this holds for any realization, it also holds in expectation: $n = \sum_{a \in \mathcal{A}} \mathbb{E}[T_a(n)]$.

$$n\mu^* = \left(\sum_{a \in \mathcal{A}} \mathbb{E}[T_a(n)]\right) \mu^* = \sum_{a \in \mathcal{A}} \mu^* \mathbb{E}[T_a(n)]$$

Also we can rewrite the second term like this:

$$\sum_{t=1}^{n} \mu_{a_t} = \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{I}\{a_t = a\}\mu_a = \sum_{a \in \mathcal{A}} \mu_a \sum_{t=1}^{n} \mathbb{I}\{a_t = a\} = \sum_{a \in \mathcal{A}} \mu_a T_a(n)$$

Taking the expectation gives:

$$\mathbb{E}\left[\sum_{t=1}^{n} \mu_{a_t}\right] = \mathbb{E}\left[\sum_{a \in \mathcal{A}} \mu_a T_a(n)\right] = \sum_{a \in \mathcal{A}} \mu_a \mathbb{E}[T_a(n)]$$

Now, we substitute these expressions back into the regret formula:

$$R_n = \sum_{a \in \mathcal{A}} \mu^* \mathbb{E}[T_a(n)] - \sum_{a \in \mathcal{A}} \mu_a \mathbb{E}[T_a(n)]$$

Factoring out the common term:

$$R_n = \sum_{a \in \mathcal{A}} (\mu^* - \mu_a)\mathbb{E}[T_a(n)] = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)]$$

This completes the proof. □

2.1.2 b)[4-points]

Solution. If we assume δ is a fixed constant C, the UCB exploration bonus $\sqrt{2\log(1/C)/T_i(t-1)}$ has a constant numerator. This is problematic over a long horizon.

A bad event occurs when the optimal arm a^*, is underestimated. Imagine this scenario: Due to a sequence of unlikely low rewards, the empirical mean of the optimal arm, $\hat{\mu}_{a^*}$, becomes very low. This causes its UCB score to fall below the true mean of a suboptimal arm, μ_j:

$$\hat{\mu}_{a^*}(t-1) + \sqrt{\frac{2\log(1/C)}{T_{a^*}(t-1)}} < \mu_j$$

Once this occurs, the algorithm will stop pulling the optimal arm a^*. It will persistently select the suboptimal arm j (or another arm it now deems better), causing the number of pulls for that arm, $T_j(K)$, to be proportional to the total steps K. This results in linear regret.

With a fixed $\delta = C$, the probability of such a concentration failure is a non-zero constant for each evaluation. Over many steps, it becomes certain that this failure will occur at some point.

How to Choose δ to Avoid This: To prevent linear regret, the confidence parameter δ must not be fixed. It should be a decaying function of time.

We must choose δ to decrease as the number of rounds t (or the total horizon n) increases. For example, one could set $\delta_t = 1/t^c$ for some $c > 1$. We'll later choose $\delta = 1/n^2$ for the final proof.

By making δ time-dependent, we ensure that our confidence in our estimates grows over time. Using a union bound over all arms and all time steps, the total probability of any confidence bound failing throughout the entire run can be kept small and bounded. This choice is what guarantees the desirable sub-linear regret of the UCB algorithm. □

2.1.3 c)[4-points]

Solution. We prove this statement by contradiction. Assume that event G_i occurs, but arm i is played more than u_i times, i.e., $T_i(n) > u_i$.

If arm i is played more than u_i times, then there must exist a time step t_0 at which arm i is chosen for the (u_i+1)-th time. At the moment just before this selection (at time $t_0 - 1$), arm i had been played exactly u_i times. Thus, $T_i(t_0 - 1) = u_i$.

According to the UCB algorithm, arm i was chosen at t_0 because its UCB index was the maximum. This implies its UCB was at least as large as the UCB of the optimal arm:

$$UCB_i(t_0 - 1, \delta) \geq UCB_1(t_0 - 1, \delta) \tag{1}$$

Now we use the two conditions of event G_i. From first condition, we know that:

$$UCB_1(t_0 - 1, \delta) > \mu_1 \tag{2}$$

Chaining the inequalities from (1) and (2), we get:

$$UCB_i(t_0 - 1, \delta) > \mu_1$$

However, second condition of G_i states that the UCB of arm i after u_i pulls is strictly less than μ_1. Since $T_i(t_0 - 1) = u_i$, the UCB for arm i at that specific moment is:

$$UCB_i(t_0 - 1, \delta) = \hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} < \mu_1$$

We have arrived at a contradiction: $\mu_1 > UCB_i(t_0 - 1, \delta) > \mu_1$. Therefore our initial assumption must be false. Therefore, if event G_i occurs, it must be that $T_i(n) \leq u_i$. □

2.1.4 d)[4-points]

Solution. We'll condition the expected number of pulls of a suboptimal arm, on whether the good event G_i occurs.

We begin by applying the law of total expectation to the random variable $T_i(n)$, the number of times arm i is pulled, conditioning on the event G_i and its complement G_i^c:

$$\mathbb{E}[T_i(n)] = \mathbb{E}[T_i(n)|G_i]P(G_i) + \mathbb{E}[T_i(n)|G_i^c]P(G_i^c)$$

We now bound the two conditional expectations from the formula above.

- **Case 1 (G_i occurs):** The lemma from part (c) states that if G_i occurs, then $T_i(n) \leq u_i$. Therefore:

$$\mathbb{E}[T_i(n)|G_i] \leq u_i$$

- **Case 2 (G_i^c occurs):** If G_i does not occur, we must resort to a trivial (worst-case) bound. The number of times any arm is pulled cannot exceed the total number of rounds, n:

$$\mathbb{E}[T_i(n)|G_i^c] \leq n$$

Substituting these bounds back into the main equation:

$$\mathbb{E}[T_i(n)] \leq u_i \cdot P(G_i) + n \cdot P(G_i^c)$$

Now note that $P(G_i) \leq 1$. This completes the proof:

$$\mathbb{E}[T_i(n)] \leq u_i + nP(G_i^c)$$

□

2.1.5 e)[6-points]

Solution. The proof relies on rearranging the event inside the probability and then applying the inequality of problem 1.2.c for the mean of sub-Gaussian variables.

The event is:

$$\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1$$

By subtracting μ_i from both sides and rearranging, we get an equivalent event:

$$\hat{\mu}_{iu_i} - \mu_i \geq \mu_1 - \mu_i - \sqrt{\frac{2\log(1/\delta)}{u_i}}$$

Let's define the deviation threshold ϵ as the term on the right-hand side. Using the definition $\Delta_i = \mu_1 - \mu_i$, we have:

$$\epsilon = \Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}}$$

The event is now in the standard form $\mathbb{P}(\hat{\mu}_{iu_i} - \mu_i \geq \epsilon)$. Since the problem states the rewards are from a 1-sub-Gaussian distribution, we can use inequality from problem 1.2c with variance proxy $\sigma^2 = 1$:

$$\mathbb{P}(\hat{\mu}_{iu_i} - \mu_i \geq \epsilon) \leq \exp\left(-\frac{u_i\epsilon^2}{2}\right)$$

The problem provides the assumption that u_i is chosen such that $\epsilon \geq c\Delta_i$. As $c > 0$ and $\Delta_i > 0$, we can square both sides to get $\epsilon^2 \geq c^2\Delta_i^2$.

We substitute this into the exponent of our bound. The negative sign flips the inequality direction:

$$-\frac{u_i\epsilon^2}{2} \leq -\frac{u_i(c\Delta_i)^2}{2} = -\frac{u_i c^2\Delta_i^2}{2}$$

Plugging this into the probability bound gives:

$$\mathbb{P}(\hat{\mu}_{iu_i} - \mu_i \geq \epsilon) \leq \exp\left(-\frac{u_i c^2\Delta_i^2}{2}\right)$$

Since the event $\{\hat{\mu}_{iu_i} - \mu_i \geq \epsilon\}$ is identical to the original event in the proposition, we have proven the desired inequality:

$$\mathbb{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right) \leq \exp\left(-\frac{u_i c^2\Delta_i^2}{2}\right)$$

□

2.1.6 f)[4-points]

Solution. We'll decompose the event G_i^c and then bound the probability of each component part.

The event G_i is the intersection of two "good" events, $G_i = A \cap B$, where:

- $A = \left\{\mu_1 < \min_{t\in[n]} UCB_1(t,\delta)\right\}$
- $B = \left\{\hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} < \mu_1\right\}$

The complement event is $G_i^c = (A \cap B)^c$. By De Morgan's laws, this becomes $G_i^c = A^c \cup B^c$. Using the union bound for probabilities, we can write:

$$\mathbb{P}(G_i^c) = \mathbb{P}(A^c \cup B^c) \leq \mathbb{P}(A^c) + \mathbb{P}(B^c)$$

We now bound the probability of each of these "bad" events, A^c and B^c, separately.

The event B^c is the complement of B:

$$B^c = \left\{ \hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1 \right\}$$

This is precisely the event whose probability we bounded in part (e) of this problem. we have:

$$\mathbb{P}(B^c) \leq \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$$

The event A^c is that the UCB of the optimal arm fails at some point during the n rounds:

$$A^c = \{\exists t \in [n] \text{ such that } \mu_1 \geq UCB_1(t, \delta)\}$$

A failure of the UCB for arm 1 after it has been pulled exactly s times is the event $E_s = \left\{ \mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}} \right\}$. By the construction of the UCB confidence interval from the sub-Gaussian concentration inequality, we know that $\mathbb{P}(E_s) \leq \delta$. A^c is a subset of the union of all possible such failure events:

$$A^c \subseteq \bigcup_{s=1}^{n-1} E_s$$

Using the union bound on this set of events:

$$\mathbb{P}(A^c) \leq \mathbb{P}\left(\bigcup_{s=1}^{n-1} E_s\right) \leq \sum_{s=1}^{n-1} \mathbb{P}(E_s)$$

Since $\mathbb{P}(E_s) \leq \delta$ for each of the $n-1$ possible values of s:

$$\mathbb{P}(A^c) \leq \sum_{s=1}^{n-1} \delta = (n-1)\delta < n\delta$$

Finally, we substitute the bounds for $\mathbb{P}(A^c)$ and $\mathbb{P}(B^c)$ into our original inequality:

$$\mathbb{P}(G_i^c) \leq \mathbb{P}(A^c) + \mathbb{P}(B^c) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$$

This completes the proof. □

2.1.7 g)[6-points]

Solution. From part (d), we have $\mathbb{E}[T_i(n)] \leq u_i + nP(G_i^c)$. Substituting the bound for $P(G_i^c)$ from part (f) gives:

$$\begin{aligned}\mathbb{E}[T_i(n)] &\leq u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right) \\ &= u_i + n^2\delta + n\exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\end{aligned}$$

To simplify the expression, we make the following standard choices:

- Let $\delta = 1/n^2$. This makes the term $n^2\delta = n^2(1/n^2) = 1$.
- Let $c = 1/2$. This is a valid choice in $(0, 1)$ that we will show works for all conditions.

With these choices, the bound on the expected pulls becomes:

$$\mathbb{E}[T_i(n)] \leq u_i + 1 + n \exp\left(-\frac{u_i(1/2)^2\Delta_i^2}{2}\right) = u_i + 1 + n \exp\left(-\frac{u_i\Delta_i^2}{8}\right)$$

Now we must choose u_i large enough to satisfy the prerequisite from part (e) and to make the final exponential term small. The condition from part (e) is $\Delta_i - \sqrt{2\log(1/\delta)/u_i} \geq c\Delta_i$. With our choices, this becomes:

$$\Delta_i - \sqrt{2\log(n^2)/u_i} \geq \frac{1}{2}\Delta_i \implies \frac{1}{2}\Delta_i \geq \sqrt{\frac{4\log(n)}{u_i}}$$

Squaring both sides and solving for u_i gives the condition:

$$\frac{\Delta_i^2}{4} \geq \frac{4\log(n)}{u_i} \implies u_i \geq \frac{16\log(n)}{\Delta_i^2}$$

To satisfy this, let's choose $u_i = \left\lceil \frac{16\log(n)}{\Delta_i^2} \right\rceil$. Since $u_i \geq \frac{16\log(n)}{\Delta_i^2}$:

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

We have chosen parameters such that our main inequality is now:

$$\mathbb{E}[T_i(n)] \leq u_i + 1 + \frac{1}{n}$$

Also we have:

$$u_i = \left\lceil \frac{16\log(n)}{\Delta_i^2} \right\rceil \leq \frac{16\log(n)}{\Delta_i^2} + 1$$

Substituting this in:

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

For any horizon $n \geq 1$, we have $1/n \leq 1$. Therefore, we can get the final desired bound:

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16\log(n)}{\Delta_i^2}$$

□

2.1.8 h)[5-points]

Solution. From the regret decomposition lemma in part (a), the total regret R_n is:

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

where k is the number of arms and $\Delta_i = \mu^* - \mu_i$ is the reward gap.

The terms in the sum corresponding to the optimal arms are zero, since for any optimal arm i^*, $\Delta_{i^*} = 0$. Therefore, we only need to sum over the set of suboptimal arms:

$$R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)]$$

In part (g), we showed that for any arm i:

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16\log(n)}{\Delta_i^2}$$

We substitute this inequality into our regret expression:

$$\begin{aligned} R_n &\leq \sum_{i:\Delta_i>0} \Delta_i \left(3 + \frac{16\log(n)}{\Delta_i^2}\right) \\ &= \sum_{i:\Delta_i>0} \left(3\Delta_i + \Delta_i \cdot \frac{16\log(n)}{\Delta_i^2}\right) \\ &= \sum_{i:\Delta_i>0} \left(3\Delta_i + \frac{16\log(n)}{\Delta_i}\right) \end{aligned}$$

We can split the sum into two parts:

$$R_n \leq \sum_{i:\Delta_i>0} 3\Delta_i + \sum_{i:\Delta_i>0} \frac{16\log(n)}{\Delta_i}$$

The first term, $\sum_{i:\Delta_i>0} 3\Delta_i$, can be rewritten as a sum over all arms $i = 1, \ldots, k$, since the added terms for the optimal arms are zero. This gives us the final form of the bound:

$$R_n \leq 3\sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log(n)}{\Delta_i}$$

This completes the proof. □

2.1.9 i)[5-points]

Solution. We begin with the regret decomposition from part (a), summed over suboptimal arms:

$$R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)]$$

Following the hint, we introduce a gap threshold $\Delta > 0$ and split the sum into two groups: arms with small gaps ($0 < \Delta_i \leq \Delta$) and arms with large gaps ($\Delta_i > \Delta$).

$$R_n = \underbrace{\sum_{i:0<\Delta_i\leq\Delta} \Delta_i \mathbb{E}[T_i(n)]}_{\text{Small-gap arms}} + \underbrace{\sum_{i:\Delta_i>\Delta} \Delta_i \mathbb{E}[T_i(n)]}_{\text{Large-gap arms}}$$

We now bound each component separately.

Bounding the Regret from Small-Gap Arms For this sum, we use the fact that $\Delta_i \leq \Delta$. The sum of expected pulls over this subset of arms is at most n.

$$\sum_{i:0<\Delta_i\leq\Delta} \Delta_i \mathbb{E}[T_i(n)] \leq \sum_{i:0<\Delta_i\leq\Delta} \Delta \cdot \mathbb{E}[T_i(n)] = \Delta \sum_{i:0<\Delta_i\leq\Delta} \mathbb{E}[T_i(n)] \leq n\Delta$$

Bounding the Regret from Large-Gap Arms For arms with $\Delta_i > \Delta$, we use the bound from part (g):

$$\begin{aligned}\sum_{i:\Delta_i>\Delta} \Delta_i \mathbb{E}[T_i(n)] &\leq \sum_{i:\Delta_i>\Delta} \Delta_i \left(3 + \frac{16\log(n)}{\Delta_i^2}\right) \\ &= \sum_{i:\Delta_i>\Delta} 3\Delta_i + \sum_{i:\Delta_i>\Delta} \frac{16\log(n)}{\Delta_i}\end{aligned}$$

Since $\Delta_i > \Delta$ for this sum, we have $1/\Delta_i < 1/\Delta$. The number of arms in the sum is at most k.

$$\sum_{i:\Delta_i>\Delta} \frac{16\log(n)}{\Delta_i} < \sum_{i:\Delta_i>\Delta} \frac{16\log(n)}{\Delta} \leq k \cdot \frac{16\log(n)}{\Delta}$$

So the regret from large-gap arms is bounded by $\sum_{i:\Delta_i>\Delta} 3\Delta_i + \frac{16k\log(n)}{\Delta}$.

Combining the bounds for both parts gives a total regret bound. We can bound the sum $\sum 3\Delta_i$ by extending it to all arms.

$$R_n \leq n\Delta + \frac{16k\log(n)}{\Delta} + 3\sum_{i=1}^{k} \Delta_i$$

To get the tightest bound, we choose Δ to minimize the expression $f(\Delta) = n\Delta + \frac{16k\log(n)}{\Delta}$. Taking the derivative with respect to Δ and setting it to zero yields the optimal choice:

$$\Delta = \sqrt{\frac{16k\log(n)}{n}} = 4\sqrt{\frac{k\log(n)}{n}}$$

Substituting this optimal Δ back into $f(\Delta)$:

$$n\left(4\sqrt{\frac{k\log(n)}{n}}\right) + \frac{16k\log(n)}{4\sqrt{\frac{k\log(n)}{n}}} = 4\sqrt{nk\log(n)} + 4\sqrt{nk\log(n)} = 8\sqrt{nk\log(n)}$$

Plugging this minimized value back into our full regret inequality gives the final result:

$$R_n \leq 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k} \Delta_i$$

□

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

Solution. The algorithm proceeds in phases $l = 1, 2, \ldots, L$, where phase l has length 2^{l-1}. The total number of steps is $n \approx \sum_{l=1}^{L} 2^{l-1} = 2^L - 1$, so the number of phases L is approximately $\log_2(n)$. The total regret is given by the decomposition lemma:

$$R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}[T_i(n)] = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)]$$

The core of the proof is to find an upper bound for $\mathbb{E}[T_i(n)]$ for any suboptimal arm i.

Bounding the Expected Pulls of a Suboptimal Arm Let's analyze the condition for choosing a suboptimal arm i at the start of phase l. This occurs if $UCB_i(t_l - 1) \geq UCB_{a^*}(t_l - 1)$. As shown in the main UCB proof, for this to happen under a high-probability "good event", the number of previous pulls of arm i, $T_i(t_l - 1)$, must be small. With confidence parameter $\delta = 1/n^2$, this condition is:

$$T_i(t_l - 1) \leq \frac{16 \log n}{\Delta_i^2} =: N_i^*$$

This implies that once the total number of pulls for arm i surpasses the threshold N_i^*, it will not be selected again. Let l_{last} be the last phase in which arm i was chosen. The total number of pulls is given by the sum of pulls from all chosen phases:

$$T_i(n) = 1 + \sum_{l \in \{\text{phases where } i \text{ was chosen}\}} 2^{l-1} = T_i(t_{l_{last}} - 1) + 2^{l_{last}-1}$$

For arm i to have been chosen at phase l_{last}, its prior pull count must have satisfied $T_i(t_{l_{last}} - 1) \leq N_i^*$.

To obtain a bound on $T_i(n)$, we need to bound the length of the last phase, $2^{l_{last}-1}$. The number of pulls before phase l_{last} is a sum of distinct powers of 2 (plus one for initialization), so $T_i(t_{l_{last}} - 1) \leq 1 + \sum_{j=1}^{l_{last}-1} 2^{j-1} = 2^{l_{last}-1}$. The condition $T_i(t_{l_{last}} - 1) \leq N_i^*$ must hold. A simple but valid upper bound on the phase length is $2^{l_{last}-1} \leq T_i(t_{l_{last}} - 1) \cdot 2$ if we assume the pulls form a contiguous block of powers of two, or more generally, $2^{l_{last}-1} \leq N_i^* + 1$. A slightly looser argument that holds is that the sum of pulls will be at most twice the next power of two, leading to $T_i(n) \leq 2(N_i^* + O(1))$. For simplicity, we establish the bound:

$$\mathbb{E}[T_i(n)] \leq O\left(\frac{\log n}{\Delta_i^2}\right)$$

Using the constants from our analysis, this is approximately $\mathbb{E}[T_i(n)] \leq \frac{32 \log n}{\Delta_i^2} + O(1)$.

Deriving the Final Regret Bound We now use this bound on expected pulls to derive the overall regret.

$$R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)] \leq \sum_{i:\Delta_i>0} \Delta_i \cdot O\left(\frac{\log n}{\Delta_i^2}\right) = O\left(\log n \sum_{i:\Delta_i>0} \frac{1}{\Delta_i}\right)$$

We use the same splitting-sum technique as in part (i), with a threshold Δ.

- Regret from small-gap arms ($\Delta_i \leq \Delta$) is bounded by $n\Delta$.

- Regret from large-gap arms ($\Delta_i > \Delta$) is bounded by:

$$\sum_{i:\Delta_i>\Delta} \Delta_i \left(\frac{32 \log n}{\Delta_i^2} + O(1) \right) < \frac{32k \log n}{\Delta} + O(k\Delta_{max})$$

The total regret is thus bounded by $R_n \leq n\Delta + \frac{32k \log n}{\Delta} + O(k)$. We find the tightest bound by choosing Δ to minimize $f(\Delta) = n\Delta + \frac{32k \log n}{\Delta}$. The optimal choice is $\Delta = \sqrt{\frac{32k \log n}{n}}$. Substituting this back gives:

$$f(\Delta_{opt}) = 2\sqrt{32nk \log n} = 2 \cdot 4\sqrt{2}\sqrt{nk \log n} = 8\sqrt{2}\sqrt{nk \log n}$$

Therefore the regret for the Power-of-2 UCB algorithm is bounded by:

$$R_n \leq 8\sqrt{2}\sqrt{nk \log n} + O(k)$$

This demonstrates that the algorithm achieves the same sub-linear asymptotic regret profile of $O(\sqrt{nk \log n})$ as the standard UCB algorithm. The constant factor is, however, larger.

3 Online Learning[50-points]

3.1 Randomized Weighted Majority Algorithm[35-points]

3.1.1 a)[5-points]

Solution. Let $S_t = \sum_{i=1}^{N} w_i(t)$ be the total weight at round t. The total weight at round $t+1$ is:

$$\begin{aligned} S_{t+1} &= \sum_{i=1}^{N} w_i(t+1) = \sum_{i \notin M_t} w_i(t) + \sum_{i \in M_t} w_i(t)(1-\epsilon) \\ &= \left(\sum_{i \notin M_t} w_i(t) + \sum_{i \in M_t} w_i(t) \right) - \epsilon \sum_{i \in M_t} w_i(t) \\ &= S_t - \epsilon \sum_{i \in M_t} w_i(t) \end{aligned}$$

Where M_t is the set of indices of experts those make the wrong choice at time t.

The learner makes a mistake (the event $\tilde{m}_t = 1$) if its chosen expert, i_t, is in the set M_t. The probability of this, conditioned on the weights at time t, is the sum of the probabilities of choosing any of the wrong experts:

$$P(\tilde{m}_t = 1 | \{w_i(t)\}) = \sum_{i \in M_t} P(\text{choose } i) = \sum_{i \in M_t} \frac{w_i(t)}{S_t} = \frac{\sum_{i \in M_t} w_i(t)}{S_t}$$

We can rearrange this to express the sum of the weights of the mistaken experts:

$$\sum_{i \in M_t} w_i(t) = S_t \cdot P(\tilde{m}_t = 1 | \{w_i(t)\})$$

Substituting this back into our exact update rule for S_{t+1}:

$$S_{t+1} = S_t - \epsilon \left(S_t \cdot P(\tilde{m}_t = 1 | \{w_i(t)\}) \right) = S_t \left(1 - \epsilon \cdot P(\tilde{m}_t = 1 | \{w_i(t)\}) \right)$$

This gives an exact relationship between the random variables at each step. Taking the expectation over the history of weights:

$$\mathbb{E}[S_{t+1}] = \mathbb{E}\left[S_t \left(1 - \epsilon \cdot P(\tilde{m}_t = 1) \right) \right]$$

The proof is completed. □

3.1.2 b)[8-points]

Solution. From part (a), we have:

$$\mathbb{E}[S_{t+1}] = \mathbb{E}[S_t] \cdot (1 - \epsilon \cdot P(\tilde{m}_t = 1))$$

Using the inequality $1 - x \leq e^{-x}$, which holds for all real x, we can bound the update rule:

$$\mathbb{E}[S_{t+1}] \leq \mathbb{E}[S_t] \cdot e^{-\epsilon \cdot P(\tilde{m}_t = 1)}$$

We can apply this inequality recursively from the final step $t = T$ back to the first step $t = 1$. Let $P_t = P(\tilde{m}_t = 1)$.

$$\begin{aligned}
\mathbb{E}[S_{T+1}] &\leq \mathbb{E}[S_T] \cdot e^{-\epsilon P_T} \\
&\leq \left(\mathbb{E}[S_{T-1}] \cdot e^{-\epsilon P_{T-1}}\right) \cdot e^{-\epsilon P_T} \\
&\leq \cdots \\
&\leq \mathbb{E}[S_1] \cdot \prod_{t=1}^{T} e^{-\epsilon P_t}
\end{aligned}$$

The initial weights are all set to 1, so the total weight before the first round is $S_1 = \sum_{i=1}^{N} w_i(1) = N$. Substituting the initial weight gives the final result:

$$\begin{aligned}
\mathbb{E}[S_{T+1}] &\leq N \cdot \prod_{t=1}^{T} e^{-\epsilon P(\tilde{m}_t = 1)} \\
&= N \cdot e^{-\epsilon \sum_{t=1}^{T} P(\tilde{m}_t = 1)}
\end{aligned}$$

□

3.1.3 c)[15-points]

Solution. The total weight at the end of T rounds, S_{T+1}, must be at least as large as the weight of any individual expert i.

$$S_{T+1} \geq w_i(T + 1)$$

An expert's weight begins at $w_i(1) = 1$ and is multiplied by $(1 - \epsilon)$ each time it makes a mistake. If expert i makes M_i mistakes, its final weight is $w_i(T + 1) = (1 - \epsilon)^{M_i}$. Therefore:

$$\mathbb{E}[S_{T+1}] \geq (1 - \epsilon)^{M_i}$$

From part (b), we have an upper bound on the final weight that depends on the learner's expected number of mistakes, $\mathbb{E}[M] = \sum_{t=1}^{T} P(\tilde{m}_t = 1)$:

$$\mathbb{E}[S_{T+1}] \leq N \cdot e^{-\epsilon \mathbb{E}[M]}$$

We can now combine the lower and upper bounds on $\mathbb{E}[S_{T+1}]$:

$$(1 - \epsilon)^{M_i} \leq \mathbb{E}[S_{T+1}] \leq N \cdot e^{-\epsilon \mathbb{E}[M]}$$

Taking the logarithm of the two expressions:

$$\begin{aligned}
\ln\left((1 - \epsilon)^{M_i}\right) &\leq \ln\left(N \cdot e^{-\epsilon \mathbb{E}[M]}\right) \\
M_i \ln(1 - \epsilon) &\leq \ln(N) - \epsilon \mathbb{E}[M]
\end{aligned}$$

Now, we rearrange to solve for $\mathbb{E}[M]$:

$$\epsilon\mathbb{E}[M] \leq \ln(N) - M_i \ln(1-\epsilon)$$
$$\mathbb{E}[M] \leq \frac{\ln(N)}{\epsilon} + M_i \left(\frac{-\ln(1-\epsilon)}{\epsilon}\right)$$

We use the standard inequality for logarithms, $-\ln(1-x) \leq x + x^2$ for $x \in [0, 1/2)$. Applying this with $x = \epsilon$:

$$\mathbb{E}[M] \leq \frac{\ln(N)}{\epsilon} + M_i \left(\frac{\epsilon + \epsilon^2}{\epsilon}\right) = \frac{\ln(N)}{\epsilon} + M_i(1+\epsilon)$$

Rearranging the terms gives the final inequality, which holds for any expert i:

$$\mathbb{E}[M] \leq (1+\epsilon)M_i + \frac{\ln N}{\epsilon}$$

□

3.1.4 d)[7-points]

Proposition 1.

Solution. From part (c), the learner's expected mistakes $\mathbb{E}[M]$ are bounded relative to the best expert, who makes $M_{min} = \min_i M_i$ mistakes:

$$\mathbb{E}[M] \leq (1+\epsilon)M_{min} + \frac{\ln N}{\epsilon}$$

The bound on the right-hand side depends on our choice of ϵ. To get the tightest bound, we would need to choose ϵ to minimize the expression. The optimal choice, $\epsilon_{opt} = \sqrt{\ln N / M_{min}}$, depends on M_{min}, which is unknown in advance.

Instead, we set a practical learning rate using the hint that the number of mistakes is always less than the number of rounds, $M_{min} \leq T$. We set:

$$\epsilon = \sqrt{\frac{\ln N}{T}}$$

We plug this choice of ϵ into the inequality from part (c):

$$\mathbb{E}[M] \leq \left(1 + \sqrt{\frac{\ln N}{T}}\right) M_{min} + \frac{\ln N}{\sqrt{\frac{\ln N}{T}}}$$
$$= M_{min} + M_{min}\sqrt{\frac{\ln N}{T}} + \sqrt{T \ln N}$$

Now, use the fact that $M_{min} \leq T$:

$$M_{min}\sqrt{\frac{\ln N}{T}} \leq T\sqrt{\frac{\ln N}{T}} = \sqrt{T^2 \cdot \frac{\ln N}{T}} = \sqrt{T \ln N}$$

Substituting this back into the expression gives the final bound:

$$\mathbb{E}[M] \leq M_{min} + \sqrt{T \ln N} + \sqrt{T \ln N}$$
$$= \min_{i \in [N]} M_i + 2\sqrt{T \ln N}$$

Relation to Regret In the context of online learning with expert advice, regret is the difference between the number of mistakes made by our algorithm and the number of mistakes made by the best single expert in hindsight. The inequality we proved provides a direct bound on the expected regret, $\mathbb{E}[\text{Regret}_T]$:

$$\mathbb{E}[\text{Regret}_T] = \mathbb{E}[M] - \min_{i \in [N]} M_i \leq 2\sqrt{T \ln N}$$

This is considered a very good regret bound. A bound is considered good if it is **sub-linear** in the time horizon T, meaning that the average regret per round approaches zero as T grows large. Let's verify this condition:

$$\text{Average Regret} = \frac{\mathbb{E}[\text{Regret}_T]}{T} \leq \frac{2\sqrt{T \ln N}}{T} = \frac{2\sqrt{\ln N}}{\sqrt{T}}$$

As $T \to \infty$, the average regret $\frac{2\sqrt{\ln N}}{\sqrt{T}} \to 0$. This sub-linear regret of $O(\sqrt{T})$ demonstrates that the algorithm is effectively learning as its performance, converges to the best possible expert.
