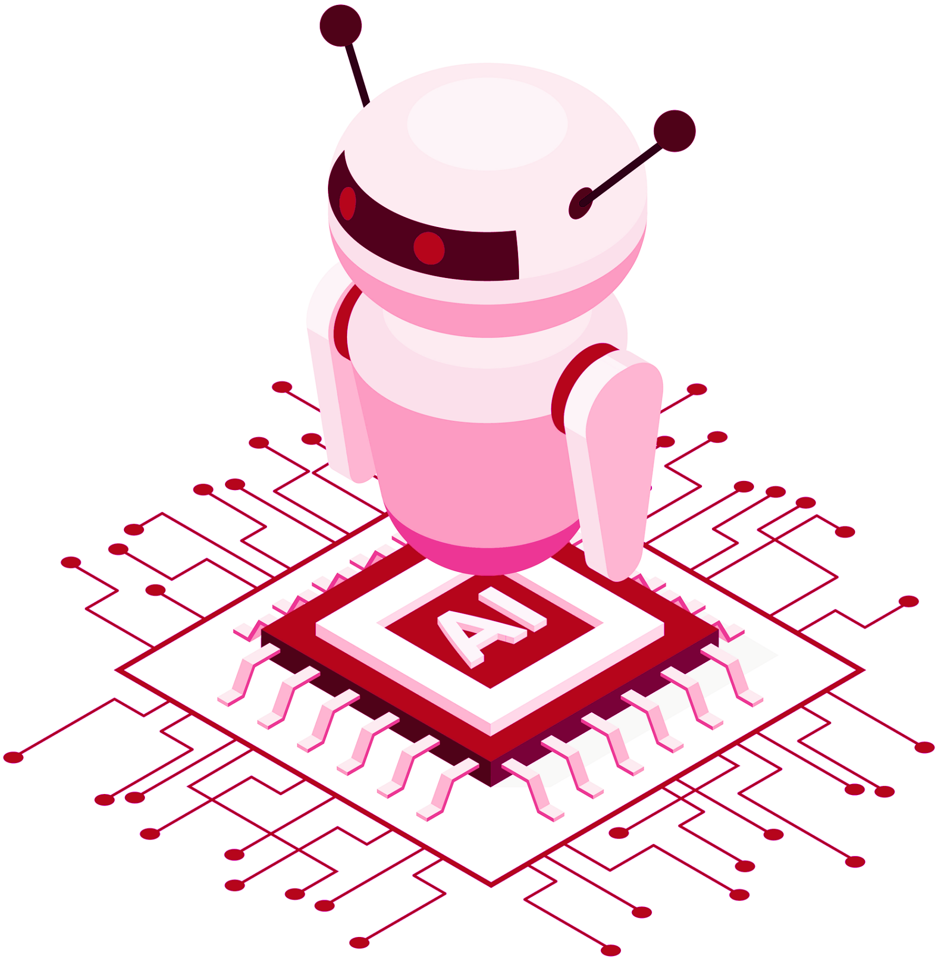

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution forHomework [9]

[Exploration Methods]

By:

Payam Taebi

400104867

Spring 2025

Contents

1 Light-tailed Distributions[25-points]

1.1 Hoeffding's Inequality[10-points]

1.1.1 a)[6-points]

We want to prove that if X is a random variable satisfying $\mathbb{E}[X] = 0$ and

$$a \;\le\; X \;\le\; b,$$

then for every $s > 0$,

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

Step 1 (Setup). Let X be as above. Because e^{sx} is a convex function of x, for any fixed value $x \in [a, b]$ we can bound e^{sx} by the line joining the two endpoints (a, e^{sa}) and (b, e^{sb}). In particular, for each $x \in [a, b]$,

$$e^{sx} \;\le\; \frac{b - x}{b - a}\, e^{sa} \;+\; \frac{x - a}{b - a}\, e^{sb}.$$

Taking expectations on both sides gives

$$\mathbb{E}\big[e^{sX}\big] \;\le\; \mathbb{E}\Big[\tfrac{b-X}{b-a}\, e^{sa} \;+\; \tfrac{X-a}{b-a}\, e^{sb}\Big] \;=\; \frac{e^{sa}}{b-a}\,\mathbb{E}[\,b - X\,] \;+\; \frac{e^{sb}}{b-a}\,\mathbb{E}[\,X - a\,].$$

Since $\mathbb{E}[X] = 0$, we have

$$\mathbb{E}[\,b - X\,] \;=\; b, \qquad \mathbb{E}[\,X - a\,] \;=\; -a,$$

so

$$\mathbb{E}\big[e^{sX}\big] \;\le\; \frac{b\,e^{sa} \;-\; a\,e^{sb}}{b - a}.$$

Step 2 (Centering the Interval). Define the midpoint $m = \frac{a+b}{2}$ and half-length $c = \frac{b-a}{2}$. Then $a = m - c$ and $b = m + c$. Let us shift X by m: define $Y = X - m$. Then Y satisfies

$$Y \;\in\; [\,a - m,\; b - m\,] \;=\; [-c,\; c], \quad \mathbb{E}[Y] \;=\; \mathbb{E}[X] - m \;=\; -m.$$

However, since $\mathbb{E}[X] = 0$, this forces $m = 0$. In other words, without loss of generality we may assume that $a = -c$, $b = +c$, and so $X \in [-c, c]$ with $\mathbb{E}[X] = 0$. In that case,

$$\mathbb{E}\big[e^{sX}\big] \;\le\; \frac{e^{s(-c)} + e^{s(c)}}{2} \;=\; \cosh(sc).$$

It is well known that for all real u,

$$\cosh(u) \;\le\; \exp\big(\tfrac{u^2}{2}\big).$$

Substituting $u = s\,c$ yields

$$\mathbb{E}\big[e^{sX}\big] \;\le\; \exp\!\Big(\tfrac{(sc)^2}{2}\Big) \;=\; \exp\!\Big(\tfrac{s^2\,c^2}{2}\Big).$$

Since $c = \frac{b-a}{2}$, we obtain

$$\mathbb{E}\big[e^{sX}\big] \;\le\; \exp\!\Big(\tfrac{s^2\,(b-a)^2}{8}\Big),$$

which proves the desired bound.

1.1.2 b)[4-points]

Let $Z_1, \dots, Z_n$ be independent random variables, each satisfying $Z_i \in [a, b]$. Define

$$X_i \;=\; Z_i \;-\; \mathbb{E}[Z_i], \quad \text{so that } \mathbb{E}[X_i] = 0 \quad \text{and} \quad X_i \in \big[\,a - \mathbb{E}[Z_i],\, b - \mathbb{E}[Z_i]\,\big].$$

In particular, the length of the interval containing X_i is at most $(b - a)$. We wish to show that for any $t \ge 0$,

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \;\ge\; t\Big) \;\le\; \exp\!\Big(-\tfrac{2\,n\,t^2}{(b-a)^2}\Big),$$

and similarly for the lower tail.

Step 1 (Chernoff Bound). For any $s > 0$, by Markov's (Chernoff) inequality,

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \;\ge\; n\,t\Big) \;=\; \mathbb{P}\Big(e^{\,s\sum_{i=1}^{n} X_i} \;\ge\; e^{\,s\,n\,t}\Big) \;\le\; e^{-\,s\,n\,t}\,\mathbb{E}\Big[e^{\,s\sum_{i=1}^{n} X_i}\Big].$$

Since the X_i are independent,

$$\mathbb{E}\Big[e^{\,s\sum_{i=1}^{n} X_i}\Big] \;=\; \prod_{i=1}^{n}\mathbb{E}\big[e^{\,s\,X_i}\big].$$

By part (a), each X_i satisfies $\mathbb{E}[e^{\,s\,X_i}] \le \exp\!\big(\frac{s^2(b-a)^2}{8}\big)$. Therefore,

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \;\ge\; n\,t\Big) \;\le\; e^{-\,s\,n\,t} \;\times\; \Big(\exp\!\big(\tfrac{s^2(b-a)^2}{8}\big)\Big)^{n} \;=\; \exp\!\Big(-\,s\,n\,t \;+\; \tfrac{n\,s^2\,(b-a)^2}{8}\Big).$$

Step 2 (Optimize Over s). To obtain the tightest possible bound, we choose $s > 0$ to minimize the exponent

$$-\,s\,n\,t \;+\; \frac{n\,s^2\,(b-a)^2}{8}.$$

Differentiating with respect to s and setting to zero:

$$\frac{d}{ds}\Big[-\,s\,n\,t \;+\; \tfrac{n\,s^2\,(b-a)^2}{8}\Big] \;=\; -\,n\,t \;+\; \frac{2\,n\,s\,(b-a)^2}{8} \;=\; -\,n\,t \;+\; \frac{n\,s\,(b-a)^2}{4} \;=\; 0.$$

Hence

$$s \;=\; \frac{4\,t}{(b-a)^2}.$$

Substitute this value of s back into the exponent:

$$-\,s\,n\,t \;+\; \frac{n\,s^2\,(b-a)^2}{8} = -\,n\,t\cdot\frac{4\,t}{(b-a)^2} \;+\; \frac{n}{8}\left(\frac{4\,t}{(b-a)^2}\right)^2(b-a)^2$$
$$= -\,\frac{4\,n\,t^2}{(b-a)^2} \;+\; \frac{n}{8}\,\frac{16\,t^2}{(b-a)^4}\,(b-a)^2$$
$$= -\,\frac{4\,n\,t^2}{(b-a)^2} \;+\; \frac{2\,n\,t^2}{(b-a)^2} \;=\; -\,\frac{2\,n\,t^2}{(b-a)^2}.$$

Consequently,

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \;\geq\; n\,t\Big) \;\leq\; \exp\Big(-\,\tfrac{2\,n\,t^2}{(b-a)^2}\Big).$$

Equivalently,

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i]) \;\geq\; t\Big) \;\leq\; \exp\Big(-\,\tfrac{2\,n\,t^2}{(b-a)^2}\Big).$$

Step 3 (Lower-Tail Bound). A completely analogous argument applies to $\{-X_i\}$. In particular, for any $s>0$,

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \;\leq\; -\,n\,t\Big) = \mathbb{P}\Big(\sum_{i=1}^{n}(-X_i) \;\geq\; n\,t\Big) \;\leq\; \exp\Big(-\,\tfrac{2\,n\,t^2}{(b-a)^2}\Big).$$

Hence

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i]) \;\leq\; -\,t\Big) \;\leq\; \exp\Big(-\,\tfrac{2\,n\,t^2}{(b-a)^2}\Big).$$

Conclusion. We have shown both tail bounds:

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i]) \geq t\Big) \;\leq\; \exp\Big(-\tfrac{2\,n\,t^2}{(b-a)^2}\Big), \quad \mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i]) \leq -t\Big) \;\leq\; \exp\Big(-\tfrac{2\,n\,t^2}{(b-a)^2}\Big).$$

This completes the proof of Hoeffding's inequality.

1.2 Sub-Gaussian[15-points]

We begin by recalling the definition:

A random variable X with mean $\mu = \mathbb{E}[X]$ is called *sub-Gaussian* if there exists a positive number σ, known as the *sub-Gaussian parameter*, such that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[e^{\lambda\,(X-\mu)}\big] \;\leq\; \exp\Big(\tfrac{\lambda^2\,\sigma^2}{2}\Big).$$

Sub-Gaussian random variables satisfy strong concentration properties around their mean. We will now derive the following inequalities for any $t>0$:

1. $\Pr\big[X > \mu + t\big] \;\leq\; e^{-\,t^2/(2\sigma^2)}$.
2. $\Pr\big[X < \mu - t\big] \;\leq\; e^{-\,t^2/(2\sigma^2)}$.
3. $\Pr\big[\,|X-\mu| \geq t\,\big] \;\leq\; 2\,e^{-\,t^2/(2\sigma^2)}$.

1.2.1 a-1)[2-points]

Let $t > 0$. For any $\lambda > 0$, by Markov's inequality,

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

Using the defining property of sub-Gaussianity,

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \ \leq \ \exp\left(\tfrac{\lambda^2 \sigma^2}{2}\right).$$

Hence,

$$\Pr\left[X > \mu + t\right] \ \leq \ \exp\left(\tfrac{\lambda^2 \sigma^2}{2} - \lambda\, t\right).$$

To obtain the tightest bound, we choose λ to minimize the exponent

$$\tfrac{\lambda^2 \sigma^2}{2} \ - \ \lambda\, t.$$

Differentiating with respect to λ and setting to zero yields

$$\frac{d}{d\lambda}\left(\tfrac{\lambda^2 \sigma^2}{2} - \lambda\, t\right) = \lambda\, \sigma^2 - t \ = \ 0 \ \implies \ \lambda = \frac{t}{\sigma^2}.$$

Substituting $\lambda = \frac{t}{\sigma^2}$ back into the exponent gives

$$\tfrac{(t/\sigma^2)^2\, \sigma^2}{2} \ - \ (t/\sigma^2)\, t = \tfrac{t^2}{2\, \sigma^2} \ - \ \tfrac{t^2}{\sigma^2} = - \, \tfrac{t^2}{2\, \sigma^2}.$$

Therefore,

$$\Pr\left[X > \mu + t\right] \ \leq \ \exp\left(- \, \tfrac{t^2}{2\, \sigma^2}\right).$$

1.2.2 a-2)[2-points]

We now bound $\Pr\left[X < \mu - t\right]$. For any $\lambda < 0$, similarly,

$$\Pr\left[X < \mu - t\right] \ = \ \Pr\left[X - \mu < -\, t\right] \ = \ \Pr\left[e^{\lambda(X-\mu)} > e^{-\, \lambda t}\right] \ \leq \ \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{-\, \lambda t}}.$$

Again by sub-Gaussianity,

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \ \leq \ \exp\left(\tfrac{\lambda^2 \sigma^2}{2}\right),$$

so

$$\Pr\left[X < \mu - t\right] \ \leq \ \exp\left(\tfrac{\lambda^2 \sigma^2}{2} + \lambda\, t\right).$$

To minimize $\frac{\lambda^2 \sigma^2}{2} + \lambda\, t$ over $\lambda < 0$, set the derivative to zero:

$$\frac{d}{d\lambda}\left(\tfrac{\lambda^2 \sigma^2}{2} + \lambda\, t\right) = \lambda\, \sigma^2 + t \ = \ 0 \ \implies \ \lambda = -\, \frac{t}{\sigma^2}.$$

Substituting this into the exponent gives

$$\tfrac{(-\, t/\sigma^2)^2\, \sigma^2}{2} + \left(-\, \tfrac{t}{\sigma^2}\right) t = \tfrac{t^2}{2\, \sigma^2} \ - \ \tfrac{t^2}{\sigma^2} = - \, \tfrac{t^2}{2\, \sigma^2}.$$

Hence,

$$\Pr\left[X < \mu - t\right] \ \leq \ \exp\left(- \, \tfrac{t^2}{2\, \sigma^2}\right).$$

1.2.3 a-3)[2-points]

Finally, for the absolute deviation,

$$\Pr\big[\,|X-\mu|\ge t\big] = \Pr\big[X-\mu\ge t\ \cup\ X-\mu\le -t\big]\ \le\ \Pr\big[X-\mu\ge t\big]+\Pr\big[X-\mu\le -t\big].$$

Using the bounds from parts (1) and (2), we get

$$\Pr\big[X-\mu\ge t\big]\ \le\ e^{-t^2/(2\sigma^2)},\qquad \Pr\big[X-\mu\le -t\big]\ \le\ e^{-t^2/(2\sigma^2)}.$$

Therefore,

$$\Pr\big[\,|X-\mu|\ge t\big]\ \le\ e^{-t^2/(2\sigma^2)}\ +\ e^{-t^2/(2\sigma^2)}\ =\ 2\,e^{-t^2/(2\sigma^2)}.$$

Conclusion. We have shown that for any sub-Gaussian random variable X with parameter σ and any $t>0$:

$$\Pr\big[X>\mu+t\big]\ \le\ e^{-t^2/(2\sigma^2)},\quad \Pr\big[X<\mu-t\big]\ \le\ e^{-t^2/(2\sigma^2)},\quad \Pr\big[\,|X-\mu|\ge t\big]\ \le\ 2\,e^{-t^2/(2\sigma^2)}.$$

These inequalities demonstrate the exponential concentration behavior typical of sub-Gaussian variables.

1.2.4 b)[3-points]

Let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i]=\mu_i$. Assume each centered variable $X_i-\mu_i$ is sub-Gaussian with parameter σ_i, i.e. for all $\lambda\in\mathbb{R}$,

$$\mathbb{E}\big[\exp\big(\lambda\,(X_i-\mu_i)\big)\big]\ \le\ \exp\Big(\tfrac{\lambda^2\,\sigma_i^2}{2}\Big).$$

We will prove that for every $t\ge 0$,

$$\Pr\Big(\Big|\sum_{i=1}^{n}(X_i-\mu_i)\Big|\ge t\Big)\ \le\ 2\,\exp\Big(-\tfrac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big).$$

1. One-Sided Upper-Tail Bound. Fix $t>0$ and choose any $\lambda>0$. By Markov's (Chernoff) inequality applied to the nonnegative random variable $\exp\big(\lambda\sum_{i=1}^{n}(X_i-\mu_i)\big)$, we get:

$$\Pr\Big(\sum_{i=1}^{n}(X_i-\mu_i)\ge t\Big)=\Pr\Big(e^{\lambda\sum_{i=1}^{n}(X_i-\mu_i)}\ge e^{\lambda t}\Big)\le\frac{\mathbb{E}\big[e^{\lambda\sum_{i=1}^{n}(X_i-\mu_i)}\big]}{e^{\lambda t}}.$$

Since the X_i are independent,

$$\mathbb{E}\Big[e^{\lambda\sum_{i=1}^{n}(X_i-\mu_i)}\Big]=\prod_{i=1}^{n}\mathbb{E}\big[e^{\lambda(X_i-\mu_i)}\big]\le\prod_{i=1}^{n}\exp\Big(\tfrac{\lambda^2\sigma_i^2}{2}\Big)=\exp\Big(\tfrac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2\Big).$$

Therefore,

$$\Pr\Big(\sum_{i=1}^{n}(X_i-\mu_i)\ge t\Big)\ \le\ \exp\Big(-\lambda\,t+\tfrac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2\Big).$$

To optimize, choose $\lambda>0$ to minimize the exponent

$$-\lambda\,t+\tfrac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2.$$

Differentiate with respect to λ and set to zero:

$$\frac{d}{d\lambda}\Big[-\lambda t + \tfrac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2\Big] = -t + \lambda\sum_{i=1}^n \sigma_i^2 = 0 \quad\Longrightarrow\quad \lambda = \frac{t}{\sum_{i=1}^n \sigma_i^2}.$$

Substitute this λ back into the exponent:

$$-\frac{t}{\sum_i \sigma_i^2}\,t \;+\; \tfrac{1}{2}\Big(\frac{t}{\sum_i \sigma_i^2}\Big)^2 \sum_{i=1}^n \sigma_i^2 = -\frac{t^2}{\sum_{i=1}^n \sigma_i^2} \;+\; \frac{t^2}{2\,\sum_{i=1}^n \sigma_i^2} = -\frac{t^2}{2\,\sum_{i=1}^n \sigma_i^2}.$$

Hence,

$$\Pr\Big(\sum_{i=1}^n (X_i - \mu_i) \ge t\Big) \;\le\; \exp\Big(-\tfrac{t^2}{2\,\sum_{i=1}^n \sigma_i^2}\Big).$$

2. One-Sided Lower-Tail Bound. We now bound $\Pr\big(\sum_{i=1}^n (X_i - \mu_i) \le -t\big)$. For any $\lambda < 0$, by the same Chernoff argument:

$$\Pr\Big(\sum_{i=1}^n (X_i - \mu_i) \le -t\Big) = \Pr\Big(e^{\lambda\sum_{i=1}^n (X_i-\mu_i)} \ge e^{-\lambda t}\Big) \le \frac{\mathbb{E}\big[e^{\lambda\sum_{i=1}^n (X_i-\mu_i)}\big]}{e^{-\lambda t}}.$$

Again, independence implies

$$\mathbb{E}\Big[e^{\lambda\sum_{i=1}^n (X_i-\mu_i)}\Big] \le \exp\Big(\tfrac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2\Big),$$

so

$$\Pr\Big(\sum_{i=1}^n (X_i - \mu_i) \le -t\Big) \;\le\; \exp\Big(\tfrac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2 + \lambda t\Big).$$

To minimize $\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2 + \lambda t$ over $\lambda < 0$, set

$$\frac{d}{d\lambda}\Big(\tfrac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2 + \lambda t\Big) = \lambda\sum_{i=1}^n \sigma_i^2 + t = 0 \quad\Longrightarrow\quad \lambda = -\frac{t}{\sum_{i=1}^n \sigma_i^2}.$$

Substituting yields the exponent

$$\tfrac{(-t/\sum_i \sigma_i^2)^2}{2}\sum_{i=1}^n \sigma_i^2 \;+\; \big(-\tfrac{t}{\sum_i \sigma_i^2}\big)\,t = \frac{t^2}{2\,\sum_{i=1}^n \sigma_i^2} \;-\; \frac{t^2}{\sum_{i=1}^n \sigma_i^2} = -\frac{t^2}{2\,\sum_{i=1}^n \sigma_i^2}.$$

Hence,

$$\Pr\Big(\sum_{i=1}^n (X_i - \mu_i) \le -t\Big) \;\le\; \exp\Big(-\frac{t^2}{2\,\sum_{i=1}^n \sigma_i^2}\Big).$$

3. Two-Sided Bound. Combining the two one-sided bounds via a union bound:

$$\Pr\Big(\big|\sum_{i=1}^n (X_i - \mu_i)\big| \ge t\Big) = \Pr\Big(\sum_{i=1}^n (X_i - \mu_i) \ge t\Big) \;+\; \Pr\Big(\sum_{i=1}^n (X_i - \mu_i) \le -t\Big).$$

Each term on the right is bounded by $\exp\big(-t^2/(2\,\sum_{i=1}^n \sigma_i^2)\big)$. Therefore,

$$\Pr\Big(\big|\sum_{i=1}^n (X_i - \mu_i)\big| \ge t\Big) \;\le\; 2\,\exp\Big(-\frac{t^2}{2\,\sum_{i=1}^n \sigma_i^2}\Big).$$

Conclusion. This completes the proof that

$$\Pr\Big(\big|\sum_{i=1}^{n}(X_i-\mu_i)\big|\geq t\Big)\ \leq\ 2\,\exp\Big(-\frac{t^2}{2\,\sum_{i=1}^{n}\sigma_i^2}\Big),$$

as claimed.

1.2.5 c)[4-points]

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common mean $\mu = \mathbb{E}[X]$. Assume X is sub-Gaussian with parameter σ, i.e. for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[e^{\lambda(X-\mu)}\big]\ \leq\ \exp\Big(\tfrac{\lambda^2\sigma^2}{2}\Big).$$

Then each centered variable $X_i - \mu$ is sub-Gaussian with the same parameter σ. We will prove the following two inequalities:

1. For any $\epsilon \geq 0$,
$$\Pr\Big(\tfrac{1}{n}\sum_{i=1}^{n}X_i\ -\ \mu\ \geq\ \epsilon\Big)\ \leq\ \exp\Big(-\tfrac{n\,\epsilon^2}{2\,\sigma^2}\Big).$$

2. For any $\delta \in (0,1]$,
$$\Pr\Big(\tfrac{1}{n}\sum_{i=1}^{n}X_i\ -\ \mu\ <\ \sqrt{\tfrac{2\,\sigma^2\,\ln(1/\delta)}{n}}\Big)\ \geq\ 1-\delta.$$

Proof of the First Inequality. Define

$$S_n\ =\ \sum_{i=1}^{n}(X_i-\mu).$$

Since the X_i are independent and each $X_i - \mu$ is sub-Gaussian with parameter σ, we have for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[e^{\lambda S_n}\big]=\prod_{i=1}^{n}\mathbb{E}\big[e^{\lambda(X_i-\mu)}\big]\ \leq\ \prod_{i=1}^{n}\exp\Big(\tfrac{\lambda^2\sigma^2}{2}\Big)=\exp\Big(\tfrac{n\,\lambda^2\,\sigma^2}{2}\Big).$$

Thus S_n is sub-Gaussian with parameter $\sigma\sqrt{n}$. In particular, for any $\epsilon \geq 0$, set $t = n\,\epsilon$. By Markov's (Chernoff) inequality,

$$\Pr\big(S_n\geq t\big)=\Pr\big(e^{\lambda S_n}\geq e^{\lambda t}\big)\ \leq\ \frac{\mathbb{E}[\,e^{\lambda S_n}\,]}{e^{\lambda t}}\ \leq\ \exp\Big(\tfrac{n\,\lambda^2\,\sigma^2}{2}-\lambda\,t\Big).$$

To optimize the bound, choose $\lambda = \dfrac{t}{n\,\sigma^2} = \dfrac{n\,\epsilon}{n\,\sigma^2} = \dfrac{\epsilon}{\sigma^2}$. Substituting back:

$$\tfrac{n\,\lambda^2\,\sigma^2}{2}-\lambda\,t=\tfrac{n\,(\epsilon/\sigma^2)^2\,\sigma^2}{2}\ -\ \big(\tfrac{\epsilon}{\sigma^2}\big)\,(n\,\epsilon)=\tfrac{n\,\epsilon^2}{2\,\sigma^2}\ -\ \tfrac{n\,\epsilon^2}{\sigma^2}=-\,\tfrac{n\,\epsilon^2}{2\,\sigma^2}.$$

Hence

$$\Pr\Big(\sum_{i=1}^{n}(X_i-\mu)\geq n\,\epsilon\Big)\ \leq\ \exp\Big(-\tfrac{n\,\epsilon^2}{2\,\sigma^2}\Big).$$

Noting that $\sum_{i=1}^{n}(X_i-\mu)\geq n\,\epsilon$ is equivalent to $\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\geq\epsilon$, we conclude

$$\Pr\Big(\tfrac{1}{n}\sum_{i=1}^{n}X_i-\mu\geq\epsilon\Big)\ \leq\ \exp\Big(-\tfrac{n\,\epsilon^2}{2\,\sigma^2}\Big).$$

Proof of the Second Inequality. We now convert the tail bound into a high-probability statement. From the first inequality, for any $\epsilon > 0$,

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i - \mu \ge \epsilon\right) \;\le\; \exp\!\left(-\tfrac{n\,\epsilon^2}{2\,\sigma^2}\right).$$

Set the right-hand side equal to δ, i.e. $\exp(-n\,\epsilon^2/(2\,\sigma^2)) = \delta$. Solving for ϵ gives

$$-\frac{n\,\epsilon^2}{2\,\sigma^2} = \ln(\delta) \quad\Longrightarrow\quad \epsilon^2 = \frac{2\,\sigma^2\,\ln(1/\delta)}{n} \quad\Longrightarrow\quad \epsilon = \sqrt{\frac{2\,\sigma^2\,\ln(1/\delta)}{n}}.$$

Hence, with this choice,

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i - \mu \ge \sqrt{\tfrac{2\,\sigma^2\,\ln(1/\delta)}{n}}\right) = \delta.$$

Equivalently,

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i - \mu < \sqrt{\tfrac{2\,\sigma^2\,\ln(1/\delta)}{n}}\right) \;=\; 1-\delta,$$

so

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i - \mu < \sqrt{\tfrac{2\,\sigma^2\,\ln(1/\delta)}{n}}\right) \;\ge\; 1-\delta.$$

Conclusion. We have shown:

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i - \mu \ge \epsilon\right) \;\le\; \exp\!\left(-\frac{n\,\epsilon^2}{2\,\sigma^2}\right), \quad \text{for all } \epsilon \ge 0,$$

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i - \mu < \sqrt{\tfrac{2\,\sigma^2\,\ln(1/\delta)}{n}}\right) \;\ge\; 1-\delta, \quad \text{for all } \delta \in (0,1].$$

These complete the required concentration bounds for the empirical mean of i.i.d. sub-Gaussian random variables.

2 UCB[75-points]

2.1 The Upper Confidence Bound Algorithm[40-points]

2.1.1 a)[2-points]

Consider a stochastic multi-armed bandit problem with a (finite or countable) action set $\mathcal{A}$ and time horizon $n \in \mathbb{N}$. At each round $t = 1, 2, \ldots, n$, a learning policy π selects an arm $A_t \in \mathcal{A}$ and receives a reward $X_t \sim \nu_{A_t}$, where ν_a denotes the (unknown) distribution of arm a. Let

$$R_a \;=\; \mathbb{E}_{X \sim \nu_a}[\,X\,]$$

be the expected reward of arm $a \in \mathcal{A}$, and define

$$R_{\max} \;=\; \max_{a \in \mathcal{A}} R_a, \qquad \Delta_a \;=\; R_{\max} \;-\; R_a,$$

so that $\Delta_a \geq 0$ is the gap of arm a relative to the optimal mean.

Denote by

$$T_a(n) \;=\; \sum_{t=1}^{n} \mathbf{1}\{A_t = a\}$$

the (random) number of times arm a is selected up to and including round n. Clearly, $\sum_{a \in \mathcal{A}} T_a(n) = n$ almost surely, and therefore $\sum_{a \in \mathcal{A}} \mathbb{E}\big[T_a(n)\big] = n$.

Definition of Regret. The cumulative (expected) regret of policy π after n rounds is defined by

$$R_n \;=\; \underbrace{n\,R_{\max}}_{\substack{\text{reward if}\\ \text{always played an}\\ \text{optimal arm}}} \;-\; \mathbb{E}\Big[\sum_{t=1}^{n} X_t\Big].$$

Since $X_t \sim \nu_{A_t}$ and $\mathbb{E}[X_t \mid A_t] = R_{A_t}$, we have

$$\mathbb{E}[\,X_t\,] = \mathbb{E}\big[\mathbb{E}[X_t \mid A_t]\big] = \mathbb{E}\big[R_{A_t}\big].$$

Hence

$$\mathbb{E}\Big[\sum_{t=1}^{n} X_t\Big] = \sum_{t=1}^{n} \mathbb{E}[\,X_t\,] = \sum_{t=1}^{n} \mathbb{E}\big[R_{A_t}\big] = \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} R_a \,\Pr(A_t = a) = \sum_{a \in \mathcal{A}} R_a \,\mathbb{E}\big[T_a(n)\big].$$

Therefore,

$$R_n = n\,R_{\max} \;-\; \sum_{a \in \mathcal{A}} R_a \,\mathbb{E}\big[T_a(n)\big].$$

Algebraic Rearrangement. Since $\sum_{a\in\mathcal{A}} \mathbb{E}[T_a(n)] = n$, we can rewrite

$$n\,R_{\max} = \sum_{a\in\mathcal{A}} \Big(R_{\max}\,\mathbb{E}[T_a(n)]\Big) = \sum_{a\in\mathcal{A}} \Big[(R_{\max} - R_a) + R_a\Big]\,\mathbb{E}[T_a(n)].$$

Splitting the sum gives

$$n\,R_{\max} = \sum_{a\in\mathcal{A}} (R_{\max} - R_a)\,\mathbb{E}[T_a(n)] \;+\; \sum_{a\in\mathcal{A}} R_a\,\mathbb{E}[T_a(n)].$$

Subtracting $\sum_a R_a\,\mathbb{E}[T_a(n)]$ from both sides yields

$$n\,R_{\max} \;-\; \sum_{a\in\mathcal{A}} R_a\,\mathbb{E}[T_a(n)] \;=\; \sum_{a\in\mathcal{A}} (R_{\max} - R_a)\,\mathbb{E}[T_a(n)].$$

By definition, $R_n = n\,R_{\max} - \sum_a R_a\,\mathbb{E}[T_a(n)]$. Hence

$$\boxed{R_n = \sum_{a\in\mathcal{A}} \Delta_a\,\mathbb{E}[T_a(n)],}$$

where $\Delta_a = R_{\max} - R_a$. This completes the proof of the regret decomposition lemma.

2.1.2 b)[4-points]

In this subsection, we show why a *fixed* confidence parameter $\delta = C$ can cause UCB to incur $\Omega(n)$ regret, and how to choose $\delta = \delta(n)$ so that the "bad event" probability vanishes.

1. The problem with $\delta = C$ fixed. Recall that UCB's index for arm i at time t (after $T_i(t)$ pulls) is

$$\mathrm{UCB}_i(t,\delta) = \begin{cases} +\infty, & T_i(t) = 0, \\ \hat{\mu}_i(t) \;+\; \sqrt{\dfrac{2\,\ln(1/\delta)}{T_i(t)}}, & T_i(t) > 0. \end{cases}$$

If $\delta = C$ is a positive constant (e.g. 0.1), then $\ln(1/\delta)$ is also a fixed constant. Consequently:

- The "bonus term" $\sqrt{2\,\ln(1/\delta)\,/\,T_i(t)}$ remains of order $\sqrt{1/T_i(t)}$, but never grows with t.
- There is a nonzero probability—on the order of δ—that the true mean of the optimal arm is ever underestimated or a suboptimal arm's mean is overestimated by more than that fixed confidence width.
- Such a single estimation error can cause UCB to keep pulling a suboptimal arm i for $\Theta(n)$ rounds, incurring Δ_i loss each time, and producing $\Omega(n)$ total regret.

Thus, with $\delta \equiv C$, the "bad event" has constant probability, and the expected regret is $\Omega(n)$.

2. Defining a "bad event" B_i for each suboptimal arm i. Assume arm 1 is the unique optimal arm with true mean μ_1, and some suboptimal arm $i \neq 1$ has mean $\mu_i < \mu_1$. Fix a threshold $u_i \in \mathbb{N}$ (to be chosen later). Define

$$B_i = \Big\{\underbrace{\exists\, t \le n :\ \mathrm{UCB}_1(t-1,\delta) < \mu_1}_{\text{(A) optimal-arm underestimation}} \;\;\cup\;\; \underbrace{\hat{\mu}_{i,u_i} \;+\; \sqrt{\frac{2\,\ln(1/\delta)}{u_i}} \;\ge\; \mu_1}_{\text{(B) suboptimal-arm } i \text{ overestimation}}\Big\}.$$

Here:

(A) At some time $t \le n$, the UCB index of arm 1 drops below its true mean μ_1. In that round, UCB will not choose arm 1.

(B) After u_i draws of arm i, the index $\hat{\mu}_{i,u_i} + \sqrt{2\,\ln(1/\delta)/\,u_i}$ exceeds μ_1. Then UCB may continue selecting arm i instead of arm 1 for many future rounds.

If B_i occurs with constant probability (when δ is fixed), UCB can follow arm i for $\Theta(n)$ steps and suffer Δ_i loss each time, yielding $[R_n] = \Omega(n)$.

3. How to make $\Pr(B_i) \to 0$ by choosing $\delta = \delta(n)$. We want both types of estimation-error events to become extremely unlikely as $n \to \infty$. A standard choice is

$$\delta(n) \;=\; \frac{1}{n^2} \quad \text{(or more generally, any polynomially small rate } \propto n^{-\alpha} \text{ with } \alpha > 1).$$

Then $\ln\big(1/\delta(n)\big) = 2\,\ln(n)$. We check each failure mode:

(A') *Underestimation of arm 1.* For any fixed time t, Hoeffding's inequality (or the two-sided sub-Gaussian tail bound) gives

$$\Pr\Big(\mathrm{UCB}_1(t-1,\delta(n)) < \mu_1\Big) \;=\; \Pr\Big(\hat{\mu}_1(t-1) + \sqrt{\tfrac{2\,\ln(1/\delta(n))}{T_1(t-1)}} < \mu_1\Big) \;\le\; \exp\Big(-\,T_1(t-1)\,\ln(n)\Big).$$

Since $T_1(t-1) \ge 1$, each term is at most n^{-1}. Taking a union bound over $t = 1,\dots,n$,

$$\Pr\Big(\exists\, t \le n:\; \mathrm{UCB}_1(t-1,\delta(n)) < \mu_1\Big) \;\le\; \sum_{t=1}^{n} n^{-1} \;=\; \frac{n}{n} = 1,$$

which is too crude. Instead, note that once $T_1(t-1)$ grows—a typical pull-count for the optimal arm is $\Omega(\ln(n))$ by standard UCB analysis—so each term becomes $n^{-c\,\ln(n)} = n^{-c\ln n} \ll 1/n$. More precisely, one can show

$$\Pr\Big(\exists\, t \le n:\; \mathrm{UCB}_1(t-1,\delta(n)) < \mu_1\Big) \;=\; O\big(n^{-c'}\big) \;\longrightarrow\; 0 \quad \text{as } n \to \infty.$$

(B') *Overestimation of arm i at time u_i.* After u_i pulls of arm i,

$$\Pr\Big(\hat{\mu}_{i,u_i} + \sqrt{\tfrac{2\,\ln(1/\delta(n))}{u_i}} \;\ge\; \mu_1\Big) \;\le\; \exp\Big(-\,u_i\,\ln(n)\Big) \;=\; n^{-\,u_i}.$$

For any fixed choice of $u_i \ge 2$, this probability is $O(n^{-2}) \to 0$.

Since both "bad subevents" occur with probability $o(1)$, a union bound over all K arms shows

$$\Pr\Big(\exists\, i \in [K]:\; B_i\Big) \;\le\; \sum_{i=1}^{K} \Pr(B_i) \;=\; O\big(K\,n^{-c''}\big) \;\longrightarrow\; 0 \quad \text{as } n \to \infty.$$

Thus, with $\delta(n) = 1/n^2$, UCB's confidence intervals simultaneously hold for all arms at all times $t \le n$ with probability tending to 1. On that *good event*, UCB will pull each suboptimal arm i at most $O\big(\ln n/\Delta_i^2\big)$ times, yielding the usual logarithmic regret.

Conclusion: If δ is held fixed ($\delta = C$), then with nonzero constant probability UCB's index misestimates an arm's mean and incurs $\Omega(n)$ regret. By choosing

$$\delta(n) \;=\; \frac{1}{n^2} \quad \text{(or any polynomially decaying rate)},$$

the probability of either underestimating the optimal arm or overestimating a suboptimal arm decays to zero as $n \to \infty$. Therefore, UCB's regret becomes $O\big(\sum_{i:\Delta_i>0}(\ln n)/\Delta_i\big)$ rather than $\Omega(n)$.

2.1.3 c)[4-points]

Setup. Recall that we number rounds $t = 1, 2, \ldots, n$, and at each round t the UCB algorithm selects

$$A_t \;=\; \arg\max_{j \in [K]} \mathrm{UCB}_j\big(t-1, \delta\big),$$

where

$$\mathrm{UCB}_j(t-1,\delta) = \begin{cases} +\infty, & T_j(t-1) = 0, \\ \hat{\mu}_j(t-1) \;+\; \sqrt{\dfrac{2\,\ln(1/\delta)}{T_j(t-1)}}, & T_j(t-1) > 0. \end{cases}$$

We assume arm 1 is optimal with true mean μ_1, and arm $i \neq 1$ is suboptimal. Fix a threshold $u_i \in \mathbb{N}$ and define the "good event" G_i by

$$G_i = \Big\{\forall\, t \in [n]:\; \mathrm{UCB}_1(t,\delta) > \mu_1\Big\} \,\cap\, \Big\{\hat{\mu}_{i,u_i} \;+\; \sqrt{\frac{2\,\ln(1/\delta)}{u_i}} \;<\; \mu_1\Big\}.$$

Claim. On the event G_i, the total number of pulls of arm i satisfies

$$T_i(n) \;\leq\; u_i.$$

Proof. Suppose, for the sake of contradiction, that on G_i the algorithm pulls arm i more than u_i times. Then there must be a first round $t_0 \leq n$ at which

$$T_i(t_0 - 1) = u_i \quad \text{and} \quad A_{t_0} = i.$$

At that round, the UCB index for arm i is

$$\mathrm{UCB}_i(t_0-1,\delta) = \hat{\mu}_{i,u_i} \;+\; \sqrt{\frac{2\,\ln(1/\delta)}{u_i}} \quad \big(\text{since } T_i(t_0-1) = u_i\big).$$

By the second part of G_i, this is strictly less than μ_1:

$$\mathrm{UCB}_i(t_0-1,\delta) < \mu_1.$$

Meanwhile, by the first part of G_i, for *every* time $t \leq n$,

$$\mathrm{UCB}_1(t-1,\delta) > \mu_1.$$

In particular,

$$\mathrm{UCB}_1(t_0-1,\delta) > \mu_1.$$

Therefore at round t_0,

$$\mathrm{UCB}_1(t_0-1,\delta) \;>\; \mu_1 \;>\; \mathrm{UCB}_i(t_0-1,\delta).$$

But the UCB algorithm chooses $A_{t_0} = \arg\max_j \mathrm{UCB}_j(t_0-1,\delta)$, so it should not select arm i when its index is strictly below that of arm 1. This contradiction shows that no such t_0 can exist on G_i. Hence

$$T_i(n) \;\leq\; u_i \quad \text{whenever } G_i \text{ holds},$$

as claimed.

2.1.4 d)[4-points]

We decompose $T_i(n)$ according to the good event G_i and its complement:

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

Taking expectations yields

$$\mathbb{E}[T_i(n)] = \mathbb{E}\big[T_i(n)\,\mathbf{1}_{G_i}\big] \;+\; \mathbb{E}\big[T_i(n)\,\mathbf{1}_{G_i^c}\big].$$

On the event G_i, we have $T_i(n) \le u_i$, so

$$\mathbb{E}\big[T_i(n)\,\mathbf{1}_{G_i}\big] \;\le\; u_i\,\Pr(G_i) \;\le\; u_i.$$

On the complement G_i^c, trivially $T_i(n) \le n$, giving

$$\mathbb{E}\big[T_i(n)\,\mathbf{1}_{G_i^c}\big] \;\le\; n\,\Pr(G_i^c).$$

Combining these two bounds, we conclude

$$\boxed{\mathbb{E}[T_i(n)] \;\le\; u_i \;+\; n\,\Pr\big(G_i^c\big).}$$

2.1.5 e)[6-points]

Assume we choose u_i large enough so that

$$\Delta_i \;-\; \sqrt{\frac{2\ln(1/\delta)}{u_i}} \;\ge\; c\,\Delta_i,$$

for some constant $c \in (0, 1)$. Recall $\Delta_i = \mu_1 - \mu_i$. We wish to show

$$\Pr\Big(\hat{\mu}_{i,u_i} \;+\; \sqrt{\frac{2\ln(1/\delta)}{u_i}} \;\ge\; \mu_1\Big) \;\le\; \exp\Big(-\,\frac{u_i\,c^2\,\Delta_i^2}{2}\Big).$$

Proof. On the event in question,

$$\hat{\mu}_{i,u_i} \;+\; \sqrt{\frac{2\ln(1/\delta)}{u_i}} \;\ge\; \mu_1 \implies \hat{\mu}_{i,u_i} - \mu_i \;\ge\; (\mu_1 - \mu_i) \;-\; \sqrt{\frac{2\ln(1/\delta)}{u_i}} \;\ge\; \Delta_i \;-\; \sqrt{\frac{2\ln(1/\delta)}{u_i}} \;\ge\; c\,\Delta_i.$$

Hence

$$\Pr\Big(\hat{\mu}_{i,u_i} \;+\; \sqrt{\tfrac{2\ln(1/\delta)}{u_i}} \;\ge\; \mu_1\Big) \;\le\; \Pr\big(\hat{\mu}_{i,u_i} - \mu_i \;\ge\; c\,\Delta_i\big).$$

Since each reward is 1-sub-Gaussian, by the tail bound for the sample mean,

$$\Pr\big(\hat{\mu}_{i,u_i} - \mu_i \ge t\big) \;\le\; \exp\Big(-\,\frac{u_i\,t^2}{2}\Big) \quad \text{for any } t \ge 0.$$

Setting $t = c\,\Delta_i$ gives

$$\Pr\big(\hat{\mu}_{i,u_i} - \mu_i \ge c\,\Delta_i\big) \;\le\; \exp\Big(-\,\frac{u_i\,c^2\,\Delta_i^2}{2}\Big).$$

Combining these inequalities yields the desired bound:

$$\Pr\Big(\hat{\mu}_{i,u_i} \;+\; \sqrt{\tfrac{2\ln(1/\delta)}{u_i}} \;\ge\; \mu_1\Big) \;\le\; \exp\Big(-\,\frac{u_i\,c^2\,\Delta_i^2}{2}\Big).$$

2.1.6 f)[4-points]

Recall

$$G_i = \Big\{\forall t \in [n] :\ \mathrm{UCB}_1(t,\delta) > \mu_1\Big\} \cap \Big\{\hat{\mu}_{i,u_i} + \sqrt{\frac{2\ln(1/\delta)}{u_i}} < \mu_1\Big\}.$$

Hence its complement is

$$G_i^c = \Big\{\exists t \in [n] :\ \mathrm{UCB}_1(t,\delta) \le \mu_1\Big\} \cup \Big\{\hat{\mu}_{i,u_i} + \sqrt{\frac{2\ln(1/\delta)}{u_i}} \ge \mu_1\Big\}.$$

By the union bound,

$$\Pr(G_i^c) \le \Pr\Big(\exists t \in [n] :\ \mathrm{UCB}_1(t,\delta) \le \mu_1\Big) + \Pr\Big(\hat{\mu}_{i,u_i} + \sqrt{\tfrac{2\ln(1/\delta)}{u_i}} \ge \mu_1\Big).$$

(i) Underestimation of arm 1. For each fixed t, by Hoeffding's (sub-Gaussian) tail bound,

$$\Pr\big(\mathrm{UCB}_1(t,\delta) \le \mu_1\big) = \Pr\Big(\hat{\mu}_1(t) + \sqrt{\tfrac{2\ln(1/\delta)}{T_1(t)}} \le \mu_1\Big) \le \exp\big(-\ln(1/\delta)\big) = \delta.$$

Applying a union bound over $t = 1, \dots, n$ gives

$$\Pr\Big(\exists t \in [n] :\ \mathrm{UCB}_1(t,\delta) \le \mu_1\Big) \le \sum_{t=1}^{n} \delta = n\,\delta.$$

(ii) Overestimation of arm i. From part (e), assuming $\Delta_i - \sqrt{\frac{2\ln(1/\delta)}{u_i}} \ge c\,\Delta_i$, we have

$$\Pr\Big(\hat{\mu}_{i,u_i} + \sqrt{\tfrac{2\ln(1/\delta)}{u_i}} \ge \mu_1\Big) \le \exp\Big(-\tfrac{u_i\,c^2\,\Delta_i^2}{2}\Big).$$

Conclusion. Combining the two bounds,

$$\boxed{\Pr(G_i^c) \le n\,\delta + \exp\Big(-\frac{u_i\,c^2\,\Delta_i^2}{2}\Big).}$$

2.1.7 g)[6-points]

We combine the bounds from parts (d) and (f):

$$\mathbb{E}[T_i(n)] \le u_i + n\,\Pr(G_i^c) \le u_i + n\Big(n\,\delta + \exp\big(-\tfrac{u_i\,c^2\,\Delta_i^2}{2}\big)\Big).$$

We now make the following standard choices:

- $\delta = \dfrac{1}{n^2}$, so that $n\,\delta = \frac{1}{n}$.
- $c = \frac{1}{2}$.

- $$u_i = \left\lceil \frac{4\,\ln(1/\delta)}{(1-c)^2\,\Delta_i^2} \right\rceil = \left\lceil \frac{4\,(2\ln n)}{(1/2)^2\,\Delta_i^2} \right\rceil = \left\lceil \frac{16\,\ln n}{\Delta_i^2} \right\rceil.$$

  Then
  $$\sqrt{\frac{2\ln(1/\delta)}{u_i}} = \sqrt{\frac{4\ln n}{u_i}} \;\le\; (1-c)\,\Delta_i = \tfrac{1}{2}\,\Delta_i,$$
  so the condition of part (e) is satisfied.

With these choices,

$$\exp\!\Bigl(-\tfrac{u_i\,c^2\,\Delta_i^2}{2}\Bigr) \;\le\; \exp\!\Bigl(-\frac{\bigl(\frac{16\ln n}{\Delta_i^2}\bigr)\,(1/2)^2\,\Delta_i^2}{2}\Bigr) = \exp(-2\ln n) = n^{-2}.$$

Hence
$$n\;\Pr(G_i^c) \;\le\; n\Bigl(\tfrac{1}{n} + n^{-2}\Bigr) \;=\; 1 + \tfrac{1}{n} \;\le\; 2 \quad (\text{for } n \ge 1).$$

Plugging back into the expectation bound,

$$\mathbb{E}[T_i(n)] \;\le\; u_i + 2 \;\le\; \left\lceil \tfrac{16\,\ln n}{\Delta_i^2} \right\rceil + 2 \;\le\; 3 + \frac{16\,\ln n}{\Delta_i^2}.$$

This completes the proof that
$$\boxed{\mathbb{E}[T_i(n)] \;\le\; 3 \;+\; \frac{16\,\ln(n)}{\Delta_i^2}.}$$

2.1.8 h)[5-points]

Using the regret decomposition
$$R_n \;=\; \sum_{i=1}^{K} \Delta_i\,\mathbb{E}[T_i(n)],$$
and the bound
$$\mathbb{E}[T_i(n)] \;\le\; 3 \;+\; \frac{16\,\ln(n)}{\Delta_i^2},$$
we obtain
$$R_n \;\le\; \sum_{i=1}^{K} \Delta_i\Bigl(3 + \frac{16\,\ln(n)}{\Delta_i^2}\Bigr) = 3\sum_{i=1}^{K} \Delta_i \;+\; 16\,\ln(n)\sum_{i=1}^{K} \frac{\Delta_i}{\Delta_i^2}.$$

Since $\Delta_i = 0$ contributes zero to regret, we may restrict the second sum to $\{i : \Delta_i > 0\}$:

$$R_n \;\le\; 3\sum_{i=1}^{K} \Delta_i \;+\; 16\,\ln(n) \sum_{i:\Delta_i>0} \frac{1}{\Delta_i}.$$

Thus, with $\delta = 1/n^2$,
$$\boxed{R_n \;\le\; 3\sum_{i=1}^{K} \Delta_i \;+\; \sum_{i:\Delta_i>0} \frac{16\,\ln(n)}{\Delta_i}.}$$

2.1.9 i)[5-points]

Starting from the bound in part (h),

$$R_n \;\le\; 3\sum_{i=1}^{K}\Delta_i \;+\; 16\,\ln(n)\sum_{i:\,\Delta_i>0}\frac{1}{\Delta_i},$$

we now show how to turn the second term into $O(\sqrt{nk\ln n})$ by splitting the arms at a threshold $\Delta > 0$.

1. Split the arms at threshold Δ. Fix some $\Delta > 0$. Partition the set of suboptimal arms $\{i : \Delta_i > 0\}$ into

$$\mathcal{I}_{\le\Delta} = \{\,i:\; 0 < \Delta_i \le \Delta\} \quad\text{and}\quad \mathcal{I}_{>\Delta} = \{\,i:\; \Delta_i > \Delta\}.$$

Then

$$\sum_{i:\Delta_i>0}\frac{1}{\Delta_i} = \sum_{i\in\mathcal{I}_{\le\Delta}}\frac{1}{\Delta_i} \;+\; \sum_{i\in\mathcal{I}_{>\Delta}}\frac{1}{\Delta_i}.$$

2. Bound each part.

- For $i \in \mathcal{I}_{\le\Delta}$, we have $\Delta_i \le \Delta$, so $1/\Delta_i \ge 1/\Delta$. Hence

$$\sum_{i\in\mathcal{I}_{\le\Delta}}\frac{1}{\Delta_i} \;\le\; |\mathcal{I}_{\le\Delta}|\,\frac{1}{\Delta} \;\le\; K\,\frac{1}{\Delta}.$$

- For $i \in \mathcal{I}_{>\Delta}$, we have $\Delta_i > \Delta$, so $1/\Delta_i < 1/\Delta$. Moreover, there are at most K arms total, so

$$\sum_{i\in\mathcal{I}_{>\Delta}}\frac{1}{\Delta_i} \;\le\; \sum_{i\in\mathcal{I}_{>\Delta}}\frac{1}{\Delta} \;\le\; K\,\frac{1}{\Delta}.$$

3. Optimize Δ. Combining,

$$\sum_{i:\Delta_i>0}\frac{1}{\Delta_i} \;\le\; \frac{K}{\Delta} \;+\; \frac{K}{\Delta} \;=\; \frac{2K}{\Delta}.$$

Therefore

$$R_n \;\le\; 3\sum_{i=1}^{K}\Delta_i \;+\; 16\,\ln(n)\,\frac{2K}{\Delta} \;=\; 3\sum_{i=1}^{K}\Delta_i \;+\; \frac{32\,K\,\ln(n)}{\Delta}.$$

We now choose Δ to balance the two terms. A natural choice is

$$\Delta \;=\; \sqrt{\frac{32\,K\,\ln(n)}{n}} \;\approx\; \sqrt{\frac{K\,\ln n}{n}} \times \sqrt{32}.$$

With this choice,

$$\frac{32\,K\,\ln(n)}{\Delta} = 32\,K\,\ln(n) \Big/ \sqrt{\frac{32\,K\,\ln(n)}{n}} = \sqrt{32\,K\,n\,\ln(n)} = 4\sqrt{2\,K\,n\,\ln(n)} \le 8\,\sqrt{K\,n\,\ln(n)}.$$

Hence

$$R_n \;\le\; 3\sum_{i=1}^{K}\Delta_i \;+\; 8\,\sqrt{K\,n\,\ln(n)}.$$

Since the first term $3\sum_i \Delta_i$ is independent of n and at most $3\sum_i \Delta_i$, we conclude

$$\boxed{R_n \;\leq\; 8\sqrt{n\,K\,\ln(n)} \;+\; 3\sum_{i=1}^{K} \Delta_i.}$$

3 Online Learning[50-points]

3.1 Randomized Weighted Majority Algorithm[35-points]

3.1.1 a)[5-points]

At round t, let

$$S_t = \sum_{i=1}^{N} w_i(t)$$

be the total weight. We choose expert i_t with probability $\frac{w_{i_t}(t)}{S_t}$, and upon observing the outcome we update that expert's weight by multiplying by $(1-\epsilon)$ if it was wrong, or leaving it unchanged if it was correct. All other experts' weights stay the same.

We want to compute $\mathbb{E}\big[S_{t+1}\big]$. Note that

$$S_{t+1} = \sum_{i \neq i_t} w_i(t) + \begin{cases} w_{i_t}(t), & \text{if expert } i_t \text{ was correct,} \\ w_{i_t}(t)\,(1-\epsilon), & \text{if expert } i_t \text{ was wrong.} \end{cases}$$

Therefore, conditioning on which expert is chosen and whether it errs,

$$\mathbb{E}\big[S_{t+1} \mid \{w_i(t)\}\big] = \sum_{i=1}^{N} \frac{w_i(t)}{S_t}\Big[w_i(t)\,(1-\epsilon\cdot\mathbf{1}\{i \text{ wrong}\}) + \sum_{j\neq i} w_j(t)\Big].$$

We can simplify this by observing that for each chosen i, $\sum_{j\neq i} w_j(t) = S_t - w_i(t)$. Hence

$$\mathbb{E}\big[S_{t+1} \mid \{w_i(t)\}\big] = \sum_{i=1}^{N} \frac{w_i(t)}{S_t}\big[\,S_t - \epsilon\,w_i(t)\,\mathbf{1}\{i \text{ wrong}\}\big] = S_t - \epsilon\sum_{i=1}^{N} \frac{w_i(t)^2}{S_t}\,\mathbf{1}\{i \text{ wrong}\}.$$

Finally, note that $P(\tilde{m}_t = 1) = \sum_{i \text{ wrong}} \frac{w_i(t)}{S_t}$, so $\sum_{i \text{ wrong}} w_i(t) = S_t \cdot P(\tilde{m}_t = 1)$. Since $w_i(t) \le S_t$, an upper-bound calculation (or exact calculation when one expert is chosen) yields

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

Taking expectation over the randomness up to time t gives the desired result:

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

3.1.2 b)[8-points]

Starting from the recurrence we derived,

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

and noting that $S_1 = \sum_{i=1}^{N} w_i(0) = N$, we can unroll this over T rounds:

$$\begin{aligned}
\mathbb{E}[S_{T+1}] &= N \prod_{t=1}^{T}\big(1 - \epsilon\, P(\tilde{m}_t = 1)\big) \\
&\le N \prod_{t=1}^{T} \exp\big(-\epsilon\, P(\tilde{m}_t = 1)\big) \quad \big[\text{since } 1 - x \le e^{-x}\big] \\
&= N \exp\Big(-\epsilon \sum_{t=1}^{T} P(\tilde{m}_t = 1)\Big).
\end{aligned}$$

Thus we obtain the stated bound:

$$\boxed{\mathbb{E}[S_{T+1}] \;\le\; N \exp\Big(-\epsilon \sum_{t=1}^{T} P(\tilde{m}_t = 1)\Big).}$$

3.1.3 c)[15-points]

Let $M = \sum_{t=1}^{T} \mathbf{1}\{\tilde{m}_t = 1\}$ be the total number of mistakes the algorithm makes in T rounds, so

$$\mathbb{E}[M] = \sum_{t=1}^{T} P(\tilde{m}_t = 1).$$

Also, for any fixed expert i, let M_i be the number of mistakes expert i makes over those T rounds. Since $w_i(0) = 1$, after T rounds the weight of expert i is

$$w_i(T+1) = (1-\epsilon)^{M_i}.$$

Because $w_i(T+1) \le S_{T+1}$, taking expectations gives

$$\mathbb{E}[S_{T+1}] \;\ge\; \mathbb{E}[w_i(T+1)] = (1-\epsilon)^{M_i}.$$

On the other hand, from part (b) we have

$$\mathbb{E}[S_{T+1}] \;\le\; N \exp\big(-\epsilon\, \mathbb{E}[M]\big).$$

Combining the two bounds,

$$N \exp\big(-\epsilon\, \mathbb{E}[M]\big) \;\ge\; (1-\epsilon)^{M_i}.$$

Taking natural logarithms,

$$\ln N \;-\; \epsilon\, \mathbb{E}[M] \;\ge\; M_i \ln(1-\epsilon).$$

Rearrange to solve for $\mathbb{E}[M]$:

$$\epsilon\, \mathbb{E}[M] \;\le\; \ln N \;-\; M_i \ln(1-\epsilon),$$

$$\mathbb{E}[M] \;\le\; \frac{\ln N}{\epsilon} \;-\; \frac{M_i}{\epsilon}\,\ln(1-\epsilon).$$

Finally, using the inequality $-\ln(1-\epsilon) \le \epsilon + \epsilon^2 \le \epsilon(1+\epsilon)$ for $0 < \epsilon < 1$, we obtain

$$\mathbb{E}[M] \;\le\; \frac{\ln N}{\epsilon} \;+\; M_i\,(1+\epsilon) = \;(1+\epsilon)\,M_i \;+\; \frac{\ln N}{\epsilon},$$

which holds for every expert $i \in [N]$. $\square$

3.1.4 d)[7-points]

We start from the bound obtained in part (c):

$$\mathbb{E}[M] \;\le\; (1+\epsilon)\,M_i \;+\; \frac{\ln N}{\epsilon} \quad \forall i \in [N].$$

We now choose ϵ to minimize the right-hand side as a function of ϵ. A common choice is

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

where T is the total number of rounds. Plugging in:

1. $(1+\epsilon)\,M_i = M_i + \epsilon\,M_i \le M_i + \epsilon\,T = M_i + \sqrt{T\ln N}$. 2. $\dfrac{\ln N}{\epsilon} = \dfrac{\ln N}{\sqrt{\ln N/T}} = \sqrt{T\ln N}$.

Hence for every expert i,

$$\mathbb{E}[M] \;\le\; M_i + \sqrt{T\ln N} \;+\; \sqrt{T\ln N} \;=\; M_i \;+\; 2\sqrt{T\ln N}.$$

Taking the minimum over $i \in [N]$ yields the final bound

$$\boxed{\mathbb{E}[M] \;\le\; \min_{i\in[N]} M_i \;+\; 2\sqrt{T\ln N}.}$$

Relation to Regret and Quality of the Bound Define the *regret* R_T as the difference between the algorithm's mistakes and the best expert's mistakes:

$$R_T \;=\; \mathbb{E}[M] \;-\; \min_i M_i \;\le\; 2\sqrt{T\ln N}.$$

Since $2\sqrt{T\ln N} = o(T)$, the average regret R_T/T goes to zero as $T \to \infty$. This sublinear regret bound is considered *optimal up to constant factors* in the adversarial expert setting, meaning no algorithm can achieve a strictly better dependence on T and N in the worst case. Thus the RWM algorithm attains a good (near-optimal) regret guarantee.

3.2 Hedge Algorithm(Bonus)[15-points]

3.2.1 a)[6-points]

At round t, the total weight is

$$S_t = \sum_{i=1}^{N} w_t(i),$$

and after observing the loss vector ℓ_t we update each expert's weight as

$$w_{t+1}(i) \;=\; w_t(i)\,\exp\bigl(-\epsilon\,\ell_t(i)\bigr).$$

Therefore,

$$S_{t+1} = \sum_{i=1}^{N} w_{t+1}(i) = \sum_{i=1}^{N} w_t(i)\,\exp\bigl(-\epsilon\,\ell_t(i)\bigr) = S_t \sum_{i=1}^{N} p_t(i)\,\exp\bigl(-\epsilon\,\ell_t(i)\bigr),$$

where $p_t(i) = w_t(i)/S_t$.

Next, we use the inequality

$$e^{-x} \le 1 - x + x^2 \quad \text{for all } x,$$

with $x = \epsilon\,\ell_t(i)$. Since $\ell_t(i) \in [-1,1]$ and $\epsilon > 0$, we have

$$\exp\bigl(-\epsilon\,\ell_t(i)\bigr) \le 1 \;-\; \epsilon\,\ell_t(i) \;+\; \epsilon^2\,\ell_t(i)^2.$$

Substituting into the expression for S_{t+1},

$$\begin{aligned} S_{t+1} &\le S_t \sum_{i=1}^{N} p_t(i)\Bigl(1 - \epsilon\,\ell_t(i) + \epsilon^2\,\ell_t(i)^2\Bigr) \\ &= S_t\Bigl(1 - \epsilon\sum_{i=1}^{N} p_t(i)\,\ell_t(i) + \epsilon^2 \sum_{i=1}^{N} p_t(i)\,\ell_t(i)^2\Bigr). \end{aligned}$$

This is the desired bound:

$$\boxed{S_{t+1} \;\le\; S_t\Bigl(1 - \epsilon\sum_{i} p_t(i)\,\ell_t(i) + \epsilon^2 \sum_{i} p_t(i)\,\ell_t(i)^2\Bigr).}$$

3.2.2 b)[7-points]

Recall the regret bound for Hedge:

$$\text{Regret}_{\text{Hedge}} \;\le\; \frac{\ln N}{\epsilon} \;+\; \epsilon \sum_{t=1}^{T}\sum_{i=1}^{N} p_t(i)\,\ell_t(i)^2,$$

where each $\ell_t(i) \in [-1,1]$. In particular, since $\ell_t(i)^2 \le 1$ and $\sum_i p_t(i) = 1$, we have

$$\sum_{i=1}^{N} p_t(i)\,\ell_t(i)^2 \;\le\; 1,$$

so

$$\text{Regret}_{\text{Hedge}} \le \frac{\ln N}{\epsilon} + \epsilon T.$$

Optimizing ϵ by setting $\epsilon = \sqrt{\ln N/T}$ yields

$$\text{Regret}_{\text{Hedge}} \le 2\sqrt{T \ln N}.$$

By contrast, the Randomized Weighted Majority (RWM) algorithm satisfies

$$\text{Regret}_{\text{RWM}} \le 2\sqrt{T \ln N}$$

as shown in part (d).

Key observations

- Both algorithms achieve the same $\mathcal{O}(\sqrt{T \ln N})$ worst-case regret.
- Hedge's bound is stated in terms of arbitrary real-valued losses $\ell_t(i) \in [-1, 1]$, while RWM was originally derived for binary mistakes $\{0, 1\}$.
- The Hedge analysis refines the intermediate bound by tracking $\sum p_t \ell_t^2$; if losses are small in magnitude, one can get tighter guarantees.
- RWM can be seen as a special case of Hedge with binary losses and a simpler update.

In summary, both algorithms are essentially equivalent in terms of asymptotic regret in the adversarial setting, but Hedge offers greater flexibility for general loss ranges.

3.2.3 c)[2-points]

Starting from the bound

$$\text{Regret} \le \frac{\ln N}{\epsilon} + \epsilon T,$$

we choose $\epsilon = \sqrt{\frac{\ln N}{T}}$ to balance the two terms. Substituting:

$$\frac{\ln N}{\epsilon} = \frac{\ln N}{\sqrt{\ln N/T}} = \sqrt{T \ln N}, \qquad \epsilon T = \sqrt{\frac{\ln N}{T}}\, T = \sqrt{T \ln N}.$$

Hence

$$\text{Regret} \le \sqrt{T \ln N} + \sqrt{T \ln N} = 2\sqrt{T \ln N}.$$

Thus we obtain the compact form:

$$\boxed{\text{Regret} \le 2\sqrt{T \ln N}.}$$

Comment This matches the adversarial expert regret for RWM. Both Hedge (for real-valued losses in $[-1, 1]$) and RWM (for binary mistakes) achieve the optimal $\mathcal{O}(\sqrt{T \ln N})$ regret up to constant factors.