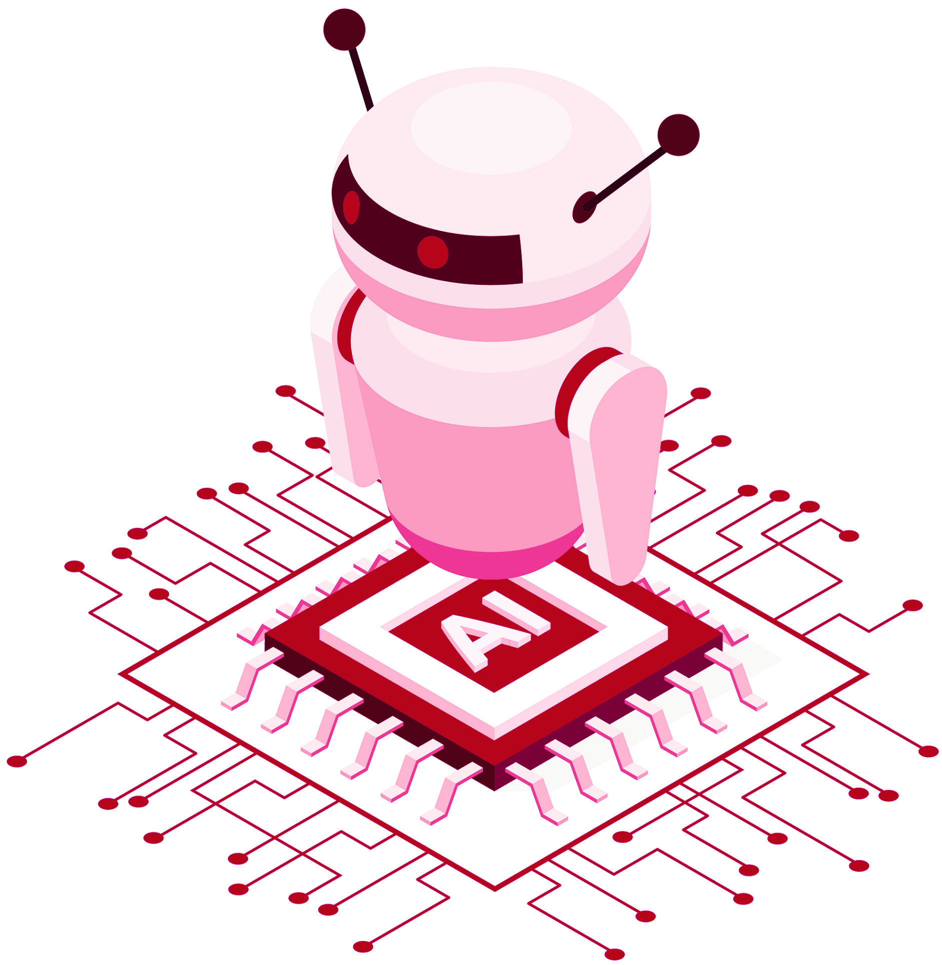

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution forHomework [9]

Advanced Theory

By:

Amir Kooshan Fattah Hesari

401102191

RIML

Spring 2025

Contents

1 Light-tailed Distributions[25-points]

1.1 Hoeffding's Inequality[10-points]

1.1.1 a)[6-points]

If we have a random variable X with $\mathbf{E}[X] = 0$ and we know that $a \leq X \leq b$, we will prove the desired relation using the definition below.
First, we define the function $\phi(s)$ as follows:

$$\phi(s) = \log \mathbf{E}\left[e^{sX}\right]$$

$$\phi'(s) = \frac{\partial \log \mathbf{E}\left[e^{sX}\right]}{\partial s} = \frac{\frac{\partial \mathbf{E}\left[e^{sX}\right]}{\partial s}}{\mathbf{E}\left[e^{sX}\right]} = \frac{\mathbf{E}\left[X \cdot e^{sX}\right]}{\mathbf{E}\left[e^{sX}\right]} = \int X \cdot \underbrace{\frac{e^{sX} d\mathbf{P}(X)}{\mathbf{E}\left[e^{sX}\right]}}_{d\mathbf{P_s}(X)}$$

$$= \mathbf{E}_{x\sim\mathbf{P}_s(X)}\left[X\right]$$

$$\phi''(s) = \frac{\partial}{\partial s}\mathbf{E}_{x\sim\mathbf{P_s}(X)}\left[X\right] = \frac{\mathbf{E}\left[X^2 \cdot e^{sX}\right]\mathbf{E}\left[e^{sX}\right] - \left(\mathbf{E}\left[X \cdot e^{sX}\right]\right)^2}{(\mathbf{E}\left[e^{sX}\right])^2}$$

$$= \frac{\mathbf{E}\left[X^2 \cdot e^{sX}\right]}{\mathbf{E}\left[e^{sX}\right]} - \left(\frac{\mathbf{E}\left[X \cdot e^{sX}\right]}{\mathbf{E}\left[e^{sX}\right]}\right)^2$$

$$\phi''(s) = \mathbf{E}_{x\sim\mathbf{P_s}(X)}\left[X^2\right] - \left(\mathbf{E}_{x\sim\mathbf{P_s}(X)}\left[X\right]\right)^2 = \mathbb{V}_{x\sim\mathbf{P_s}(X)}\left[X\right]$$

We also know that:

$$\mathbb{V}\left[X\right] = \mathbb{V}\left[X - (\frac{a+b}{2})\right] = \mathbf{E}\left[(X - (\frac{a+b}{2}))^2\right] - \left(\mathbf{E}\left[X - (\frac{a+b}{2})\right]\right)^2$$

$$\leq \mathbf{E}\left[\left(X - (\frac{a+b}{2})\right)^2\right]$$

The final expected value in the above expression reaches its maximum when the random variable X takes one of the extreme values (i.e., either a or b). In that case, we have:

$$X = a \Rightarrow (a - \frac{a+b}{2})^2 = (b - \frac{a+b}{2})^2 = (\frac{b-a}{2})^2$$

$$\mathbf{E}\left[\left(X - \frac{a+b}{2}\right)^2\right] \leq \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{4} \Rightarrow \mathbb{V}\left[X\right] \leq \frac{(b-a)^2}{4}$$

And finally we have: [1]

Hoeffding's Lemma

$$\phi(s) = \int\int \phi''(s) = \int_0^s \int_0^\mu \mathbb{V}_{x\sim \mathbf{P_q}(X)}[X]\, dq\, d\mu \leq \int_0^s \int_0^\mu \frac{(b-a)^2}{4} dq\, d\mu$$
$$= \int_0^s \frac{\mu(b-a)^2}{4} d\mu$$
$$= \frac{s^2(b-a)^2}{8}$$

$$\phi(s) = \log \mathbf{E}\left[e^{sX}\right] \leq \frac{s(b-a)^2}{8} \Rightarrow \mathbf{E}\left[e^{sX}\right] \leq \exp(\frac{s^2(b-a)^2}{8})$$

1.1.2 b)[4-points]

From the definition of subgaussian functions with parameter σ^2, we know that they satisfy the following inequality:

$$\mathbf{E}\left[e^{\lambda x}\right] \leq \exp(\frac{\lambda^2\sigma^2}{2})$$

Now using Markov's inequality and the Cramer–Chernoff method[3], we prove the following relation which will be used in the proof of Hoeffding's inequality:

$$\mathbf{P}(X \geq \epsilon) \leq \exp(\frac{-\epsilon^2}{2\sigma^2})$$

$$\mathbf{P}(X \geq \epsilon) = \mathbf{P}(\lambda X \geq \lambda\epsilon) = \mathbf{P}(\exp(\lambda X) \geq \exp(\lambda\epsilon)) \leq \frac{\mathbf{E}\left[\exp(\lambda X)\right]}{\exp(\lambda\epsilon)} \quad \text{Markov}$$
$$\leq \exp(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon) \quad \text{Def. of subgaussianity}$$

Now if we choose $\lambda = \frac{\epsilon}{\sigma^2}$, then:

$$\exp(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon) = \exp(\frac{\epsilon^2\sigma^2}{2\sigma^4} - \frac{\epsilon^2}{\sigma^2}) = \exp(-\frac{\epsilon^2}{2\sigma^2})$$
$$\Rightarrow \mathbf{P}(X \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}) \qquad \blacksquare$$

Also, if we have independent subgaussian random variables X_i with parameter σ_i, then:

$$q = \sum_i X_i$$

$$\mathbf{E}\left[e^{\lambda q}\right] \overset{\text{independent}}{=\!=} \prod_i \mathbf{E}\left[e^{\lambda X_i}\right] \leq \exp(\frac{\lambda^2 \sum \sigma_i^2}{2}) \Rightarrow q \sim \mathcal{SG}(\sqrt{\sum_i \sigma_i^2})$$

[1] The inequality we used above to find the upper bound for the variance is known as Popoviciu's inequality on variances.
[2] σ-subgaussian
[3] Cramer–Chernoff method

Also, cX is subgaussian with parameter $|c|\sigma$.
Now if the random variables $Z_i \, \forall i \in \{1, 2, ..., n\}$ are bounded and lie within $[a, b]$, then:

$$\mathbf{P}(Z_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$$

From the first part, we know such random variables are subgaussian with parameter $\frac{(b-a)}{2}$. So we have:

Hoeffding's Inequality (Right Tail)

$$\mathbf{P}(Z_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}) = \exp(-\frac{\epsilon^2 \cdot 2}{(b-a)^2})$$

$$\mathbf{P}(\sum_i \overbrace{(Z_i - \mathbf{E}[Z_i])}^{q_i} \geq \epsilon) \;\; = \mathbf{P}(\sum_i q_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2 \cdot (\sqrt{\sum_i \sigma_i^2})^2}) = \exp(-\frac{4 \cdot \epsilon^2}{2 \cdot n(b-a)^2})$$

$$\mathbf{P}(\frac{1}{n}\sum_i q_i \geq \epsilon) \leq \exp(-\frac{n^2 \cdot 4 \cdot \epsilon^2}{2 \cdot n \cdot (b-a)^2}) = \exp(-\frac{n \cdot 2 \cdot \epsilon^2}{(b-a)^2})$$

$$\boxed{\mathbf{P}(\frac{1}{n}\sum_i (Z_i - \mathbf{E}[Z_i]) \geq \epsilon) \leq \exp(-\frac{2 \cdot n \cdot \epsilon^2}{(b-a)^2})}$$

For the left tail of the inequality, we proceed with the following variable change:

$$Y_i = -Z_i \, ; \quad -b \leq Y_i \leq -a \, ; \quad \mathbf{E}[Y_i] = -\mathbf{E}[Z_i]$$

$$\frac{1}{n}\sum_i Z_i - \mathbf{E}[Z_i] \leq -\epsilon \Rightarrow \frac{1}{n}\sum_i Y_i - \mathbf{E}[Y_i] \geq \epsilon$$

$$\mathbf{P}(\frac{1}{n}\sum_i Z_i - \mathbf{E}[Z_i] \leq -\epsilon) = \mathbf{P}(\frac{1}{n}\sum_i Y_i - \mathbf{E}[Y_i] \geq \epsilon) \leq \exp(-\frac{2 \cdot n \cdot \epsilon^2}{(b-a)^2})$$

Hoeffding's Inequality (Left Tail)

$$\mathbf{P}(\frac{1}{n}\sum_i Z_i - \mathbf{E}[Z_i] \leq -\epsilon) \leq \exp(-\frac{2 \cdot n \cdot \epsilon^2}{(b-a)^2})$$

1.2 Sub-Gaussian[15-points]

1.2.1 a-1)[2-points]

We prove the following three inequalities for subgaussian random variables:
For the first inequality, we use the Cramer–Chernoff method, which was also introduced in the previous

section:

$$\mathbf{P}(\underbrace{X - \mathbf{E}[X]}_{Z} \geq \epsilon) \leq \exp(\frac{-\epsilon^2}{2\sigma^2})$$

$$\begin{aligned}\mathbf{P}(Z \geq \epsilon) = \mathbf{P}(\lambda Z \geq \lambda\epsilon) = \mathbf{P}(\exp(\lambda Z) \geq \exp(\lambda\epsilon)) &\leq \frac{\mathbf{E}[\exp(\lambda Z)]}{\exp(\lambda\epsilon)} \quad \text{Markov}\\ &\leq \exp(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon) \quad \text{Def. of subgaussianity}\end{aligned}$$

Right Tail

$$\lambda = \frac{\epsilon}{\sigma^2}\,; \forall\,\epsilon > 0$$
$$\exp(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon) = \exp(\frac{\epsilon^2\sigma^2}{2\sigma^4} - \frac{\epsilon^2}{\sigma^2}) = \exp(-\frac{\epsilon^2}{2\sigma^2})$$
$$\Rightarrow \mathbf{P}(X - \mathbf{E}[X] \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}) \quad \blacksquare$$

1.2.2 a-2)[2-points]

For the second inequality, which concerns the left tail of the distribution, we have:

$$\mathbf{P}(X < \mathbf{E}[X] - \epsilon) = \mathbf{P}(\underbrace{X - \mathbf{E}[X]}_{Z} < -\epsilon) = \mathbf{P}(Z < -\epsilon)$$

$$\begin{aligned}\mathbf{P}(Z < -\epsilon) \xrightarrow{\lambda<0} \mathbf{P}(\lambda Z > -\lambda\epsilon) = \mathbf{P}(\exp(\lambda Z) > \exp(-\lambda\epsilon)) &\leq \frac{\mathbf{E}[\exp(\lambda Z)]}{\exp(-\lambda\epsilon)} \quad \text{(Markov)}\\ \text{(Def. of Subgaussianity)} \quad &\leq \frac{\exp(\frac{\lambda^2\sigma^2}{2})}{\exp(-\lambda\epsilon)} = \exp(\frac{\lambda^2\sigma^2}{2} + \lambda\epsilon)\end{aligned}$$

Left Tail

$$\lambda = \frac{-\epsilon}{\sigma^2};\, \forall\,\epsilon > 0$$
$$\exp\left(\frac{\epsilon^2\sigma^2}{2\sigma^4} - \frac{\epsilon^2}{\sigma^2}\right) = \exp(-\frac{\epsilon^2}{2\sigma^2})$$

$$\Rightarrow \mathbf{P}(Z < -\epsilon) = \mathbf{P}(X < \mathbf{E}[X] - \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}) \quad \blacksquare$$

1.2.3 a-3)[2-points]

For the third inequality, we use the union bound theorem[4]. We have:

$$\mathbf{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbf{P}(A_i)$$

$$\begin{aligned}\mathbf{P}(|X - \mathbf{E}[X]| \geq \epsilon) &= \mathbf{P}\left(X - \mathbf{E}[X] \geq \epsilon \bigcup -X + \mathbf{E}[X] \geq \epsilon\right) \\ &= \mathbf{P}\left(X - \mathbf{E}[X] \geq \epsilon \bigcup X - \mathbf{E}[X] \leq -\epsilon\right)\end{aligned}$$

Union of Two Tails

$$\mathbf{P}\left(X - \mathbf{E}[X] \geq \epsilon \bigcup X - \mathbf{E}[X] \leq -\epsilon\right) \leq \mathbf{P}(X - \mathbf{E}[X] \geq \epsilon) + \mathbf{P}(X - \mathbf{E}[X] \leq -\epsilon)$$
$$\Rightarrow \mathbf{P}(|X - \mathbf{E}[X]| \geq \epsilon) \leq 2\exp(-\frac{\epsilon^2}{2\sigma^2})$$

1.2.4 b)[3-points]

As we saw in the proof of Hoeffding's inequality in the first question, and using the third relation from the first part of this question, we have:

$$\mathbf{P}(|X_i - \mathbf{E}[X_i]| \geq \epsilon) \leq 2\exp(-\frac{\epsilon^2}{2\sigma_i^2})$$

And we know that the sum of n subgaussian variables with parameters σ_i is itself a subgaussian variable with parameter $\sqrt{\sum_i \sigma_i^2}$. Using this fact, we get:

Hoeffding's Inequality – Question 2

$$\mathbf{P}\left(\sum_i |X_i - \mu_i| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_i \sigma_i^2}\right)$$

[4] Union bound

1.2.5 c)[4-points]

If the variables X_i follow a subgaussian distribution with parameter σ^2, then:

$$q = \sum_i X_i;\ \mathbf{E}\left[e^{\lambda q}\right] \overset{\text{independent}}{=\!=} \prod_i \mathbf{E}\left[e^{\lambda X_i}\right] \leq \exp(\frac{\lambda^2 \sum \sigma_i^2}{2}) \Rightarrow q \sim \mathcal{SG}(\sqrt{\sum_i \sigma_i^2})$$

$$\mathbf{P}(X_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$$

$$\mathbf{P}(\sum_i \overbrace{(X_i - \mathbf{E}\left[X_i\right])}^{q_i} \geq \epsilon) = \mathbf{P}(\sum_i q_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2 \cdot (\sqrt{\sum_i \sigma_i^2})^2}) = \exp(-\frac{\epsilon^2}{2 \cdot \sigma^2 \cdot n})$$

$$\mathbf{P}(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) \geq \epsilon) \xrightarrow{cX \sim \mathcal{SG}(|c|\sigma)} \mathbf{P}(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2 \cdot \frac{\sigma^2}{n^2} \cdot n})$$

Proof of Part Three

$$\mathbf{P}\left(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$$

Now, if we set this upper bound equal to δ, we get:

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

$$\log\delta = -\frac{n\epsilon^2}{2\sigma^2} \Rightarrow 2\sigma^2 \log\left(\frac{1}{\delta}\right) = n\epsilon^2$$

$$\epsilon = \sqrt{\frac{2\sigma^2 \log\left(\frac{1}{\delta}\right)}{n}}$$

$$\Rightarrow \mathbf{P}\left(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) \geq \sqrt{\frac{2\sigma^2 \log\left(\frac{1}{\delta}\right)}{n}}\right) \leq \delta$$

Part Three – Second Equation

The probability that the quantity $\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right])$ exceeds $\sqrt{\frac{2\sigma^2 \log\left(\frac{1}{\delta}\right)}{n}}$ is less than δ. Therefore, the probability that it is **less** than $\sqrt{\frac{2\sigma^2 \log\left(\frac{1}{\delta}\right)}{n}}$ is **greater than** $1 - \delta$:

$$\mathbf{P}\left(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) < \sqrt{\frac{2\sigma^2 \log\left(\frac{1}{\delta}\right)}{n}}\right) > 1 - \delta$$

2 UCB[75-points]

2.1 The Upper Confidence Bound Algorithm[40-points]

2.1.1 a)[2-points]

First, we introduce a few definitions. At each step, one action from the n available actions can be selected. Thus, at each time step t, the following sum is equal to one:

$$\sum_{a\in\mathcal{A}} \mathbb{I}\{A_t = a\} = 1$$

For each action taken at each step, we receive a reward of amount X_t. The total reward after n steps is denoted by S_n:

$$S_n = \sum_t X_t = \sum_t \sum_{a\in\mathcal{A}} X_t \cdot \mathbb{I}\{A_t = a\}$$

Also, the difference between the received reward and the optimal reward at each step is defined as:

$$\Delta_a = \mu^* - \mu_a(V)$$

Here, μ_a is the average reward received from arm_a in environment v, and μ^* is the best expected reward among all arms.
With these definitions, we can express the overall regret after n steps as:

$$R_n = n\mu^* - \mathbf{E}[S_n] = \sum_t \mu^* - \sum_t \mathbf{E}[X_t] = \sum_t \mathbf{E}[\mu^* - X_t]$$
$$= \sum_t \sum_{a\in\mathcal{A}} \mathbf{E}[(\mu^* - X_t)\mathbb{I}\{A_t = a\}]$$

If at step t the action $A_t = a$ is taken, then the reward X_t equals $\mu_a(v)$. Therefore, we have:

Decomposition Lemma

$$\sum_t \sum_{a\in\mathcal{A}} \mathbf{E}[(\mu^* - \mu_a(v))\mathbb{I}\{A_t = a\}] = \sum_{a\in\mathcal{A}} \underbrace{(\mu^* - \mu_a(v))}_{\Delta_a} \mathbf{E}\left[\overbrace{\sum_t \mathbb{I}\{A_t = a\}}^{T_a(n)}\right]$$

$$\Rightarrow R_n = \sum_a \Delta_a \mathbf{E}[T_a(n)]$$

2.1.2 b)[4-points]

If, for a chosen value of δ, our confidence interval given by $\sqrt{\frac{2\log(\frac{1}{\delta})}{T_i(t-1)}}$ is no longer useful and the algorithm's index (i.e., $UCB(T_i(t),\delta)$) falls below the true mean of that arm, the algorithm will no longer select the optimal arm and will incur linear regret.
To avoid this issue, δ can be chosen to decrease over time, which in turn reduces the confidence interval and lowers the probability that the index falls below the true mean. Choices such as $\delta = \frac{1}{n^2}, \frac{1}{n}, \ldots$ can be good options.

2.1.3 c)[4-points]

Suppose we are only dealing with favorable events (G_i). If we have that $u_i < T_i(n)$, where an arm number i has been pulled more than u_i times, then in these n rounds, there must exist a time t such that:

$T_i(t-1) = u_i \Rightarrow A_t = i$

Contradiction

Using the definition of G_i, we have:

$$UCB_i(t-1,\delta) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(\frac{1}{\delta})}{T_i(t-1)}} = \hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}} < \mu_1 < UCB_1(t-1,\delta)$$

At each step, the action taken in the next round is determined by:

$A_t = \text{argmax}_j \, \text{UCB}_j(t-1,\delta)$

Since we saw that the UCB of arm i at step $t-1$ is less than that of arm 1, it will definitely not be selected thus it's impossible for it to be played more than u_i times.
Therefore, by contradiction, we conclude that arm i is pulled at most u_i times in total.

2.1.4 d)[4-points]

By rewriting the expression $T_i(n)$ as follows, we get:

$\mathbf{E}\left[T_i(n)\right] = \mathbf{E}\left[\mathbb{I}\{G_i\}T_i(n)\right] + \mathbf{E}\left[\mathbb{I}\{G_i^c\}T_i(n)\right]$

That is, the expected value split into the contribution from **good events** and from **bad events**:

$\mathbf{E}\left[\mathbb{I}\{G_i\}T_i(n)\right] < u_i$ (Maximum number of T_i in Good events)
$\mathbf{E}\left[\mathbb{I}\{G_i^c\}T_i(n)\right] < n\mathbb{P}(G_i^c)$ (At worst all n times the Bad events happen)

$*\mathbf{E}\left[\mathbb{I}\{G_i^c\}\right] = \int_w \mathbb{I}\{G_i^c\}\, d\mathbb{P}(w) = \mathbb{P}(G_i^c)$

$\Rightarrow \mathbf{E}\left[T_i(n)\right] \leq u_i + n\mathbb{P}(G_i^c)$

2.1.5 e)[6-points]

We rewrite the given expression as follows:

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

$$\underbrace{\mathbf{P}(\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}})}_{\text{the area in red}} \leq \underbrace{\mathbf{P}(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i)}_{\text{the area in red + blue}}$$

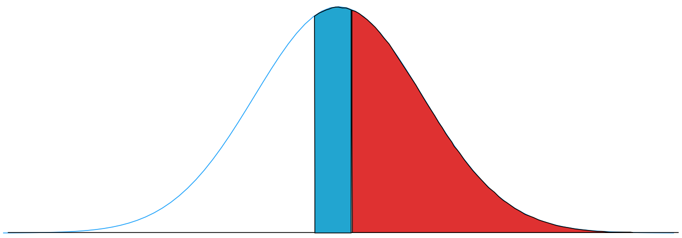

Figure 1: The probability of $(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i)$ occurring is greater.

Now using the Cramer–Chernoff method, we have:

$$\mathbf{P}(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i) \leq \exp\left(-\frac{(c\Delta_i)^2 n}{2\sigma^2}\right)$$
$$\xrightarrow{\sigma=1,\ n=u_i} \mathbf{P}(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i) \leq \exp\left(-\frac{c^2\Delta_i^2 u_i}{2}\right)$$

2.1.6 f)[4-points]

For G_i^c, which is the complement of G_i, we have:

$$G_i = \left\{\mu_1 \leq \min_{t\in[n]} UCB_1(t,\delta)\right\} \bigcap \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} \leq \mu_1\right\}$$

$$G_i^c = \left\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\right\} \bigcup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\right\}$$

We know that when a set's element is greater than the minimum of another set, it must be greater than at least one of its elements. So the first event above can be written as:

$$\left\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\right\} \subseteq \bigcup_t \{\mu_1 > UCB_1(t,\delta)\} \quad (**)$$

$$\Rightarrow \mathbf{P}\left(\left\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\right\}\right) \leq$$

$$\mathbf{P}\left(\left\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\right\}\right) + \mathbf{P}\left(\left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\right\}\right) \quad \text{(Union Bound)}$$

Now we compute the probability of the first event:

$$\begin{aligned}\mathbf{P}\left(\left\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\right\}\right) &\leq \mathbf{P}\left(\bigcup_t \{\mu_1 > UCB_1(t,\delta)\}\right) \quad (**)\\ &\leq \sum_{t=0}^{n} \mathbf{P}\left(\mu_1 > \hat{\mu}_{1t} + \sqrt{\frac{2\log(\frac{1}{\delta})}{t}}\right)\end{aligned}$$

The upper bound for this sum is derived using the Cramer–Chernoff method:

$$\mathbf{P}(X - \hat{X} \geq \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right), \quad \text{where } \epsilon = \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{t}}, \quad \sigma = 1 \text{ (assumption)}$$

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

$$\therefore \sum_t \mathbf{P}\left(\mu_1 > \hat{\mu}_{1t} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{t}}\right) \leq n\delta$$

The second probability, $\mathbf{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\right)$, was already computed in Part 5. Therefore, we conclude:

Probability of Bad Event

$$\begin{aligned}\mathbf{P}(G_i^c) = \mathbf{P}\left(\left\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\right\}\right) &+ \mathbf{P}\left(\left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} > \mu_1\right\}\right)\\ \leq n\delta \quad &+ \exp\left(-\frac{c^2\Delta_i^2 u_i}{2}\right)\end{aligned}$$

$$\boxed{\Rightarrow \mathbf{P}(G_i^c) \leq n\delta + \exp\left(-\frac{c^2\Delta_i^2 u_i}{2}\right)}$$

2.1.7 g)[6-points]

From Section 6, where we computed the probability $\mathbf{P}(G_i^c)$, we substitute it into the bound we had previously obtained for the expectation of $T_i(n)$:

$$\mathbf{E}[T_i(n)] \le u_i + n\mathbb{P}(G_i^c)$$
$$\mathbf{E}[T_i(n)] \le u_i + n(n\delta + \exp(-\frac{c^2\Delta_i^2 u_i}{2}))$$

The value of u_i we choose must also satisfy the inequality from Section 5: $\Delta_i - \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}} \ge c\Delta_i$. A typical choice is to pick the **smallest** u_i that satisfies this:

$$\Delta_i - \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}} = c\Delta_i$$
$$\Delta_i(1-c) = \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}}$$
$$(\Delta_i(1-c))^2 = \frac{2\log(\frac{1}{\delta})}{u_i} \Rightarrow u_i = \frac{2\log(\frac{1}{\delta})}{(\Delta_i(1-c))^2}$$
$$u_i = \left\lceil \frac{2\log(\frac{1}{\delta})}{(\Delta_i(1-c))^2} \right\rceil \quad \text{(Because } u_i \text{ is integer)}$$

Now, setting $\delta = \frac{1}{n^2}$ and substituting u_i into the expectation bound:

$$\mathbf{E}[T_i(n)] \le u_i + 1 + n^{1-\frac{2c^2}{(1-c)^2}} = \left\lceil \frac{2\log(n^2)}{(1-c)^2\Delta_i^2} \right\rceil + 1 + n^{1-\frac{2c^2}{(1-c)^2}}$$

If we set $c = \frac{1}{2}$, then:

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

Final Regret Bound

We use the upper bound $n^{-1} \le 1$, and $\left\lceil \frac{16\log(n)}{\Delta_i^2} \right\rceil \le \frac{16\log(n)}{\Delta_i^2} + 1$. Substituting these gives:

$$\left\lceil \frac{16\log(n)}{\Delta_i^2} \right\rceil + 1 + n^{-1} \le \frac{16\log(n)}{\Delta_i^2} + 1 + 1 + 1 = \frac{16\log(n)}{\Delta_i^2} + 3$$
$$\Rightarrow \mathbb{E}[T_i(n)] \le \frac{16\log(n)}{\Delta_i^2} + 3$$

2.1.8 h)[5-points]

By substituting the final bound from Section 7 into the regret decomposition formula from Section 1, we get:

$$R_n = \sum_a \Delta_a \mathbf{E}\left[T_a(n)\right] \leq \sum_a \Delta_a \left(\frac{16\log(n)}{\Delta_a^2} + 3\right) = \sum_{a:\Delta_a \neq 0} \frac{16\log(n)}{\Delta_a} + 3\sum_a \Delta_a$$

$$\Rightarrow R_n \leq \sum_{a:\Delta_a \neq 0} \frac{16\log(n)}{\Delta_a} + 3\sum_a \Delta_a$$

2.1.9 i)[5-points]

By choosing $\Delta = \sqrt{\frac{16k\log(n)}{n}}$, we have:

$$\begin{aligned}
R_n &= \sum_{i=1}^{k} \Delta_i \mathbb{E}[T_i(n)] \\
&= \sum_{i:\Delta_i < \Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i \geq \Delta} \Delta_i \mathbb{E}[T_i(n)] \\
&\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \left(3\Delta_i + \frac{16\log(n)}{\Delta_i}\right) \\
&\leq n\Delta + \frac{16k\log(n)}{\Delta} + 3\sum_{i=1}^{k} \Delta_i \\
&\leq 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k} \Delta_i
\end{aligned}$$

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]

$$\mathbf{P}(X = i) = \frac{w_i(t)}{S_t}$$
$$w_i(t+1) = w_i(t)\left(1 - \epsilon \cdot \mathbb{I}(\hat{m}_t = 1)\right)$$
$$S_{t+1} = \sum_i w_i(t+1) = \sum_i w_i(t)\left(1 - \epsilon \cdot \mathbb{I}(\hat{m}_t = 1)\right)$$

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

3.1.2 b)[8-points]

$$\begin{aligned}
\mathbf{E}\left[S_{T+1}\right] &= \mathbf{E}\left[S_T\right](1 - \epsilon \cdot \mathbf{P}(\hat{m}_T = 1)) \\
&= \mathbf{E}\left[S_{T-1}\right](1 - \epsilon \cdot \mathbf{P}(\hat{m}_T = 1))(1 - \epsilon \cdot \mathbf{P}(\hat{m}_{T-1} = 1)) \\
&\vdots \\
&= \mathbf{E}\left[S_0\right]\prod_{t=1}^{T}(1 - \epsilon \cdot \mathbf{P}(\hat{m}_t = 1)) \\
&= N \cdot \prod_{t=1}^{T}(1 - \epsilon \cdot \mathbf{P}(\hat{m}_t = 1)) \quad \text{(At time } t = 0\text{, all weights are 1, i.e., } w_i(0) = 1\text{)}
\end{aligned}$$

We also know from the Taylor expansion of e^{-x} that:

$$e^{-x} = 1 - x + \frac{x^2}{2} + \text{HOT (higher-order terms)}$$
$$\Rightarrow e^{-x} > 1 - x$$
$$\Rightarrow \prod_{t=1}^{T}(1 - \epsilon \cdot \mathbf{P}(\hat{m}_t = 1)) \leq \prod_{t=1}^{T} e^{-\epsilon \cdot \mathbf{P}(\hat{m}_t = 1)} = e^{-\epsilon \sum_{t=1}^{T} \mathbf{P}(\hat{m}_t = 1)}$$

$$\Rightarrow \mathbf{E}\left[S_{T+1}\right] = N \cdot \prod_{t=1}^{T}(1-\epsilon \cdot \mathbf{P}(\hat{m}_t = 1)) \leq N \cdot e^{-\epsilon \sum_{t=1}^{T} \mathbf{P}(\hat{m}_t=1)}$$

3.1.3 c)[15-points]

$$S_{t+1} = S_t(1-\epsilon \cdot \mathbf{P}(\hat{m}_t = 1))$$
$$P_t(i) = \frac{w_i(t)}{S_t} \Rightarrow w_i(t) = S_t \cdot P_t(i)$$
$$\mathbf{E}[\hat{m}_t] = \sum_i P_i(t)\hat{m}_t(i)$$

The last line gives us the expected number of mistakes from expert i at time t.

$$S_{t+1} = S_t - \epsilon \cdot S_t \cdot \mathbf{P}(\hat{m}_t = 1)$$
$$S_t \cdot \mathbf{P}(\hat{m}_t = 1) = \sum_i w_i(t) \cdot \mathbf{P}(\hat{m}_t = 1) = \sum_i S_t P_i(t) \cdot \mathbf{P}(\hat{m}_t = 1) = S_t \sum_i P_i(t) \cdot \mathbf{P}(\hat{m}_t = 1)$$
$$= S_t \cdot \mathbf{E}[\hat{m}_t]$$
$$\Rightarrow S_{t+1} = S_t(1-\epsilon \cdot \mathbf{E}[\hat{m}_t]) \leq S_t \cdot e^{-\epsilon \cdot \mathbf{E}[\hat{m}_t]}$$
$$w_i(T) = \prod_{t=1}^{T}(1-\epsilon \cdot \mathbf{P}(\hat{m}_t = 1))\,; \quad w_i(T) \leq S_T = \sum_{i=1}^{N} w_i(T)$$

Expected Mistake Bound

If we slightly rephrase and express the weight update after T rounds as:

$$w_T(i) = (1-\epsilon)^{M_T(i)}$$
$$(1-\varepsilon)^{M_T(i)} \leq N \cdot \exp\left(-\varepsilon \cdot \mathbb{E}\left[\sum_{t=1}^{T} \tilde{m}_t\right]\right) = N \cdot \exp\left(-\varepsilon \cdot \mathbb{E}[M_T]\right)$$

Taking logarithms:

$$M_T(i) \cdot \log(1-\varepsilon) \leq \log N - \varepsilon \cdot \mathbb{E}[M_T]$$

Using the inequality $\log(1-x) \leq -x - x^2$ near zero:

$$-M_T(i)(\varepsilon + \varepsilon^2) \leq \log N - \varepsilon \cdot \mathbb{E}[M_T]$$

Rearranging gives the final desired form:

$$\mathbb{E}[M_T] \leq (1+\varepsilon)M_T(i) + \frac{\log N}{\varepsilon}$$

3.1.4 d)[7-points]

From the inequality derived in Section 3, we know that for all i, the following holds:

$$\mathbb{E}[M_T] \leq (1+\varepsilon)M_T(i) + \frac{\log N}{\varepsilon}$$

Since this holds for all i, we can bound it over the minimum number of mistakes:

$$\mathbb{E}[M_T] \leq \min_i \{(1+\varepsilon)M_T(i)\} + \frac{\log N}{\varepsilon} \quad (*)$$

Now we simplify the multiplicative term:

$$(1+\varepsilon)M_i = M_i + \varepsilon M_i \leq M_i + \varepsilon T \quad \text{(since } M_i \leq T\text{)}$$

Plugging this into (*), we get:

$$\Rightarrow \mathbb{E}[M_T] \leq \min_i \{M_i + \varepsilon T\} + \frac{\log N}{\varepsilon}$$
$$\Rightarrow \mathbb{E}[M_T] \leq \min_i M_i + \varepsilon T + \frac{\log N}{\varepsilon}$$

To minimize the bound, take the derivative of the sum term:

$$\frac{\partial}{\partial \varepsilon}\left(\varepsilon T + \frac{\log N}{\varepsilon}\right) = T - \frac{\log N}{\varepsilon^2} = 0$$
$$\Rightarrow \varepsilon = \sqrt{\frac{\log N}{T}}$$

Substitute back in:

$$\varepsilon T + \frac{\log N}{\varepsilon} = \sqrt{\frac{\log N}{T}} \cdot T + \frac{\log N}{\sqrt{\frac{\log N}{T}}} = 2\sqrt{T \log N}$$

Final Mistake Bound

$$\mathbb{E}[M_T] \le \min_i M_i + 2\sqrt{T \log N}$$

This is a good bound because it grows sublinearly in T. If it can be improved from $\Omega(\sqrt{T \ln N})$ needs more checking.

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

3.2.1 a)[6-points]

$$\exp(-x) = \sum_{n=0}^{\infty} \frac{(-x)^n}{n!} \Rightarrow \exp(-x) \le 1 - x + \frac{x^2}{2}$$

$$S_{t+1} = \sum_i w_{t+1}(i) = \sum_i w_t(i) \cdot \exp(-\epsilon l_{ti}) \le \sum_i w_t(i) \left(1 - \epsilon l_{ti} + \frac{(\epsilon l_{ti})^2}{2}\right)$$

$$w_t(i) = p_t(i) \cdot \sum_i w_t(i) = p_t(i) S_t$$

$$\Rightarrow S_{t+1} \le \sum_i p_t(i) S_t \left(1 - \epsilon l_{ti} + \frac{(\epsilon l_{ti})^2}{2}\right) = S_t \left(1 - \epsilon \sum_i p_t(i) l_t(i) + \frac{\epsilon^2}{2} \sum_i p_t(i) l_t(i)^2\right)$$

we can drop the 1/2 from the epsilon

$$S_{t+1} \le S_t \left(1 - \epsilon \sum_i p_t(i) l_t(i) + \epsilon^2 \sum_i p_t(i) l_t(i)^2\right)$$

Loss-Based Upper Bound on S_{t+1}

$$S_{t+1} \le S_t \left(1 - \epsilon \sum_i p_t(i) l_t(i) + \epsilon^2 \sum_i p_t(i) l_t(i)^2\right)$$

3.2.2 b)[7-points]

$$\begin{aligned} S_{t+1} &\le S_t \left(\sum_i p_t(i) - \varepsilon \sum_i p_t(i) \ell_t(i) + \varepsilon^2 \sum_i p_t(i) \ell_t(i)^2\right) \\ &\le S_t \left(1 - \varepsilon p_t^\top \ell_t + \varepsilon^2 p_t^\top \ell_t^2\right) \\ &\le S_t \exp\left(-\varepsilon p_t^\top \ell_t + \varepsilon^2 p_t^\top \ell_t^2\right). \end{aligned}$$

$$S_T \leq S_1 \exp\left(-\varepsilon \sum_{t=1}^{T} p_t^\top \ell_t + \varepsilon^2 \sum_{t=1}^{T} p_t^\top \ell_t^2\right)$$
$$\leq N \exp\left(-\varepsilon \sum_{t=1}^{T} p_t^\top \ell_t + \varepsilon^2 \sum_{t=1}^{T} p_t^\top \ell_t^2\right)$$
$$S_T \geq \exp\left(-\varepsilon \sum_{t=1}^{T} \ell_t(i)\right)$$
$$-\varepsilon \sum_{t=1}^{T} \ell_t(i) \leq \ln(N) - \varepsilon \sum_{t=1}^{T} p_t^\top \ell_t + \varepsilon^2 \sum_{t=1}^{T} p_t^\top \ell_t^2$$
$$\sum_{t=1}^{T} p_t^\top \ell_t - \sum_{t=1}^{T} \ell_t(i) \leq \varepsilon \sum_{t=1}^{T} p_t^\top \ell_t^2 + \frac{\ln(N)}{\varepsilon}$$

Upper Bound on Regret

$$R_T \leq \varepsilon \sum_{t=1}^{T} p_t^\top \ell_t^2 + \frac{\ln(N)}{\varepsilon}$$

This is somewhat similar to the bound for the RWM algorithm, that it may be slightly, but both have the same $2\sqrt{T \ln N}$ upper bound in general.

3.2.3 c)[2-points]

Given that $l_t(i)$ lies between -1 and 1, we have:

$$\epsilon \sum_t p_t^\top \ell_t^2 \leq \epsilon T$$
$$\Rightarrow R_T \leq \epsilon T + \frac{\ln(N)}{\epsilon}$$

Now, choosing the optimal value for ϵ:

$$\epsilon = \sqrt{\frac{2\ln(N)}{T}} \Rightarrow R_T \leq \sqrt{2\ln(N)T} + \sqrt{\frac{\ln(N)T}{2}} \leq 2\sqrt{\ln(N)T}$$