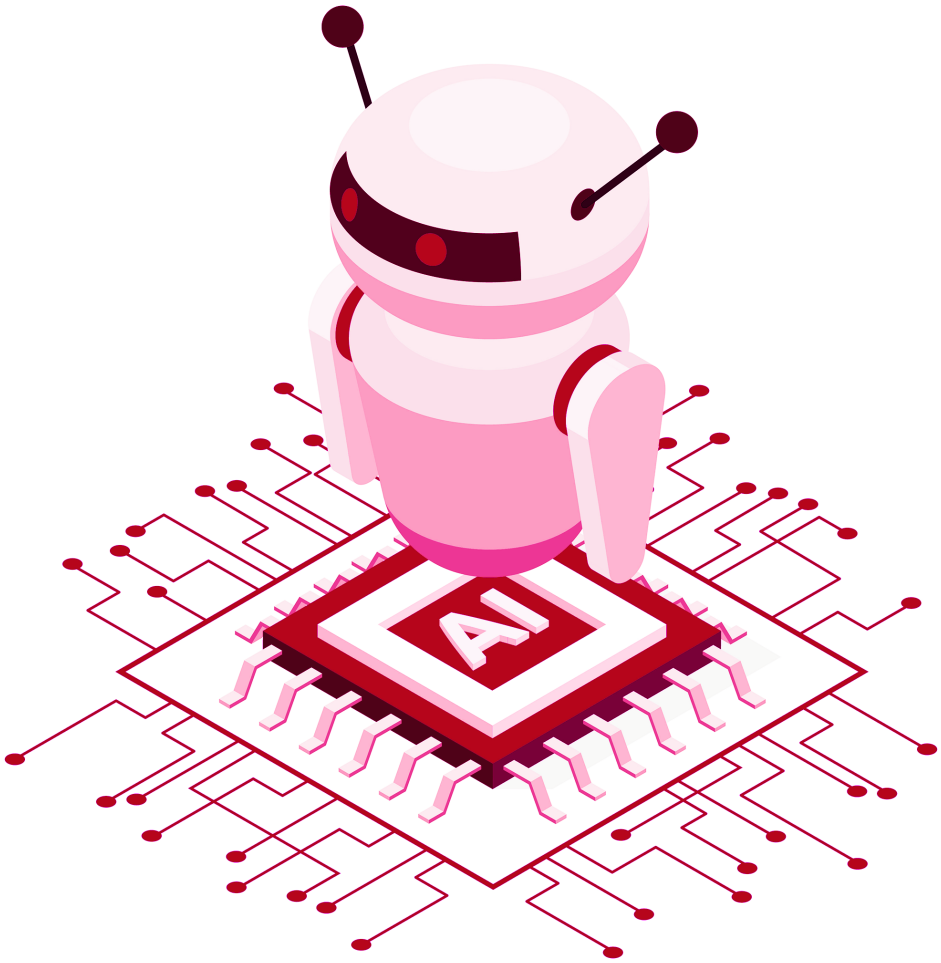

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution forHomework [9]

[Exploration Methods]

By:

[Asemaneh Nafe]

[400105285]

RIML

Spring 2025

Contents

1 Light-tailed Distributions[25-points]

1.1 Hoeffding's Inequality[10-points]

1.1.1 a)[6-points]

Let X be a random variable such that $\mathbb{E}[X] = 0$ and $a \leq X \leq b$. We aim to prove that for all $s > 0$:

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

Since the function $f(x) = e^{sx}$ is convex, for all $x \in [a, b]$, we have:

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

Substitute $V = X$, and use $\mathbb{E}[V] = 0$. We get:

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

Let us define:

$$p := \frac{b}{b-a}, \quad 1-p = -\frac{a}{b-a}, \quad u := s(b-a).$$

Note that:

$$\mathbb{E}[e^{sV}] \leq pe^{sa} + (1-p)e^{sb} = e^{sa}\left(p + (1-p)e^{s(b-a)}\right).$$

So taking the logarithm:

$$\log \mathbb{E}[e^{sV}] \leq sa + \log\left(p + (1-p)e^{u}\right) =: \varphi(u),$$

where

$$\varphi(u) := sa + \log\left(p + (1-p)e^{u}\right) = (p-1)u + \log\left(p + (1-p)e^{u}\right).$$

We expand $\varphi(u)$ using Taylor's theorem:

$$\varphi(u) = \varphi(0) + \varphi'(0)u + \frac{1}{2}\varphi''(\xi)u^2,$$

for some $\xi \in [0, u]$. Now compute:

$$\varphi(0) = 0,$$
$$\varphi'(u) = (p-1) + \frac{(1-p)e^u}{p+(1-p)e^u},$$
$$\varphi'(0) = (p-1) + \frac{(1-p)}{p+(1-p)} = (p-1) + 1 - \frac{p}{p+(1-p)} = 0,$$
$$\varphi''(u) = \frac{p(1-p)e^u}{\left(p+(1-p)e^u\right)^2}.$$

Using calculus (or known inequality), it can be shown that for all $u \in \mathbb{R}$, the second derivative satisfies:

$$\varphi''(u) \leq \frac{1}{4}.$$

Therefore:

$$\log \mathbb{E}[e^{sV}] \leq \frac{1}{2} \cdot \frac{1}{4} \cdot u^2 = \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

Exponentiating both sides gives:

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

Bounding $\varphi''(u) \leq \frac{1}{4}$: We want to show that:

$$\varphi''(u) = \frac{p(1-p)e^u}{\left(p+(1-p)e^u\right)^2} \leq \frac{1}{4}, \quad \text{for all } u \in \mathbb{R}.$$

Let us define:

$$z := e^u > 0.$$

Then,

$$\varphi''(u) = \frac{p(1-p)z}{\left(p+(1-p)z\right)^2} := f(z).$$

We now study the function:

$$f(z) = \frac{Az}{(p+Bz)^2}, \quad \text{where } A = p(1-p), \quad B = 1-p.$$

We will find the maximum of $f(z)$ for $z > 0$. Take the derivative:

$$f'(z) = \frac{A(p+Bz)^2 - 2ABz(p+Bz)}{(p+Bz)^4} = \frac{A(p+Bz)[(p+Bz) - 2Bz]}{(p+Bz)^4} = \frac{A(p+Bz)(p-Bz)}{(p+Bz)^4}.$$

So $f'(z) = 0$ when $z = \frac{p}{B} = \frac{p}{1-p}$. Since this is the only critical point and the function $f(z)$ is positive and vanishes as $z \to 0$ and $z \to \infty$, the maximum occurs at this point.

Compute $f(z)$ at this critical point:

$$z = \frac{p}{1-p} \quad \Rightarrow \quad p + (1-p)z = p + (1-p) \cdot \frac{p}{1-p} = p + p = 2p.$$

So:

$$\varphi''(u) = f(z) = \frac{p(1-p) \cdot \frac{p}{1-p}}{(2p)^2} = \frac{p^2}{4p^2} = \frac{1}{4}.$$

Hence, for all $u \in \mathbb{R}$,

$$\varphi''(u) \leq \frac{1}{4}.$$

1.1.2 b)[4-points]

(b)

Let $Z_1, \ldots, Z_n$ be independent random variables with $Z_i \in [a, b]$ and finite expectations $\mathbb{E}[Z_i]$. Then, for all $t \geq 0$, we claim:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t\right) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right),$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \leq -t\right) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right).$$

We prove this using the Chernoff bounding method.

Chernoff Bound (General Form): Let X be a real-valued random variable and $\lambda > 0$. Then for any $t \in \mathbb{R}$ (we have prooved this in the end):

$$\mathbb{P}(X \geq t) \leq \mathbb{E}[e^{\lambda X}] \cdot e^{-\lambda t}.$$

Step 1: Apply Chernoff Bound to the sum. We apply the bound to the sum of centered variables:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t\right) = \mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq nt\right).$$

Using Chernoff's bound with $\lambda > 0$:

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq nt\right) \leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right)\right] \cdot e^{-\lambda nt}.$$

Step 2: Use independence and Hoeffding's Lemma. Since the Z_i are independent, the exponentiated expectation factorizes:

$$= \left(\prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda(Z_i - \mathbb{E}[Z_i])} \right] \right) \cdot e^{-\lambda n t}.$$

By Hoeffding's Lemma (proved in part (a)), for $Z_i \in [a, b]$, we have:

$$\mathbb{E}\left[e^{\lambda(Z_i - \mathbb{E}[Z_i])} \right] \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8} \right).$$

Thus,

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \geq t \right) \leq \exp\left(n \cdot \frac{\lambda^2 (b-a)^2}{8} - \lambda n t \right).$$

Step 3: Optimize over λ. Minimize the upper bound over $\lambda > 0$. The expression:

$$\exp\left(\frac{n\lambda^2 (b-a)^2}{8} - \lambda n t \right)$$

is minimized when the exponent is minimized. Setting the derivative with respect to λ to zero:

$$\frac{d}{d\lambda}\left(\frac{n\lambda^2 (b-a)^2}{8} - \lambda n t \right) = \frac{n\lambda (b-a)^2}{4} - nt = 0,$$

so the optimal $\lambda = \frac{4t}{(b-a)^2}$. Substituting:

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \geq t \right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2} \right).$$

Step 4: Lower Tail. To bound the lower tail, apply the same argument to $-Z_i$, which also lies in $[-b, -a]$, so the width $b - a$ is preserved. Then:

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \leq -t \right) &= \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (-Z_i + \mathbb{E}[Z_i]) \geq t \right) \\ &= \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (-Z_i - \mathbb{E}[-Z_i]) \geq t \right) \\ &\leq \exp\left(-\frac{2nt^2}{(b-a)^2} \right). \end{aligned}$$

Combining both tails, we obtain Hoeffding's inequality:

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right| \geq t \right) \leq 2 \exp\left(-\frac{2nt^2}{(b-a)^2} \right). \square$$

Proof of Chernoff Bound: We begin with Markov's inequality, which states that for any non-negative random variable Y and any $\alpha > 0$,

$$\mathbb{P}(Y \geq \alpha) \leq \frac{\mathbb{E}[Y]}{\alpha}.$$

Proof. Let A denote the event $\{X \geq a\}$. Then:

$$\mathbb{E}(X) = \sum_{s \in S} p(s)X(s) = \sum_{s \in A} p(s)X(s) + \sum_{s \in \bar{A}} p(s)X(s).$$

As X is non-negative, we have $\sum_{s \in \bar{A}} p(s)X(s) \geq 0$. Hence:

$$\mathbb{E}(X) \geq \sum_{s \in A} p(s)X(s) \geq a \sum_{s \in A} p(s) = a \cdot \mathbb{P}(A).$$

Apply Markov's Inequality to the non-negative random variable $Y = e^{\lambda X}$, and choose $\alpha = e^{\lambda t}$. Then:

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

This completes the proof. ■

1.2 Sub-Gaussian[15-points]

1.2.1 a-1)[2-points]

We begin by applying the Chernoff bound to the centered variable $X - \mu$. For any $\lambda > 0$, we have:

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

Using the sub-Gaussian property:

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

we get:

$$\mathbb{P}(X > \mu + t) \leq \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda t\right).$$

We now minimize the right-hand side over $\lambda > 0$. The expression:

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

is minimized at $\lambda = \frac{t}{\sigma^2}$. Substituting this value:

$$\mathbb{P}(X > \mu + t) \leq \exp\left(\frac{t^2}{2\sigma^2} - \frac{t^2}{\sigma^2}\right) = \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

1.2.2 a-2)[2-points]

Let X be a sub-Gaussian random variable with mean $\mu = \mathbb{E}[X]$ and sub-Gaussian parameter σ. We aim to bound the lower tail probability:

$$\mathbb{P}(X < \mu - t) = \mathbb{P}(\mu - X > t).$$

Let $Y = -X$. Then Y is also sub-Gaussian with mean $-\mu$ and the same sub-Gaussian parameter σ, because:

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

which still satisfies the sub-Gaussian condition.

Now apply the upper tail bound to Y:

$$\mathbb{P}(-X > -\mu + t) = \mathbb{P}(X < \mu - t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

This completes the proof of the lower tail bound. □

1.2.3 a-3)[2-points]

We begin by expressing the two-sided event as the union of the upper and lower tail events:

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) = \mathbb{P}(X > \mathbb{E}[X] + t) + \mathbb{P}(X < \mathbb{E}[X] - t).$$

Since X is sub-Gaussian with parameter σ, we apply the upper and lower tail bounds derived previously:

$$\mathbb{P}(X > \mathbb{E}[X] + t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

$$\mathbb{P}(X < \mathbb{E}[X] - t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Adding the two:

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

This completes the proof. □

1.2.4 b)[3-points]

Proof. We show that $Z = \sum_{i=1}^n X_i$ is sub-Gaussian with parameter $\sigma_Z^2 = \sum_{i=1}^n \sigma_i^2$.

$$\mathbb{E}\left(e^{\lambda(Z-\mathbb{E}[Z])}\right) = \mathbb{E}\left(e^{\lambda(\sum X_i - \mathbb{E}[\sum X_i])}\right) = \prod_{i=1}^n \mathbb{E}\left(e^{\lambda(X_i-\mathbb{E}[X_i])}\right) \quad \text{(independence)} \leq \prod_{i=1}^n e^{\sigma_i^2\lambda^2/2} = e^{\left(\sum\sigma_i^2\right)\lambda^2/2}.$$

The Theorem now follows from Lemma 3.(a-3)

1.2.5 c)[4-points]

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We aim to bound

$$\mathbb{P}(\bar{X} - \mu \geq \epsilon) = \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq n\epsilon\right).$$

Using the Chernoff bound, for any $\lambda > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq n\epsilon\right) \leq \exp(-\lambda n\epsilon) \cdot \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n}(X_i - \mu_i)}\right].$$

Since the X_i's are independent,

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

So we obtain:

$$\mathbb{P}(\bar{X} - \mu \geq \epsilon) \leq \exp\left(\frac{n\lambda^2\sigma^2}{2} - \lambda n\epsilon\right).$$

To minimize this bound, choose $\lambda = \frac{\epsilon}{\sigma^2}$, which gives:

$$\mathbb{P}(\bar{X} - \mu \geq \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right).$$

To derive the second inequality, we invert the tail bound. Let:

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

Then:

$$\Pr\left(\bar{X}_n - \mu \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) = \delta.$$

Therefore, with probability at least $1 - \delta$, we have:

$$\bar{X}_n - \mu < \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}.$$

This completes the proof. □

2 UCB[75-points]

2.1 The Upper Confidence Bound Algorithm[40-points]

2.1.1 a)[2-points]

Since R_n is based on summing over rounds, and the right-hand side of the lemma statement is based on summing over actions, to convert one sum into the other one, we introduce indicators. In particular, note that for any fixed t we have $\sum_{a \in \mathcal{A}} 1\{A_t = a\} = 1$. Hence $S_n = \sum_t X_t = \sum_t \sum_a X_t 1\{A_t = a\}$, and thus

$$\begin{aligned} R_n &= n\mu^* - \mathbb{E}[S_n] \\ &= \sum_{a \in \mathcal{A}} \sum_{t=1}^{n} \mathbb{E}[(\mu^* - X_t) 1\{A_t = a\}]. \end{aligned} \tag{1}$$

The expected reward in round t conditioned on A_t is μ_{A_t}, which means that

$$\begin{aligned} \mathbb{E}[(\mu^* - X_t) 1\{A_t = a\} \mid A_t] &= 1\{A_t = a\} \mathbb{E}[\mu^* - X_t \mid A_t] && (2) \\ &= 1\{A_t = a\}(\mu^* - \mu_{A_t}) && (3) \\ &= 1\{A_t = a\}(\mu^* - \mu_a) && (4) \\ &= 1\{A_t = a\}\Delta_a. && (5) \end{aligned}$$

and using the definition of $T_a(n)$:

$$\sum_{t=1}^{n} \mathbb{E}\left[(\mu^* - X_t) 1\{A_t = a\} \mid A_t\right] = \Delta_a \sum_{t=1}^{n} \mathbb{E}[1\{A_t = a\}] = \Delta_a \cdot \mathbb{E}[T_a(n)]$$

Substituting into Eq. (1) gives:

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \cdot \mathbb{E}[T_a(n)] \tag{6}$$

Equation (6) is a standard decomposition of expected regret in stochastic multi-armed bandits, where each suboptimal arm contributes to regret proportionally to how often it is selected and how suboptimal it is.

2.1.2 b)[4-points]

Assume δ is a fixed constant, say $\delta = 0.05$. Consider a scenario where due to randomness in the reward samples, the estimated mean $\hat{\mu}_1(t)$ of the optimal arm (say arm 1 with true mean μ_1) is significantly underestimated at some time step t.

In this case, the upper confidence bound of the optimal arm becomes:

$$\text{UCB}_1(t, \delta) = \hat{\mu}_1(t) + \sqrt{\frac{2\log(1/\delta)}{T_1(t)}}$$

which may fall *below* the upper confidence bounds of some suboptimal arms. This event may persist for a large number of rounds (especially for large horizon n), causing the algorithm to select a suboptimal arm $a \neq 1$ repeatedly.

Since regret accumulates as Δ_a for every selection of a suboptimal arm a, if this bad event lasts for $\Theta(n)$ rounds, the total regret becomes $\Theta(n)$, i.e., linear in n.

for example Suppose we choose a relatively large value for δ, say $\delta = 0.1$. Then with probability 0.1, the confidence interval for an optimal arm may fail to contain its true mean. Now imagine the following scenario:

- Arm 1 is optimal with $\mu_1 = 0.9$.
- Arm 2 is suboptimal with $\mu_2 = 0.6$.
- Due to randomness in the first few samples, $\hat{\mu}_1(t)$ is underestimated, and $\text{UCB}_1(t, \delta)$ drops below μ_2.
- The algorithm stops playing arm 1 and keeps choosing arm 2.

Since this failure event can persist over many rounds, the algorithm may suffer *linear regret*. That is, the expected regret grows proportionally with n:

$$\mathbb{E}[R(n)] = \Omega(n)$$

the best choise is $\delta = 1/n$

To mitigate this, we can choose $\delta = 1/n$. Then the probability that the confidence interval for an optimal arm fails in any single round is at most $1/n$, and by applying a union bound over all n rounds, the *total* probability that we ever make such a failure is at most 1. In fact, more commonly we set $\delta = 1/n^2$ or even $\delta_t = 1/t^2$ to ensure:

- With high probability, the optimal arm's true mean remains inside the confidence interval at all times.
- Regret due to incorrect confidence intervals becomes negligible.

This way, the algorithm continues to explore the optimal arm enough to maintain sublinear regret:

$$\mathbb{E}[R(n)] = \mathcal{O}(\log n)$$

if we do not acces to n To avoid this, we should ensure that the confidence intervals tighten over time, thereby reducing the chance of persistent underestimation of the optimal arm. This can be achieved by choosing a time-dependent δ such as:

$$\delta_t = \frac{1}{t^2}$$

This guarantees that the probability of a large deviation decays sufficiently fast, making the total probability of failure over n rounds summable:

$$\sum_{t=1}^{n} \delta_t = \sum_{t=1}^{n} \frac{1}{t^2} < \infty$$

Therefore, the total regret remains sublinear.

Conclusion Choosing δ as a fixed constant can lead to catastrophic outcomes (linear regret) due to a fixed probability of estimation error. To avoid this, we should choose δ as a function of time, typically $\delta_t = \frac{1}{t^2}$ or $\delta_t = \frac{1}{n^\alpha}$ for some $\alpha > 1$. The choice of δ is critical. If too large, the algorithm risks eliminating the optimal arm early, leading to linear regret. Choosing $\delta = 1/n$ (or smaller) ensures that such failures are rare and keeps regret low. This provides a practical guideline when tuning UCB algorithms in real-world settings.

2.1.3 c)[4-points]

we do by contradiction. Suppose that $T_i(n) > u_i$. Then arm i was played more than u_i times over the n rounds, and so there must exist a round $t \in [n]$ where $T_i(t-1) = u_i$ and $A_t = i$. Using the definition of G_i,

$$\begin{aligned}
\mathsf{UCB}_i(t-1,\delta) &= \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} && \text{(definition of } \mathsf{UCB}_i(t-1,\delta)) \\
&= \hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} && \text{(since } T_i(t-1) = u_i) \\
&< \mu_1 && \text{(definition of } G_i) \\
&< \mathsf{UCB}_1(t-1,\delta). && \text{(definition of } G_i) \qquad (7)
\end{aligned}$$

Hence $A_t = \arg\max_j \mathsf{UCB}_j(t-1,\delta) \neq i$, which is a contradiction. Therefore if G_i occurs, then $T_i(n) \leq u_i$.

2.1.4 d)[4-points]

Consider decomposing the expectation based on whether the good event G_i occurs or not:

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

From part (c), we know that if G_i occurs, then $T_i(n) \leq u_i$. Hence:

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

On the other hand, since arm i can be pulled at most n times:

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

Combining the two:

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

2.1.5 e)[6-points]

Recall that $\mu_1 - \mu_i = \Delta_i$, so $\mu_1 = \mu_i + \Delta_i$. Then,

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

since $\Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq c\Delta_i$:

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

Since $X_1, \ldots, X_n$ are i.i.d. 1-sub-Gaussian random variables, the last probability can be bounded as

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

where the inequality follows from the result in Question 1.2 part (c).

2.1.6 f)[4-points]

Let us now turn to upper bounding $\mathbb{P}(G_i^c)$. By its definition,

$$G_i^c = \left\{\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right\}. \tag{8}$$

The first of these sets is decomposed using the definition of $\mathrm{UCB}_1(t,\delta)$,

$$\left\{\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \subseteq \left\{\mu_1 \geq \min_{s\in[n]} \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\} = \bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}. \tag{9}$$

from Question 1.2 part (c) we have:

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

so

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}}\right) \leq \delta. \tag{10}$$

Then using a union bound and the concentration bound for sums of independent sub-Gaussian random variables in (10), we obtain:

$$\begin{aligned}\mathbb{P}\left(\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right) &\leq \mathbb{P}\left(\bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}\right) \\ &\leq \sum_{s=1}^{n} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right) \leq n\delta. \end{aligned} \tag{11}$$

Taking part e together with (11) and (8), we have

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

2.1.7 g)[6-points]

by substituting last part into part d, we obtain

$$\mathbb{E}[T_i(n)] \leq u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right) \tag{12}$$

we should choose $u_i \in [n]$ satisfying $\Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq c\Delta_i$. A natural choice is the smallest integer for which it holds, which is

$$u_i = \left\lceil \frac{2\log(1/\delta)}{(1-c)^2\Delta_i^2} \right\rceil.$$

This choice of u_i can be larger than n, but in this case Eq. (12) holds trivially since $T_i(n) \leq n$. Then, using the assumption that $\delta = 1/n^2$ and this choice of u_i leads via (12) to

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

All that remains is to choose $c \in (0,1)$. The second term will contribute a polynomial dependence on n unless $2c^2/(1-c)^2 \geq 1$. However, if c is chosen too close to 1, then the first term blows up. Somewhat arbitrarily we choose $c = 1/2$, which leads to

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16\log(n)}{\Delta_i^2}. \tag{13}$$

2.1.8 h)[5-points]

The result follows by substituting the above display in (6).

From the regret expression:

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \cdot \mathbb{E}[T_a(n)] \tag{14}$$

and the bound on the expected number of times suboptimal arm i is pulled:

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

we conclude:

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

Or more compactly:

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

2.1.9 i)[5-points]

Let $\Delta > 0$ be some value to be tuned subsequently, and recall from the last proof that for each suboptimal arm i, we can bound

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

Therefore, using the basic regret decomposition (6) again, we have

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

$$\leq n\Delta + \sum_{i:\Delta_i\geq\Delta} \left(3\Delta_i + \frac{16\log(n)}{\Delta_i}\right) \leq n\Delta + \frac{16k\log(n)}{\Delta} + 3\sum_i \Delta_i$$

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

where the first inequality follows because $\sum_{i:\Delta_i<\Delta} T_i(n) \leq n$ and the last line by choosing $\Delta = \sqrt{\frac{16k\log(n)}{n}}$.

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

3 Online Learning[50-points]

3.1 Randomized Weighted Majority Algorithm[35-points]

3.1.1 a)[5-points]

Let us compute the total weight at round $t+1$. It is the sum of updated weights:

$$S_{t+1} = \sum_{\text{correct } i} w_i(t) + \sum_{\text{incorrect } i} w_i(t)(1-\epsilon)$$

This simplifies to:

$$\begin{aligned} S_{t+1} &= \sum_{i \in [N]} w_i(t) - \epsilon \sum_{\text{incorrect } i} w_i(t) \\ &= S_t - \epsilon \cdot \left(\sum_{\text{incorrect } i} w_i(t) \right) \end{aligned}$$

Taking expectation on both sides:

$$\mathbb{E}[S_{t+1}] = \mathbb{E}[S_t] - \epsilon \cdot \mathbb{E}\left[\sum_{\text{incorrect } i} w_i(t) \right]$$

Note that $\sum_{\text{incorrect } i} w_i(t) = S_t \cdot \mathbb{P}(\tilde{m}_t = 1)$, since $\mathbb{P}(\tilde{m}_t = 1)$ is the probability mass on incorrect experts. Substituting:

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

3.1.2 b)[8-points]

We begin with the recurrence relation derived in part (a):

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

Apply this recursively from $t = 1$ to $t = T$:

$$\mathbb{E}[S_{T+1}] = S_1 \cdot \prod_{t=1}^{T} (1 - \epsilon \cdot \mathbb{P}(\tilde{m}_t = 1))$$

Since initially all weights are $w_i(0) = 1$, and there are N experts, we have:

$$S_1 = \sum_{i=1}^{N} w_i(0) = N$$

We now apply the inequality $1 - x \leq e^{-x}$ for all $x \in [0, 1]$, which holds since $\epsilon \cdot \mathbb{P}(\tilde{m}_t = 1) \in [0, \epsilon] \subset [0, 1]$:

$$\prod_{t=1}^{T} (1 - \epsilon \cdot \mathbb{P}(\tilde{m}_t = 1)) \leq \prod_{t=1}^{T} e^{-\epsilon \cdot \mathbb{P}(\tilde{m}_t = 1)} = e^{-\epsilon \sum_{t=1}^{T} \mathbb{P}(\tilde{m}_t = 1)}$$

Therefore,

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

3.1.3 c)[15-points]

Let $M = \sum_{t=1}^{T} \tilde{m}_t$ be the total number of expected mistakes made by the algorithm, where $\tilde{m}_t \in \{0, 1\}$ indicates whether the chosen expert made a mistake at round t.

Let M_i denote the number of mistakes made by expert i over T rounds.

We define the weight update rule as:

$$w_i(t+1) = \begin{cases} w_i(t) & \text{if expert } i \text{ is correct at round } t \\ w_i(t) \cdot (1 - \epsilon) & \text{if expert } i \text{ is wrong at round } t \end{cases}$$

So after T rounds, the weight of expert i becomes:

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

On the other hand, the total weight satisfies from part (b):

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

Since $\mathbb{E}[w_i(T+1)] \leq \mathbb{E}[S_{T+1}]$, we have:

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

Take the natural logarithm of both sides:

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

so,

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

Now use the Taylor expansion of $\ln(1 - \epsilon)$ around $\epsilon = 0$:

$$\ln(1 - \epsilon) = -\epsilon - \frac{\epsilon^2}{2} - \frac{\epsilon^3}{3} - \cdots \Rightarrow -\ln(1 - \epsilon) = \epsilon + \frac{\epsilon^2}{2} + \cdots$$

Thus,

$$\frac{-\ln(1-\epsilon)}{\epsilon} = 1 + \frac{\epsilon}{2} + \cdots$$

Substitute this into the inequality:

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

Dropping higher-order terms gives the approximate bound:

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

3.1.4 d)[7-points]

We are given the general bound from previous parts:

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

We aim to minimize the right-hand side over all $i \in [N]$, i.e.:

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

Now suppose that we know $M_i \leq T$, since the best expert cannot make more than T mistakes in T rounds. Let:

$$M^* = \min_{i \in [N]} M_i$$

We choose the learning rate ϵ to optimize the bound. One standard approach is to balance the two terms by setting:

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

Then:

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

Since $M^* \leq T$, we get:

$$\epsilon M^* \leq \sqrt{T \ln N}, \quad \frac{\ln N}{\epsilon} = \sqrt{T \ln N}$$

Thus:

$$\mathbb{E}[M] \leq M^* + 2\sqrt{T \ln N}$$

Final bound:

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

Interpretation:

This is a *regret bound* of the form:

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

This means that the expected number of additional mistakes made by the randomized weighted majority (RWM) algorithm, compared to the best fixed expert in hindsight, grows sublinearly with T. Specifically, the average regret per round:

$$\frac{\text{Regret}_T}{T} \leq 2\sqrt{\frac{\ln N}{T}} \to 0 \quad \text{as } T \to \infty$$

This is a *very good regret bound* in online learning — it is sublinear in T and logarithmic in the number of experts N, meaning the algorithm becomes nearly as good as the best expert over time.

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

3.2.1 a)[6-points]

We are given the weight update rule:

$$w_{t+1}(i) = w_t(i) \cdot e^{-\epsilon \ell_{ti}},$$

and the total weight at round $t+1$ is:

$$S_{t+1} = \sum_{i=1}^{N} w_{t+1}(i) = \sum_{i=1}^{N} w_t(i) \cdot e^{-\epsilon \ell_{ti}}.$$

Let $p_t(i) = \frac{w_t(i)}{S_t}$, where $S_t = \sum_{j=1}^{N} w_t(j)$. Then we can write:

$$S_{t+1} = \sum_{i=1}^{N} S_t \cdot p_t(i) \cdot e^{-\epsilon \ell_{ti}} = S_t \sum_{i=1}^{N} p_t(i) e^{-\epsilon \ell_{ti}}.$$

Now, we apply the inequality $e^{-x} \leq 1 - x + x^2$ for $x \in [-1, 1]$ (which holds since $\ell_{ti} \in [-1, 1]$ and $\epsilon > 0$):

$$e^{-\epsilon \ell_{ti}} \leq 1 - \epsilon \ell_{ti} + \epsilon^2 \ell_{ti}^2.$$

Substitute this into the sum:

$$S_{t+1} \leq S_t \sum_{i=1}^{N} p_t(i) \left(1 - \epsilon \ell_{ti} + \epsilon^2 \ell_{ti}^2\right).$$

Distribute the sum:

$$S_{t+1} \leq S_t \left(\sum_{i=1}^{N} p_t(i) - \epsilon \sum_{i=1}^{N} p_t(i)\ell_{ti} + \epsilon^2 \sum_{i=1}^{N} p_t(i)\ell_{ti}^2 \right).$$

Since $\sum_{i=1}^{N} p_t(i) = 1$, we get:

$$S_{t+1} \leq S_t \left(1 - \epsilon \sum_{i=1}^{N} p_t(i)\ell_{ti} + \epsilon^2 \sum_{i=1}^{N} p_t(i)\ell_{ti}^2 \right).$$

Thus, the inequality is proven.

3.2.2 b)[7-points]

Lower Bound on S_{T+1} Using Jensen's Inequality

Recall that the weight update rule in the Hedge algorithm is:

$$w_{T+1}(i) = \exp\left(-\epsilon \sum_{t=1}^{T} \ell_{ti} \right).$$

Thus, the total weight at time $T+1$ is:

$$S_{T+1} = \sum_{i=1}^{N} w_{T+1}(i) = \sum_{i=1}^{N} \exp\left(-\epsilon \sum_{t=1}^{T} \ell_{ti} \right).$$

Now, fix any distribution $p \in \Delta^N$ (i.e., any fixed convex combination over experts). By Jensen's inequality, which states that for any convex function f and probability distribution $p(i)$,

$$f\left(\sum_i p(i)x_i \right) \leq \sum_i p(i) f(x_i),$$

we get the reverse inequality for concave functions. Since $f(x) = \exp(-\epsilon x)$ is a **concave** function (because the exponential of a negative linear function is concave), we have:

$$\sum_{i=1}^{N} p(i) \exp\left(-\epsilon \sum_{t=1}^{T} \ell_{ti} \right) \geq \exp\left(-\epsilon \sum_{i=1}^{N} p(i) \sum_{t=1}^{T} \ell_{ti} \right) = \exp\left(-\epsilon \sum_{t=1}^{T} \ell_t \cdot p \right).$$

Since $S_{T+1} \geq \sum_{i=1}^{N} p(i) w_{T+1}(i)$, we conclude:

$$S_{T+1} \geq \exp\left(-\epsilon \sum_{t=1}^{T} \ell_t \cdot p \right).$$

This inequality provides a lower bound on S_{T+1} in terms of the cumulative loss of any fixed distribution p. Taking logarithms on both sides, we obtain:

$$\log S_{T+1} \geq -\epsilon \sum_{t=1}^{T} \ell_t \cdot p.$$

Also from earlier:

$$S_{t+1} \leq S_t \left(1 - \epsilon \ell_t \cdot p_t + \epsilon^2 \ell_t^2 \cdot p_t\right),$$

and using the inequality $\log(1 - x + x^2) \leq -x + x^2$, we obtain:

$$\log S_{T+1} \leq \log N - \epsilon \sum_{t=1}^{T} \ell_t \cdot p_t + \epsilon^2 \sum_{t=1}^{T} \ell_t^2 \cdot p_t.$$

Combining both sides:

$$-\epsilon \sum_{t=1}^{T} \ell_t \cdot p \leq \log N - \epsilon \sum_{t=1}^{T} \ell_t \cdot p_t + \epsilon^2 \sum_{t=1}^{T} \ell_t^2 \cdot p_t.$$

Rearranging:

$$\sum_{t=1}^{T} \ell_t \cdot p_t - \sum_{t=1}^{T} \ell_t \cdot p \leq \frac{\log N}{\epsilon} + \epsilon \sum_{t=1}^{T} \ell_t^2 \cdot p_t.$$

Since the above holds for any $p \in \Delta^N$, we conclude:

$$\text{Regret} \leq \frac{\log N}{\epsilon} + \epsilon \sum_{t=1}^{T} \ell_t^2 \cdot p_t.$$

The regret of the Hedge algorithm is defined as:

$$\text{Regret} = \sum_{t=1}^{T} \ell_t \cdot p_t - \min_{p \in \Delta^N} \sum_{t=1}^{T} \ell_t \cdot p,$$

and under the assumptions $\ell_{ti} \in [-1, 1]$, the regret bound becomes:

$$\text{Regret} \leq \frac{\log N}{\epsilon} + \epsilon \sum_{t=1}^{T} \sum_{i=1}^{N} p_t(i) \ell_{ti}^2.$$

Since $\ell_{ti} \in [-1, 1]$, we have $\ell_{ti}^2 \leq 1$, and therefore:

$$\sum_{i=1}^{N} p_t(i) \ell_{ti}^2 \leq 1.$$

Thus, the regret bound simplifies to:

$$\text{Regret} \leq \frac{\log N}{\epsilon} + \epsilon T.$$

Now, compare this with the regret bound of the Randomized Weighted Majority (RWM) algorithm, which deals with binary losses $\ell_{ti} \in \{0, 1\}$ and has regret:

$$\text{Regret}_{\text{RWM}} \leq 2\sqrt{T \log N}.$$

To match this bound with Hedge, choose $\epsilon = \sqrt{\frac{\log N}{T}}$, which gives:

$$\text{Regret}_{\text{Hedge}} \leq \sqrt{T \log N} + \sqrt{T \log N} = 2\sqrt{T \log N}.$$

Conclusion:

Both Hedge and RWM algorithms achieve the same regret bound of:

$$\mathcal{O}(\sqrt{T \log N}),$$

when the learning rate ϵ is optimally chosen. However, Hedge is more general:

- Hedge handles real-valued losses in $[-1, 1]$,
- RWM is restricted to binary losses $\{0, 1\}$.

Therefore, Hedge provides a broader and more flexible framework for online learning scenarios.

another way to proof: We begin by bounding S_{t+1} from above:

$$S_{t+1} = \sum_{i=1}^{N} w_{t+1}(i) = \sum_{i=1}^{N} w_t(i) e^{-\epsilon \ell_t(i)}$$

like part a Using the convexity of the exponential and the fact that $\sum_i p_t(i) = 1$, we apply the inequality $e^{-x} \leq 1 - x + \frac{x^2}{2}$ (for $x \in [-1, 1]$):

$$S_{t+1} \leq \sum_{i=1}^{N} w_t(i) \left(1 - \epsilon \ell_t(i) + \frac{\epsilon^2}{2} \ell_t(i)^2\right)$$

$$= S_t \left(1 - \epsilon \sum_{i=1}^{N} \frac{w_t(i)}{S_t} \ell_t(i) + \frac{\epsilon^2}{2} \sum_{i=1}^{N} \frac{w_t(i)}{S_t} \ell_t(i)^2\right)$$

$$= S_t \left(1 - \epsilon \mathbf{p}_t^\top \boldsymbol{\ell}_t + \frac{\epsilon^2}{2} \mathbf{p}_t^\top \boldsymbol{\ell}_t^2\right)$$

Iterating this bound over $t = 1$ to T, we get:

$$S_{T+1} \leq S_1 \cdot \exp\left(-\epsilon \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t + \frac{\epsilon^2}{2} \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t^2\right)$$

Since $S_1 = \sum_{i=1}^{N} w_1(i) = N$, we have:

$$S_{T+1} \leq N \cdot \exp\left(-\epsilon \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t + \frac{\epsilon^2}{2} \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t^2\right)$$

Fix any expert $i \in \{1, \ldots, N\}$. Then:

$$w_{T+1}(i) = \exp\left(-\epsilon \sum_{t=1}^{T} \ell_t(i)\right) \quad \Rightarrow \quad S_{T+1} \geq w_{T+1}(i) = \exp\left(-\epsilon \sum_{t=1}^{T} \ell_t(i)\right)$$

Combining upper and lower bounds on S_{T+1}:

$$\exp\left(-\epsilon \sum_{t=1}^{T} \ell_t(i)\right) \leq N \cdot \exp\left(-\epsilon \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t + \frac{\epsilon^2}{2} \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t^2\right)$$

Taking logs on both sides:

$$-\epsilon \sum_{t=1}^{T} \ell_t(i) \leq \log N - \epsilon \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t + \frac{\epsilon^2}{2} \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t^2$$

Rearranging:

$$\sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t \leq \sum_{t=1}^{T} \ell_t(i) + \frac{\epsilon}{2} \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t^2 + \frac{\log N}{\epsilon}$$

Since this holds for any expert i, it also holds for the best expert in hindsight:

$$\sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t \leq \min_i \sum_{t=1}^{T} \ell_t(i) + \frac{\epsilon}{2} \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t^2 + \frac{\log N}{\epsilon}$$

Therefore, the regret is bounded as:

$$\text{Regret}_T = \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t - \min_i \sum_{t=1}^{T} \ell_t(i) \leq \frac{\log N}{\epsilon} + \frac{\epsilon}{2} \sum_{t=1}^{T} \mathbf{p}_t^\top \boldsymbol{\ell}_t^2$$

3.2.3 c)[2-points]

From part (b), we have the following regret bound for the Hedge algorithm:

$$\text{Regret} \leq \frac{\log N}{\epsilon} + \epsilon \sum_{t=1}^{T} \ell_t^2 \cdot p_t.$$

Since the losses are bounded in $[-1, 1]$, we know that $\ell_{ti}^2 \leq 1$, so:

$$\ell_t^2 \cdot p_t = \sum_{i=1}^{N} p_t(i) \ell_{ti}^2 \leq \sum_{i=1}^{N} p_t(i) = 1.$$

Thus, we can upper bound the second term:

$$\sum_{t=1}^{T} \ell_t^2 \cdot p_t \leq T.$$

Substituting into the regret bound gives:

$$\text{Regret} \leq \frac{\log N}{\epsilon} + \epsilon T.$$

Now, choose ϵ to minimize the right-hand side. Setting: We aim to minimize this expression with respect to ϵ. Define:

$$f(\epsilon) = \frac{a}{\epsilon} + b\epsilon, \quad \text{where } a = \log N,\ b = T.$$

To find the minimum value of $f(\epsilon)$, we compute the derivative:

$$f'(\epsilon) = -\frac{a}{\epsilon^2} + b.$$

Setting $f'(\epsilon) = 0$ gives:

$$-\frac{a}{\epsilon^2} + b = 0 \quad \Rightarrow \quad \epsilon^2 = \frac{a}{b} \quad \Rightarrow \quad \epsilon = \sqrt{\frac{\log N}{T}}.$$

we substitute this value into the expression:

$$\text{Regret} \leq \frac{\log N}{\sqrt{\frac{\log N}{T}}} + \sqrt{\frac{\log N}{T}} \cdot T = \sqrt{T \log N} + \sqrt{T \log N} = 2\sqrt{T \log N}.$$

Therefore, we conclude:

$$\text{Regret} \leq 2\sqrt{T \log N}.$$

More generally, for any $x > 0$, the following inequality holds due to the Arithmetic Mean–Geometric Mean (AM–GM) inequality:

$$\frac{a}{x} + bx \geq 2\sqrt{ab},$$

with equality if and only if $x = \sqrt{\frac{a}{b}}$.

Thus, for the Hedge regret bound:

$$\text{Regret} \leq \frac{\log N}{\epsilon} + \epsilon T \geq 2\sqrt{T \log N},$$

and the tightest bound is achieved by choosing $\epsilon = \sqrt{\frac{\log N}{T}}$.
