Reinforcement Learning

Assignment 1 (Theoretical Questions)

Group Members

Matheus da Silva Araujo – ID!!!

Miguel Ángel Carrillo – ID!!!

January 26, 2026

Problem Statement

Application of the Lai–Robbins bound. The asymptotic lower bound on the total regret L_T for any consistent bandit algorithm is given by the Lai–Robbins bound:

$$\liminf_{T\to\infty} \frac{\mathbb{E}[L_T]}{\ln T} \geq \sum_{a:\,\Delta_a>0} \frac{\Delta_a}{D_{\mathrm{KL}}(P_a \,\|\, P_*)},$$

where D_{KL} is the Kullback–Leibler divergence between the distribution of a suboptimal arm a (P_a) and the optimal arm (P_*), and Δ_a is the gap in expected reward between the optimal arm and arm a.

Question 1

Derive the explicit formula for the KL-divergence between two Bernoulli distributions with parameters p and q:

$$D_{\mathrm{KL}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)).$$

Derivation:

General expression for KL-divergence between distributions $r(x)$ and $s(x)$ with discrete random variables:

$$D_{\mathrm{KL}}(r(x) \,\|\, s(x)) = \sum_{x\in X} r(x) \log\left(\frac{r(x)}{s(x)}\right) \tag{1}$$

Expression for a Bernoulli distribution with parameter p:

$$P(X=x) = \begin{cases} 1-p, & \text{if } X=0 \\ p, & \text{if } X=1 \end{cases}$$

Combining both expressions:

$$\begin{aligned}
D_{\mathrm{KL}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)) &= \sum_{x \in X = \{0,1\}} P(X=x) \log\left(\frac{P(X=x)}{Q(X=x)}\right) \\
&= P(X=0) \log\left(\frac{P(X=0)}{Q(X=0)}\right) + P(X=1) \log\left(\frac{P(X=1)}{Q(X=1)}\right) \\
&= (1-p) \log\left(\frac{1-p}{1-q}\right) + p \log\left(\frac{p}{q}\right).
\end{aligned}$$

Final Answer

$$D_{\mathrm{KL}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)) = (1-p) \log\left(\frac{1-p}{1-q}\right) + p \log\left(\frac{p}{q}\right).$$

Question 2

Same question for two Gaussian distributions sharing the same variance.

Derivation:

General expression for Gaussian distribution with variance σ^2 and mean μ:

$$\mathcal{N}(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Assuming without loss of generality that the two Gaussian distributions have different means, $P(X, \mu_1, \sigma), Q(X, \mu_2, \sigma)$.

General expression for KL-divergence between distributions $R(x)$ and $S(x)$ with continuous random variables:

$$D_{\mathrm{KL}}(R(x) \,\|\, S(x)) = \int_{-\infty}^{+\infty} R(x) \log\left(\frac{R(x)}{S(x)}\right) dx = \mathbb{E}_{X \sim R}\left[\log\left(\frac{R(X)}{S(X)}\right)\right] \tag{2}$$

,

where $\mathbb{E}$ is the expected value.

Analyzing the log-term isolated first:

$$\begin{aligned}
\log \frac{P(x)}{Q(x)} &= \log P(x) - \log Q(x) \\
&= \log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)\right] - \log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)\right] \\
&= \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)\right)\right) - \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)\right)\right) \\
&= \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x-\mu_1)^2}{2\sigma^2}\right) - \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x-\mu_2)^2}{2\sigma^2}\right) \\
&= \frac{(x-\mu_2)^2 - (x-\mu_1)^2}{2\sigma^2}
\end{aligned}$$

$$\rightarrow \log \frac{P(x)}{Q(x)} = \frac{(x-\mu_2)^2 - (x-\mu_1)^2}{2\sigma^2} \tag{3}$$

Analyzing the numerator of Equation 3:

$$\begin{aligned}(x-\mu_2)^2-(x-\mu_1)^2 &= (x^2-2x\mu_2+\mu_2^2)-(x^2-2x\mu_1+\mu_1^2)\\ &= 2x(\mu_1-\mu_2)+\mu_2^2-\mu_1^2\\ &= 2x(\mu_1-\mu_2)+(\mu_2-\mu_1)(\mu_2+\mu_1)\\ &= (\mu_1-\mu_2)(2x-\mu_2-\mu_1)\end{aligned}$$

$$\rightarrow (x-\mu_2)^2-(x-\mu_1)^2 = (\mu_1-\mu_2)(2x-\mu_2-\mu_1) \tag{4}$$

Substituting Equations 4 into the numerator of 3:

$$\log\frac{P(x)}{Q(x)} = \frac{(\mu_1-\mu_2)(2x-\mu_1-\mu_2)}{2\sigma^2} \tag{5}$$

Then, from the expected value definition of KL-divergence and Equation 5, it follows that

$$\begin{aligned}D_{\mathrm{KL}}(P(x)\,\|\,Q(x)) &= \mathbb{E}_{X\sim P}\left[\log\left(\frac{P(X)}{Q(X)}\right)\right]\\ &= \mathbb{E}_{X\sim P}\left[\frac{(\mu_1-\mu_2)(2X-\mu_1-\mu_2)}{2\sigma^2}\right]\\ &= \frac{(\mu_1-\mu_2)}{2\sigma^2}\mathbb{E}_{X\sim P}\left[2X-\mu_1-\mu_2\right].\end{aligned} \tag{6}$$

Finally, from the linearity of expectation ($\mathbb{E}\left[aX+b\right]=a\mathbb{E}\,]X]+b$) and the information that $P(X)$ is a Gaussian distribution (implying that $\mathbb{E}_{X\sim P}=\mu_1$), applied to Equation 6:

$$\begin{aligned}&D_{\mathrm{KL}}(P(x)\,\|\,Q(x))\\ &\qquad= \frac{(\mu_1-\mu_2)}{2\sigma^2}\mathbb{E}_{X\sim P}\left[(2X-\mu_1-\mu_2)\right]\\ &\qquad= \frac{(\mu_1-\mu_2)}{2\sigma^2}\left(2\mu_1-\mu_1-\mu_2\right)\\ &\qquad= \frac{(\mu_1-\mu_2)}{2\sigma^2}\left(\mu_1-\mu_2\right)\\ &\qquad= \frac{(\mu_1-\mu_2)^2}{2\sigma^2}\end{aligned}$$

$$\rightarrow D_{\mathrm{KL}}(\mathrm{P}(\mathrm{X},\mu_1,\sigma)(X)\,\|\,\mathrm{Q}(\mathrm{X},\mu_2,\sigma)(X)) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2}$$

Final Answer

$$D_{\mathrm{KL}}(\mathrm{P}(\mathrm{X},\mu_1,\sigma)(X)\,\|\,\mathrm{Q}(\mathrm{X},\mu_2,\sigma)(X)) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2} \tag{7}$$

Question 3

Show that for the Bernoulli bandit, it is "easier" (i.e., theoretically implies lower regret) to distinguish an arm with mean $p = 0.9$ from an optimal arm with $p_* = 0.99$ than it is to distinguish an arm with $p = 0.55$ from an optimal arm with $p_* = 0.64$, even though the difference in means is identical ($\Delta = 0.09$) in both cases. What about the Gaussian case?

Answer

Bernoulli case:

From Question 1 final answer:

$$D_{\mathrm{KL}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)) = (1-p)\log\left(\frac{1-p}{1-q}\right) + p\log\left(\frac{p}{q}\right).$$

For $p = 0.9, p_* = 0.99$ using the distributions P_{a}, P_*:

$$D_{\mathrm{KL}}(\mathrm{P_a}(p) \,\|\, \mathrm{P}_*(p_*)) = D_{\mathrm{KL}}(\mathrm{P_a}(0.9) \,\|\, \mathrm{P}_*(0.99)) = (1-p)\log\left(\frac{1-p}{1-p_*}\right) + p\log\left(\frac{p}{p_*}\right) \approx 0.1445$$

By applying the previous value and $\Delta_{\mathrm{a}} = 0.09$ in the Lai-Robbins bound, it is obtained

$$\frac{\Delta_a}{D_{\mathrm{KL}}(P_a \,\|\, P_*)} \approx \frac{0.09}{0.1445} \approx 0.623 \tag{8}$$

Likewise, for $p = 0.55, p_* = 0.64$ using the distributions P_{a}, P_*:

$$D_{\mathrm{KL}}(\mathrm{P_a}(p) \,\|\, \mathrm{P}_*(p_*)) = D_{\mathrm{KL}}(\mathrm{P_a}(0.55) \,\|\, \mathrm{P}_*(0.64)) = (1-p)\log\left(\frac{1-p}{1-p_*}\right) + p\log\left(\frac{p}{p_*}\right) \approx 0.0171$$

Again, by applying the previous value and $\Delta_{\mathrm{a}} = 0.09$ in the Lai-Robbins bound, it is obtained

$$\frac{\Delta_a}{D_{\mathrm{KL}}(P_a \,\|\, P_*)} \approx \frac{0.09}{0.0171} \approx 5.26 \tag{9}$$

Comparing Equations 8 and 9, the conclusion is that the theoretical lower bound for the regret is smaller in the first case ($p = 0.9, p_* = 0.99$) than in the second case ($p = 0.55, p_* = 0.64$), which means that the first case is "easier".

Gaussian case:

From Question 2 final answer:

$$D_{\mathrm{KL}}(\mathrm{P}(\mathrm{X}, \mu_1, \sigma)(X) \,\|\, \mathrm{Q}(\mathrm{X}, \mu_2, \sigma)(X)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2} \tag{10}$$

For $p = \mu_1 = 0.9, p_* = \mu_2 = 0.99$ using the Gaussian distributions P_{a}, P_*:

$$D_{\mathrm{KL}}(\mathrm{P_a}(p,\sigma) \,\|\, \mathrm{P}_*(p_*,\sigma)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2} = \frac{(0.9 - 0.99)^2}{2\sigma^2} = \frac{0.09^2}{2\sigma^2}$$

For $p = \mu_1 = 0.55, p_* = \mu_2 = 0.64$ using the Gaussian distributions P_{a}, P_*:

$$D_{\mathrm{KL}}(\mathrm{P_a}(p,\sigma) \,\|\, \mathrm{P}_*(p_*,\sigma)) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2} = \frac{(0.55-0.64)^2}{2\sigma^2} = \frac{0.09^2}{2\sigma^2}$$

Since the gaps are the same and the KL-divergence is the same in both cases, this means that one does not have a theoretically lower bound than the other, meaning that one is not easier than the other.

Conclusion: the theoretical lower bound depends only on the gap and, as such, both cases have the same theoretical lower bound and are equally "easy".

Proof of simplified Lai-Robbins bound

Consider two Bernoulli arms:

- *Arm 1 has a known success probability* $p_1 = 0.5$.
- *Arm 2 has an unknown success probability* $p_2 = 0.5 + \Delta$.

You want to determine if Arm 2 is better than Arm 1 with high confidence. You collect n *samples from Arm 2 and compute the empirical mean* $\hat{p}_n$*. You decide Arm 2 is "Better" if* $\hat{p}_n > 0.5$.

1 Question 4

Suppose the truth is that Arm 2 is actually worse ($\Delta < 0$)*. Use Hoeffding's inequality to find an upper bound on the probability that you incorrectly classify it as "Better" (i.e.,* $P(\hat{p}_n > 0.5)$*) after* n *samples.*

Hoeffding's Inequality states that for independent identically distributed random variables $X_1, \ldots, X_n$ with expected value μ:

$$P(\bar{X}_n - \mu \geq t) \leq \exp\left(-2nt^2\right)$$

Let $X_1, \ldots, X_n$ be the n samples from Arm 2, where each $X_i \sim \mathrm{Ber}(p_2)$ with $p_2 = 0.5 + \Delta$ and $\Delta < 0$.

The empirical mean is:

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Since it is a Bernoulli distribution, the expected value $\mu = p_2$.

The question asks to bound the expression

$$P(\hat{p}_n > 0.5) = P(\hat{p}_n - p_2 > 0.5 - p_2)$$

Note that $p_2 = 0.5 + \Delta \rightarrow 0.5 - p_2 = -\Delta$

Combining the last two expressions, the required expression is equivalent to bound

$$P(\hat{p}_n - p_2 > -\Delta)$$

Note that $\Delta < 0 \rightarrow -\Delta > 0$, which is a deviation above the mean of the distribution.

Finally, set $t = -\Delta$ (which is positive because $\Delta < 0$) in the Hoeffdings Inequality to obtain:

$$P(\hat{p}_n > 0.5) = P(\hat{p}_n - p_2 > -\Delta) \leq \exp(-2n\Delta^2)$$

This is an upper bound on the probability required in the question.

Final Answer

$$\boxed{P(\hat{p}_n > 0.5) \leq \exp(-2n\Delta^2)}$$

Question 5

Set this error probability to be at most δ (e.g., $\delta = 1/T$). Rearrange your bounds to show that the number of samples n required to avoid this error must be at least:

$$n \geq \frac{\ln(1/\delta)}{2\Delta^2}.$$

In Question 4 final answer it was obtained:

$$P(\hat{p}_n > 0.5) \leq \exp(-2n\Delta^2)$$

The question asks that the probability of misclassification to be at most δ i.e.,

$$P(\hat{p}_n > 0.5) \leq \delta$$

Using this upper bound:

$$\begin{aligned} P(\hat{p}_n > 0.5) \leq \exp(-2n\Delta^2) &\leq \delta \\ \implies \quad -2n\Delta^2 &\leq \ln(\delta) \\ \implies \quad 2n\Delta^2 &\geq \ln(1/\delta) \\ \implies \quad n &\geq \frac{\ln(1/\delta)}{2\Delta^2}. \end{aligned}$$

Final Answer

$$\boxed{n \geq \frac{\ln(1/\delta)}{2\Delta^2}}$$

Question 6

For small gaps, $D_{\mathrm{KL}} \approx 2\Delta^2$. Substitute D_{KL} into the inequality above. Explain how this explains the Lai–Robbins term

$$\frac{\ln T}{D_{\mathrm{KL}}}.$$

From Question 4, it was determined that the number of samples needed is:

$$n \geq \frac{\ln(1/\delta)}{2\Delta^2}$$

Using the approximation for small gaps specified in the problem statement $D_{KL} \approx 2\Delta^2$ and substituting into the previous bound:

$$n \geq \frac{\ln(1/\delta)}{D_{KL}}$$

Setting $\delta = 1/T$ (the probability of error tolerated over T rounds):

$$n \geq \frac{\ln(T)}{D_{KL}}$$

This shows that, in order to distinguish a suboptimal Bernoulli arm from the optimal arm with error probability at most $1/T$, the suboptimal arm must be sampled at least on the order of

$$\frac{\ln T}{D_{\mathrm{KL}}(P_a \| P_*)}$$

times. Since each such sample incurs regret Δ, this directly explains the Lai–Robbins bound term

$$\frac{\Delta \ln T}{D_{\mathrm{KL}}(P_a \| P_*)}.$$