Quiz #1
DATA 37200: Learning, Decisions, and Limits (Winter 2025)

February 16, 2025

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.

Name: ______________________________

Mathematical facts (for your convenience)

This quiz is closed-note. But we provide here some standard facts which might be useful for the problems. (You can also ignore them if you don't need them.)

- Hoeffding's inequality: if $X_1, \dots, X_n$ are independent such that $|X_i| \le 1$ for all i, then

$$\mathbf{Pr}\left(\left|\sum_i X_i - \sum_i \mathbb{E}\, X_i\right| > t\sqrt{n}\right) \le 2e^{-t^2/2}.$$

- Azuma-Hoeffding: the above statement is also true (more generally) if $Y_t = \sum_{i=1}^t X_i$ is a martingale.
- Cauchy-Schwarz inequality: for any vectors a, b we have $|a \cdot b| \le \|a\|_2 \|b\|_2$ where $\|a\|_2 = \sqrt{\sum_i a_i^2}$ is the ℓ_2 norm.
- AM-GM inequality: for any numbers $a, b \in \mathbb{R}$, $|ab| \le a^2/2 + b^2/2$.
- Facts about Gaussians: if $Z \sim N(0,1)$, then $\mathbb{E}\, Z = 0, \mathbb{E}\, Z^2 = 1$, and $\mathbb{E}\, |Z| = \sqrt{2/\pi} < 1$.

Problem 1 (ETC Revisited, AS STATED, UNINTENDED)

Explaining the motivation: In class, we saw the Explore Then Commit algorithm (ETC) for the stochastic multi-armed bandit. For other stochastic settings like MDPs/RL and contextual bandits, it is also possible to apply this principal, with the caveat that the exploration procedure often needs to be more complex than in the bandit case.

We consider a general game between the Agent and Nature played over T rounds of the following structure. Each round t, the Agent selects a "policy" π_t, which is an element of a set called Π, based off their experience in previous rounds. Nature samples, independently of everything else, a random instance X_t which is an element of a set K and such that $X \sim \mathcal{D}$. (I.e. $X_1, \dots, X_T \sim \mathcal{D}$ are i.i.d. and $\mathcal{D}$ is a distribution over a set K). $r : \Pi \times K \to [0, 1]$ is a function which given a policy π and instance x, tells us the reward $r(\pi, x)$ gained in this round. To summarize, the game works as follows.

For $t = 1$ to T:

1. Player selects a policy $\pi_t \in \Pi$ based off their past experience.
2. Nature samples $X_t \sim \mathcal{D}$ independently of everything else.
3. Player gains (deterministic) reward $r(\pi_t, X_t)$.

Define $\pi^* \in \Pi$ to be the fixed policy with optimal expected reward, so

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{X' \sim \mathcal{D}}\, r(\pi, X').$$

Define the total reward over all rounds to be $\sum_{t=1}^{T} r(\pi_t, X_t)$ and the regret (compared to π^*, and as a random variable) is

$$r(\pi^*, X_t) - \sum_{t=1}^{T} r(\pi_t, X_t).$$

Suppose we already have an exploration strategy which after n rounds of play can achieve the following guarantee: if the player follows this exploration strategy for n rounds, they can then compute a policy $\widehat{\pi}_n$ which is a function of $X_1, \dots, X_n$ and the observed rewards, such that for some fixed $\alpha > 0$, it is guaranteed that

$$\mathbb{E}_{X_1,\dots,X_n} \mathbb{E}_{X' \sim \mathcal{D}}\, r(\widehat{\pi}_n, X') \geq \mathbb{E}_{X' \sim \mathcal{D}}\, r(\pi^*, X') - 1/n^{\alpha}$$

i.e. in expectation over the dataset, it performs well on average on a fresh training sample X'.

Prove that there exists $C_\alpha > 0$ (depending possibly on α but not on T), such that for all sufficiently large T,

$$\mathbb{E} \sum_{t=1}^{T} r(\pi^*, X_t) - \min_{n : 1 \leq n \leq T} \mathbb{E}\left[n + \sum_{t=n+1}^{T} r(\widehat{\pi}_n, X_t) \right] \leq C_\alpha T^{1/(1+\alpha)}$$

(Motivation/interpretation: the left hand side is an upper bound on the expected regret no matter what our reward is in the first n rounds, if we switch to always playing $\widehat{\pi}_n$ afterward.)

PROBLEM (as pointed out by a student): It's easy to see the minimum is attained at $n = 1$ so the inequality can fail to hold. The problem is that while this is an upper bound, it is not a very good one.

Problem 1 (ETC Revisited, AS INTENDED)

Explaining the motivation: In class, we saw the Explore Then Commit algorithm (ETC) for the stochastic multi-armed bandit. For other stochastic settings like MDPs/RL and contextual bandits, it is also possible to apply this principal, with the caveat that the exploration procedure often needs to be more complex than in the bandit case.

We consider a general game between the Agent and Nature played over T rounds of the following structure. Each round t, the Agent selects a "policy" π_t, which is an element of a set called Π, based off their experience in previous rounds. Nature samples, independently of everything else, a random instance X_t which is an element of a set K and such that $X \sim \mathcal{D}$. (I.e. $X_1, \dots, X_T \sim \mathcal{D}$ are i.i.d. and $\mathcal{D}$ is a distribution over a set K). $r : \Pi \times K \to [0, 1]$ is a function which given a policy π and instance x, tells us the reward $r(\pi, x)$ gained in this round. To summarize, the game works as follows.

For $t = 1$ to T:

1. Player selects a policy $\pi_t \in \Pi$ based off their past experience.
2. Nature samples $X_t \sim \mathcal{D}$ independently of everything else.
3. Player gains (deterministic) reward $r(\pi_t, X_t)$.

Define $\pi^* \in \Pi$ to be the fixed policy with optimal expected reward, so

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{X' \sim \mathcal{D}}\, r(\pi, X').$$

Define the total reward over all rounds to be $\sum_{t=1}^{T} r(\pi_t, X_t)$ and the regret (compared to π^*, and as a random variable) is

$$\sum_{t} r(\pi^*, X_t) - \sum_{t=1}^{T} r(\pi_t, X_t).$$

Suppose we already have an exploration strategy which after n rounds of play can achieve the following guarantee: if the player follows this exploration strategy for n rounds, they can then compute a policy $\widehat{\pi}_n$ which is a function of $X_1, \dots, X_n$ and the observed rewards, such that for some fixed $\alpha > 0$, it is guaranteed that

$$\mathbb{E}_{X_1, \dots, X_n} \mathbb{E}_{X' \sim \mathcal{D}}\, r(\widehat{\pi}_n, X') \geq \mathbb{E}_{X' \sim \mathcal{D}}\, r(\pi^*, X') - 1/n^{\alpha}$$

i.e. in expectation over the dataset, it performs well on average on a fresh training sample X'.

Prove that there exists $C_\alpha > 0$ (depending possibly on α but not on T), such that for all sufficiently large T,

$$\mathbb{E} \sum_{t=1}^{T} r(\pi^*, X_t) - \max_{n : 1 \leq n \leq T} \mathbb{E}\left[\sum_{t=n+1}^{T} r(\widehat{\pi}_n, X_t)\right] \leq C_\alpha T^{1/(1+\alpha)}$$

(Motivation/interpretation: the left hand side is an upper bound on the expected regret no matter what our reward is in the first n rounds, if we switch to always playing $\widehat{\pi}_n$ afterward.)

Problem 1 Solution (AS INTENDED)

First, we analyze the upper bound on the expected regret after switching to $\widehat{\pi}$ in the nth round. Using our lower bound on the reward of $\widehat{\pi}$ ($\mathbb{E}[r(\widehat{\pi}, X')] \geq \mathbb{E}[r(\pi^*, X')] - 1/n^\alpha$):

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} r(\pi^*, X_t)\right] - \max_{1\leq n\leq T} \mathbb{E}\left[\sum_{t=n+1}^{T} r(\widehat{\pi}, X_t)\right] \\
&\leq \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} r(\pi^*, X_t)}_{\text{cancel last } [n+1, T] \text{ terms}}\right] - \max_{1\leq n\leq T} \mathbb{E}\left[\sum_{t=n+1}^{T}\left(r(\pi^*, X_t) - \frac{1}{n^\alpha}\right)\right] \\
&= \max_{1\leq n\leq T}\left(\mathbb{E}\left[\sum_{t=1}^{n} r(\pi^*, X_t)\right] - \mathbb{E}\left[\sum_{t=n+1}^{T}\frac{1}{n^\alpha}\right]\right) \\
&= \max_{1\leq n\leq T} \mathbb{E}\left[\sum_{t=1}^{n}\underbrace{r(\pi^*, X_t)}_{r\leq 1, \forall t} - \left[-\frac{T-n}{n^\alpha}\right]\right] \\
&\leq \max_{1\leq n\leq T} \mathbb{E}\left[n + \frac{T}{n^\alpha}\right]
\end{aligned}$$

Now, we solve for the maximizing n^* by setting the derivative with respect to n to 0:

$$\nabla\left(n + \frac{T}{n^\alpha}\right) = 1 - \alpha\frac{T}{n^{\alpha+1}} = 0 \implies n^* = (\alpha T)^{\frac{1}{\alpha+1}}$$

Plugging n^* back into the upper bound:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} r(\pi^*, X_t)\right] - \max_{1\leq n\leq T} \mathbb{E}\left[\sum_{t=n+1}^{T} r(\widehat{\pi}, X_t)\right] &\leq \max_{1\leq n\leq T} \mathbb{E}\left[n + \frac{T}{n^\alpha}\right] \\
&\leq (\alpha T)^{\frac{1}{\alpha+1}} + \frac{T}{(\alpha T)^{\frac{\alpha}{\alpha+1}}} \\
&\leq C_\alpha T^{\frac{1}{\alpha+1}}
\end{aligned}$$

As desired. □

Problem 2 (iterated coin flipping game)

We consider the following (online) gambling game which occurs over T rounds, where a gambler is betting on the outcome of random coin flips.

For $t = 1$ to T:

1. Gambler predicts a number $a_t \in [-1, 1]$. (They are allowed to base their guess on everything that happened in the past, i.e. before time t, but not on anything in the future.)

2. Nature samples $Y_t \in Uni\{\pm 1\}$ independently of everything that happened before. (I.e. Nature flips a fair coin which gives us $+1$ with probability 50 percent and -1 with probability 50 percent.)

3. Gambler suffers loss $(Y_t - a_t)^2$.

We will consider analyzing the regret of the gambler with various strategies. (Here a strategy means an algorithm by which the gambler selects a_t in each round. In other contexts like RL, this would be called a policy.) Define the regret R (as a random variable) by

$$R = \sum_{t=1}^{T} (Y_t - a_t)^2 - \min_{a \in [-1,1]} \sum_{t=1}^{T} (Y_t - a)^2.$$

This measures the regret of the Gambler compared to offline strategies which always play the same number a at every step.

Answer the following questions:

1. Prove that the regret admits the following alternative formula:

$$R = \sum_{t=1}^{T} (a_t^2 - 2Y_t a_t) - \min_{a \in [-1,1]} \left(a^2 T - 2a \sum_{t=1}^{T} Y_t \right)$$

2. Let

$$a^* = \arg \min_{a \in [-1,1]} \left(a^2 T - 2a \sum_{t=1}^{T} Y_t \right).$$

   Write down a closed form solution for a^* in terms of $Y_1, \ldots, Y_T$. Compute the total expected loss of a^*, i.e.

$$\mathbb{E} \left[\min_{a \in [-1,1]} \left(a^2 T - 2a \sum_{t=1}^{T} Y_t \right) \right]$$

3. Specify a strategy for the gambler which minimizes their expected regret, compute its expected regret, and prove that it is optimal (no other strategy has strictly smaller expected regret).

Problem 2 Solutions

1. We can expand the regret term:

$$\begin{aligned}
R &= \sum_{t=1}^{T}(Y_t - a_t)^2 - \min_{a\in[-1,1]} \sum_{t=1}^{T}(Y_t - a)^2 \\
&= \sum_{t=1}^{T}\left(Y_t^2 - 2Y_t a_t + a_t^2\right) - \min_{a\in[-1,1]} \sum_{t=1}^{T}\left(Y_t^2 - 2Y_t a + a^2\right) \\
&= \sum_{t=1}^{T}\left(Y_t^2 - 2Y_t a_t + a_t^2\right) - \sum_{t=1}^{T} Y_t^2 - \min_{a\in[-1,1]} \sum_{t=1}^{T}\left(a^2 - 2Y_t a\right) \qquad (1) \\
&= \sum_{t=1}^{T}\left(a_t^2 - 2Y_t a_t\right) - \min_{a\in[-1,1]} \left(a^2 T - 2a\sum_{t=1}^{T} Y_t\right) \qquad (2)
\end{aligned}$$

We extract the $\sum_t Y_t$ term out of the $\min$ in line 1 since it does not depend on a. Then, we can factor the a and the $2a$ terms in line 2 to obtain the desired form. □

2. To find the minimizing a^*, set the derivative with respect to a equal to 0:

$$\begin{aligned}
&\nabla\left(a^2 T - 2a\sum_{t=1}^{T} Y_t\right) = 0 \\
&\implies 2a^*T - 2\sum_{t=1}^{T} Y_t = 0 \\
&\implies 2a^*T = 2\sum_{t=1}^{T} Y_t \implies a^* = \frac{\sum_{t=1}^{T} Y_t}{T}
\end{aligned}$$

To find the expected loss of a^*, simply plug it back into the original expression and take the expectation:

$$\begin{aligned}
\mathbb{E}\left[(a^*)^2 T - 2a^*\sum_{t=1}^{T} Y_t\right] &= \mathbb{E}\left[\frac{(\sum_{t=1}^{T} Y_t)^2}{T} - \frac{2(\sum_{t=1}^{T} Y_t)^2}{T}\right] \\
&= \mathbb{E}\left[-\frac{(\sum_{t=1}^{T} Y_t)^2}{T}\right] \\
&= -\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \overbrace{(Y_t)^2}^{\text{always 1}} + \sum_{t_1\neq t_2} Y_{t_1} Y_{t_t}\right] \\
&= -\frac{1}{T}\left(T + \underbrace{\mathbb{E}\left[\sum_{t_1\neq t_2} Y_{t_1} Y_{t_t}\right]}_{\text{Expected value is 0}}\right) = -1 \qquad (3)
\end{aligned}$$

Note that the final expectation in line 3 is always zero by case work: $Y_{t_1}Y_{t_2} = 1$ when $Y_1 = Y_2$ (combined 50% probability) and $Y_{t_1}Y_{t_2} = -1$ otherwise. □

3. First, we expand the expected regret term (substituting in for the optimal a^* from part 2):

$$\begin{aligned}
\mathbb{E}[R] &= \mathbb{E}\left[\sum_{t=1}^{T}\left(a_t^2 - 2Y_t a_t\right) - \min_{a\in[-1,1]}\left(a^2T - 2a\sum_{t=1}^{T} Y_t\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left(a_t^2 - 2Y_t a_t\right) - (-1) \\
&= \sum_{t=1}^{T} \mathbb{E}[a_t^2] - \underbrace{\sum_{t=1}^{T} \mathbb{E}[2Y_t a_t]}_{\text{Extract } 2a_t \text{ from each term}} +1 \\
&= \sum_{t=1}^{T} \mathbb{E}[a_t^2] - \sum_{t=1}^{T} 2a_t \underbrace{\mathbb{E}[Y_t]}_{0} +1 \\
&= \sum_{t=1}^{T} a_t^2 + 1
\end{aligned}$$

So expected regret is clearly minimized when $a_t = 0, \forall t \in [T]$, which defines our optimal strategy. □

Problem 3 (random walk game)

We study the following game, which is similar but different to the previous game in Problem 2.

For $t = 1$ to T:

1. Gambler predicts a number $a_t \in [-1, 1]$. (They are allowed to base their guess on everything that happened in the past, i.e. before time t, but not on anything in the future.)

2. Nature samples $X_t \sim N(0, 1)$ independently of everything that happened before.

3. Gambler suffers loss $a_t X_t$. (Note: the "loss" can be negative in this game.)

Similar to the previous problem, the regret R is defined as

$$R = \sum_{t=1}^{T} a_T X_t - \min_{a \in [-1,1]} \sum_{t=1}^{T} a X_t$$

Answer the following questions:

1. Prove that the expected regret of the Gambler is the same no matter what strategy they choose.

2. Compute the expected regret of the Gambler.

(problem 3 answer space)

Problem 3 Solutions

1. We take the expected value of the regret term.

$$\begin{aligned}
\mathbb{E}[R] &= \mathbb{E}\left[\sum_{t=1}^{T} a_t X_t - \min_{a\in[-1,1]} \sum_{t=1}^{T} a X_t\right] \\
&= \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[a_t X_t\right]}_{X_t \text{ is independent}} - \mathbb{E}\left[\min_{a\in[-1,1]} \sum_{t=1}^{T} a X_t\right] \\
&= \sum_{t=1}^{T} \underbrace{\mathbb{E}[a_t]\mathbb{E}[X_t]}_{0} - \mathbb{E}\left[\min_{a\in[-1,1]} \sum_{t=1}^{T} a X_t\right] \\
&= -\mathbb{E}\left[\min_{a\in[-1,1]} \sum_{t=1}^{T} a X_t\right]
\end{aligned}$$

So we can see that the expected regret term is independent of a_t, so all strategies lead to the same expected regret. □

2. First, we reason about the regret term:

$$\mathbb{E}[R] = -\mathbb{E}\left[\min_{a\in[-1,1]} \sum_{t=1}^{T} a X_t\right] = -\mathbb{E}\left[\min_{a\in[-1,1]} a \sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[\max_{a\in[-1,1]} -a \sum_{t=1}^{T} X_t\right]$$

The regret is linear in a, we know that only setting a to the extremes (-1 or 1) will maximize the $-a\sum_t X_t$ term. Specifically, if $\sum_t X_t < 0$, then setting $a = 1$ will maximize the term, and if $\sum_t X_t > 0$, setting $a = -1$ will maximize the term. In both cases $-a\sum_t X_t = |\sum_t X_t|$. So, we can reason that:

$$\mathbb{E}[R] = \mathbb{E}\left[\max_{a\in[-1,1]} -a \sum_{t=1}^{T} X_t\right] = E\left[\left|\overbrace{\sum_{t=1}^{T} X_t}^{\sim N(0,T)}\right|\right] = \sqrt{T}\sqrt{\frac{2}{\pi}}$$

Where we use the fact that the $\mathbb{E}[|X|] = \sigma\sqrt{\frac{2}{\pi}}$ when X is drawn from $N(0, \sigma^2)$. □