Information and Coding Theory **Winter 2021**

Lecture 5: January 26, 2021

Lecturer: Madhur Tulsiani

1 Convexity of KL-divergence

Before we consider applications, let us prove an important property of KL-divergence. We prove below that $D(P \parallel Q)$, when viewed as a function of the inputs P and Q, is jointly convext in both it's inputs i.e., it is convex in the input (P, Q) when viewed as a tuple.

Proposition 1.1. *Let P_1, P_2, Q_1, Q_2 be distributions on a finite universe $\mathcal{X}$, and let $\alpha \in [0,1]$. Then,*

$$D\left(\alpha \cdot P_1 + (1-\alpha) \cdot P_2 \parallel \alpha \cdot Q_1 + (1-\alpha) \cdot Q_2\right) \;\le\; \alpha \cdot D\left(P_1 \parallel Q_1\right) + (1-\alpha) \cdot D\left(P_2 \parallel Q_2\right) .$$

Proof: For this proof, we will use an inequality called the log-sum inequality, the proof of which is left is an exercise. The inequality states that for $a_1, a_2, b_1, b_2 \ge 0$

$$(a_1 + a_2) \cdot \log\left(\frac{a_1 + a_2}{b_1 + b_2}\right) \;\le\; a_1 \cdot \log\left(\frac{a_1}{b_1}\right) + a_2 \cdot \log\left(\frac{a_2}{b_2}\right)$$

Using the above inequality, we can bound the LHS as

$$\begin{aligned}
& D\left(\alpha \cdot P_1 + (1-\alpha) \cdot P_2 \parallel \alpha \cdot Q_1 + (1-\alpha) \cdot Q_2\right) \\
= \;& \sum_{x \in \mathcal{X}} \left(\alpha \cdot p_1(x) + (1-\alpha) \cdot p_2(x)\right) \cdot \log\left(\frac{\alpha \cdot p_1(x) + (1-\alpha) \cdot p_2(x)}{\alpha \cdot q_1(x) + (1-\alpha) \cdot q_2(x)}\right) \\
\le \;& \sum_{x \in \mathcal{X}} \alpha \cdot p_1(x) \cdot \log\left(\frac{\alpha \cdot p_1(x)}{\alpha \cdot q_1(x)}\right) + (1-\alpha) \cdot p_2(x) \cdot \log\left(\frac{(1-\alpha) \cdot p_2(x)}{(1-\alpha) \cdot q_2(x)}\right) \\
= \;& \alpha \cdot D\left(P_1 \parallel Q_1\right) + (1-\alpha) \cdot D\left(P_2 \parallel Q_2\right) .
\end{aligned}$$

■

Exercise 1.2 (Log-sum inequality). *Prove that for $a_1, a_2, b_1, b_2 \ge 0$*

$$(a_1 + a_2) \cdot \log\left(\frac{a_1 + a_2}{b_1 + b_2}\right) \;\le\; a_1 \cdot \log\left(\frac{a_1}{b_1}\right) + a_2 \cdot \log\left(\frac{a_2}{b_2}\right) .$$

2 Distinguishing two coins

We will now use Pinsker's inequality to derive a lower bound on the number of samples needed to distinguish two coins with slightly differing biases. You can use Chernoff bounds to see that this bound is optimal. The optimality will also follow from a much more general result known as Sanov's theorem which we will derive later. Suppose we are given one of the following two coins (think of 1 as "heads" and 0 as "tails"):

$$P = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases} \qquad \text{and} \qquad Q = \begin{cases} 1 & \text{w.p. } \frac{1}{2} + \varepsilon \\ 0 & \text{w.p. } \frac{1}{2} + \varepsilon \end{cases}$$

Suppose we have an algorithm $T(x_1, x_2, ...x_n) \to \{0, 1\}$ that takes the output of n independent coin tosses, and makes a decision about which coin the tosses came from. Suppose that T outputs 0 to indicate the coin with distribution P and 1 to indicate the coin with distribution Q. Let us say that T identifies both coins with probability at least $9/10$, i.e.,

$$\mathop{\mathbb{P}}_{x \in P^n} [T(x) = 0] \;\geq\; \frac{9}{10} \quad \text{and} \quad \mathop{\mathbb{P}}_{x \in Q^n} [T(x) = 1] \;\geq\; \frac{9}{10}$$

The goal is to derive a lower bound for n. We will be able to derive a lower bound without knowing anything about T. We first rewrite the above conditions as

$$\mathop{\mathbb{E}}_{x \in P^n} [T(x)] \;\leq\; \frac{1}{10} \quad \text{and} \quad \mathop{\mathbb{E}}_{x \in Q^n} [T(x)] \;\geq\; \frac{9}{10},$$

which gives

$$\mathop{\mathbb{E}}_{x \in Q^n} [T(x)] - \mathop{\mathbb{E}}_{x \in P^n} [T(x)] \;\geq\; \frac{8}{10} \quad \Rightarrow \quad \|P^n - Q^n\|_1 \;\geq\; \frac{8}{5},$$

using the fact that the total variation distance upper bounds the distinguishing probability of the best distinguisher. Using the chain rule for KL-divergence and Pinsker's inequality, we get

$$n \cdot D(P \,\|\, Q) \;=\; D(P^n \,\|\, Q^n) \;\geq\; \frac{1}{2 \ln 2} \cdot \left(\frac{8}{5}\right)^2 \quad \Rightarrow \quad n \;\geq\; \frac{1}{2 \ln 2 \cdot D(P \,\|\, Q)} \cdot \left(\frac{8}{5}\right)^2$$

Finally, it remains to give an upper bound on $D(P \,\|\, Q)$, which can be obtained by writing it out as

$$\begin{aligned}
D\left(P \parallel Q\right) &= \left(\frac{1}{2}\right) \cdot \log\left(\frac{1/2}{1/2+\varepsilon}\right) + \left(\frac{1}{2}\right) \cdot \log\left(\frac{1/2}{1/2-\varepsilon}\right) \\
&= \left(\frac{1}{2}\right) \cdot \log\left(\frac{1}{1-4\varepsilon^2}\right) \\
&= \frac{1}{2\ln 2} \cdot \ln\left(1 + \frac{4\varepsilon^2}{1-4\varepsilon^2}\right) \\
&\le \frac{1}{2\ln 2} \cdot \frac{4\varepsilon^2}{1-4\varepsilon^2} \le \frac{8\varepsilon^2}{2\ln 2} && \left(\text{using } 1+z \le e^z,\ \varepsilon \le \frac{1}{4}\right)
\end{aligned}$$

Plugging in this upper bound, we get

$$n \ \ge\ \frac{1}{2\ln 2 \cdot D(P\|Q)} \cdot \left(\frac{8}{5}\right)^2 \ \ge\ \frac{1}{8\varepsilon^2} \cdot \left(\frac{8}{5}\right)^2 \ \ge\ \frac{8}{25\varepsilon^2}\,.$$

Exercise 2.1. *Prove using Chernoff bounds that $O(1/\varepsilon^2)$ samples are enough to distinguish the two coins.*

Exercise 2.2. *How many samples are needed in the case when one coin comes up heads with probability $p = \varepsilon$ and the other with probability $q = 2\varepsilon$?*

Note that while in the above application, we chose to use $D\left(P \parallel Q\right)$ to bound $\|P-Q\|_1$, we could also have used $D\left(Q \parallel P\right)$ instead, since $\|P-Q\|_1$ is a symmetric distance function. You can check that in the above case, the two bounds are quite similar. In general, we can always use the stronger bound

$$\min\left\{D\left(P \parallel Q\right), D\left(Q \parallel P\right)\right\} \ \ge\ \frac{1}{2\ln 2} \cdot \|P-Q\|_1^2\,.$$

3 Lower bounds for bandit problems

Bandit problems are a common way of modeling decision making under partial information. The problem is specified in terms of K "arms", each of which generates a (possibly random) "reward" at time t. As the player, we get to make a choice $C_t \in [K]$ at time t for which arm to choose, and we get to see the reward generated by the arm C_t. After making choices for times $t = 1, \ldots, n$, we compare the reward earned, against the best arm in hindsight. Denoting the reward for arm i at time t by $X_{i,t}$, the goal is to optimize

$$\min_{C_1,\ldots,C_n}\left(\max_{i\in[K]} \sum_{t=1}^{n} X_{i,t} \ -\ \sum_{t=1}^{n} X_{C_t,t}\right)\,.$$

There are several variants of the problem, where are rewards for each arm at each time t are selected randomly from an unknown distribution, or chosen adversarially based on the players choices $C_1, \ldots, C_{t-1}$. Also, while comparison with a single arm is the most common way to define regret, one can consider other models where the comparison is again "low complexity" choices in hindsight. See the book by Cesa-Bianchi and Lugosi [CBL06] and the survey by Bubeck and Cesa-Bianchi [BCB12] for a detailed discussion of various models and results.

When the rewards are bounded (say) in the range $[0,1]$, known algorithms can achieve an upper bound of $O(\sqrt{nK \log K})$ even in the adversarial case. The lower bound construction we discuss below, which was given by Auer et al. [ACBFS02], yields a lower bound of $\Omega(\sqrt{nK})$ even in the case where the rewards are random and independent of the player's choices. The construction uses the following distribution: the rewards for one of the K arms, chosen uniformly at random at the beginning, are generated according to a biased coin, which is 1 with probability $1/2 + \varepsilon$ and 0 with probability $1/2 - \varepsilon$. The rewards for all other arms are chosen according to a fair coin, are $0/1$ with probability$1/2$ each. The authors prove that

$$\mathbb{E}\left[\text{regret}\right] \;\geq\; \varepsilon \cdot n \cdot \left(1 - \frac{1}{K} - c_0\sqrt{\frac{\varepsilon^2 \cdot n}{K}}\right) .$$

The result extends the bound we proved for distinguishing biased coins. One shows that with a small number of samples, not only is it not possible to distinguish the distributions, but any algorithm to guess the right distribution makes about as many errors as random guessing. The above expression can be interpreted in this way, thinking of the cost of each incorrect guess as ε, and the number of mistakes made by random guessing, as being $n\left(1 - \frac{1}{K}\right)$. The regret bound follows from the above estimate by choosing $\varepsilon^2 = \Theta(K/n)$.

We will analyze a toy version of the above problem, with $K = 2$. We will argue that any guessing algorithm must make a mistake on about 1/2 the guesses. For the case of $K = 2$, we will prove that

$$\mathbb{E}\left[\text{regret}\right] \;\geq\; \varepsilon \cdot n \cdot \left(\frac{1}{2} - c_0 \cdot \sqrt{\varepsilon^2 \cdot n}\right) .$$

This captures many of the ideas in the proof of the general case, while avoiding some notational difficulties.

3.1 Lower bound for two-armed bandits

We consider a case with only two arms, labelled ℓ and r, with rewards $X_{\ell,t}$ and $X_{r,t}$ at time t. We consider a random variable H which equals ℓ or r with probability 1/2 each, and chooses the location of the biased coin (this is chosen and fixed at the beginning). If $H = \ell$, rewards are sampled according to the distribution P_ℓ, where each $X_{\ell,t}$ is independently 1 with probability $1/2 + \varepsilon$ and 0 with probability $1/2 - \varepsilon$, and each $X_{r,t}$ is 0/1 with probability $1/2$ each. Likewise, we denote by P_r the distribution in the case when $H = r$. In this case, each $X_{\ell,t}$ is $0/1$ with probability $1/2$ each, and the rewards $X_{r,t}$ are $0/1$ with probabilities $1/2 - \varepsilon$ and $1/2 + \varepsilon$.

We first change the goal from the analysis of the expected regret, to analyzing the deviation from the choice with the best *expected* reward. This quantity, which involves a switch of $\mathbb{E}[\max\{\cdot\}]$ and $\max\{\mathbb{E}[\cdot]\}$ is referred to as the "pseudo-regret".

Proposition 3.1. *For any player strategy, we have that*

$$\mathbb{E}[\mathit{regret}] \;\geq\; \left(\frac{1}{2} + \varepsilon\right) \cdot n - \mathbb{E}\left[\sum_{t=1}^{n} X_{C_t,t}\right] .$$

Proof: We have

$$\begin{aligned}
\mathbb{E}[\text{regret}] &= \mathbb{E}\left[\max_{i \in \{\ell,r\}} \sum_{t=1}^{n} X_{i,t} \;-\; \sum_{t=1}^{n} X_{C_t,t}\right] \\
&\geq \max_{i \in \{\ell,r\}} \cdot\, \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t}\right] \;-\; \mathbb{E}\left[\sum_{t=1}^{n} X_{C_t,t}\right] \\
&= \left(\frac{1}{2} + \varepsilon\right) \cdot n \;-\; \mathbb{E}\left[\sum_{t=1}^{n} X_{C_t,t}\right] ,
\end{aligned}$$

using the fact that the maximum of the two expectations is $\left(\frac{1}{2} + \varepsilon\right) \cdot n$. ■

Randomness and determinism. Note that if R_C denotes the randomness used in the strategy of the player, and R_X denotes the randomness in the rewards, then *using the fact that the distribution of all the rewards is independent of the player's strategy*, we can write the second term (expected reward) above as

$$\mathbb{E}\left[\sum_{t=1}^{n} X_{C_t,t}\right] \;=\; \mathop{\mathbb{E}}_{R_C} \mathop{\mathbb{E}}_{R_X} \left[\sum_{t=1}^{n} X_{C_t,t}\right] .$$

Thus, if the expectation is at least θ, then there exists a value of R_C which achieves a value of at least θ. Using the contrapositive, it suffices to prove an upper bound on the expected reward (and hence a lower bound on regret) only against deterministic strategies of the player. Note that this only means that the choices C_t are *deterministic functions* of the rewards seen by the player. Viewed in isolation, there is still randomness in the choice C_t, which comes from the randomness of the rewards $\{X_{\ell,1}, \ldots, X_{\ell,t-1}\}$ and $\{X_{r,1}, \ldots, X_{r,t-1}\}$

We can formalize this as follows. Let $\Omega = \{0,1\}^n \times \{0,1\}^n$ be the outcome space containing outcomes of all coin tosses (rewards) $\{X_{\ell,t}, X_{r,t}\}_{t \in [n]}$. Then each choice C_t is a function

$C_t : \Omega \to \{\ell, r\}$. We also define random variables $Z_t = X_{C_t,t}$ capturing the rewards actually seen by the player. Notice again that $Z_t : \Omega \to \{0,1\}$ is a function completely determined by the outcomes of the tosses. Since P_ℓ and P_r are distributions for all the coin tosses (rewards), we can also think of them as giving distributions for the views $Z_1, \ldots, Z_n$.

Another crucial property is that since the strategy of the player *only depends on the rewards they actually see*, we can actually think of each C_t as a deterministic function of the *values* of $Z_1, \ldots, Z_{t-1}$. This can seem confusing since it seems what we see depends on the choices C_t, but remember that we have already fixed a deterministic strategy for the player, and are only claiming that the choices C_t depend on the $0/1$ *values* seen by the player in the previous steps (these values may correspond to $X_{\ell,j}$ or $X_{r,j}$ depending on C_j). Given the fixed strategy, C_1 is already determined and we already know if the player is going to see the value of $X_{\ell,1}$ or $X_{r,1}$. Now, given this value Z_1, C_2 is determined, which determines if the player is going to see $X_{\ell,2}$ or $X_{r,2}$. Thus, given Z_1 and Z_2, C_3 is determined an so on.

We can now resume our analysis of the expected regret. We first relate it to the number of "wrong choices" where $C_t \neq H$.

Proposition 3.2.

$$\mathbb{E}\left[\textit{regret}\right] \;\geq\; \varepsilon \cdot \mathbb{E}\left[\left|\{t \mid C_t \neq H\}\right|\right] .$$

Proof: Note that $\mathbb{E}\left[X_{C_t,t}\right] = \left(\frac{1}{2} + \varepsilon\right)$ if $C_t = H$, and $\mathbb{E}\left[X_{C_t,t}\right] = \frac{1}{2}$ if $C_t \neq H$. Using Proposition 3.1, we have

$$\mathbb{E}\left[\text{regret}\right] \;\geq\; \left(\frac{1}{2} + \varepsilon\right) \cdot n - \mathbb{E}\left[\sum_{t=1}^{n} X_{C_t,t}\right] \;=\; \varepsilon \cdot \mathbb{E}\left[\left|\{t \mid C_t \neq H\}\right|\right] .$$

■

We can now relate the number of mistakes to the statistical distance between the distributions of the rewards seen by the player, in the cases when $H = \ell$ and when $H = r$.

Proposition 3.3.

$$\mathbb{E}\left[\left|\{t \mid C_t \neq H\}\right|\right] \;\geq\; \frac{n}{2} \cdot \left(1 - \frac{1}{2} \cdot \left\|P_\ell(Z_1, \ldots, Z_n) - P_r(Z_1, \ldots, Z_n)\right\|_1\right) .$$

Proof: Computing the expectation conditioned on the value of H, we get

$$\begin{aligned}\mathbb{E}\left[\left|\{t \mid C_t \neq H\}\right|\right] &= \frac{1}{2} \cdot \mathbb{E}\left[\left|\{t \mid C_t = r\}\right| \;\middle|\; H = \ell\right] + \frac{1}{2} \cdot \mathbb{E}\left[\left|\{t \mid C_t = \ell\}\right| \;\middle|\; H = r\right] \\ &= \frac{1}{2} \cdot \left(n - \mathop{\mathbb{E}}_{P_\ell}\left[\left|\{t \mid C_t = \ell\}\right|\right] + \mathop{\mathbb{E}}_{P_r}\left[\left|\{t \mid C_t = \ell\}\right|\right]\right)\end{aligned}$$

Since the choices of the player are functions of the view $(Z_1, \ldots, Z_n)$, we can write $|\{t \mid C_t = \ell\}| = f(Z_1, \ldots, Z_n)$ as a function which takes values between 0 and n. Thus, we get

$$\begin{aligned}\mathop{\mathbb{E}}_{P_\ell}[|\{t \mid C_t = \ell\}|] - \mathop{\mathbb{E}}_{P_r}[|\{t \mid C_t = \ell\}|] &= \mathop{\mathbb{E}}_{P_\ell}[f(Z_1, \ldots, Z_n)] - \mathop{\mathbb{E}}_{P_r}[f(Z_1, \ldots, Z_n)] \\ &\leq \frac{n}{2} \cdot \|P_\ell(Z_1, \ldots, Z_n) - P_r(Z_1, \ldots, Z_n)\|_1 .\end{aligned}$$

Substituting this bound in the above equality then proves the claim. ■

As before, we can now bound the statistical distance using Pinsker's inequality.

Proposition 3.4. *There exists a constant $c > 0$ such that*

$$\|P_\ell(Z_1, \ldots, Z_n) - P_r(Z_1, \ldots, Z_n)\|_1^2 \leq c \cdot \varepsilon^2 \cdot n .$$

Proof: As before, we use Pinsker's inequality and bound the KL-divergence. We have

$$D\left(P_\ell(Z_1, \ldots, Z_n) \parallel P_r(Z_1, \ldots, Z_n)\right) = \sum_{t=1}^{n} D\left(P_\ell(Z_t \mid Z_1, \ldots, Z_{t-1}) \parallel P_r(Z_t \mid Z_1, \ldots, Z_{t-1})\right) .$$

We now use the fact that C_t is determined by the values of $Z_1, \ldots, Z_{t-1}$, which we denote below by $Z_{<t}$ for short. Also, since the rewards $X_{\ell,t}$ and $X_{r,t}$ are independent of the history (in both P_ℓ and P_r), we get

$$D\left(P_\ell(Z_t \mid Z_{<t}) \parallel P_r(Z_t \mid Z_{<t})\right) = \begin{cases} D\left(P_\ell(X_{\ell,t}) \parallel P_r(X_{\ell,t})\right) & \text{if } C_t(Z_{<t}) = \ell \\ D\left(P_\ell(X_{r,t}) \parallel P_r(X_{r,t})\right) & \text{if } C_t(Z_{<t}) = r \end{cases}$$

Finally, using $D(p \parallel q)$ to denote the KL-divergence of distributions P and Q on $\{0,1\}$ with $p(1) = p$ and $q(1) = q$, we get that

$$D\left(P_\ell(X_{\ell,t}) \parallel P_r(X_{\ell,t})\right) = D\left(\frac{1}{2} + \varepsilon \parallel \frac{1}{2}\right) \quad \text{and} \quad D\left(P_\ell(X_{r,t}) \parallel P_r(X_{r,t})\right) = D\left(\frac{1}{2} \parallel \frac{1}{2} + \varepsilon\right) .$$

Since both the divergences above are bounded by $c' \cdot \varepsilon^2$ for some constant c' (check!) we get using Pinsker's inequality that

$$\|P_\ell(Z_1, \ldots, Z_n) - P_r(Z_1, \ldots, Z_n)\|_1^2 \leq 2 \ln 2 \cdot n \cdot c' \varepsilon^2 ,$$

which proves the claim. ■

Combining the bounds from Proposition 3.2, Proposition 3.3 and Proposition 3.4, we now get that

$$\begin{aligned}\mathbb{E}[\text{regret}] &\geq \frac{\varepsilon n}{2} \cdot \left(1 - \frac{1}{2} \cdot \|P_\ell(Z_1, \ldots, Z_n) - P_r(Z_1, \ldots, Z_n)\|_1\right) \\ &\geq \frac{\varepsilon n}{2} \cdot \left(1 - \sqrt{c \cdot \varepsilon^2 \cdot n}\right) = \varepsilon n \cdot \left(\frac{1}{2} - c_0 \cdot \sqrt{\varepsilon^2 \cdot n}\right) ,\end{aligned}$$

for $c_0 = \sqrt{c}/2$.

References

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire, *The nonstochastic multiarmed bandit problem*, SIAM journal on computing **32** (2002), no. 1, 48–77. 4

[BCB12] Sébastien Bubeck and Nicolò Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, Foundations and Trends® in Machine Learning **5** (2012), no. 1, 1–122. 4

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi, *Prediction, learning, and games*, Cambridge University Press, 2006. 4