Information and Coding Theory **Autumn 2022**

Lecture 6: October 13, 2022

Lecturer: Omar Montasser

1 Total variation distance and Pinsker's inequality

We can relate KL-divergence to some other notions of distance between two probability distributions.

Definition 1.1. *Let P and Q be two distributions on a finite universe $\mathcal{X}$. Then the* total-variation distance *or* statistical distance *between P and Q is defined as*

$$\delta_{TV}(P,Q) \;=\; \frac{1}{2}\cdot\|P-Q\|_1 \;=\; \frac{1}{2}\cdot\sum_{x\in\mathcal{X}}|p(x)-q(x)|\;.$$

The quantity $\|P-Q\|_1$ is referred to as the ℓ_1-distance *between P and Q.*

The total variation distance of P and Q represents the maximum probability with which any test can distinguish between the two distributions *given one random sample*. It may seem that the restriction to one sample severely limits the class of tests, but we can always think of an n-sample test for P and Q as getting one sample from one of the product distributions P^n or Q^n.

Let $f : \mathcal{X} \to \{0,1\}$ be any classifier, which given one sample $x \in \mathcal{X}$, outputs 1 if the guess is that the sample came from P, and 0 if the guess is that it came from Q. The difference in its behavior over the two distributions can be measured by the quantity (which can be thought of as the rate of true positive minus the rate of false positive) $|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]|$. The following lemma bounds this in terms of the total variation distance.

Lemma 1.2. *Let P, Q be any distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to [0,B]$. Then*

$$\left|\mathop{\mathbb{E}}_{P}[f(x)] - \mathop{\mathbb{E}}_{Q}[f(x)]\right| \;\le\; \frac{B}{2}\cdot\|P-Q\|_1 \;=\; B\cdot\delta_{TV}(P,Q)\,.$$

Proof:

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{P}[f(x)] - \mathop{\mathbb{E}}_{Q}[f(x)]\right| &= \left|\sum_{x\in\mathcal{X}} p(x)\cdot f(x) - \sum_{x\in\mathcal{X}} q(x)\cdot f(x)\right| \\
&= \left|\sum_{x\in\mathcal{X}} (p(x)-q(x))\cdot f(x)\right| \\
&= \left|\sum_{x\in\mathcal{X}} (p(x)-q(x))\cdot \left(f(x)-\frac{B}{2}\right) + \frac{B}{2}\cdot\left(\sum_{x\in\mathcal{X}} p(x)-q(x)\right)\right| \\
&\le \sum_{x\in\mathcal{X}} |p(x)-q(x)|\cdot\left|f(x)-\frac{B}{2}\right| \\
&\le \frac{B}{2}\cdot\|P-Q\|_1
\end{aligned}$$

■

Exercise 1.3. *Prove that the above inequality is tight. What is the optimal classifier f?*

In many applications, we want to actually bound the ℓ_1-distance between P and Q but it's easier to analyze the KL-divergence. The following inequality helps relate the two.

Lemma 1.4 (Pinsker's inequality)**.** *Let P and Q be two distributions defined on a universe $\mathcal{X}$. Then*

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

We will prove the inequality in two steps. Let us first consider a special case when $\mathcal{X} = \{0,1\}$ and P, Q are distributions as below

$$P \;=\; \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \qquad \text{and} \qquad Q \;=\; \begin{cases} 1 & \text{w.p. } q \\ 0 & \text{w.p. } 1-q \end{cases}$$

In this case, we have

$$D(P\|Q) \;=\; p\cdot\log\left(\frac{p}{q}\right) + (1-p)\cdot\log\left(\frac{1-p}{1-q}\right) \qquad \text{and} \qquad \|P-Q\|_1 \;=\; 2\cdot|p-q|\,.$$

We will first prove Pinsker's inequality for this special case.

Proposition 1.5 (Pinsker's inequality for $\mathcal{X} = \{0,1\}$)**.** *Let P and Q be distributions as above. Then,*

$$p\cdot\log\left(\frac{p}{q}\right) + (1-p)\cdot\log\left(\frac{1-p}{1-q}\right) \;\ge\; \frac{2}{\ln 2}\cdot(p-q)^2\,.$$

Proof: Let

$$f(p,q) \;:=\; p \cdot \log\left(\frac{p}{q}\right) + (1-p)\cdot \log\left(\frac{1-p}{1-q}\right) - \frac{2}{\ln 2}\cdot (p-q)^2 .$$

We have,

$$\frac{\partial f}{\partial q} = -\frac{(p-q)}{\ln 2}\left(\frac{1}{q(1-q)} - 4\right) .$$

Since $\frac{1}{q(1-q)} - 4 \geq 0$ for all q, we have that $\frac{\partial f}{\partial q} \leq 0$ when $q \leq p$ and $\frac{\partial f}{\partial q} \geq 0$ when $q \geq p$. Moreover, $f(p,q) = \infty$ when $q = 0$ and $f(p,q) = 0$ when $q = p$. Thus, the function achieves its minimum value at $q = p$ and is always non-negative, which proves the desired inequality. ■

We can now reduce the general case of Pinsker's inequality, to the case of $\mathcal{X} = \{0,1\}$ considered above.

Proposition 1.6. *Let P and Q be distributions on a finite set $\mathcal{X}$. Then, there exist distributions P', Q' on $\{0,1\}$ such that*

$$\left\|P' - Q'\right\|_1 \;=\; \|P - Q\|_1 \quad \textit{and} \quad D(P\|Q) \;\geq\; D(P'\|Q')$$

Proof: Let $A \subset \mathcal{X}$ be

$$A = \{x \mid p(x) \geq q(x)\} .$$

and P' and Q' be

$$P' \;:=\; \begin{cases} 1 & \text{w.p. } \sum\limits_{x\in A} p(x) \\ 0 & \text{w.p. } \sum\limits_{x\notin A} p(x) \end{cases} \qquad \text{and} \qquad Q' \;:=\; \begin{cases} 1 & \text{w.p. } \sum\limits_{x\in A} q(x) \\ 0 & \text{w.p. } \sum\limits_{x\notin A} q(x) \end{cases}$$

Then,

$$\begin{aligned} \|P - Q\|_1 &= \sum_{x\in\mathcal{X}} |p(x) - q(x)| \\ &= \sum_{x\in A} (p(x) - q(x)) + \sum_{x\notin A} (q(x) - p(x)) \\ &= \left|\sum_{x\in A} p(x) - \sum_{x\in A} q(x)\right| + \left|\left(1 - \sum_{x\in A} p(x)\right) - \left(1 - \sum_{x\in A} q(x)\right)\right| \\ &= \left\|P' - Q'\right\|_1 \end{aligned}$$

To calculate the KL-divergence, we define a random variable Z (which is a function of X) as

$$Z \;=\; \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} .$$

Since Z is a function of X, we can also think of the two distributions P and Q as joint distributions for the random variables (X, Z). Also, note that the marginal distributions of Z are P' and Q'. Applying the chain rule for KL-divergence gives

$$\begin{aligned} D(P\|Q) &= D(P(X,Z) \parallel Q(X.Z)) \\ &= D(P(Z) \parallel Q(Z)) + D(P(X|Z) \parallel Q(X|Z) \\ &\geq D(P(Z) \parallel Q(Z)) \\ &= D\left(P'\|Q'\right) \end{aligned}$$

which completes the proof. ■

Finally, we can complete the proof of Pinkser's inequality for the general case, by noting that

$$D(P\|Q) \;\geq\; D(P'\|Q') \;\geq\; \frac{1}{2\ln 2} \cdot \left\|P' - Q'\right\|_1^2 \;=\; \frac{1}{2\ln 2} \cdot \|P - Q\|_1^2 .$$

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 $T(x_1, x_2, ...x_n) \rightarrow \{0, 1\}$ 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.,

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

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

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

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

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

Plugging in this upper bound, we get

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

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 $D(P \,\|\, Q)$ to bound $\|P - Q\|_1$, we could also have used $D(Q \,\|\, P)$ 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

$$\min\{D(P \,\|\, Q), D(Q \,\|\, P)\} \;\geq\; \frac{1}{2\ln 2} \cdot \|P - Q\|_1^2.$$