

Lecture 5: January 21, 2025

Lecturer: Madhur Tulsiani

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 $f : \mathcal{X} \rightarrow \{0, 1\}$ 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 $f : \mathcal{X} \rightarrow [0, B]$. Then

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

Proof:

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

■

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

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

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

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

In this case, we have

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

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,

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

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. \blacksquare

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*

$$\|P' - Q'\|_1 = \|P - Q\|_1 \quad \text{and} \quad D(P\|Q) \geq D(P'\|Q')$$

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

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

and P' and Q' be

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

Then,

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

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

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

which completes the proof. ■

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

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

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"):

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

Suppose we have an algorithm $T(x_1, x_2, \dots, x_n) \rightarrow \{0, 1\}$ 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.,

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

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

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

which gives

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

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

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

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

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

Plugging in this upper bound, we get

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

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

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

3 Dealing with infinite universes

So far, we have only considered random variables taking values over a finite universe. We now consider how to define the various information theoretic quantities, when the set of possible values is not finite.

3.1 Countable universes

When the universe is countable, various information theoretic quantities such as entropy and KL-divergence can be defined essentially as before. Of course, since we now have infinite sums in the definitions, these should be treated as limits of the appropriate series. Hence, all quantities are defined as limits of the corresponding series, *when the limit exists*.

Convergence is usually not a problem, but it is possible to construct examples where the entropy is infinite. Consider the case of $U = \mathbb{N}$, and a probability distribution P satisfying $\sum_{x \in \mathbb{N}} p(x) = 1$. Since the sequence $\sum_x p(x)$ converges, usually the terms of $\sum_x p(x) \cdot \log(1/p(x))$ are not much larger. However, we can construct an example using the fact that $\sum_{n \geq 2} 1/(k \cdot (\log k)^\alpha)$ converges if and only if $\alpha > 1$. Define

$$p(x) = \frac{C}{x \cdot (\log x)^2} \quad \forall x \geq 2 \quad \text{where} \quad \lim_{n \rightarrow \infty} \sum_{2 \leq x \leq n} \frac{1}{x \cdot (\log x)^2} = \frac{1}{C}.$$

Then, for a random variable X distributed according to P ,

$$H(X) = \sum_{x \geq 2} \frac{C}{x \cdot (\log x)^2} \cdot \log \left(\frac{x \cdot (\log x)^2}{C} \right) = \infty.$$

Exercise 3.1. Calculate $H(X)$ when X be a geometric random variable with

$$\mathbb{P}[X = n] = (1 - p)^{n-1} \cdot p \quad \forall n \geq 1$$

3.2 Uncountable universes

When the universe is not countable, one has to use measure theory to define the appropriate information theoretic quantities (actually, it is the KL-divergence which is defined this way). However, we will mostly consider the special case of distributions with a probability density function. Such random variables are referred to as continuous random variables. Given a random variable X taking values in (say) \mathbb{R}^n with associated density function $p(x)$, we have the property that for any “box” $B = I_1 \times \dots \times I_n$, where I_1, \dots, I_n are (open or closed) intervals, we have

$$\mathbb{P}[X \in B] = \int_B p(x) \cdot dx.$$

A common example is the Gaussian distribution. The distribution of a one-dimensional Gaussian random variable X with mean $\mathbb{E}[X] = \mu$ and variance $\mathbb{E}[(X - \mu)^2] = \sigma^2$ is denoted by $N(\mu, \sigma^2)$ and has the associated density function

$$p(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

Similarly, for a Gaussian random variable taking values in \mathbb{R}^n with mean vector $\mathbb{E}[X] = \mu$ and covariance matrix $\mathbb{E}[(X - \mu)(X - \mu)^\top] = \Sigma$, we denote the distribution as $N(\mu, \Sigma)$ and have the density function

$$p(x) = \frac{1}{(2\pi)^{n/2} \cdot |\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (x - \mu)^\top \Sigma^{-1} (x - \mu)\right),$$

where $|\Sigma|$ denotes $\log(|\det(\Sigma)|)$ for the positive definite matrix Σ .

3.3 Differential entropy

A commonly used definition in the case of continuous random variables, is that of differential entropy.

Definition 3.2. *Let X be a random variable taking values in \mathbb{R}^n , with density p . Then the differential entropy of X is defined to be the following integral (if it exists)*

$$h(X) := \int p(x) \cdot \log\left(\frac{1}{p(x)}\right) dx.$$

Although the expression for differential entropy looks syntactically similar to that of entropy in the finite case, $p(x)$ appearing in the expression above is a probability density function *and not a probability!* In fact, it is problematic to think of $h(X)$ as a measure of uncertainty or “randomness content” for a random variable as illustrated by the following example.

Example 3.3. *Consider X to be uniform on $[0, 1]$. Then*

$$h(X) = \int_0^1 1 \cdot \log(1) dx = 0.$$

Thus, the differential entropy for X is 0 even though it non-trivial random variable! Even more troublingly, for $Y = X/2$, which is now uniform in $[0, 1/2]$, we have

$$h(Y) = \int_0^{1/2} 2 \cdot \log(1/2) dy = -1.$$

Thus, $h(Y)$ is non even a non-negative quantity! Finally, consider $Z = X^2$, where X is uniform in $[0, 1]$. One can check that the density function is now $p(z) = \frac{1}{2\sqrt{z}}$, which gives

$$h(Z) = \int_0^1 \frac{1}{2\sqrt{z}} \cdot \log(2\sqrt{z}) dz = 1 - \frac{1}{\ln 2}.$$

As the above example shows, the differential entropy is not always a non-negative quantity, and depends on how we parametrize a distribution. A uniform distribution on disks with diameters in $[0, 1]$ can be parametrized in terms of the diameters, radii, or area. The above example shows that we will obtain different values for differential entropy in each of these cases.

Relating differential entropy to the limit of a sum. One way of trying to understand the above behavior is to consider the derivation of entropy for a continuous random variables, using the limit of a sum. Let P be such that both $p(x)$ and $p(x) \cdot \log(1/p(x))$ are Riemann integrable. If we divide the real line into intervals of length ε , using the mean value theorem, we can find a point x_k for each interval $[k \cdot \varepsilon, (k + 1) \cdot \varepsilon]$ (where $k \in \mathbb{Z}$) such that

$$\varepsilon \cdot p(x_k) = \int_{k \cdot \varepsilon}^{(k+1) \cdot \varepsilon} p(x) dx.$$

Consider the random variable X' taking values in the countable set $\{x_k\}_{k \in \mathbb{Z}}$ such that

$$\mathbb{P}[X' = x_k] = \varepsilon \cdot p(x_k).$$

Then, we have

$$H(X') = \sum_{k \in \mathbb{Z}} \varepsilon \cdot p(x_k) \cdot \log\left(\frac{1}{\varepsilon \cdot p(x_k)}\right) = \sum_{k \in \mathbb{Z}} \varepsilon \cdot p(x_k) \cdot \log\left(\frac{1}{p(x_k)}\right) + \frac{1}{\varepsilon}$$

Note that the definition of differential entropy is the limit of the first sum, as $\varepsilon \rightarrow 0$. However, this is *not* the limit of $H(X')$, which is actually infinite. Hence, the concept of differential entropy is not a measure of the randomness content of a random variable and one should be careful about how to interpret it.

Since differential entropy is the limit up to the discretization factor of $\log(1/\varepsilon)$, it also changes when we scale the random variable. Let X be any random variable with the density p and let $Y = \alpha \cdot X$. Then, Y has the density $q(y) = (1/\alpha) \cdot p(y/\alpha)$ and

$$h(Y) = \int q(y) \cdot \log\left(\frac{1}{q(y)}\right) dy = \int \frac{1}{\alpha} \cdot p(y/\alpha) \cdot \log\left(\frac{\alpha}{p(y/\alpha)}\right) = h(X) + \log(\alpha).$$

Thus, in general it is problematic to compare the values differential entropy for two random variables, without controlling for the scale. Occasionally, we will see a comparison between two random variables once we restrict them to having the same values for some moments (which fixes a scale). See the introduction by Marsh [Mar13] on how to work with the notion of differential entropy.

References

[Mar13] Charles Marsh, *Introduction to continuous entropy*, 2013. [8](#)