Information and Coding Theory **Autumn 2017**

Lecture 5: October 11, 2017

Lecturer: Madhur Tulsiani

1 Pinsker's inequality and its applications to lower bounds

We first prove Pinsker's inequality for the general case, extending the proof from the last lecture for the case of the binary universe $U = \{0, 1\}$. Recall the statement of the inequality.

1.1 Pinsker's inequality

Lemma 1.1 (Pinsker's inequality) *Let P and Q be two distributions defined on a universe U. Then*

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

We proved the above inequality in the last lecture, for the special case of $U = \{0, 1\}$. We now show how one can prove the general case, by reducing it to the binary case. We use the optimal binary classifier between P and Q (recall the exercise from the last lecture or see below!) to reduce to the case of distributions simply over the output of the classifier (which is binary). It will follow easily that this preserves the total variation distance, while the KL-divergence only decreases. Combining these two will yield the proof.

Proof: Let P and Q be distributions on U. Let $A \subset U$ be

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

and P_A and Q_A be

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

Then,

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

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). Applying the chain rule for KL-divergence gives

$$\begin{aligned}
D(P\|Q) &= D(P(X,Z) \parallel Q(X.Z)) \\
&= D(P(Z) \parallel Q(Z)) + D(P(X|Z) \parallel Q(X|Z) \\
&\geq D(P(Z) \parallel Q(Z)) \\
&= D\left(P_A \| Q_A\right) \\
&\geq \frac{1}{2 \ln 2} \cdot \|P_A - Q_A\|_1^{\,2} \\
&= \frac{1}{2 \ln 2} \cdot \|P - Q\|_1^{\,2},
\end{aligned}$$

which completes the proof. ∎

1.2 Distinguishing two coins

We will now use Pinsker's inequality to derive a lower bound on the number of samples neede 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 in the next lecture. Suppose we are given one of the following two coins (think of 1 as "heads" and 0 as "tails"):

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

Suppose we have an algorithm $T(x_1, x_2, ...x_m) \to \{0, 1\}$ that takes the output of m independent coin tosses, and makes a decision about which coint 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 AT identifies both coins with probability at least $9/10$, i.e.,

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

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

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

which gives

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

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

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

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

$$\begin{aligned}
D(P\|Q) \;&=\; \left(\frac{1}{2}-\varepsilon\right)\log\left(\frac{\frac{1}{2}-\varepsilon}{\frac{1}{2}}\right) + \left(\frac{1}{2}+\varepsilon\right)\log\left(\frac{\frac{1}{2}+\varepsilon}{\frac{1}{2}}\right) \\
&=\; \frac{1}{2}\log\left((1-2\varepsilon)(1+2\varepsilon)\right) + \varepsilon\log\left(\frac{1+2\varepsilon}{1-2\varepsilon}\right) \\
&\leq\; \frac{\varepsilon}{\ln 2}\ln\left(1+\frac{4\varepsilon}{1-2\varepsilon}\right) \\
&\leq\; \frac{4\varepsilon^2}{\ln 2}\frac{1}{1-2\varepsilon} && (\text{used } \ln 1 + x \leq e^x) \\
D(P\|Q) \;&\leq\; \frac{8\varepsilon^2}{\ln 2} && \left(\text{assumed } \varepsilon \leq \frac{1}{4}\right)
\end{aligned}$$

Plugging in this upper bound, we get

$$m \;\geq\; \frac{1}{2\ln 2\cdot D(P\|Q)}\cdot\left(\frac{8}{5}\right)^2 \;\geq\; \frac{4}{25\varepsilon^2}.$$

Exercise 1.2 *Prove using Chernoff bounds that $O(1/\varepsilon^2)$ samples are enough to distinguish the two coins.*

Exercise 1.3 *How many samples are needed in the case when one coin comes up heads with probability $p = \varepsilon$ and the other with probability $q = 2\varepsilon$?*

2 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.

2.1 Countable universes

When the universe is countable, various information theoretic quantities such as entropy an KL-divergence can be defined essentialy 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 an only if $\alpha > 1$. Define

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

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 2.1 *Calculate $H(X)$ when X be a geometric random variable with*

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

2.2 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 first consider a commonly used definition for the special case of distributions with a probability density function.

Definition 2.2 *Let X be a real-valued random variable, with density p. Then the* differential entropy *of X is defined to be the following integral (if it exists)*

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

Similarly, if P and Q are two distributions with densities p and q, then their KL-divergence if defined by the integral

$$D(P\|Q) \;:=\; \int p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) dx\,.$$

Differential entropy

It is problematic to think of $h(X)$ as a measure of uncertainity or "randomness content" for a random variable. 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 X uniform in $[0,1/2]$, we have

$$h(X) \;=\; \int_0^{1/2} 2 \cdot \log(1/2) dx \;=\; -1\,.$$

Thus, $h(X)$ is non even a non-negative quantity! One way of trying to understand this is consider the derivation using the approximation 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, w 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

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

Then, we have

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

Note that the definition of differential entropy is the limit of the first sum, as $\varepsilon \to 0$. However, this is *not* the limit of $H(X')$, which is actually infinite. Hence, the concept of defferential 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 upto 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. Occassionally, we will see a comaprison between two random variables once we restrict them to having the same values for some moments. See the introduction by Marsh [Mar13] on how to work with the notion of differential entropy.

KL-divergence

Unlike the concept of differential entropy, that of KL-divergence is a direct generalization of KL-divergence for distributions on finite universes. A measure-theoretic definition of KL-divergence was developed in the works of Kolmogorov and Pinsker. A detailed treatment can be found in Chapter 7 of the book by Gray [Gra11] (Chapter 5 of the older edition linked from the aurhor's webpage).

In general, consider any two probability measures P, Q on a space Ω with underlying σ-algebra $\mathcal{F} \subseteq 2^{\Omega}$ (defining the notion of "valid events" which one can talk about). A random variable X taking values in a finite set $[n]$ is defined to be a *measurable function* $X : \Omega \to [n]$ i.e., we require $X^{-1}(S)$ to be a valid event in $\mathcal{F}$, for all subsets $S \subseteq [n]$. Then, the KL-divergence of P and Q is defined to be

$$D(P\|Q) \;=\; \sup_{X,n} D(P(X)\|Q(X))\,,$$

for X and n as above. When P and Q have densities p and q, this definition can be shown to converge to the one defined above.

Note that the measure-theoretic definition reduces the infinite case to the (supremum over) finite cases. Since mutual information can be defined in terms of the KL-divergence (see

Homework 1), this also gives a measure-theoretic definition for mutual information. Also, since $D(P(X)\|Q(X)) \geq 0$ for each of the finite cases, we still have $D(P\|Q)$ for any two distribution. Thus, any inequalities between entropies which were derived using the non-negativity of KL-divergence are still valid. These include the non-negativity of mutual information or (equivalently) the fact that conditioning reduces entropy, the sub-additivity of entropy and also Shearer's lemma. In addition, Pinsker's inequality also holds for the infinite setting, since the total variation distance can also be defined by a similar expression in terms of finite distributions.

Gaussian distributions

We conclude with some simple calculations for the Gaussian distribution, with $X \sim N(\mu, \sigma^2)$ having the density

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

We can then calculate the differential entropy of a Gaussian random variable as

$$\begin{aligned} h(X) \;&=\; \int p(x)\cdot\frac{1}{\ln 2}\cdot\left(\frac{(x-\mu)^2}{2\sigma^2}+\frac{1}{2}\ln(2\pi\sigma^2)\right)dx \\ &=\; \frac{1}{\ln 2}\cdot\left(\frac{1}{2}+\frac{1}{2}\ln(2\pi\sigma^2)\right) \\ &=\; \frac{1}{2}\cdot\log(2\pi\cdot e\cdot\sigma^2)\,. \end{aligned}$$

Exercise 2.3 *Let P and Q be Gaussian distributions with means μ_1 and μ_2 respectively, and variance σ^2 in both cases, Show that*

$$D(P\|Q) \;=\; \frac{(\mu_1-\mu_2)^2}{2\ln 2\cdot\sigma^2}\,.$$

Using Pinsker's inequality, this can be used to show that

$$\|P-Q\|_1 \;\leq\; \frac{|\mu_1-\mu_2|}{\sigma}\,.$$

The above is a common way of showing that changing the parameters of a Gaussian distribution does not alter the behavior of an algorithm using the corresponding random variable as input, by too much.

References

[Gra11] Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011. URL: https://ee.stanford.edu/~gray/it.html. 6

[Mar13] Charles Marsh. Introduction to continuous entropy, 2013. URL: http://www.crmarsh.com/static/pdf/Charles_Marsh_Continuous_Entropy.pdf. 6