Information and Coding Theory **Winter 2021**

Lecture 6: January 28, 2021

Lecturer: Madhur Tulsiani

1 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.

1.1 Countable universes

When the universe is countable, various information theoretic quantities such as entropy an 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 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 1.1. *Calculate $H(X)$ when X be a geometric random variable with*

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

1.2 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" $B = I_1 \times \times \cdots \times I_n$, where $I_1, \ldots, I_n$ are (open or closed) intervals, we have

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

A common example is the Gaussian distribution. The distribution of a one-dimensional Gaussian random variable X with mean $\mathbb{E}\left[X\right] = \mu$ and variance $\mathbb{E}\left[(X-\mu)^2\right] \;=\; \sigma^2$ 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 $\mathbb{E}\left[X\right] = \mu$ and covariance matrix $\mathbb{E}\left[(X-\mu)(X-\mu)^{\mathsf{T}}\right] \;=\; \Sigma$, we denote the distribution as $N(\mu, \Sigma)$ and have the density function

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

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

1.3 Differential entropy

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

Definition 1.2. *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 1.3. *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, 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 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.

1.4 KL-divergence

We define KL-divergence for two distributions analogously, when both distributions have associated density functions.

Definition 1.4. *If P and Q are two distributions with densities p and q, then their KL-divergence if defined by the integral*

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

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 author'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 of two random variables X, Y can be defined in terms of the KL-divergence as (see Homework 1)

$$I(X;Y) \;=\; D\left(P(X,Y) \,\|\, P(X)P(Y)\right) ,$$

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 distributions over $\mathbb{R}^n$. 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.

2 Gaussian computations

We now derive the expressions for entropy and KL-divergence of Gaussian distributions, which often come in handy.

2.1 Differential entropy

For a one-dimensional Gaussian $X \sim N(\mu, \sigma^2)$ we can calculate the differential entropy as

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

For the n-dimensional case, we first consider a Gaussian variable X with mean 0 and covariance I_n, which means that we can think of $X = (X_1, \ldots, X_n)$ where each X_i is a one-dimensional Gaussian with mean 0 and variance 1. Using the chain-rule for differential entropy (check that it holds) we get

$$h(X) \;=\; h(X_1) + \cdots + h(X_n) \;=\; \frac{n}{2} \cdot \log(2\pi \cdot e) \,.$$

Before computing the entropy of a general Gaussian variables, it is helpful to consider the following rule for change of variables.

Exercise 2.1 (Change of variables). *Let X be a random variable over $\mathbb{R}^n$ with associated density function p_X. Using the Jacobian for change of variables in integrals, check that*

1. *If $c \in \mathbb{R}^n$ is a fixed vector, then the density function for $Y = X + c$ is given by $p_Y(y) = p_X(y - c)$.*
2. *If $A \in \mathbb{R}^{n \times n}$ is a nonsingular matrix, then the density function for $Y = AX$ is given by $p_Y(y) = \frac{p_X(A^{-1}y)}{|A|}$, where $|A|$ denotes $|det(A)|$.*

Using the above, we can derive how the differential entropy of a random variable changes due to translation and scaling.

Proposition 2.2. *Let X be a continuous random variable over $\mathbb{R}^n$. Let $c \in \mathbb{R}^n$ and let $A \in \mathbb{R}^{n \times n}$ be a non-singular matrix. Then*

1. *$h(X + c) = h(X)$.*
2. *$h(AX) = h(X) + \log |A|$.*

Proof: Let p_X be the density function for X. For $Y = X + c$, we have

$$\begin{aligned} h(Y) &= \int_{\mathbb{R}^n} p_Y(y) \cdot \log\left(\frac{1}{p_Y(y)}\right) dy \\ &= \int_{\mathbb{R}^n} p_X(y-c) \cdot \log\left(\frac{1}{p_X(y-c)}\right) dy \\ &= \int_{\mathbb{R}^n} p_X(x) \cdot \log\left(\frac{1}{p_X(x)}\right) dx && \text{(substituting } x = y - c\text{)} \\ &= h(X) \end{aligned}$$

Similarly, for $Y = AX$, we have

$$\begin{aligned} h(Y) &= \int_{\mathbb{R}^n} p_Y(y) \cdot \log\left(\frac{1}{p_Y(y)}\right) dy \\ &= \int_{\mathbb{R}^n} \frac{p_X(A^{-1}y)}{|A|} \cdot \log\left(\frac{|A|}{p_X(A^{-1}y)}\right) dy \\ &= \int_{\mathbb{R}^n} \frac{p_X(x)}{|A|} \cdot \log\left(\frac{|A|}{p_X(x)}\right) |A|\, dx && \text{(substituting } x = A^{-1}y\text{)} \\ &= h(X) + \log(|A|)\,. \end{aligned}$$

■

Using the fact that $Y \sim N(\mu, \Sigma)$ can be written as $Y = \Sigma^{1/2}X + \mu$, where $X = N(0, I_n)$ (check this!) we get that

$$h(Y) \;=\; h(X) + \log(\left|\Sigma^{1/2}\right|) \;=\; \frac{n}{2} \cdot \log(2\pi \cdot e) + \frac{1}{2} \cdot \log |\Sigma|\;.$$

2.2 KL-divergence

We can compute the KL-divergence of two Gaussian distributions $P = N(\mu_1, \sigma_1^2)$ and $Q = N(\mu_2, \sigma_2^2)$ as

$$\begin{aligned}
D\left(P \parallel Q\right) &= \int_{\mathbb{R}} p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) dx \\
&= \mathop{\mathbb{E}}_{x \sim P}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] \\
&= \mathop{\mathbb{E}}_{x \sim P}\left[\frac{1}{\ln 2} \cdot \ln\left(\frac{\exp\left(-(x-\mu_1)^2/2\sigma_1^2\right)}{\sqrt{2\pi}\sigma_1} \cdot \frac{\sqrt{2\pi}\sigma_2}{\exp\left(-(x-\mu_2)^2/2\sigma_2^2\right)}\right)\right] \\
&= \frac{1}{\ln 2} \cdot \mathop{\mathbb{E}}_{x \sim P}\left[\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \ln\left(\frac{\sigma_2}{\sigma_1}\right)\right] \\
&= \frac{1}{\ln 2} \cdot \left(\frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2} + \ln\left(\frac{\sigma_2}{\sigma_1}\right)\right) \\
&= \frac{1}{\ln 2} \cdot \left(\frac{\sigma_1^2 - \sigma_2^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} \; + \; \ln\left(\frac{\sigma_2}{\sigma_1}\right)\right) .
\end{aligned}$$

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

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. Use Pinsker's inequality to show that*

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

Exercise 2.4. *Compute $D\left(P \parallel Q\right)$ for the n-dimension Gaussian distributions $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$.*

References

[Gra11] Robert M Gray, *Entropy and information theory*, Springer Science & Business Media, 2011. 4

[Mar13] Charles Marsh, *Introduction to continuous entropy*, 2013. 4