Electrical Engineering 229A Lecture 2 Notes

Daniel Raban

August 31, 2021

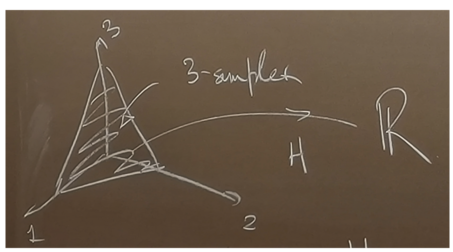
1 Entropic Quantities Relating Random Variables

1.1 The binary entropy function

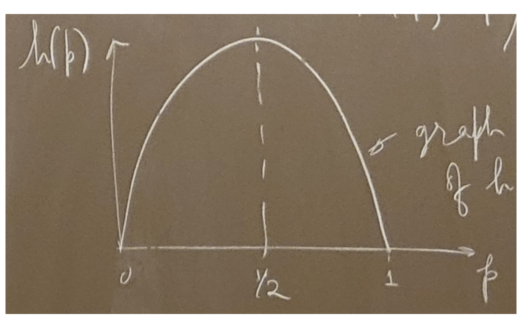
Suppose we have a probability distribution $p = (p_1, \dots, p_d)$ on a finite set of $\mathscr{X}$ of size d, say $\mathscr{X} = \{1, \dots, d\}$. We will use the notation $[d] = \{1, \dots, d\}$. The function

$$H(p_1, \dots, p_d) = -\sum_{j=1}^{d} p_i \log p_i$$

is called the entropy of the distribution p. Last lecture we saw that $H \geq 0$ and $H(p_1, \dots, p_d) \leq H(1/d, \dots, 1/d) = \log d$ as a consequence of the *concavity* of H as a function on the unit d-simplex. Concavity of H means that for $\lambda \in [0, 1]$, $H(\lambda p^{(1)} + (1 - \lambda)p^{(0)}) \geq \lambda H(p^{(1)}) + (1 - \lambda)H(p^{(0)})$.



Example 1.1. For $d = 2$, $H(p, 1 - p) = -p \log p - (1 - p) \log(1 - p)$. We denote this as $h(p)$.



The function $h(p)$ is known as the **binary entropy function**. The graph is very steep near 0; all the derivatives approach ∞. $h(1/2) = 1$, and $h(p) = h(1-p)$. We can calculate

$$\begin{aligned} h'(p) &= \log_2 e(-\log_e p - 1 + \log_e(1-p) + 1) \\ &= \log \frac{1-p}{p}, \end{aligned}$$

which is $+\infty$ at $p = 0$ and $-\infty$ at $p = 1$. We can check

$$h''(p) = \log_2 e \left(-\frac{1}{1-p} - \frac{1}{p} \right),$$

which is $-\infty$ at $p = 0$ and $p = 1$.

1.2 Convexity and Jensen's inequality

Definition 1.1. A set $D \subseteq \mathbb{R}^n$ is **convex** if when $\lambda \in [0,1]$ and $x^{(0)}, x^{(1)} \in D$, $\lambda x^{(0)} + (1-\lambda)x^{(1)} \in D$, as well.

Definition 1.2. A function $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^n$ is a convex set is called a **convex function** if for all $\lambda \in [0,1]$ and $x^{(0)}, x^{(1)} \in D$, we have

$$f(\lambda x^{(1)} + (1-\lambda)x^{(0)}) \leq \lambda f(x^{(1)}) + (1-\lambda) f(x^{(0)}).$$

This implies that if for any $m \geq 1$, $x^{(1)}, x^{(2)}, \ldots, x^{(m)} \in D$ and any probability distribution $(\lambda_1, \ldots, \lambda_m)$ on $[m]$, we have

$$f\left(\sum_{i=1}^m \lambda_i x^{(i)} \right) \leq \sum_{i=1}^m \lambda_i f(x^{(i)}).$$

More generally, we have the following:

Theorem 1.1 (Jensen's inequality)**.** *For any random variable Z taking values in a convex set $D \subseteq \mathbb{R}^n$,*

$$f(\mathbb{E}[Z]) \leq \mathbb{E}[f(Z)].$$

1.3 Joint and conditional entropy

If X is a random variable taking values in $[d]$, we write $H(X)$ for $H(p_1, \ldots, p_d)$, where $p_i := \mathbb{P}(X = i)$. If X takes values in $\mathscr{X}$, then $H(X)$ denotes $H(p(x), x \in \mathscr{X})$, where $p(x) := \mathbb{P}(X = x)$. Now suppose X takes values in $\mathscr{X}$ and Y takes values in $\mathscr{Y}$, where $\mathscr{X}, \mathscr{Y}$ are finite sets. They have a joint probability distribution $(p(x,y), (x,y) \in \mathscr{X} \times \mathscr{Y})$.

Definition 1.3. The **joint entropy** of the pair (X, Y), which is just a random variable taking values in $\mathscr{X} \times \mathscr{Y}$, is denoted $H(X, Y)$ and equals

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y).$$

Definition 1.4. The difference $H(X, Y) - H(X)$, denoted $H(Y \mid X)$, is called the **conditional entropy** of Y given X.

Recall that the entropy is $H(X) = \mathbb{E}[\log 1/p(X)]$. The joint entropy can be written similarly:

$$H(X, Y) = \mathbb{E}\left[\log \frac{1}{p(X, Y)}\right].$$

We can also write the conditional entropy as

$$\begin{aligned} H(Y \mid X) &= \mathbb{E}\left[\log \frac{1}{p(Y \mid X)}\right] \\ &= \sum_{x,y} p(x, y) \log \frac{1}{p(y \mid x)} \\ &= \sum_{x} p(x) \sum_{y} p(y \mid x) \log \frac{1}{p(y \mid x)}. \end{aligned}$$

For each fixed $x \in \mathscr{X}$, $\sum_y p(y \mid x) \log \frac{1}{p(y|x)}$ is denoted $H(Y \mid X = x)$. It is the entropy of the conditional distribution of Y given that $X = x$. With this notation,

$$H(Y \mid X) = \sum_{x} p(x) H(Y \mid X = x).$$

Remark 1.1. This notation is not consistent with the rest of probability notation. $H(Y \mid X)$ is a number, rather than a random variable. This notation is widespread in information theory, however, because it was introduced by Shannon himself.

From this formula, we can see that $H(Y \mid X) \geq 0$.

1.4 Mutual information

We might hope that we "learn" about Y from observing X, i.e. the uncertainty in Y is reduced. That is, we hope that $H(Y) \geq H(Y \mid X)$. This is true.

Definition 1.5. $H(Y) - H(Y \mid X)$ is denoted $I(X; Y)$ (or sometimes denoted as $I(X \wedge Y)$) and is called the **mutual information** between X and Y.

We have

$$
\begin{aligned}
I(X;Y) &= \mathbb{E}\left[\log \frac{1}{p(Y)}\right] - \mathbb{E}\left[\log \frac{1}{p(Y \mid X)}\right] \\
&= \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.
\end{aligned}
$$

This is symmetric when X and Y are interchanged. That is, $I(X;Y) = I(Y;X)$.

1.5 Relative entropy

$I(X,Y) \geq 0$ because it is a relative entropy.

Definition 1.6. Given two probability distributions $(p(z), z \in \mathscr{Z})$ and $(q(z), z \in \mathscr{Z})$, we write

$$
D(p \mid\mid q) = \sum_{z \in \mathscr{Z}} p(z) \log \frac{p(z)}{q(z)},
$$

which is called the **relative entropy** of p with respect to q. It is also called the **information distance/divergence** of p from q or the **Kullback-Leibler divergence**.

Remark 1.2. The relative entropy is *not* a distance; it is not symmetric in p and q and does not satisfy the triangle inequality.

We want to show that $D(p \mid\mid q) \geq 0$. Note that

$$
I(X;Y) = D(p(x,y) \mid\mid p(x)p(y)),
$$

where $p(x,y)$ is the joint distribution of (X,Y) and $p(x)p(y)$ is the distribution of $(\widetilde{X}, \widetilde{Y})$, where $\widetilde{X} \overset{d}{=} X$, $\widetilde{Y} \overset{d}{=} Y$, and $\widetilde{X}, \widetilde{Y}$ are independent. So we will get $I(X;Y) \geq 0$ if we can prove $D(p \mid\mid q) \geq 0$ in general.

The relative entropy is a natural statistical quantity that measures how far p is from q. So the conceptual meaning of $I(X;Y)$ is that it measures how far apart the joint distribution of (X,Y) is from being a product distribution of independent X, Y.

Proposition 1.1. *$D(p \mid\mid q) \geq 0$.*

Proof. Write

$$
D(p \mid\mid q) = \sum_{z \in \mathscr{Z}} q(z) \frac{p(z)}{q(z)} \log \frac{p(z)}{q(z)}
$$

$$= \sum_{z \in \mathscr{Z}} q(z) \phi \left(\frac{p(z)}{q(z)} \right),$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is given by $\phi(u) = u \log u$, which is convex (checked below). Using Jensen's inequality,

$$\geq \phi \left(\sum_{z \in \mathscr{Z}} q(z) \frac{p(z)}{q(z)} \right)$$
$$= \phi(1)$$
$$= 0.$$

To check that ϕ is convex, we have $\phi'(u) = \log_2 e(\log_e u + 1)$, so $\phi''(u) = \log_2 e \cdot \frac{1}{u} \geq 0$. $\square$

Corollary 1.1. $I(X;Y) \geq 0$.