Electrical Engineering 229A Lecture 1 Notes

Daniel Raban

August 26, 2021

1 Introduction to Shannon Entropy

1.1 Shannon entropy

Information theory is unusual in that it originated from the work of one person, Claude Elwood Shannon, in the late 1950s.[1] Shannon's idea was how to numerically measure the "amount of (statistical) uncertainty" inherent in a probabilistic experiment.

Example 1.1 (Coin flipping)**.** The "uncertainty" in $(1/2, 1/2)$ is "more" than in $(3/4, 1/4)$, which is "more" than in $(99/100, 1/100)$.

Shannon developed a calculus to work with such quantities. This notion is called *entropy*.

Definition 1.1. Consider a probability distribution $(p(1), \dots, p(d))$ on $\{1, \dots, d\}$. The **Shannon entropy** of p is

$$H(p) = -\sum_{i=1}^{d} p(i) \log p(i).$$

Here, the log is base 2, which was Shannon's convention and the convention for engineers. In mathematics and statistical mechanics, the natural logarithm is used. We take the convention that $0 \log 0 = 0$ (which is $\lim_{x \downarrow 0} x \log x$).

Example 1.2. Note that

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = \log 2 = 1.$$

This is a kind of normalization.

[1]Shannon lived from 1916-2001. His master's thesis is also considered a landmark. It introduced the boolean circuit view of computing. There is a 2017 movie about Shannon called *The Bit Player* and a book called *A Mind at Play*.

1.2 Motivation for the formula of entropy

To motivate the actual formula, consider $d = 2$ and n independent copies of $\{1, 2\}$-valued random variables with probability distribution p. For a sequence x^n of 1s and 2s,

$$\begin{aligned} p(x^n) &= \prod_{i=1}^{n} p(x_i) \\ &= p(1)^{N(1|x^n)} p(2)^{N(2|x^n)} \\ &= 2^{n(N(1|x^n)/n \log p(1) + N(2|x^n)/n \log p(2))}, \end{aligned}$$

where $N(i \mid x^n)$ is the number of times i appears in x^n. But by the strong law of large numbers, $\frac{N(i|x^n)}{n} \to p(1)$ almost surely as $n \to \infty$. So

$$p(x^n) \approx (2^{p(1) \log p(1) + p(2) \log p(2)})^n.$$

This suggests that $-p(1) \log p(1) - p(2) \log p(2)$ represents the "uncertainty" in one toss.

1.3 Expectation formulation of entropy

If X is a random variable taking calues in $\{1, \dots, d\}$ with probability distribution p, i.e. $\mathbb{P}(x = i) = p(i)$ for $1 \le i \le d$, we write $H(X)$ for $H(p)$. With this notation,

$$H(X) = \sum_{i=1}^{d} \mathbb{P}(X = i) \log \frac{1}{\mathbb{P}(X = i)} = \mathbb{E}[\log 1/p(X)].$$

1.4 Concavity of Shannon entropy and entropy of uniform distributions

Fix $d \ge 2$. The set of probability distributions on $\{1, \dots, d\}$ is called the **unit d-simplex** in $\mathbb{R}^d$. We can write it as $\{(p(1), \dots, p(n)) : p(i) \ge 0, \sum_{i=1}^{d} p(i) = 1\}$. This is a **convex** set, and H can be viewed as a function on this set.

Proposition 1.1. *H is a **concave function** on the (unit) d-simplex for each fixed d. That is, for all $p_0, p_1 \in \{1, \dots, d\}$ and $\lambda \in [0, 1]$, if p_α denotes $\lambda p_1 + (1 - \lambda) p_0$, then $p_\lambda(i)$, then*

$$H(p_\lambda) \ge \lambda H(p_1) + (1 - \lambda) H(p_0).$$

Proof. Because $H(p) = -\sum_{i=1}^{d} p(i) \log p(i)$, we want to check that $x \log x$ is convex. This is twice differentiable, so it suffices to show that the second derivative is ≥ 0. Write

$$\begin{aligned} (x \log x)'' &= (\log_2 e)(x \log_e x)'' \\ &= (\log_2 e)(\log_e x + 1)' \\ &= (\log_2 e) \frac{1}{x} \\ &\ge 0. \end{aligned}$$

□

Corollary 1.1. *The uniform distribution on* $\{1, \ldots, d\}$ *has the largest entropy among probability distributions on* $\{1, \ldots, d\}$.

Proof. Let S_d denote the set of permutations of $\{1, \ldots, d\}$. Then

$$(1/d, \ldots, 1/d) = \frac{1}{d!} \sum_{\sigma \in S_d} (p(\sigma(1)), p(\sigma(2)), \ldots, p(\sigma(d))),$$

so by the concavity of H,

$$\begin{aligned} H(1/d, \ldots, 1/d) &\geq \frac{1}{d!} \sum_{\sigma \in S_d} H(p(\sigma(1)), p(\sigma(2)), \ldots, p(\sigma(d))) \\ &= H(p). \end{aligned}$$

□

1.5 Conditional entropy

The entropy calculus starts with the definition of "conditional entropy." Given a pair of random variables (X, Y), we consider $H(X, Y) - H(Y)$ and denote this $H(X \mid Y)$. This is known as the **conditional entropy of** X **given** Y. Next time, we will consider the information $I(X; Y) := H(X) - H(X \mid Y)$ and see that this is actually symmetric in X and Y.