Learner's Space 2025
Information Theory and Coding
Week 1, Handout 2

Suketu Patni

June 13, 2025

Abstract

Throughout this document we will be restricting ourselves to discrete random variables, that is, those with a finite or countable image. Once we have a firm footing in many of the information-theoretic concepts defined for discrete random variables, we will later go on to work with continuous random variables. This handout will teach you the basics of Shannon information content and entropy, along with some of their important properties and constructs.
This document will give you very little if you are unwilling to work with it. Do not shy away from filling in the proofs and exercises.
I will also upload some references for information theory on Moodle soon.

1 Light revision

A discrete random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is a measurable function $\Omega \to \mathbb{R}$ where its image $\mathcal{X}$ is finite or countable. As we have seen, $p(a) = P(X = a)$ is shorthand for $P(\{w \in \Omega : X(w) = a\})$. $p(\cdot)$ is called the probability mass function of X. The support of X, denoted as $\operatorname{supp} X$ is defined as $\{x \in \mathbb{R} : P(X = x) > 0\}$.

2 Shannon information content and entropy

It was Claude Elwood Shannon's seminal paper, titled "A Mathematical Theory of Communication", that is responsible for birthing as well as stating several results in this field of information theory.

Definition 2.1. *Consider a discrete random variable X on the probability space $(\Omega, \mathcal{F}, P)$. Let $a \in \operatorname{supp} X$. The **Shannon information content (SIC)** of a is defined as $-\log_2(p(a))$ and is measured in bits.*

Definition 2.2. *The **entropy** of X, denoted as $H(X)$, is its expected SIC i.e.*

$$H(X) = \mathbb{E}[-\log_2(p(X))] = -\sum_{x \in \operatorname{supp} X} p(x)\log_2(p(x))$$

Note that $\lim_{t\to 0^+} t\log_2 t = 0$, so at the expense of angering the pedant, we often relax the condition $x \in \operatorname{supp} X$ and write $x \in \mathcal{X}$, since adding 0s does not change the summation. Importantly, note that the entropy **has nothing to do with** $\mathcal{X}$. It does not depend on the actual values that the random variable takes, and only depends on the function $p(\cdot)$. Due to this reason, in case $\mathcal{X}$ is finite (with n elements, say), we denote $H(X) = H(p_1, p_2, \ldots, p_n)$ where $\{p_i : 1 \leq i \leq n\} = p(\mathcal{X})$. For conciseness, in case $\mathcal{X}$ is finite (and equal to, say, n), we define a "probability vector" $\vec{p} = (p_1, p_2, \ldots, p_n) \in \mathbb{R}^n$ and write $H(X) = H(\vec{p})$.
The key thing to realize here is that we know how to find expectations of functions of random variables. So given X, we can find $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\mathbb{E}[\log_2(X)]$ etc, since X^2 and $\log_2(X)$ are random variables in their own right. But so is $p(X)$! Indeed, for instance we may write

$$\mathbb{E}[p(X)] = \sum_{x\in\mathcal{X}} p(x)p(x)$$

So it does make sense to write $\mathbb{E}[-\log_2(p(X))]$.

2.1 Binary entropy function

Let us take an example and calculate the entropy of a random variable with $|\mathcal{X}| = 2$. Say $\vec{p} = (p, 1-p)$. Then,

$$H(p, 1-p) = -p\log_2(p) - (1-p)\log_2(1-p)$$

This function, also written as $H_2(p)$, has a special name - it is called the **binary entropy function**. Clearly, the domain of $H_2(p)$ is $[0, 1]$.

Exercise 2.1. *Prove that $H_2(p)$ has a global maximum at $p = 1/2$. Prove that it is concave in p. Hence prove that it is strictly increasing in $[0, 1/2]$ and strictly decreasing in $[1/2, 1]$. (all easily shown using elementary calculus)*

Here we may assign some intuition to our seemingly arbitrary definitions of entropy and SIC. Considering $H_2(p)$ as the entropy of a random variable denoting the outcome of a biased coin, with $P(\text{heads}) = p$, the graph of $H_2(p)$ (don't be lazy, draw it) starts to show two qualitative meanings of entropy :-

1. The **uncertainty** of the outcome: If p were 0 (or 1), we would have no uncertainty about the outcome, as it would always be tails (or heads). $H_2(0) = H_2(1) = 0$. At $p = 0.5$, we are most unsure of what the outcome will be.

2. The **information** we gain after knowing the outcome of the experiment: If p were close to 0 or 1, we'd be almost certain of the outcome. So once the coin was tossed, knowing the outcome would very likely give us very little information. But when $p = 0.5$, because we are most unsure of what the outcome will be, we will gain the maximum information on knowing it.

Improbable outcomes convey more information than probable outcomes. This convinces us that the usage of the so-defined entropy as a measure of information is not a bad one. In fact, **the outcome of a random experiment is guaranteed to be most informative if the probability distribution over outcomes is uniform** (we will prove this soon). For a development of why it makes sense to have a logarithm in the definition of entropy, we readily refer the reader to the first chapter in the course notes of Stefan M. Moser, which have been uploaded on Moodle.

2.2 Decomposability

Theorem 2.1. *Let* $|\operatorname{supp} X| = n$, $p(\operatorname{supp} X) = \{p_1, p_2, \dots p_n\}$. *Put* $\sum_{i=1}^{m} p_i = \alpha$ *and* $\sum_{i=m+1}^{n} p_i = \beta$. *Then*

$$H(p_1, p_2, \dots, p_n) = H(\alpha, \beta) + \alpha H\left(\frac{p_1}{\alpha}, \frac{p_2}{\alpha}, \dots, \frac{p_m}{\alpha}\right) + \beta H\left(\frac{p_{m+1}}{\beta}, \frac{p_{m+2}}{\beta}, \dots, \frac{p_n}{\beta}\right)$$

Proof. Trivial, and left to the reader. Simply expand the right hand side. In fact, decomposability is one of the few properties used in axiomatically defining entropy. □

Khinchin had proved that given that entropy would satisfy a few simple properties, including continuity in argument, symmetry, decomposability, and some maximization criteria, that it could only have the form as we have specified or a scaled version of that (nevertheless, the argument is long and contrived). To digress, although it is a very unconventional usage, if we replace the $\log_2$ in the definition of entropy with ln, the entropy so calculated is said to be in "nats".

2.3 Bounds

Theorem 2.2. *For any discrete random variable* X *with finite image* $\mathcal{X}$,

$$0 \leq H(X) \leq \log_2 |\mathcal{X}|$$

with the lower bound achieved when $\exists$ *exactly one* $a \in \mathcal{X}$ *such that* $p(a) = 1$ *and upper bound achieved when* $p(\mathcal{X}) = \left\{\frac{1}{|\mathcal{X}|}\right\}$.

In other words, entropy is maximum when the probability distribution is uniform. To prove this, we will first need Jensen's inequality.

Lemma 2.1. *(Jensen's inequality) For a strictly convex function f on an appropriate interval, $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ for any discrete random variable X with finite image, with equality iff the random variable is constant.*

In fact it is true for random variables with infinite images also, but we restrict for convenience.
To recollect, a function f is said to be strictly convex over an interval (a, b) iff $\forall x_1, x_2 \in (a, b)$, and $\forall \lambda \in [0, 1]$, $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$, with equality iff $\lambda \in \{0, 1\}$.

Proof. This proof is instructive. We induct on $|\mathcal{X}|$. For $|\mathcal{X}| = 2$, say $\mathcal{X} = \{a_1, a_2\}$, the assertion is that $p_1 f(a_1) + p_2 f(a_2) \geq f(p_1 a_1 + p_2 a_2)$, which follows directly from the convexity of f since $p_2 = 1 - p_1$.
Now let the inequality be true for $|\mathcal{X}| = n$. Then,

$$\begin{aligned}
\sum_{i=1}^{n+1} f(a_i)p_i &= f(a_{n+1})p_{n+1} + (1-p_{n+1}) \sum_{i=1}^{n} f(a_i) \left(\frac{p_i}{1-p_{n+1}} \right) \\
&\geq f(a_{n+1})p_{n+1} + (1-p_{n+1}) f\left(\sum_{i=1}^{n} a_i \left(\frac{p_i}{1-p_{n+1}} \right) \right) \quad \text{(Induction hypothesis)} \\
&\geq f\left(a_{n+1}p_{n+1} + (1-p_{n+1}) \sum_{i=1}^{n} a_i \left(\frac{p_i}{1-p_{n+1}} \right) \right) \qquad (f \text{ is convex}) \\
&= f\left(\sum_{i=1}^{n+1} a_i p_i \right)
\end{aligned}$$

So it is true for $|\mathcal{X}| = n + 1$. We are done. □

Exercise 2.2. *Prove the condition for equality in Jensen's inequality.*

Now we may prove the upper bound for entropy.

Proof. We use Jensen's inequality. Let $p(\mathcal{X}) = \{p_1, p_2, \ldots, p_n\}$. Construct a random variable Y which takes values $g(p_i)$ with probability p_i where

$$g(p_i) = \begin{cases} \frac{1}{p_i}, & p_i > 0 \\ 1, & p_i = 0 \end{cases}$$

and let the function be $f(x) = -\log_2(x)$, which is indeed strictly convex. Then,

$$\begin{aligned}
\mathbb{E}[f(Y)] &\geq f(\mathbb{E}[Y]) \\
\Rightarrow \sum_{i=1}^{n} p_i(-\log_2(g(p_i)) &\geq -\log_2\left(\sum_{i=1}^{n} p_i g(p_i) \right) \\
\Rightarrow \sum_{x \in \operatorname{supp} X} p(x)\log_2(p(x)) &\geq -\log_2(|\operatorname{supp} X|) \\
\Rightarrow H(X) \leq \log_2(|\operatorname{supp} X|) &\leq \log_2(|\mathcal{X}|)
\end{aligned}$$

And if $p(a) = \frac{1}{|\mathcal{X}|} \forall a \in \mathcal{X}$, the random variable is constant. In such a case, $\operatorname{supp} X = \mathcal{X}$ and $H(X) = \log_2(|\mathcal{X}|)$. □

Exercise 2.3. *Prove the lower bound for entropy.*

3 Asymptotic Equipartition Property

We have already seen that we can treat $p(X)$ as a random variable. We can talk about joint probability mass functions also in a similar fashion. Now we see an important result called the Shannon–McMillan–Breiman theorem or the asymptotic equipartition property.

Theorem 3.1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables with $X_1 \sim X$ (the symbol "$\sim$" means that X_1 is equal in distribution to X) where X is a discrete random variable with finite image. Then,*

$$-\frac{1}{n} \log_2 \left(p(X_1, X_2, \ldots, X_n) \right) \xrightarrow{a.s.} H(X)$$

Proof. Let $Y_i = -\log_2 p(X_i)$, so that $Y_1, Y_2, \ldots$ are i.i.d. random variables with $\mathbb{E}[Y_i] = H(X)$ by definition. Also,

$$-\log_2 p(X_1, X_2, \ldots, X_n) = -\sum_{i=1}^{n} \log_2 p(X_i) = \sum_{i=1}^{n} Y_i.$$

since all the X_i are independent, so their joint probability mass function is the product of their individual probability mass functions. The required result is now immediate using the strong law of large numbers. □

Exercise 3.1. *Think about what $p(X_1, X_2, \ldots, X_n)$ means qualitatively.*

In the ensuing week 1 assignment, we shall see some important uses of the AEP.