Note 09. Information theory I

Alex Fu

Fall 2022

1 Introduction

The field of information theory is perhaps unusual in that it is often considered to be founded by one person, in one paper — Claude Shannon, in his seminal 1948 paper "A Mathematical Theory of Communication." There are two questions that lie at the heart of information theory:

1. **Source coding**. How many bits does one need to losslessly represent an observation?
2. **Channel coding**. How reliably and quickly can one send a message over a noisy channel?

In this note, we will examine basic information measures related to *entropy*. Two famous results of Shannon, the *source coding theorem* and *channel coding theorem*, are given in the next note. For readers with further interest, see "A Mathematical Theory of Communication," EE 229A, and *Elements of Information Theory* by Cover and Thomas.

2 Preliminaries

Definition 1 (Linearity; convexity; concavity).

A real-valued function $f : X \to \mathbb{R}$ is

- **linear** if for every $x, y \in X$ and $a, b \in \mathbb{R}$, $f(ax + by) = af(x) + bf(y)$.
- **convex** if for every $x, y \in X$ and $t \in [0, 1]$, $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$.
- **concave** if for every $x, y \in X$ and $t \in [0, 1]$, $f(tx + (1-t)y) \geq tf(x) + (1-t)f(y)$.

Convex and concave are also called *concave up* and *concave down* respectively. f is **strictly** convex or concave if the inequality is strict. Equivalently, f is convex if for every $x \in X$,

$$f(x) = \sup \{\ell(x) = ax + b : \ell \leq f\}.$$

Dually, f is concave if f is the pointwise infimum of linear functions which upper bound it.

The *convex combination* $tf(x) + (1-t)f(y)$ traces out the line segment from $f(y)$ to $f(x)$ as the time parameter t varies from 0 to 1. Consider the trajectory given by another "particle" along $f(tx+(1-t)y)$. If f is convex or concave, how do these two trajectories compare? How are the two sets of definitions equivalent?

Proposition 1 (Properties of convex and concave functions).

a. f is (strictly) convex iff $-f$ is (strictly) concave.

b. f is affine iff it is both convex and concave: $f(tx+(1-t)y) = tf(x)+(1-t)f(y)$ for all $t \in [0,1]$ and $x, y \in X$.

c. Linear combinations of convex functions are convex if the coefficients are nonnegative. The same holds for concave functions.

d. If f is twice-differentiable, then f is convex iff $f''(x) \geq 0$, concave iff $f''(x) \leq 0$, and strictly so if the inequalities are strict.

e. Optionally, a region A is convex if the line segment connecting any two points $x, y \in A$ lies entirely inside A. Then a function f is convex iff its epigraph is convex, concave iff its hypograph is convex, and linear iff both are convex, where

$$\begin{aligned} \text{graph}(f) &:= \{(x,y) \in X \times Y \mid y = f(x)\} \\ \text{epigraph}(f) &:= \{(x,r) \in X \times \mathbb{R} \mid r \geq f(x)\} \\ \text{hypograph}(f) &:= \{(x,r) \in X \times \mathbb{R} \mid r \leq f(x)\}\,. \end{aligned}$$

Examples of functions both convex and concave, in fact the only examples, are affine functions $\ell(x) = ax + b$. The polynomials of even degree x^{2k}, or more generally $|x^n| = |x|^n$, are convex. One example of a concave function: the logarithm $\log x$, central to information theory.

Definition 2 (Logarithm).

The **logarithm** of $x > 0$ in *base* $a > 0$, $a \neq 1$, is the real number $\log_a(x)$ such that

$$a^{\log_a(x)} = x.$$

In information theory, a logarithm with an unspecified base is assumed to be in base 2.

Proposition 2 (Properties of the logarithm).

a. $\log_a\colon (0,\infty) \to \mathbb{R}$ is strictly increasing, bijective, and strictly concave.

b. **Change of base.** All logarithms are constant multiples of each other:

$$\log_a(x) = \frac{1}{\log_b(a)} \log_b(x) \propto \log_b(x).$$

We can thus leave the base of a logarithm unspecified without much issue. Special cases to note include $\log_a(a) = 1$ and $\log_a(b) \cdot \log_b(a) = 1$.

c. **Product-to-sum**. Another one of the most important properties of the logarithm:

$$\log(ab) = \log(a) + \log(b).$$

It follows that $\log(\frac{a}{b}) = \log(a) - \log(b)$ and $\log(a^b) = b\log(a)$. Special cases to note include $\log(1) = 0$; $\log(\frac{1}{b}) = \log(b^{-1})$; and $\log(\sqrt[n]{b}) = \frac{1}{n}\log(b)$.

d. Other properties.* These identities are included here for the sake of reference:

$$\log_{\frac{1}{a}}(b) = -\log_a(b)$$

$$\log_{a^m}(b^n) = \frac{n}{m}\log_b(a)$$

$$\log_a(b) \cdot \log_b(c) = \log_a(c).$$

One of the most useful results for working with information measures is a probabilistic version of **Jensen's inequality**, which is applicable to the convex function $-\log_2$.

Theorem 1 (Jensen's inequality).

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then for any real-valued random variable X,

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X))$$

with equality iff $\varphi(X)$ is not strictly convex. Dually, if φ is concave, then the inequality in the opposite direction holds.

3 Derivation of entropy*

How much information is given by an event? An information function $I \colon \mathcal{F} \to [0, \infty]$ should satisfy a few properties. The quantity of information $I(A)$ should only depend on the probability $\mathbb{P}(A)$, so we could also write $I(p)$. The reciprocal $I(A) := \mathbb{P}(A)^{-1}$ is a good starting point: more surprising events give more information.

More specifically, a certain or near-certain event, like "the sun rises in the morning," gives little to no information, while a less likely event, like "tomorrow's lottery numbers are 1–2–3–4–5," gives more information. Also, independent events should give "unrelated" information somehow. So,

Proposition 3 (Information of an event).

The information function $I(\cdot) := -\log(\mathbb{P}(\cdot))$ is the unique function, up to a constant multiple, that satisfies the following three properties:

1. If A is an event with probability 1, then $I(A) = 0$.
2. $I(\cdot)$ is nonincreasing with respect to $\mathbb{P}(\cdot)$: if $\mathbb{P}(A) \leq \mathbb{P}(B)$, then $I(A) \geq I(B)$.
3. If $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$, then $I(A \cap B) = I(A) + I(B)$.

Considering the event $\{X = x\}$ for a discrete random variable X, we obtain

Definition 3 (Surprisal).

The **surprisal** or *self-information* of the value x being taken on by X is

$$I_X(x) := -\log_2 \mathbb{P}(X = x) = \log_2 \frac{1}{p_X(x)}.$$

Base 2 is conventionally chosen so that the unit of information is the binary digit, or *bit*.

From here, it is natural to consider defining $I(X)$, the information content of a random variable. We could consider summing every surprisal $I_X(x)$, but more probable values should contribute more to the overall information. By weighting probabilistically, we obtain the expected surprisal — the *information entropy* of a random variable.

4 Information measures

Definition 4 (Entropy).

The **entropy** of a discrete random variable X with pmf p is its expected surprisal

$$H(X) := \mathbb{E}\left(\log_2 \frac{1}{p(x)}\right) = \sum_{x \in S} p(x) \log_2 \frac{1}{p(x)}.$$

$H(X)$ is also written $H(p)$, as it is determined by only the distribution. If $p(x) = 0$, we set $-p(x) \log_2 p(x) = 0$ conventionally.

We note that $x \log x$ is convex in x, so entropy is convex in the pmf p, which allows us to bring in Jensen's inequality.

Proposition 4 (Lower and upper bounds on entropy).

a. $H(X) \geq 0$, with equality iff X is degenerate: $p(x) = 1$ for some $x \in S$.

b. $H(X) \leq \log_2 |S|$, with equality iff X is uniform: $p(x) = \frac{1}{|S|}$ for all $x \in S$.

So, every discrete distribution falls along a scale between the degenerate distribution, or point mass, and the uniform distribution, in terms of their uncertainty.

A natural question to ask is why variance, or other measures of spread, are unsuitable as measures of information. X and $2X$ give the same amount of information, but their spreads are a factor of 4 apart. However, the probability values p_i of their distributions are the same. Having information independent of the type of value taken on, i.e. determined only by the distribution, also means that entropy can apply to arbitrary data, such as symbols in an alphabet.

Exercises: find and plot the entropy of a coin flip, $H_b(p) := H(\text{Bernoulli}(p))$, as a function of p. Then find the entropy of a geometric random variable, $H(\text{Geometric}(p))$.

Just as multiple random variables have joint and conditional distributions, we can define joint and conditional entropy analogously as the expected surprisal of a distribution. Their interpretations are quite natural as well — the expected information content of X and Y taken together, and the expected uncertainty remaining in Y even after X is known.

Definition 5 (Joint entropy).

The **joint entropy** of discrete random variables X and Y is

$$H(X,Y) := \mathbb{E}\left(\log_2 \frac{1}{p_{X,Y}(x,y)}\right) = \sum_{x,y} p(x,y) \log_2 \frac{1}{p(x,y)}.$$

Definition 6 (Conditional entropy).

The **conditional entropy** of Y given X is

$$\begin{aligned} H(Y \mid X) &:= \mathbb{E}\left(\log_2 \frac{1}{p_{Y|X}(y \mid x)}\right) \\ &= \sum_{x,y} p_{X,Y}(x,y) \log_2 \frac{1}{p_{Y|X}(y \mid x)}. \\ &= \sum_{x \in S} p_X(x) \cdot H(Y \mid X = x). \end{aligned}$$

Proposition 5 (Properties of joint and conditional entropy).

a. **Symmetry**. $H(X,Y) = H(Y,X)$.

b. **Conditioning**. The information content of Y can only be reduced given X:

$$H(Y \mid X) \leq H(Y)$$

with equality iff X and Y are independent. Equivalently, $H(X,Y) \leq H(X) + H(Y)$.

c. **Lower and upper bounds**. Analogously to the union bound,

$$\max_{1 \le i \le n} H(X_i) \le H(X_1, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i).$$

d. **Chain rule**. The joint information conveyed by X and Y is the sum of information in X, plus the information in Y gained even after observing X.

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y).$$

See if you can find an analogue to Bayes' rule. More generally,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, X_{i-1}).$$

Optionally, see if you notice any connections between the Gram–Schmidt procedure, the chain rule for entropy, and the disjointization of a countable union, all of which are processes which remove "redundacy" step-by-step.

Mutual information, the information about one variable gained from the other variable, is an information measure which does not correspond directly to any probability distribution, but allows us to recast the previous identities:

Definition 7 (Mutual information).

The **mutual information** of X and Y is

$$I(X; Y) := H(X) - H(X \mid Y).$$

Proposition 6 (Properties of mutual information).

a. **Symmetry**. $I(X; Y) = I(Y; X)$.

b. **Nonnegativity**. $I(X; Y) \ge 0$, with equality iff X and Y are independent.

c. **Analogue of Bayes' rule**.

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X, Y) \\ &= H(X, Y) - H(X \mid Y) - H(Y \mid X). \end{aligned}$$

d. **Upper bound**. $I(X; Y) \le \min(H(X), H(Y))$

e. **Mutual information with self**. $I(X; X) = H(X)$.

We will find it helpful to pictorially summarize the information measures of two random variables X and Y as a Venn diagram.

- The entropies $H(X)$ and $H(Y)$ are analogous to the areas of the regions X and Y, which are "sets of information."
- The joint entropy $H(X, Y)$ is analogous to the union $X \cup Y$ of the sets of information. The union is disjoint iff the information is disjoint, i.e. the random variables are independent.
- The conditional entropy $H(Y \mid X)$ is analogous to the set difference $Y \setminus X$, the information of Y that is *not* given by X.
- The mutual information $I(X; Y)$ is analogous to the intersection $X \cap Y$.
- The chain rule is analogous to the principle of inclusion-exclusion!

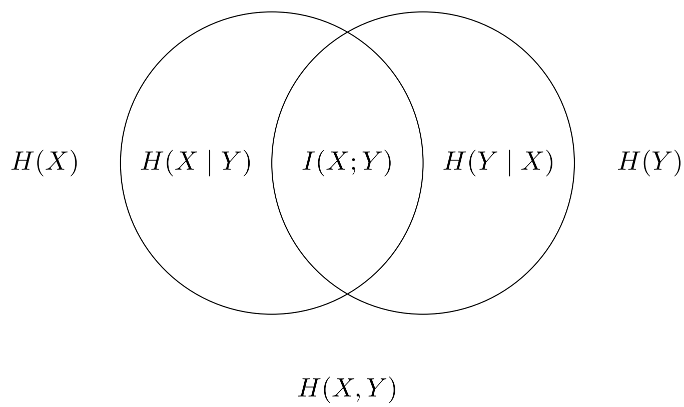

Figure 1: Entropy, joint entropy, conditional entropy, and mutual information.

5 Further information measures*

We introduce some information measures which allow us to compare the uncertainties of two distributions, and generalize the previous discussion to the continuous case.

Definition 8 (Kullback-Leibler divergence).

The **relative entropy** or **Kullback–Leibler divergence** between two discrete probability distributions p and q is

$$D_{\mathrm{KL}}(p \parallel q) := \mathbb{E}_{x \sim p}\left(\log_2 \frac{p(x)}{q(x)}\right) = \sum_{x \in S} p(x) \log_2 \frac{p(x)}{q(x)}.$$

Proposition 7 (Properties of KL divergence).

a. **Asymmetry**. $D_{\mathrm{KL}}(p \parallel q) \neq D_{\mathrm{KL}}(q \parallel p)$ in general, so KL divergence only informally describes the "distance" between two distributions.

b. **Gibbs' inequality**. $D_{\mathrm{KL}}(p \parallel q) \geq 0$, with equality iff $p = q$. A hint for the proof: Jensen's inequality will be your friend in showing $-D(p \parallel q) \leq 0$.

Definition 9 (Variation of information).

The **variation of information** or *shared information metric* is

$$H(X, Y) - I(X; Y) = H(X \mid Y) + H(Y \mid X) = H(X) + H(Y) - 2I(X; Y).$$

In the Venn diagram, variation of information is analogous to the symmetric set difference $X \triangle Y = (X \cup Y) \setminus (X \cap Y) = (X \setminus Y) \cup (Y \setminus X)$. It forms a genuine metric, or distance function, on distributions.

Definition 10 (Cross entropy).

The **cross entropy** of a discrete distribution q relative to p is

$$H(p, q) := \mathbb{E}_{x \sim p}\left(\log_2 \frac{1}{q(x)}\right) = \sum_{x \in S} p(x) \log_2 q(x).$$

Equivalently, $H(p, q) = H(p) + D_{\mathrm{KL}}(p \parallel q)$, which is asymmetric, and minimized with value $H(p)$ iff $p = q$.

Finally, the natural generalization of information measures to continuous random variables indeed works. We will elect to only give a surface-level introduction here.

Definition 11 (Differential entropy).

The **differential entropy** of a continuous random variable X with pdf f is

$$h(X) = h(f) := \mathbb{E}\left(\log_2 \frac{1}{f(x)}\right) = -\int_S f(x) \log_2 f(x)\, dx.$$

The other information measures all follow analogously by substituting the pmf for the pdf and the sum for the integral.

Proposition 8 (Properties of differential entropy).

a. **Translation-invariance**. $h(X + c) = h(X)$ for any $c \in \mathbb{R}$.

b. **Normal maximizer**. Among all continuous distributions with fixed variance σ^2, the normal distribution $\mathcal{N}(\mu, \sigma^2)$ maximizes differential entropy.

■