Information Theory Notes: Lecture III (Draft)

ChatGPT, proofread by OM

September 2nd 2025

1 Properties of Shannon's entropy function

Let X be a discrete rv on a finite alphabet $\mathcal{X} = \{1, \dots, n\}$ with pmf $p = (p_1, \dots, p_n)$. For a log base $b > 1$,

$$H_b(X) \coloneqq -\sum_{i=1}^{n} p_i \log_b p_i, \qquad 0 \log 0 := 0.$$

(Everything below is for discrete entropy.)

1) Nonnegativity

For $t \in [0, 1]$, $\log_b t \leq 0$, hence $-t \log_b t \geq 0$. Summing,

$$H_b(X) = \sum_i (-p_i \log_b p_i) \geq 0,$$

with equality iff one $p_i = 1$ (i.e., X is a.s. constant).

2) Maximized by the uniform distribution (via Jensen's inequality)

Statement of Jensen's inequality: Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a *convex function,* and let X be a random variable with finite expectation. Then

$$\varphi\big(\mathbb{E}[X]\big) \;\leq\; \mathbb{E}[\varphi(X)].$$

If φ is *concave,* the inequality is reversed:

$$\varphi\big(\mathbb{E}[X]\big) \;\geq\; \mathbb{E}[\varphi(X)].$$

Equality holds if φ is affine (linear) or if X is a constant with probability one.

Let $\varphi(t) = -t \log_b t$. Then φ is *concave* on $[0, 1]$ (see 6). Apply Jensen to the simple average over the n coordinates:

$$\frac{1}{n}\sum_{i=1}^{n}\varphi(p_i) \;\le\; \varphi\Big(\frac{1}{n}\sum_{i=1}^{n}p_i\Big) = \varphi\Big(\frac{1}{n}\cdot 1\Big) = -\frac{1}{n}\log_b\frac{1}{n}.$$

Multiplying by n,

$$H_b(X) = \sum_i \varphi(p_i) \;\le\; n\Big(-\frac{1}{n}\log_b\frac{1}{n}\Big) = \log_b n.$$

Equality in Jensen's inequality for a strictly concave φ holds iff all arguments are equal, i.e., $p_i = 1/n$ for all i. So entropy is maximized by the uniform pmf and the maximum is $\log_b n$.

3) Changing the logarithm base

For $a, b > 1$, $\log_b z = \dfrac{\log_a z}{\log_a b}$. Hence

$$H_b(X) = -\sum_i p_i \frac{\log_a p_i}{\log_a b} = \frac{1}{\log_a b}\, H_a(X).$$

Changing base just rescales by the constant $1/\log_a b$.

4) Binary entropy is concave

For $p \in [0,1]$, $H_b(p) = -p\log_b p - (1-p)\log_b(1-p)$. Using natural logs ($\log_b x = \ln x/\ln b$):

$$H_b'(p) = -\frac{\ln p}{\ln b} - \frac{1}{\ln b} + \frac{\ln(1-p)}{\ln b} + \frac{1}{\ln b} = \frac{\ln(1-p) - \ln p}{\ln b},$$

and

$$H_b''(p) = -\frac{1}{p\ln b} - \frac{1}{(1-p)\ln b} \;<\; 0 \quad (p \in (0,1)).$$

Thus, H_b is strictly concave on $(0,1)$.

5) Entropy is concave in general

Again set $\varphi(t) = -t\log_b t$, $t \ge 0$, $\varphi(0) = 0$. For $t > 0$,

$$\varphi''(t) = -\frac{1}{t\ln b} \;<\; 0,$$

so φ is concave on $[0,1]$. For pmfs p, q and $\lambda \in [0,1]$,

$$H_b\big(\lambda p + (1-\lambda)q\big) = \sum_i \varphi\big(\lambda p_i + (1-\lambda)q_i\big) \;\ge\; \lambda\sum_i \varphi(p_i) + (1-\lambda)\sum_i \varphi(q_i) = \lambda H_b(p) + (1-\lambda)H_b(q).$$

Hence $H_b(\cdot)$ is concave on the probability simplex (strict except along coordinates where p_i and q_i are simultaneously 0 or equal).

2 Joint and conditional entropy, chain rule of entropy

The joint entropy of two discrete random variables X and Y measures the total uncertainty associated with the pair (X, Y). It is defined as:

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

where $p(x, y)$ is the joint probability mass function of X and Y. It quantifies how much information is needed on average to describe both X and Y together.

The conditional entropy of Y given X, written $H(Y|X)$, measures the average uncertainty remaining about Y when X is known:

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

Equivalently, it can be expressed as an expectation:

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X=x).$$

It tells us how uncertain Y is once we already know the value of X.

Conditioning reduces entropy (proved via concavity + mixing)

Let $p^{(y)} = p\{X|Y = y\}$ be the conditional pmf of X given $Y = y$. The marginal of X equals

$$p(X) = \sum_y P(Y=y)\, p^{(y)}.$$

Concavity of $H_b(\cdot) = \sum_i \varphi(\cdot)$ gives

$$H_b(p) \;\geq\; \sum_y P(Y=y)\, H_b\big(p^{(y)}\big) \;=\; H_b(X|Y).$$

Thus $H_b(X) \geq H_b(X|Y)$. Equality holds iff all $p^{(y)}$ are identical (i.e., X is independent of Y).

Note that $H(Y|X) = H(Y)$ if X, Y are **independent random variables,** and in this case, conditioning preserves entropy.

The chain rule of entropy relates joint entropy to conditional entropy. For two variables X and Y:

$$H(X,Y) = H(X) + H(Y|X).$$

More generally, for n variables $X_1, X_2, \ldots, X_n$:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, X_2, \ldots, X_{i-1}).$$

To describe a sequence of random variables, you can describe them one at a time, with each step reducing uncertainty based on what's already known.

3 Submodular functions

Let $f : 2^N \to \mathbb{R}$ and write the marginal gain of an element $e \notin S \subseteq N$ as

$$\Delta_f(e \mid S) \;:=\; f(S \cup \{e\}) - f(S).$$

f is submodular iff for all $A, B \subseteq N$,

$$\boxed{f(A) + f(B) \;\geq\; f(A \cup B) + f(A \cap B)}$$

("submodularity inequality" over the union/intersection lattice).

A set function f is said to have diminishing returns iff for all $A \subseteq B \subseteq N$ and $e \notin B$,

$$\boxed{\Delta_f(e \mid A) \;\geq\; \Delta_f(e \mid B)}$$

i.e., the marginal gain of adding e does not increase as the set grows.

The above two definitions are equivalent.

(DR) $\implies$ (L)

Fix $A, B \subseteq N$. Let $C := A \cap B$ and $X := A \setminus B$. Enumerate $X = \{x_1, \ldots, x_k\}$ and define

$$X_i := \{x_1, \ldots, x_i\}, \qquad X_0 := \emptyset.$$

Then

$$f(A \cup B) - f(B) = f(B \cup X) - f(B) = \sum_{i=1}^{k} \big(f(B \cup X_i) - f(B \cup X_{i-1})\big) = \sum_{i=1}^{k} \Delta_f(x_i \mid B \cup X_{i-1}).$$

By $C \subseteq B$ and $x_i \notin B \cup X_{i-1}$, the DR property gives

$$\Delta_f(x_i \mid B \cup X_{i-1}) \;\leq\; \Delta_f(x_i \mid C \cup X_{i-1}).$$

Summing over i yields

$$f(A \cup B) - f(B) \;\leq\; \sum_{i=1}^{k} \big(f(C \cup X_i) - f(C \cup X_{i-1})\big) = f(C \cup X) - f(C) = f(A) - f(A \cap B).$$

Rearranging gives $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$.

(L) $\implies$ (DR)

Fix $A \subseteq B \subseteq N$ and $e \notin B$. Apply (L) to the pair of sets $A \cup \{e\}$ and B:

$$f(A \cup \{e\}) + f(B) \;\geq\; f\big((A \cup \{e\}) \cup B\big) + f\big((A \cup \{e\}) \cap B\big).$$

Since $e \notin B$ and $A \subseteq B$,

$$(A \cup \{e\}) \cup B = B \cup \{e\}, \qquad (A \cup \{e\}) \cap B = A.$$

Thus

$$f(A \cup \{e\}) + f(B) \;\geq\; f(B \cup \{e\}) + f(A),$$

which rearranges to

$$\Delta_f(e \mid A) \;\geq\; \Delta_f(e \mid B).$$

Submodularity is considered a form of "discrete concavity." In continuous analysis, concavity captures diminishing returns: the slope of a concave function decreases as you move right. For set functions, the diminishing returns formulation of submodularity

$$f(A \cup \{e\}) - f(A) \;\;\geq\;\; f(B \cup \{e\}) - f(B), \quad (A \subseteq B, e \notin B)$$

says exactly the same thing: marginal gains decrease as the "set" (the analog of the domain variable) grows. In this sense, submodular functions behave like concave functions defined on the integer lattice $\{0, 1\}^n$.