University of Houston

Classical and Quantum Information Theory

Math 6397

Li Gao

Khalid Hourani

Contents

1 Overview

Information theory studies the processing, quantification, storage, and communication of information.

- 1948 — Claude Shannon defines *Shannon Entropy* in "The Mathematical Theory of Communication." Answers questions:
  1. What is information?
  2. How do we quantify information?
  3. How do we transmit information?
- 2001 — Shannon Award is created, with Shannon the first recipient.
- 1900 — Max Plank describes Black-body Radiation
- 1920s — Heisenberg, Bohr, and Schrödinger, Matrix Mechanics
- 1930s — Hilbert, Dirac, Von Neumann describe the Hilbert Space, Mathematical foundation of Quantum Mechanics, and Von Neumann Entropy
- Interaction: Quantum Information
- 1950s – 1970s — Mathematical Quantities of Information
- 1970s
  - Information Transmission by Coherent Laser
  - Alexander Holevo — Holevo Bound
    - 1998 — Holevo et al show bound is tight (receive 2017 Shannon Award)
- 1980s — Richard Feynman: Computing with Quantum Mechanical Model
- 1990s — Peter Schor: Quantum Algorithm for Prime Factorization
  - In general, the only known algorithm for determining the prime factors of a number is naïve factorization. For example, given $n = 4801 \times 35317 = 169556917$, to retrieve the factors 4801 and 35317 requires substantially more time than to simply construct the number via multiplication.
- let's finish the rest of the trivia chapter later

2 Probability Theory

A discrete probability space $(\Omega, \mathbb{P})$ is given by

- a finite or countably infinite set Ω
  - e.g. $\{a, b, c, d\}$, $\mathbb{N} = \{0, 1, 2, \dots\}$
- a probability mass function $\mathbb{P} : \Omega \to [0, 1]$, such that
  (1) For all $\omega \in \Omega$, $\mathbb{P}(\omega) \geq 0$
  (2) $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$

  For $\omega \in \Omega$, $\mathbb{P}(\omega)$ is the probability that ω "occurs"

Definition 2.1 ▶ Event

Given a probability space $(\Omega, \mathbb{P})$, an *event* E is a subset $E \subseteq \Omega$, with corresponding probability

$$\mathbb{P}(E) = \sum_{\omega \in E} \mathbb{P}(\omega)$$

The function $\mathbb{P} : \Omega \to [0, 1]$ induces a *probability distribution*,

$$\mathbb{P} : 2^{\Omega} \to [0, 1]$$

also denoted by $\mathbb{P}$, with properties:

(1) if $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

(2) $\mathbb{P}(\Omega) = 1$

As an abuse of notation, we write $\mathbb{P}(\omega)$ and $\mathbb{P}(\{\omega\})$ interchangeably.

Example 2.1 ▶ Rolling a fair die

TBD

Definition 2.2 ▶ Conditional Probability

Let $A, B \subseteq \Omega$. The *conditional probability* of A given B, denoted by $\mathbb{P}(A \mid B)$, is defined

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Example 2.2 ▶ Fair Die Revisited

TBD

Theorem 2.1 ▶ Bayes' Rule

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\,\mathbb{P}(B)}{\mathbb{P}(A)}$$

Proof. By definition,

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$
$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

hence

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(B \mid A)\,\mathbb{P}(A) \\ &= \mathbb{P}(A \mid B)\,\mathbb{P}(B)\end{aligned}$$

from which the result follows. □

Example 2.3 ▶ Flipping a fair coin twice

TBD

Definition 2.3 ▶ Random Variable

A *Random Variable* X is a function

$$X : \Omega \to \mathcal{X}$$

from probability space $(\Omega, \mathbb{P})$ to a target space $\mathcal{X}$. We say X is discrete if $\mathcal{X}$ is discrete and call

$$\mathcal{X} = \{x_1, x_2, \dots\}$$

the *alphabet* of X.

Notice that X induces a distribution on $\mathcal{X}$. For any $x \in \mathcal{X}$

$$\mathbb{P}_X(x) = \mathbb{P}(\{\omega \mid X(\omega) = x\})$$

In many cases, $(X.\mathbb{P}_x)$ captures all information needed from random variable X. We write $X \sim \mathbb{P}_x$ to indicate that X has distribution $\mathbb{P}_x$ on $\mathcal{X}$.

Example 2.4 ▶ 52 Card Deck

TBD

Definition 2.4 ▶ Joint Distribution

Let $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ be two random variables. The *joint distribution* on $\mathcal{X} \times \mathcal{Y}$ is given by

$$\mathbb{P}_{XY}(X = x, Y = y) = \mathbb{P}(\{X(\omega) = x, Y(\omega) = y\})$$

For subsets $E_1 \subseteq \mathcal{X}$, $E_1 \subseteq \mathcal{Y}$

$$\mathbb{P}_{XY}(X \in E_1, Y \in E_2) = \mathbb{P}(\{X(\omega) \in E_1, Y(\omega) \in E_2\})$$

Notice that $\mathbb{P}_{XY}$ is a distribution on the product space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{XY})$.

Example 2.5 ▶ Fair Die Joint Distribution

TBD

Example 2.6 ▶ Flipping a fair coin twice joint distribution

TBD

Definition 2.5 ▶ Independent Random Variables

Two random variables X and Y are *independent* if, for any x, y

$$\mathbb{P}_{XY}(X = x, Y = y) = \mathbb{P}_X(X = x)\,\mathbb{P}_Y(Y = y)$$

Equivalently, if for any subsets E_1 and E_2

$$\mathbb{P}_{XY}(X \in E_1, Y \in E_2) = \mathbb{P}_X(X \in E_1)\,\mathbb{P}_Y(Y \in E_2)$$

Definition 2.6 ▶ Product Probability

Given two probability spaces $(\Omega_1, \mathbb{P}_1)$, $(\Omega_2, \mathbb{P}_2)$

$$\mathbb{P}_1 \times \mathbb{P}_2(E_1 \times E_2) = \mathbb{P}_1(E_1)\,\mathbb{P}_2(E_2)$$

is the product probability on $\Omega_1 \times \Omega_2$.

Thus, we have the property that X and Y are independent random variables if and only if $\mathbb{P}_{XY} = \mathbb{P}_X \times \mathbb{P}_Y$.

Example 2.7 ▶ Rank and Suit of a card

TBD

Definition 2.7 ▶ Real Random Variable

A *Real Random Variable* is a function

$$X : \Omega \to \mathbb{R}$$

For example, the height of a randomly sampled person, the value of a die, and the rank of a playing card (where Ace is 1, Jack is 11, Queen is 12, and King is 13) are all real random variables. On the other hand, the suit of a playing card is *not* a real random variable.

In the discrete case, if $X : \Omega \to \mathcal{X}$ is a random variable, then

$$\mathbb{P}_X : X \to [0, 1]$$

is a real random variable.

Definition 2.8 ▶ Conditional Distribution

Given two random variables X and Y, the conditional distribution is the real random variable given by

$$\mathbb{P}_{X|Y} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$$

where

$$\mathbb{P}_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y)$$

Given two real random variables X and Y, we can define

- $X + Y$
- $X \cdot Y$
- $f(X)$ (where $f : \mathbb{R} \to \mathbb{R}$)

as new random variables.

Definition 2.9 ▶ Expectation and Variance

The *expected value* (or expectation or mean) of a real random variable X is defined as the real number

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\,\mathbb{P}_X(X = x) = \sum_{\omega \in \Omega} X(\omega)\,\mathbb{P}(X = \omega)$$

The *variance* is defined as

$$\mathrm{Var}[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$$

Example 2.8 ▶ Expected Value and Variance of a Fair Die

TBD

Theorem 2.2 ▶ Linearity of Expectation

Let X and Y be real random variables and $a, b \in \mathbb{R}$. Then

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$$

Proof. By definition,

$$\begin{aligned}
\mathbb{E}[aX + bY] &= \sum_{\omega\in\Omega} (aX(\omega) + bY(\omega))\,\mathbb{P}(\omega) \\
&= \sum_{\omega\in\Omega} aX(\omega)\,\mathbb{P}(\omega) + \sum_{\omega\in\Omega} bY(\omega)\,\mathbb{P}(\omega) \\
&= a\sum_{\omega\in\Omega} X(\omega)\,\mathbb{P}(\omega) + b\sum_{\omega\in\Omega} Y(\omega)\,\mathbb{P}(\omega) \\
&= a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]
\end{aligned}$$

□

Corollary 2.3

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

Proof. By definition,

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[2X\,\mathbb{E}[X]] + \mathbb{E}\left[\mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}$$

□

If X and Y are independent, we have the following

Theorem 2.4

Let X and Y be independent real random variables. Then

(1) $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$

(2) $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

Proof. First, Item (1):

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xy\,\mathbb{P}_{XY}(X = x, Y = y) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xy\,\mathbb{P}_X(X = x)\,\mathbb{P}_Y(Y = y) \text{ since } X \text{ and } Y \text{ are independent} \\
&= \sum_{x\in\mathcal{X}} x\,\mathbb{P}_X(X = x) \sum_{y\in\mathcal{Y}} y\,\mathbb{P}_Y(Y = y) \\
&= \mathbb{E}[X]\,\mathbb{E}[Y]
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{Var}[X + Y] &= \mathbb{E}\left[(X + Y)^2\right] - \mathbb{E}[X + Y]^2 \\
&= \mathbb{E}\left[X^2 + 2XY + Y^2\right] - (\mathbb{E}[X] + expectation Y)^2 \\
&= \mathbb{E}\left[X^2\right] + 2\,\mathbb{E}[XY] + \mathbb{E}\left[Y^2\right] - \mathbb{E}[X]^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathbb{E}[XY] - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] \text{ by Item (1)}
\end{aligned}$$

□

Definition 2.10

A sequence of random variables X_1, X_2, ..., X_n is independent and identically distributed from $\mathbb{P}_X$ (i.i.d $\sim \mathbb{P}_X$) if

(1) for all i, $X_i \sim \mathbb{P}_x$

(2) X_1, X_2, ..., X_n are mutually independent, i.e., for any $\{i_1, i_2, \dots, i_k\} \subseteq \{1, 2, \dots, n\}$

$$\mathbb{P}(X_{i_1} X_{i_2} \dots X_{i_k}) = \mathbb{P}(X_{i_1})\,\mathbb{P}(X_{i_2}) \dots \mathbb{P}(X_{i_k})$$

Theorem 2.5 ▶ The Weak Law of Large Numbers (WLLN)

Let X_n be an infinite i.i.d. sequence drawn from $\mathbb{P}_X$. Write

$$\hat{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

and suppose $\mathrm{Var}[X]$ and $\mathbb{E}[X]$ are both finite. Then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\hat{X}_n - \mathbb{E}[X]\right| < \varepsilon\right) = 1$$

We first show the following two lemmas.

Lemma 2.6 ▶ Markov's Inequality

Let X be any non-negative random variable and $a > 0$. Then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof. Define the indicator random variable

$$1_{X\geq a} = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{if } X < a \end{cases}$$

and notice that $\mathbb{E}[1_{X\geq a}] = \mathbb{P}(X \geq a)$. Clearly, $X \geq a 1_{X\geq a}$, hence

$$\mathbb{E}[X] \geq a\,\mathbb{E}[1_{X\geq a}] = a\,\mathbb{P}(X \geq a)$$

from which the result follows. □

Lemma 2.7 ▶ Chebyshev's Inequality

Let X be any random variable with finite variance. Then

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \geq \varepsilon^2\big) \leq \frac{\mathrm{Var}[X]}{\varepsilon}$$

for any $\varepsilon > 0$.

Proof. Set $Y = (X - \mathbb{E}[X])^2$ and notice that $\mathbb{E}[Y] = \mathrm{Var}[X]$. Then,

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) &= \mathbb{P}\big(Y \geq \varepsilon^2\big) \\ &\leq \frac{\mathbb{E}[Y]}{\varepsilon^2} \text{ by Markov's Inequality} \\ &= \frac{\mathrm{Var}[X]}{\varepsilon^2} \end{aligned}$$ □

Now, we prove Theorem 2.5.

Proof. First, notice that

$$\begin{aligned}\mathbb{E}\left[\hat{X}_n\right] &= \mathbb{E}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] \\ &= \frac{1}{n} \cdot n\,\mathbb{E}[X] \text{ by Linearity of Expectation} \\ &= \mathbb{E}[X]\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}\left[\hat{X}_n\right] &= \operatorname{Var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] \\ &= \frac{1}{n^2}(\operatorname{Var}[X_1] + \operatorname{Var}[X_2] + \cdots + \operatorname{Var}[X_n]) \\ &= \frac{1}{n^2} \cdot n \operatorname{Var}[X] \\ &= \frac{1}{n} \operatorname{Var}[X]\end{aligned}$$

then, by Chebyshev's Inequality,

$$\begin{aligned}\mathbb{P}\left(\left|\hat{X}_n - \mathbb{E}[X]\right| \geq \varepsilon\right) &\leq \frac{\operatorname{Var} \hat{X}_n}{\varepsilon^2} \\ &= \frac{\operatorname{Var}[X]}{n\varepsilon^2} \to 0 \text{ as } n \to \infty\end{aligned}$$

hence

$$\mathbb{P}\left(\left|\hat{X}_n - \mathbb{E}[X]\right| < \varepsilon\right) = 1 - \mathbb{P}\left(\left|\hat{X}_n - \mathbb{E}[X]\right| \geq \varepsilon\right) \to 1 \text{ as } n \to \infty$$

□

Example 2.9 ▶ Bernoulli Random Variable

TBD

Definition 2.11 ▶ Vector Valued Random Variable

Let

$$X = (X_1, X_2, \ldots, X_n) : \Omega \to \mathbb{R}^n$$

finish this part — part in notes is a bit cryptic

3 Entropy

Definition 3.1 ▶ Entropy

Let $\mathbb{P}$ be a probability distribution on a discrete space Ω. The Shannon Entropy (hereby simply Entropy) of $\mathbb{P}$ is defined

$$H(\mathbb{P}) = \sum_{\omega \in \Omega} \mathbb{P}(\omega) \log \frac{1}{\mathbb{P}(\omega)}$$

If X is a discrete random variable, we define

$$\begin{aligned} H(X) &= H(\mathbb{P}_X) \\ &= \sum_{x \in X} \mathbb{P}_X(x) \log \frac{1}{\mathbb{P}_X(x)} \\ &= \mathbb{E}\left[\log \frac{1}{\mathbb{P}_X(X)}\right] \end{aligned}$$

noting that $\log \frac{1}{\mathbb{P}_X(X)}$ is a real random variable.

We can think of $\log \frac{1}{\mathbb{P}_X(x)}$ as the level of "surprise" that $X = x$ occurs and $H(X)$ as the uncertainty or randomness of $\mathbb{P}_X$.

Note that, in Definition 3.1, log refers to $\log_2$, and $\log_2(X)$ is the number of bits of X. Additionally, since a byte is 8 bits, $\log_{256}(X)$ is the number of bytes of X. Additionally, we define $0 \log \frac{1}{0} = 0$, which can be motivated by the fact that

$$\lim_{x \to 0^+} x \log \frac{1}{x} = 0$$

Example 3.1 ▶ Bernoulli Distribution

The Bernoulli Distribution is the discrete random variable

$$\begin{aligned} \mathbb{P}(X = 1) &= p \\ \mathbb{P}(X = 0) &= 1 - p \end{aligned}$$

and has entropy

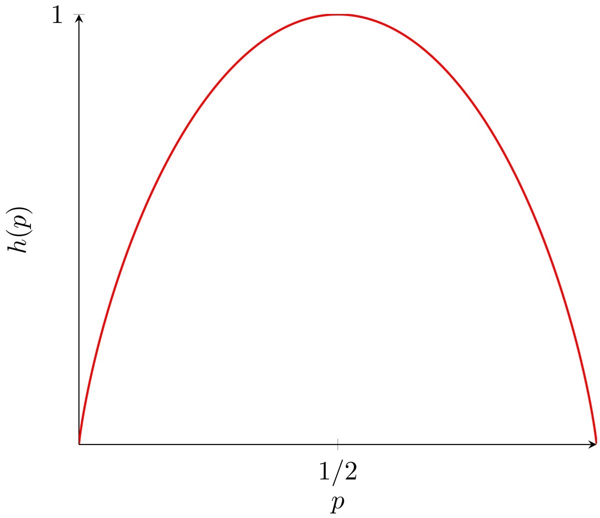
$$H(X) = p \log \frac{1}{p} + (1 - p) \log \frac{1}{1 - p}$$

Definition 3.2 ▶ Binary Entropy

The binary entropy of p, $h(p)$, is the entropy of the Bernoulli Distribution with parameter p, i.e.,

$$h(p) = p \log \frac{1}{p} + (1 - p) \log \frac{1}{1 - p}$$

Notice that $h(0) = h(1) = 0$ and $h\left(\frac{1}{2}\right) = 1$. More generally, the graph of $h(p)$ is given in Figure 1.



add some flourishes

Figure 1: Binary Entropy as a function of p. Notice that the entropy is maximized when $p = 1/2$ and 0 when $p = 0$ or $p = 1$. When $p = 0$ or $p = 1$, the Bernoulli Distribution is non-random, and thus there is no uncertainty.

Figure 2: Drawing of ant nest used to empirically verify ...

Example 3.2 ▶ Geometric Distribution

The Geometric Distribution is the positive, integer-valued random variable that describes the number of Bernoulli trials performed until a success. That is,

$$\mathbb{P}(X = k) = p(1-p)^{k-1}$$

is the probability that it will require k trials until a success.
The entropy of the Geometric Distribution is given by

$$\begin{aligned}
H(X) &= \sum_{k=1}^{\infty} p(1-p)^k \log \frac{1}{p(1-p)^k} \\
&= \sum_{k=1}^{\infty} p(1-p)^k \left(\log \frac{1}{p} + k \log \frac{1}{1-p} \right) \\
&= p \log \frac{1}{p} \sum_{k=1}^{\infty} (1-p)^k + p \log \frac{1}{1-p} \sum_{k=1}^{\infty} k(1-p)^k \\
&= p \frac{1}{p} \log \frac{1}{p} + p \log \frac{1}{1-p} \frac{1-p}{p^2} \\
&= \frac{1}{p} \left(p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \right) \\
&= \frac{h(p)}{p} \to 0 \text{ as } p \to 0^+
\end{aligned}$$

Example 3.3 ▶ Distribution with ∞ Entropy

TBD

An empirical justification for the use of $\log_2$.

Finish figure, caption, and description.

Definition 3.3 ▶ Convexity

Let $V \cong \mathbb{R}^n$ be a vector space. A subset $S \subseteq V$ is convex if, for any pair $\mathbf{x}, \mathbf{y} \in S$

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in S \text{ for all } \lambda \in [0,1]$$

figure demonstrating convexity

Example 3.4

The following are convex

(1) $\mathbb{R}^n$

(2)

(3)

Definition 3.4 ▶ Convex Function

A function $f : S \to \mathbb{R}$ is

(i) convex if $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0,1]$

(ii) *strictly* convex if $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0,1]$

Definition 3.5 ▶ Concave Function

A function $f : S \to \mathbb{R}$ is (strictly) concave if $-f$ is (strictly) convex.

Example 3.5

Notice

(1) The function $x \to x \log x$ is strictly convex

(2) The function $x \to \log x$ is strictly concave

(3) The function $X \to \mathbb{E}[X]$ is convex (but not strictly)

Theorem 3.1 ▶ Jensen's Inequality

Let X be a real vector valued random variable. Then, if f is any convex function,

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

If f is strictly convex, then $f(\mathbb{E}[X]) = \mathbb{E}[f(X)]$ if and only if $X = \mathbb{E}[X]$, i.e., X is a constant random variable.

Proof. Since f is convex,

$$\begin{aligned} f(\mathbb{E}[X]) &= f\left(\sum_{x \in X} x\, \mathbb{P}(X = x)\right) \\ &\leq \sum_{x \in X} f(x)\, \mathbb{P}(X = x) \text{ since } f \text{ is convex and } \mathbb{P}(X = x) \in [0,1] \\ &= \mathbb{E}[f(X)] \end{aligned}$$

□

Theorem 3.2 ▶ Properties of Entropy

The Entropy function satisfies

(1) $H(X) \geq 0$ with equality if and only if X is constant

(2) if $\mathcal{X}$ is finite, then $H(X) \leq \log|\mathcal{X}|$ with equality if and only if $\mathbb{P}_X$ is uniform on $\mathcal{X}$

(3) For any injective f, $H(X) = H(f(X))$

(4) $\mathbb{P} \to H(\mathbb{P})$ is strictly concave

Proof.

(1) $H(X) = \mathbb{E}\left[\log \frac{1}{\mathbb{P}_X}\right] \geq 0$ with equality if and only if $\log \frac{1}{\mathbb{P}_X} = 0$, which occurs only when $\mathbb{P}_X \equiv 1$.

(2) If $\mathcal{X}$ is finite, then

$$\begin{aligned} H(X) &= \mathbb{E}\left[\log \frac{1}{\mathbb{P}_X}\right] \\ &\leq \log \mathbb{E}\left[\frac{1}{\mathbb{P}_X}\right] \\ &= \log \sum_{x \in X} \mathbb{P}(x) \frac{1}{\mathbb{P}(X)} \\ &= \log|\mathcal{X}| \end{aligned}$$

with equality if and only if $\log \frac{1}{\mathbb{P}_x}$ is constant, which forces $\mathbb{P}(X) = \frac{1}{|\mathcal{X}|}$.

(3) If f is injective, then $\mathbb{P}_{f(X)}(f(x)) = \mathbb{P}_X(x)$, and the result follows.

(4) Take $\lambda \in [0,1]$ and write $f(x) = x \log \frac{1}{x}$, then

$$\begin{aligned} H(\lambda\mathbb{P}_1 + (1-\lambda)\mathbb{P}_2) &= \sum_{\omega \in \Omega} f(\lambda\,\mathbb{P}_1(\omega) + (1-\lambda)\,\mathbb{P}_2(\omega)) \\ &\geq \sum_{\omega \in \Omega} \lambda f(\mathbb{P}_1(\omega)) + (1-\lambda) f(\mathbb{P}_2(\omega)) \\ &= \lambda \sum_{\omega \in \Omega} f(\mathbb{P}_1(\omega)) + (1-\lambda) \sum_{\omega \in \Omega} f(\mathbb{P}_2(\omega)) \\ &= \lambda H(\mathbb{P}_1) + (1-\lambda) H(\mathbb{P}_2) \end{aligned}$$

□

4 Conditional Entropy

Definition 4.1 ▶ Joint Entropy

Given random variables X and Y, the *Joint Entropy*, $H(XY)$, is defined

$$H(XY) = \mathbb{E}\left[\log \frac{1}{\mathbb{P}_{XY}}\right] = \sum_{x \in X} \sum_{y \in Y} \mathbb{P}_{XY}(X = x, Y = y) \log \frac{1}{\mathbb{P}_{XY}(X = x, Y = y)}$$

Definition 4.2 ▶ Conditional Entropy

Let X and Y be random variables. Then

$$H(X \mid Y) = \mathop{\mathbb{E}}_{y \sim \mathbb{P}_Y}\left[H\left(\mathbb{P}_{X|Y=y}\right)\right] = \mathbb{E}\left[\log \frac{1}{\mathbb{P}_{X|Y}}\right]$$

This can be thought of as the expected uncertainty $H\left(\mathbb{P}_{X|Y=y}\right)$ over $y \sim \mathbb{P}_Y$.

Definition 4.3 ▶ Conditional Probability Notation

Some notation:

(1) $\mathbb{P}_{X|Y=y}$ is a distribution on X, with

$$\begin{aligned} \mathbb{P}_{X|Y=y}(x) &= \mathbb{P}(X = x \mid Y = y) \\ &= \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} \end{aligned}$$

(2) $\mathbb{P}_{X|Y}$ is a random variable on $\mathcal{X} \times \mathcal{Y}$ with

$$\mathbb{P}_{X|Y}(x, y) = \mathbb{P}(X = x \mid Y = y)$$

Example 4.1 ▶ Joint and Conditional Entropy of a Fair Die

TBD

Theorem 4.1 ▶ Properties of Conditional Entropy

Let X and Y be random variables. Then

(1) $H(X \mid Y) \leq H(X)$ with equality if and only if X and Y are independent

(2) $H(XY) = H(Y) + H(X \mid Y) \leq H(Y) + H(X)$ with equality if and only if X and Y are independent

(3) $H(XY) \geq \max\{H(X), H(Y)\}$

Proof.

(1)

$$\begin{aligned} H(X \mid Y) &= \mathop{\mathbb{E}}_{y \sim \mathbb{P}_Y}\left[H\left(\mathbb{P}_{X|Y=y}\right)\right] \\ &\leq H\left(\mathop{\mathbb{E}}_{y \sim \mathbb{P}_Y}\left[\mathbb{P}_{X|Y=y}\right]\right) \\ &= H(\mathbb{P}_X) \\ &= H(X) \end{aligned}$$

(2)

(3) $H(XY) = H(X) + H(Y \mid X) \geq H(X)$ The same argument shows $H(XY) \geq H(Y)$, hence it must be greater than or equal to the maximum of the two.

$\square$

Corollary 4.2

For any function f

(1) $H(X) = H(Xf(X))$

(2) $H(f(X) \mid X) = 0$

(3) $H(X) \geq H(f(X))$ with equality if and only if f is injective

Proof. $\square$