Contents

1 Probability

1.1 Probability Spaces

In an elementary level, we have been viewing probability as the quotient between the number of desired outcomes and the number of all possible outcomes. This definition, though intuitive, is not very solid when it comes to an infinite sample space. In this introductory chapter, we would establish the theories of probability using a more modern and rigorous structure.

Definition 1.1.1 ▸ Set Algebra

Let X be a set. A **set algebra** over X is a family $\mathcal{F} \subseteq \mathcal{P}(X)$ such that

- $X \backslash F \in \mathcal{F}$ for all $F \in \mathcal{F}$ (closed under complementation);
- $X \in \mathcal{F}$;
- $X_1 \cup X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$ (closed under binary union).

There are several immediate implications from the above definition.

First, by closure under complementation, we know that an algebra over any set X must contain the empty set.

Second, by De Morgan's Law, one can easily check that if the first 2 axioms hold, the closure under binary union is equivalent to

- $X_1 \cap X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$;
- $\bigcup_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$;
- $\bigcap_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$.

$(X, \mathcal{F})$ is known as a *field of sets*, where the elements of X are called *points* and those of $\mathcal{F}$, *complexes* or *admissible sets* of X.

In probability theory, what we are interested in is a special type of set algebras known as *σ-algebras*.

Definition 1.1.2 ▸ σ-Algebra

A **σ-Algebra** over a set A is a non-empty set algebra over A that is closed under countable union.

Of course, by the same argument as above, we known that any σ-algebra is closed under countable intersection as well.

Now, as we all know, we can take some set Ω as a *sample space* and denote an *event* by some subset of Ω. Roughly speaking, we could now define the probability of an event $E \subseteq \Omega$ as the ratio between the sets' volumes. The remaining question now is: how do we define the volume of a set properly?

Definition 1.1.3 ▸ Measure

Let X be a set and Σ be a σ-algebra over X. A **measure** over Σ is a function

$$\mu : \Sigma \to \mathbb{R} \cup \{-\infty, +\infty\}$$

such that

- $\mu(E) \geq 0$ for all $E \in \Sigma$ (non-negativity);
- $\mu(\varnothing) = 0$;
- $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$ for any countable collection of pairwise disjoint elements of Σ (countable additivity or σ-additivity).

The triple (X, Σ, μ) is known as a **measure space** and the pair (X, Σ), a **measurable space**.

One thing to note here is that if at least one $E \in \Sigma$ has a finite measure, then $\mu(\varnothing) = 0$ is automatically guaranteed for obvious reasons.

Definition 1.1.4 ▸ Probability Space

Let Ω be a sample space and $\mathcal{F}$ be a σ-algebra over Ω. A **probability space** is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathbb{P} : \mathcal{F} \to [0, 1]$, known as a **probability measure**, is such that $\mathbb{P}(\Omega) = 1$.

Obviously, the above definition immediately guarantees that

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;
2. $\mathbb{P}(A) \leq \mathbb{P}(B)$ if $\mathbb{P}(A) \subseteq \mathbb{P}(A)$;
3. $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

The third result follows from a direct application of the principle of inclusion and exclusion.

By induction, one can easily check that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(E_i\right)$$

for any finitely many events. The following proposition extends this result to countable collections of events:

Proposition 1.1.5 ▸ Union Bound of Countable Collections of Events

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $E_1, E_2, \cdots, E_n, \cdots \in \mathcal{F}$ is any countable sequence of events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}\left(E_i\right).$$

Proof. Define $F_1 \coloneqq E_1$ and $F_k \coloneqq E_k \backslash \bigcup_{i=1}^{k-1} E_i$ for $k \geq 2$. Clearly, the F_i's are pairwise disjoint. By Definition 1.1.2, the F_i's are elements of $\mathcal{F}$. Note that $\mathbb{P}\left(F_i\right) \leq \mathbb{E}_{\mathbb{i}}$ for all $i \in \mathbb{N}^+$, so

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} F_i\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}\left(F_i\right) \\
&\leq \sum_{i=1}^{\infty} \mathbb{P}\left(E_i\right).
\end{aligned}$$

□

Next, we will introduce the notion of *random variables* formally. For this purpose, we first establish the notion of a *Borel algebra*.

Definition 1.1.6 ▸ Borel Algebra

Let X be a topological space. A **Borel set** on X is a set which can be formed via countable union, countable intersection and relative complementation of open sets in X. The smallest σ-algebra over X containing all Borel sets on X is known as the **Borel algebra** over X.

Clearly, the Borel algebra over X contains all open sets in X according to the above axioms from Definition 1.1.2. This helps us define the following:

Definition 1.1.7 ▸ Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{X}, \mathcal{B})$ be a measurable space where $\mathcal{B}$ is the Borel algebra over $\mathcal{X}$. A **random variable** is a function $X : \Omega \to \mathcal{X}$ such that

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for all $B \in \mathcal{B}$.

Remark. Rigorously, such a random variable X is a *measurable function* or *measurable mapping* from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{B})$.

The probability measure $\mathbb{P}$ thus induces a probability measure P_X over $(\mathcal{X}, \mathcal{B})$.

Definition 1.1.8 ▸ Distribution

Let $X : \Omega \to \mathcal{X}$ be a random variable over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{B}$ be the Borel algebra over $\mathcal{X}$, the **distribution** of X is the probability measure P_X on $(\mathcal{X}, \mathcal{B})$ given by

$$P_X(B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

Remark. Often times, we write $\Pr(X \in B) = P_X(B)$.

In the context of information theory, we mostly are concerned with real-valued random variables only.

Definition 1.1.9 ▸ Real-Valued Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, a **real-valued random variable** over the space is a mapping $X : \Omega \to \mathbb{R}$ such that

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$.

Note that the Borel set over $\mathbb{R}$ is just the family of all open intervals.

Clearly, if X is a real-valued random variable, we have $\{\omega \in \Omega : X(\omega) > x\} \in \mathcal{F}$. Moreover, we claim that

$$\{\omega \in \Omega : X(\omega) < x\} = \bigcup_{y<x} \{\omega \in \Omega : X(\omega) \leq y\}.$$

The proof is quite straightforward and is left to the reader as an exercise. By Definition

1.1.2, this means that

$$\{\omega \in \Omega : X(\omega) < x\} \cup \{\omega \in \Omega : X(\omega) > x\} \in \mathcal{F}.$$

Therefore, $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$. This argument justifies the probabilities $\Pr(X < x)$ and $\Pr(X = x)$. We give a special name to the range of a random variable in computer science.

Definition 1.1.10 ▸ Alphabet

Let X be a random variable, the range of X is called an **alphabet**, denoted as $\mathcal{X}$.

Recall that we have defined expectations for discrete and continuous random variables in elementary probability theory. In terms of measure theory, the two formulae can be unified as the Lebesgue integral

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, \mathrm{d}\mathbb{P}(\omega).$$

Note that $\mathbb{E}[X]$ is a real number while $\mathbb{E}[X \mid Y]$ is a **random variable** formed as a function of Y. In a way, Y partitions the sample space into regions where $\mathbb{E}[X \mid Y = y_i]$ gives the expectation of X in the region induced by $Y = y_i$ for each $y_i \in \mathcal{Y}$. In general, the following result holds:

Theorem 1.1.11 ▸ Law of Iterated Expectations

Let X and Y be random variables, then $\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}[X]$.

The above formula can be interpreted as the fact that $\mathbb{E}[X \mid Y]$ is a best estimator for X.

1.2 Markov Chains

Recall that 2 random variables X and Z are *independent* if and only if $P_{X,Z}(x,z) = P_X(x) P_Z(z)$ or $P_{X|Z}(x \mid z) = P_X(x)$ for all $(x,z) \in \mathcal{X} \times \mathcal{Z}$. We will extend this definition with a third random variable.

Definition 1.2.1 ▸ Markov Chain

Let X, Y, Z be random variables. If

$$P_{X,Y,Z}(x,y,z) = P_X(x) P_{Y|X}(y \mid x) P_{Z|Y}(z \mid y)$$

for all $(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, then we say that X, Y, Z forms a **Markov chain** in this order, or that X and Z are conditionally independent on Y.

Recall also that the *Bayes's Rule* states the following:

Theorem 1.2.2 ▶ Bayes's Rule

For any random variables X and Y,

$$P_{X|Y}(x \mid y) = \frac{P_{Y|X}(y \mid x) P_X(x)}{\sum_{x' \in \mathcal{X}} P_{Y|X}(y \mid x') P_X(x')}.$$

Based on Theorem 1.2.2, we have

$$P_{X,Y}(x,y) = P_{X|Y}(x \mid y) P_Y(y) = P_X(x) P_{Y|X}(y \mid x).$$

By applying the formula repeatedly, we have

$$\begin{aligned} P_{X,Y,Z}(x,y,z) &= P_{X,Y}(x,y) P_{Z|X,Y}(z \mid x,y) \\ &= P_X(x) P_{Y|X}(y \mid x) P_{Z|X,Y}(z \mid x,y). \end{aligned}$$

Therefore, a Markov chain simply states that the distribution of Z is no longer dependent on X, but depends on Y solely. Therefore, this allows us to remove one condition when applying Theorem 1.2.2. Thus, it actually suffices to prove $P_{Z|X,Y} = P_{Z|Y}$ when proving that X-Y-Z forms a Markov chain.

We can denote a Markov chain by X-Y-Z. Intuitively, such a relationship should be symmetric.

Proposition 1.2.3 ▶ Symmetricity of Markov Chains

If X-Y-Z is a Markov chain, then Z-Y-X is also a Markov chain.

Proof. By Definition 1.2.1,

$$P_{X,Y,Z}(x,y,z) = P_X(x) P_{Y|X}(y \mid x) P_{Z|Y}(z \mid y).$$

By Theorem 1.2.2, we have

$$\begin{aligned} P_{X|Y}(x \mid y) &= \frac{P_X(x) P_{Y|X}(y \mid x)}{P_Y(y)} \\ &= \frac{P_{X,Y,Z}(x,y,z)}{P_Y(y) P_{Z|Y}(z \mid y)} \\ &= \frac{P_{X,Y,Z}(x,y,z)}{P_{Z,Y}(z,y)} \\ &= P_{X|Z,Y}(x \mid z,y). \end{aligned}$$

Therefore, Z-Y-X is a Markov chain. □

One obvious case where dependence exists between the random variables in a Markov chain is that one of the random variables is a function of another one.

Proposition 1.2.4 ▸ Markov Chain Involving Functions of a Random Variable

Let X and Y be any random variables and $Z \coloneqq f(Y)$ for some function f, then X-Y-Z is a Markov chain.

Proof. Notice that

$$P_{Z|X,Y}(z \mid x, y) = P_{f(Y)|X,Y}(z \mid x, y) = \begin{cases} 1 & \text{if } z = f(y) \\ 0 & \text{otherwise} \end{cases},$$

$$P_{Z|Y}(z \mid y) = P_{f(Y)|Y}(z \mid y) = \begin{cases} 1 & \text{if } z = f(y) \\ 0 & \text{otherwise} \end{cases}$$

for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Therefore, $P_{Z|X,Y} = P_{Z|Y}$ and so X-Y-Z forms a Markov chain. □

Note that if X and Z are independent, they are naturally conditionally independent given any Y. However, the inverse may not be true.

Proposition 1.2.5 ▸ Conditional Independence Does Not Imply Independence

There exists random variables X, Y, Z such that X and Z are dependent but conditionally independent given Y.

Proof. Let N_1, N_2, N_3 be pairwise independent random variables such that

$$\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{N}_3 = \{0, 1\}.$$

Take $X = N_1 + N_2$, $Y = N_2$ and $Z = N_2 + N_3$. Clearly, X and Z are dependent, but

$$\begin{aligned} P_{Z|X}(z \mid x) &= P_{N_2+N_3|N_1+N_2}(z \mid x) \\ &= P_{N_3|N_1,N_2}(z - y \mid x - y, y) \\ &= P_{N_2+N_3|N_1+N_2,N_2}(z \mid x, y) \\ &= P_{Z|X,Y}(z \mid x, y), \end{aligned}$$

which implies that X and Z are conditionally independent given Y. □

1.3 Probability Bounds

We use various bounds to make estimates and approximations for probability distributions. The first commonly used bound is *Markov's Inequality*.

Theorem 1.3.1 ▸ Markov's Inequality

If X is a non-negative random variable, then $\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$ *for all* $a > 0$.

Proof. It suffices to prove for the continuous case. Notice that

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x f_X(x)\,\mathrm{d}x \\ &\geq \int_a^\infty x f_X(x)\,\mathrm{d}x \\ &\geq a\int_0^\infty f_X(x)\,\mathrm{d}x \\ &= \Pr(X \geq a).\end{aligned}$$

Therefore, $\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$. □

Note that the bound given by Markov's inequality is a rather loose bound. The following inequality proposes a better bound:

Theorem 1.3.2 ▸ Chebyshev's Inequality

For any real-valued random variable X with finite variance,

$$\Pr\left(|X - \mathbb{E}[X]| > a\sqrt{\mathrm{Var}(X)}\right) \leq \frac{1}{a^2}$$

for all $a > 0$.

Proof. Define $g(X) : (X - \mathbb{E}[X])^2$, which is clearly non-negative. By Theorem 1.3.1, we have

$$\Pr\big(g(X) > a^2\mathrm{Var}(X)\big) \leq \frac{\mathbb{E}[g(X)]}{a^2\mathrm{Var}(X)}.$$

Note that $\mathbb{E}[g(X)] = \mathrm{Var}(X)$, so

$$\Pr\left(|X - \mathbb{E}[X]| > a\sqrt{\mathrm{Var}(X)}\right) = \Pr\big(g(X) > a^2\mathrm{Var}(X)\big) \leq \frac{1}{a^2}.$$

□

Finally, we state the following law of large numbers:

Theorem 1.3.3 ▶ Weak Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be pairwise independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2 \in \mathbb{R}$ for every $i \in \mathbb{N}^+$. For every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0.$$

Proof. Note that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$ and that

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sum_{i=1}^{n} \operatorname{Var}(X_i)}{n^2} = \frac{\sigma^2}{n}.$$

By Theorem 1.3.2, we have

$$0 \leq \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

By Squeeze Theorem, this clearly implies that

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0.$$

□

Alternatively, we may phrase Theorem 1.3.3 as "$\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ in probability". When a sequence $\{S_n\}_{n=1}^{\infty}$ converges to b in probability, we write $S_n \xrightarrow{\text{p}} b$.

> *Remark.* Essentially, what Theorem 1.3.3 says is that when n is large, the sample mean from n measurements of the same data converges to the expectation of the distribution.

Under some mild conditions, this convergence occurs exponentially fast, i.e., the probability $\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right)$ decreases at least as fast as $\exp(-ng(\epsilon))$ for some real-valued function $g : \mathbb{R}^+ \to \mathbb{R}^+$. In terms of asymptotic analysis, we write this as

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \exp(-ng(\epsilon) + o(n)).$$

Equivalently, this means that there exists a function $g : \mathbb{R} \to \mathbb{R}$ with $g(\epsilon) > 0$ for every

$\epsilon > 0$ such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \geq g(\epsilon) + o(1).$$

There is a strong version for the law, which shall be stated without proof:

Theorem 1.3.4 ▸ Strong Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be pairwise independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2 \in \mathbb{R}$ for every $i \in \mathbb{N}^+$, then

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1.$$

1.4 Convexity

Recall the following definition:

Definition 1.4.1 ▸ Convex Function

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if for any $\lambda \in [0,1]$ and any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$f\big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\big) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

From a graphical perspective, a convex function is an overestimate of all linear functions whose values are bounded above by it. The following proposition set this result in a rigorous context:

Proposition 1.4.2 ▸ Convex Functions as Overestimates for Linear Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and define

$$\mathcal{L} \coloneqq \left\{\ell \in \operatorname{Maps}(\mathbb{R}^n, \mathbb{R}) : \ell(\boldsymbol{u}) = \boldsymbol{a}^{\mathrm{T}} \cdot \boldsymbol{u} + b \leq f(\boldsymbol{u}) \text{ for all } \boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{a} \in \mathbb{R}^n, b \in \mathbb{R}\right\}$$

to be the set of all linear functions bounded above by f, then for each $\boldsymbol{x} \in \mathbb{R}^n$,

$$f(\boldsymbol{x}) = \sup_{\ell \in \mathcal{L}} \ell(\boldsymbol{x}).$$

Proof. It suffices to prove that for all $\boldsymbol{x} \in \mathbb{R}^n$, there exists some linear function $\ell \in \mathcal{L}$

such that $\ell(\boldsymbol{x}) = f(\boldsymbol{x})$. Take any $\boldsymbol{h} \in \mathbb{R}^n$. Since f is convex, we have

$$\begin{aligned} 2f(\boldsymbol{x}) &= 2f\left(\frac{1}{2}(\boldsymbol{x}+\boldsymbol{h}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{h})\right) \\ &\leq f(\boldsymbol{x}+\boldsymbol{h}) + f(\boldsymbol{x}-\boldsymbol{h}). \end{aligned}$$

Therefore, we have

$$L_1 = \lim_{\|\boldsymbol{h}\|\to 0} \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}-\boldsymbol{h})}{\|\boldsymbol{h}\|} \leq \lim_{\|\boldsymbol{h}\|\to 0} \frac{f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x})}{\|\boldsymbol{h}\|} = L_2.$$

Take some $a \in [L_1, L_2]$ and let $\ell(\boldsymbol{y}) = a\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x})$. Observe that $\ell(\boldsymbol{x}) = f(\boldsymbol{x})$. Take $\boldsymbol{h} = \boldsymbol{y} - \boldsymbol{x}$, then

$$\begin{aligned} \ell(\boldsymbol{y}) &= a\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x}) \\ &\leq \frac{f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x})}{\|\boldsymbol{h}\|}\|\boldsymbol{y}-\boldsymbol{x}\| + f(\boldsymbol{x}) \\ &= f(\boldsymbol{x}+\boldsymbol{h}) \\ &= f(\boldsymbol{y}). \end{aligned}$$

Therefore, $\ell \in \mathcal{L}$ as desired. □

The following proposition gives a simple test for convexity in one-dimensional case, which is a special case of the Hessian matrix test:

Proposition 1.4.3 ▸ Second Derivative Test for Convexity

If a real-valued function f is twice-differentiable on $[a, b]$, then it is convex if and only if $f''(x) \geq 0$ for all $x \in (a, b)$.

Convex functions produce the following interesting result regarding expectation:

Theorem 1.4.4 ▸ Jensen's Inequality

Let f be a convex function and X be a random variable, then $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

Proof. Let $\mathcal{L}$ be the set of all linear functions bounded above by f. By Proposition

1.4.2, we have

$$\begin{aligned}\mathbb{E}[f(X)] &= \mathbb{E}\left[\sup_{\ell\in\mathcal{L}} \ell(X)\right] \\ &\geq \sup_{\ell\in\mathcal{L}} \mathbb{E}[\ell(X)] \\ &= \sup_{\ell\in\mathcal{L}} \ell(\mathbb{E}[X]) \\ &= f(\mathbb{E}[X]).\end{aligned}$$

□

Remark. If f is strictly convex, the equality holds if and only if X is constant.

2

Information Theory

2.1 Entropy

In information theory, the very first question to ask is how we can measure the quantity of information contained in communication. Colloquially, we say that communication gives more information if more knowledge which has remained unknown previously is revealed.

We describe such revelation of new knowledge as the "surprise" of an event. Using probability theory, we use a random variable X to represent an event by $X = x$. Intuitively, an event is surprising if the probability of its occurrence is low. This is formally stated as follows:

Definition 2.1.1 ▸ Surprise

Let X be a random variable. The **surprise** of an event $X = x$ is defined as

$$\log_2 \frac{1}{P_X(x)} = \log_2 \frac{1}{\Pr(X = x)}.$$

Now, suppose we are **uncertain** about some event $X = x$. We may wish to measure how much uncertainty we have towards the outcome of the event, or equivalently, what is the **expected surprise** for the event. It is easy to see that if we define a random variable for surprise as a function of X, we can make use of the expectation formula to compute this quantity.

Definition 2.1.2 ▸ Entropy of Discrete Random Variables

Let X be a discrete random variable supported on a finite alphabet $\mathcal{X}$ with probability mass function p_X, then **entropy** of X is defined as

$$H(X) \coloneqq -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x).$$

It is clear that this definition can be manipulated into

$$H(X) = \mathbb{E}\left[\log_2 \frac{1}{p_X(X)}\right],$$

i.e., the entropy of X is exactly the expected surprise of X. There is a small problem, though,

which is that $\log_2 n$ is undefined when $n \leq 0$. Since $p_X(x) \geq 0$ for all $x \in \mathcal{X}$, we only need to take care of 0 as a special case. Notice that

$$\begin{aligned}\lim_{x\to 0^+} x \log_2 x &= \lim_{x\to 0^+} \frac{x \ln x}{\ln 2} \\ &= -\frac{1}{\ln 2} \lim_{x\to 0^+} \frac{-\ln x}{x^{-1}} \\ &= -\frac{1}{\ln 2} \lim_{x\to 0^+} \frac{-x^{-1}}{-x^{-2}} \\ &= 0.\end{aligned}$$

Therefore, it makes sense to set $x \log_2 x = 0$ when $x = 0$.

> *Remark.* By convention, we set $0 \log_2 0 = 0$.

We will later prove that $0 \leq H(X) \leq \log_2 |\mathcal{X}|$. Moreover, $H(X)$ is closely related with the minimal number of bits to encode X in binary number unambiguously. In particular, if we let $b(X)$ be the minimal number of bits to encode X in binary strings unambiguously, we have

$$H(X) \leq \mathbb{E}[b(X)] < H(X) + 1.$$

Moreover, if we let $q(X)$ to be the number of attempts to guess the value of X correctly, we might be surprised by the fact that

$$H(X) \leq \mathbb{E}[q(X)] < H(X) + 1,$$

i.e., it is expected to attempt at least $H(X)$ times to guess the value of X, but there is always a strategy to expect success before the $\big(H(X)+1\big)$-th attempt.

Those with prior knowledge in machine learning and decision trees might find the following special form of entropy familiar:

Definition 2.1.3 ▸ Binary Entropy

Let X be a Bernoulli random variable with parameter p. The **binary entropy** of p is defined as

$$H_b(p) = -p \log_2 p - (1-p) \log_2 (1-p).$$

With some simple computation, it is easy to check that $H_b(p)$ is maximised when $p = \frac{1}{2}$ and is zero if and only if $p = 1$ or $p = 0$.

Entropy can be defined over multiple random variables just like probability distributions.

In fact, we denote the tuple of n random variables as

$$X_1^n := (X_1, X_2, \cdots, X_n)$$

Clearly, we may view X_1^n as nothing else than a single random variable whose alphabet is just $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$.

Definition 2.1.4 ▸ Joint Entropy

Let X_1^n be a tuple of discrete random variable supported on a finite alphabet

$$\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$$

with joint probability mass function $p_{X_1^n}$. The **joint entropy** of $X_1, X_2, \cdots, X_n$ is defined as

$$H(X_1^n) := -\sum_{\boldsymbol{x} \in \mathcal{X}} p_{X_1^n}(\boldsymbol{x}) \log_2 p_{X_1^n}(\boldsymbol{x}).$$

Additionally, we can of course define the conditional entropy to measure the uncertainty of one event given the information on another event.

Definition 2.1.5 ▸ Conditional Entropy

Let (X, Y) be a pair of discrete random variables supported on an alphabet $\mathcal{X} \times \mathcal{Y}$ which is finite. Let p_X and p_Y be the probability mass functions for X and Y respectively. The **conditional entropy** of X given Y is defined as

$$H(X \mid Y) := \sum_{y \in \mathcal{Y}} p_Y(y) H(X \mid Y = y).$$

Note that here $H(X \mid Y = y)$ is also known as the *conditional entropy*, but it is different in meaning with $H(X \mid Y)$. In particular:

$$H(X \mid Y = y) = -\sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) \log_2 p_{X|Y}(x \mid y).$$

Therefore, we can expand the expression in Definition 2.1.5 into

$$\begin{aligned} H(X \mid Y) &= -\sum_{y \in \mathcal{Y}} p_Y(y) \sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) \log_2 p_{X|Y}(x \mid y) \\ &= -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{X,Y}(x, y) \log_2 p_{X|Y}(x \mid y) \\ &= \mathbb{E}\left[\log_2 \frac{1}{p_{X|Y}(X \mid Y)}\right]. \end{aligned}$$

One thing to note is that conditional entropy is **not symmetric**. We can interpret $H(X \mid Y)$ as "the remaining uncertainty of X given information on Y". Hence, it is not surprising that the following identity is true:

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

This is generalised as follows:

Proposition 2.1.6 ▸ Chain Rule of Entropy

Let X_1^n be a tuple of any n discrete random variables supported on finite alphabets, then

$$H(X_1^n) = \sum_{i=1}^{n} H\left(X_i \mid X_1^{i-1}\right).$$

Proof. The case where $n = 2$ follows directly from the result that

$$H(X_1,X_2) = H(X_1) + H(X_2 \mid X_1).$$

Suppose that there exists some integer $k \geq 2$ such that $H\left(X_1^k\right) = \sum_{i=1}^{k} H\left(X_i \mid X_1^{i-1}\right)$, consider

$$\begin{aligned} H\left(X_1^{k+1}\right) &= H\left(X_1^k, X_{k+1}\right) \\ &= H\left(X_1^k\right) + H\left(X_{k+1} \mid X_1^k\right) \\ &= \sum_{i=1}^{k} H\left(X_i \mid X_1^{i-1}\right) + H\left(X_{k+1} \mid X_1^k\right) \\ &= \sum_{i=1}^{k+1} H\left(X_i \mid X_1^{i-1}\right). \end{aligned}$$

□

A direct application of Proposition 2.1.6 yields the following result:

Corollary 2.1.7 ▸ Chain Rule of Entropy for Conditional Joint Distributions

Let X, Y, Z be discrete random variables supported on finite alphabets, then

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

Proof. Let $X_1 \coloneqq X \mid Z$ and $X_2 \coloneqq Y \mid Z$, then

$$\begin{aligned} H(X, Y \mid Z) &= H(X_1, X_2) \\ &= H(X_1) + H(X_2 \mid X_1) \\ &= H(X \mid Z) + H\big((Y \mid Z) \mid (X \mid Z)\big) \\ &= H(X \mid Z) + H(Y \mid X, Z). \end{aligned}$$

□

Given different distributions for the same random variable, we may be interested to know how much the distributions differ from one another. In other words, we wish to measure how much one distribution is different from another in terms of uncertainty.

Definition 2.1.8 ▸ Relative Entropy

Let p and q be probability mass functions for some discrete random variable X supported over an alphabet $\mathcal{X}$. The **relative entropy**, or alternatively, **Kullback-Leibler (KL) divergence**, between p and q is defined as

$$D(p \parallel q) \coloneqq \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}.$$

The above definition essentially describes the "difference" between two distributions as their expected ratio because

$$\sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} = \mathbb{E}\left[\log_2 \frac{p(X)}{q(X)}\right].$$

Remark. Using a similar argument to Definition 2.1.2, we set the following conventions:

1. $0 \log_2 \frac{0}{q} = 0$ for all $q \in \mathbb{R}$;
2. $p \log_2 \frac{p}{0} = +\infty$ for all $p \in \mathbb{R}$.

Relative entropy can be defined in a conditional context as well.

Definition 2.1.9 ▸ Conditional Relative Entropy

Let $p_{X,Y}$ and $q_{X,Y}$ be joint probability mass functions for some pair of discrete random variables (X, Y) supported over an alphabet $\mathcal{X} \times \mathcal{Y}$. The **conditional relative entropy**

between $p_{Y|X}$ and $q_{Y|X}$ averaged over X is

$$D\left(p_{Y|X} \parallel q_{Y|X} \mid p_X\right) \coloneqq \sum_{x \in \mathcal{X}} p_X(x) D\left(p_{Y|X}(\cdot \parallel x) \parallel q_{Y|X}(\cdot \parallel x)\right).$$

Notice that by applying Proposition 2.1.6, we have

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

due to $H(X, Y)$ being symmetric. This implies that

$$H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

Informally speaking, the left-hand side of the above identity is the remaining uncertainty of X after knowing Y, while the right-hand side is that of Y after knowing X. Note that this quantity can be interpreted as "the uncertain part of X and Y which cannot be reduced by knowing one of them". In other words, this remaining uncertainty is shared by both X and Y. The following notion formalises this observation:

Definition 2.1.10 ▸ Mutual Information

Let (X, Y) be a pair of discrete random variables with joint probability mass function $p_{X,Y}$. The **mutual information** between X and Y is defined as

$$I(X; Y) \coloneqq D\left(p_{X,Y} \parallel p_X \cdot p_Y\right).$$

It turns out that mutual information is symmetric, because

$$\begin{aligned} I(X; Y) &= D\left(p_{X,Y} \parallel p_X \cdot p_Y\right) \\ &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{X,Y}(x, y) \log_2 \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} \\ &= \mathbb{E}_{p_{X,Y}}\left[\log_2 \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)}\right]. \end{aligned}$$

One may check that

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

In this way, by using Proposition 2.1.6, we can also see that

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

and hence the symmetric property of mutual information.

Naturally, the mutual information between X and Y cannot exceed the entropy of either of them. Therefore, it is intuitive that

$$0 \leq I(X;Y) \leq \min\left\{\log_2|\mathcal{X}|, \log_2|\mathcal{Y}|\right\}.$$

Mutual information can be conditional as well.

Proposition 2.1.11 ▶ Chain Rule of Mutual Information

Let (X_1^n, Y) be a tuple of discrete random variables with joint probability mass function p, then

$$I(X_1^n;Y) = \sum_{i=1}^{n} I\left(X_i; Y \mid X_1^{i-1}\right).$$

Proof. By using Proposition 2.1.6,

$$\begin{aligned} I(X_1^n;Y) &= H(X_1^n) - H(X_1^n \mid Y) \\ &= \sum_{i=1}^{n} H\left(X_i \mid X_1^{i-1}\right) - \sum_{i=1}^{n} H\left(X_i \mid Y, X_i^{i-1}\right) \\ &= \sum_{i=1}^{n}\left(H\left(X_i \mid X_1^{i-1}\right) - H\left(X_i \mid Y, X_i^{i-1}\right)\right) \\ &= \sum_{i=1}^{n} I\left(X_i; Y \mid X_1^{i-1}\right). \end{aligned}$$

□

We could make some analogy between entropy and set theory. Suppose we have two random variables X and Y, we could let some sets $\mathcal{H}_X$ and $\mathcal{H}_Y$ represent $H(X)$ and $H(Y)$. It is intuitive to see that $H(X \mid Y)$ corresponds to $\mathcal{H}_X \backslash \mathcal{H}_Y$, $H(X,Y)$ corresponds to $\mathcal{H}_X \cup \mathcal{H}_Y$, and that $I(X;Y)$ corresponds to $\mathcal{H}_X \cap \mathcal{H}_Y$.

This inspires us to study mutual information between more than 2 random variables via the principle of inclusion and exclusion. However, the situation becomes problematic when we consider more random variables. It can be shown that there exist random variables X, Y, Z such that $I(X;Y;Z) < 0$, which does not make much sense in information theory.

2.2 Information Inequality

A lot of theorems in information theory are developed from inequalities. Among them, the core inequality result is known as the *information inequality* which can be used to prove a

wide range of corollaries.

Theorem 2.2.1 ▶ Information Inequality

For any probability mass functions p and q for some random variable X, $D(p \parallel q) \geq 0$. The equality is attained if and only if $p = q$.

Proof. Let $A \coloneqq \{x \in \mathcal{X} : \ p(x) > 0\}$. Take $Y \coloneqq \frac{q(X)}{p(X)}$ with support

$$\mathcal{Y} \coloneqq \left\{ \frac{q(x)}{p(x)} : \ x \in A \right\}.$$

By Theorem 1.4.4, we have

$$\sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} = \mathbb{E}_p \left[\log_2 Y \right] \leq \log_2 \mathbb{E}_p [Y] = \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)}.$$

Note that $q(x) \geq 0$ for all $x \in \mathcal{X}$ and $p(x) > 0$ for all $x \in A$. Therefore,

$$\begin{aligned}
-D(p \parallel q) &= -\sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} \\
&= \sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} \\
&\leq \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \\
&= \log_2 \sum_{x \in A} q(x) \\
&\leq \log_2 \sum_{x \in \mathcal{X}} q(x) \\
&= \log_2 1 \\
&= 0.
\end{aligned}$$

Therefore, $D(p \parallel q) \geq 0$. Clearly, the equality holds if and only if

$$\mathbb{E}_p \left[\log_2 Y \right] = \log_2 \mathbb{E}_p [Y] \quad \text{and} \quad \sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x).$$

Note that $f(x) = \log_2 x$ is strictly convex, so $\mathbb{E}_p \left[\log_2 Y \right] = \log_2 \mathbb{E}_p [Y]$ if and only if Y is constant, i.e., $\frac{q(x)}{p(x)} = c$ for some fixed $c \in \mathbb{R}$ for all $x \in A$. Notice that this is equivalent to

$$1 = \sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c,$$

i.e., $p(x) = q(x)$ for all $x \in A$. Note that in the same time, $q(x) = 0$ for all $x \in \mathcal{X} \backslash A$,

i.e., $q(x) = 0$ if and only if $p(x) = 0$. Therefore, $p = q$ as desired. □

The information inequality leads to many bounding conditions to the common quantities we have discussed so far.

Corollary 2.2.2 ▸ Mutual Information Is Non-negative

For any jointly distributed discrete random variables X and Y, $I(X;Y) \geq 0$ with equality attained if and only if X and Y are independent.

Proof. Notice that

$$\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_X(x)\, p_Y(y) = \left(\sum_{x\in\mathcal{X}} p_X(x)\right)\left(\sum_{y\in\mathcal{Y}} p_Y(y)\right) = 1,$$

so $p_X \cdot p_Y$ is a probability mass function for (X, Y). Therefore, by Theorem 2.2.1,

$$I(X;Y) = D\left(p_{X,Y} \parallel p_X \cdot p_Y\right) \geq 0,$$

where the equality is attained if and only if $p_{X,Y} = p_X \cdot p_Y$, i.e., X and Y are independent. □

Naturally, the conditional relative entropy should also be non-negative.

Corollary 2.2.3 ▸ Conditional Relative Entropy Is Non-negative

For any pair of discrete random variables (X, Y), $D\left(p_{Y|X} \parallel q_{Y|X} \mid p_X\right) \geq 0$ with equality attained if and only if $p_{Y|X}(\cdot \mid x) = q_{Y|X}(\cdot \mid x)$ for all $x \in \mathcal{X}\backslash p_X^{-1}[\{0\}]$.

Proof. By Theorem 2.2.1,

$$D\left(p_{Y|X}(\cdot \parallel x) \parallel q_{Y|X}(\cdot \parallel x)\right) \geq 0$$

for all $x \in \mathcal{X}$. Since $p_X(x) \geq 0$ for all $x \in \mathcal{X}$, clearly

$$D\left(p_{Y|X} \parallel q_{Y|X} \mid p_X\right) \geq 0,$$

where the equality is attained if and only if

$$D\left(p_{Y|X}(\cdot \parallel x) \parallel q_{Y|X}(\cdot \parallel x)\right) = 0$$

for all $x \in \mathcal{X}\backslash p_X^{-1}[\{0\}]$. This is equivalent to $p_{Y|X}(\cdot \parallel x) = q_{Y|X}(\cdot \parallel x)$ for all $x \in \mathcal{X}\backslash p_X^{-1}[\{0\}]$. □

We can do a similar argument for conditional mutual information as well.

Corollary 2.2.4 ▸ Conditional Mutual Information Is Non-negative

For any discrete random variables X, Y, Z, we have $I(X;Y \mid Z) \geq 0$ with equality attained if and only if X-Z-Y is a Markov chain.

Proof. Notice that by Theorem 2.2.1,

$$I(X;Y \mid Z) = D\left(p_{X,Y|Z} \parallel p_{X|Z} \cdot p_{Y|Z}\right) \geq 0,$$

where the equality is attained if and only if $p_{X,Y|Z} = p_{X|Z} \cdot p_{Y|Z}$. □

Recall that we previously mentioned that $0 \leq H(X) \leq \log_2 |\mathcal{X}|$. The upper bound can be derived using the information inequality as well.

Proposition 2.2.5 ▸ Upper Bound of Entropy

For any discrete random variable X supported on a finite alphabet $\mathcal{X}$, we have $H(X) \leq \log_2 |\mathcal{X}|$ with equality attained if and only if p_X is uniform on $\mathcal{X}$.

Proof. Define $u(x) \coloneqq \frac{1}{|\mathcal{X}|}$ to be the uniform distribution over $\mathcal{X}$. Consider

$$\begin{aligned} D(p_X \parallel u) &= \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{p_X(x)}{u(x)} \\ &= \sum_{x \in \mathcal{X}} p_X(x) \log_2 |\mathcal{X}| p_X(x) \\ &= \sum_{x \in \mathcal{X}} p_X(x) \log_2 |\mathcal{X}| + \sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x) \\ &= \log_2 |X| - H(X). \end{aligned}$$

By Theorem 2.2.1, we have $D(p_X \parallel u) \geq 0$ and so $H(X) \leq \log_2 |X|$. The equality is attained if and only if $p_X = u$ is uniform over $\mathcal{X}$. □

One important result derived from this upper bound is as follows:

Corollary 2.2.6 ▸ Conditioning Does Not Increase Entropy

For any pair of discrete random variables (X, Y), we have $H(X \mid Y) \leq H(X)$ with equality attained if and only if X and Y are independent.

Proof. Notice that

$$H(X) - H(X \mid Y) = I(X;Y) \geq 0$$

by Corollary 2.2.2, so $H(X \mid Y) \leq H(X)$ as desired. □

However, do note that there could exist some $y \in \mathcal{Y}$ such that $H(X \mid Y = y) > H(X)$. For example, consider the following joint distribution:

Y \ X	1	2
1	0	0.75
2	0.125	0.125

We compute the conditional entropy values:

$$H(X \mid Y = 1) = -p_{X|Y}(2 \mid 1) \log_2 p_{X|Y}(2 \mid 1) = 0,$$
$$H(X \mid Y = 2) = -p_{X|Y}(1 \mid 2) \log_2 p_{X|Y}(1 \mid 2) - p_{X|Y}(2 \mid 2) \log_2 p_{X|Y}(2 \mid 2) = 1.$$

However, $H(X \mid Y) = p_Y(2) H(X \mid Y = 2) = 0.25 < H(X \mid Y = 2)$.

Corollary 2.2.6 can be generalised further. We first introduce a preliminary definition.

Definition 2.2.7 ▸ Mutual Independence

Let $X_1, X_2, \cdots, X_n$ be any n random variables. They are said to be **mutually independent** if for any $S \subseteq [n] \setminus \{0\}$,

$$p_{X_S} = \prod_{i \in S} p_{X_i}.$$

Now, we propose the following inequality for conditional entropy:

Corollary 2.2.8 ▸ Generalised Inequality for Conditional Entropy

For any random variables $X_1, X_2, \cdots, X_n$,

$$H(X_1^n) \leq \sum_{i=1}^{n} H(X_i),$$

where the equality attained if and only if the X_i's are mutually independent.

Proof. By Corollary 2.2.6, $H\left(X_i \mid X_1^{i-1}\right) \leq H(X_i)$ for all $i = 1, 2, \cdots, n$, so

$$H(X_1^n) = \sum_{i=1}^{n} H\left(X_i \mid X_1^{i-1}\right) \leq \sum_{i=1}^{n} H(X_i).$$

The equality is attained if and only if X_i and X_j are independent whenever $i \neq j$, i.e.,

the X_i's are mutually independent. □

The next inequality is very useful tool which can be used to prove the many results derived from the information inequality so far.

Theorem 2.2.9 ▸ Log-sum Inequality

Let $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ be non-negative real-valued sequences, then

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log_2 \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

Proof. If there exists some $i \in \mathbb{N}^+$ such that $b_i = 0$, then the left-hand side is $+\infty$. If there exists some $i \in \mathbb{N}^+$ such that $a_i = 0$, then a_i contribute 0 to both side of the inequality. Therefore, without loss of generality, we can assume that $a_i, b_i > 0$ for all $i = 1, 2, \cdots, n$. Let $a = \sum_{i=1}^n a_i$ and $b = \sum_{i=1}^n b_i$. One may check that the function $f(x) = x \log_2 x$ is strictly convex, so

$$\sum_{i=1}^{n} \frac{b_i a_i}{b b_i} \log_2 \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} \frac{b_i a_i}{b b_i}\right) \log_2 \sum_{i=1}^{n} \frac{b_i a_i}{b b_i}.$$

Simplifying the inequality yields

$$\frac{1}{b} \sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geq \frac{a}{b} \log_2 \frac{a}{b},$$

and so

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geq a \log_2 \frac{a}{b}.$$

□

One result which can be proven with the aid of the log-sum inequality is the joint convexity of relative entropy.

Proposition 2.2.10 ▸ Convexity of Relative Entropy

$D(p \parallel q)$ is jointly convex.

Proof. Let p_1, p_2, q_1, q_2 be probability mass functions for the same random variable X. It suffices to prove that for any $\lambda \in [0, 1]$,

$$D\big(\lambda p_1 + (1-\lambda) p_2 \parallel \lambda q_1 + (1-\lambda) q_2\big) \leq \lambda D(p_1 \parallel q_1) + (1-\lambda) D(p_2 \parallel q_2).$$

For any $x \in \mathcal{X}$, let

$$x_{p_1} = \lambda p_1(x), \qquad x_{p_2} = (1-\lambda)\, p_2(x);$$
$$x_{q_1} = \lambda q_1(x), \qquad x_{q_2} = (1-\lambda)\, q_2(x).$$

By Theorem 2.2.9,

$$\left(x_{p_1} + x_{p_2}\right) \log_2 \frac{x_{p_1} + x_{p_2}}{x_{q_1} + x_{q_2}} \leq x_{p_1} \log_2 \frac{x_{p_1}}{x_{q_1}} + x_{p_2} \log_2 \frac{x_{p_2}}{x_{q_2}}.$$

□

We can use a similar approach to analyse the convexity of entropy

Proposition 2.2.11 ▸ Concavity of Entropy

For any discrete random variable X with a finite alphabet and distribution p, $H(p)$ is concave in p.

Proof. Let u be the uniform distribution for X, then $u(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$. Notice that

$$\begin{aligned} D(p \parallel u) &= \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{u(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{u(x)} - \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)} \\ &= \log_2 |\mathcal{X}| - H(p). \end{aligned}$$

By Proposition 2.2.10, $D(p \parallel u)$ is convex, so $H(p)$ must be concave. □

Alternatively, let $T \sim \text{Bernoulli}(\lambda)$ and define $Z \coloneqq X \mid T$, where $X \mid T = 1$ and $X \mid T = 0$ have distributions p_1 and p_2 respectively, then clearly $p_Z = \lambda p_1 + (1-\lambda)\, p_2$. Therefore,

$$\begin{aligned} H\big(\lambda p_1 + (1-\lambda)\, p_2\big) &= H(Z) \\ &\geq H(Z \mid T) \\ &= p_T(1)\, H(Z \mid T = 1) + p_T(0)\, H(Z \mid T = 0) \\ &= \lambda H(X \mid T = 1) + (1-\lambda)\, H(X \mid T = 0) \\ &= \lambda H(p_1) + (1-\lambda)\, H(p_2). \end{aligned}$$

This gives a more classical approach to proving Proposition 2.2.11.

Note that we can view $I(X;Y)$ as a function of $p_{X,Y}$, which can be further unpacked as a

function of p_X and $p_{Y|X}$. Here, p_X is known as the *input distribution* and $p_{Y|X}$ is known as the *channel*.

Theorem 2.2.12 ▸ Convexity of Mutual Information

$I(X;Y)$ is concave in p_X and convex in $p_{Y|X}$.

Proof. Fix $p_{Y|X}$, consider

$$\begin{aligned} I(X;Y) &= H(Y) - H(X \mid Y) \\ &= H(p_Y) - \sum_{x \in \mathcal{X}} p_X(x) H(Y \mid X = x). \end{aligned}$$

Note that $H(Y \mid X = x)$ only depends on $p_{Y|X}$ and so is a constant and for any $y \in \mathcal{Y}$,

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_X(x) \, p_{Y|X}(y \mid x)$$

is linear in $p_X(x)$. Therefore, $I(X;Y)$ is linear in p_X and so concave in p_X. Now, fix p_X and define $T \sim \text{Bernoulli}(\lambda)$ to be independent of X, then

$$p_{Y|X} = \lambda p_{Y|X,T=1} + (1-\lambda) \, p_{Y|X,T=0}.$$

Notice that

$$I(X;T) + I(X;Y \mid T) = I(X;Y) = I(X;Y) + I(X;T \mid Y).$$

Since $I(X;T) = 0$, this implies that $I(X;Y \mid T) = I(X;Y) + I(X;T \mid Y)$. This means that

$$\begin{aligned} I(X;Y) &\leq I(X;Y \mid T) \\ &= \lambda I(X;Y \mid T = 1) + (1-\lambda) I(X;Y \mid T = 0), \end{aligned}$$

which implies that $I(X;Y)$ is convex in $p_{Y|X}$. □

The next inequality is concerning data processing.

Theorem 2.2.13 ▸ Data Processing Inequality (DPI)

If X-Y-Z forms a Markov chain, then $I(X;Y) \geq I(X;Z)$.

Proof. Notice that

$$I(X;Z) + I(X;Y \mid Z) = I(X;Y,Z) = I(X;Y) + I(X;Z \mid Y).$$

Since X-Y-Z forms a Markov chain, $I(X;Z \mid Y) = 0$. Therefore,

$$I(X;Y) = I(X;Z) + I(X;Y \mid Z) \geq I(X;Z).$$

□

Mathematically, data processing can be described as mapping a random variable Y to some transformed image via a function g. Recall that in Proposition 1.2.4, we have shown that for any random variables X and Y, X-Y-$f(Y)$ is always a Markov chain, which motivates the following application of the DPI.

Corollary 2.2.14 ▸ Mutual Information Does Not Increase After Processing

For any random variables X and Y, we have $I(X;Y) \geq I\big(X;g(Y)\big)$ for all function g.

One implication of this is that no matter what method is used to process the information Y, the shared knowledge between Y and some other data set X can be at best retained at the same level as before.

2.3 Sufficient Statistics

Consider a parametric family of probability distributions $\{f_\theta(x) : \theta \in \Theta\}$ for some index set Θ. Let $T(X)$ be any statistic. One may check that Θ-X-$T(X)$ is a Markov chain. By Theorem 2.2.13 we know that $I(\Theta;T(X)) \leq I(\Theta;X)$. If the equality is attained, no information is lost in this statistic.

Definition 2.3.1 ▸ Sufficient Statistic

A function $T : \mathcal{X} \to \mathbb{R}$ is said to be a **sufficient statistic** relative to a parametric family $\mathcal{F}_\theta := \{f_\theta(x) : \theta \in \Theta\}$ of probability distributions if Θ-$T(X)$-X forms a Markov chain.

Let X be any random variable and Y be another random variable correlated to Y. Suppose now we wish to estimate X via observations about Y. Let $\widehat{X}(Y)$ be an estimator obtained this way about X.

If $H(X \mid Y) = 0$, one can expect that a perfect estimation is possible. On the other hand, if $H(X \mid Y) = \log_2 |\mathcal{X}|$, the the estimation is bad. Notice that this happens if and only if X is uniform and independent of Y. In reality, we may wish $H(X \mid Y)$ to be small, where the error of estimation can be small.

Theorem 2.3.2 ▶ Fano's Inequality

For any estimator $\widehat{X}$ obtained from Y, let $p_e \coloneqq \Pr\left(\widehat{X} \neq X\right)$ be the probability of error, then

$$H_b(p_e) + p_e \log_2 |\mathcal{X}| \geq H\left(X \mid \widehat{X}\right) \geq H(X \mid Y).$$

Proof. Define $E \coloneqq \mathbf{1}\left\{\widehat{X} \neq X\right\}$ to be the error random variable, then $p_e = \Pr(E = 1)$. Consider

$$H\left(E \mid \widehat{X}\right) + H\left(X \mid E, \widehat{X}\right) = H\left(E, X \mid \widehat{X}\right) = H\left(X \mid \widehat{X}\right) + H\left(E \mid X, \widehat{X}\right).$$

It is clear that $H\left(E \mid X, \widehat{X}\right) = 0$. By Corollary 2.2.6, since E is a Bernoulli random variable, we have

$$H\left(E \mid \widehat{X}\right) \leq H(E) = H_b(p_e).$$

Note that

$$H\left(X \mid E, \widehat{X}\right) = \Pr(E = 1) H\left(X \mid E = 1, \widehat{X}\right) + \Pr(E = 0) H\left(X \mid E = 0, \widehat{X}\right).$$

Clearly, $H\left(X \mid E = 0, \widehat{X}\right) = 0$ and $H\left(X \mid E = 1, \widehat{X}\right) \leq \log_2 |\mathcal{X}|$, so

$$H\left(X \mid E, \widehat{X}\right) \leq p_e \log_2 |\mathcal{X}|.$$

Therefore,

$$H\left(X \mid \widehat{X}\right) = H\left(E \mid \widehat{X}\right) + H\left(X \mid E, \widehat{X}\right) \leq H_b(p_e) + p_e \log_2 |\mathcal{X}|.$$

Note that $\widehat{X}$ is a function of Y, so by Corollary 1.2.4, X-Y-$\widehat{X}$ forms a Markov chain. By Theorem 2.2.13,

$$\begin{aligned} H(X) - H(X \mid Y) &= I(X; Y) \\ &\geq I\left(X; \widehat{X}\right) \\ &= H(X) - H\left(X \mid \widehat{X}\right), \end{aligned}$$

which reduces to $H\left(X \mid \widehat{X}\right) \geq H(X \mid Y)$. □

With some algebraic manipulations, it can be obtained from Theorem 2.3.2 that

$$p_e \geq \frac{H(X \mid Y) - 1}{\log_2 |\mathcal{X}|}.$$

In other words, this means that

$$\inf \Pr\left(\widehat{X}(Y) \neq X\right) \geq \frac{H(X \mid Y) - 1}{\log_2 |\mathcal{X}|}.$$

This is one of the few results which offer a **lower bound estimate** for probabilities. In particular, this result shows that in a non-trivial scenario ($|\mathcal{X}| > 1$), perfect estimator is attainable only when $H(X \mid Y) \leq 1$.