A Not-Too-Short, Not-Quite-Long Summary of *Information Theory*

Author: Igor L.R. Azevedo - *The University of Tokyo & University of Brasilia*
Email: *igorlima1740@gmail.com*
GitHub: https://github.com/igor17400/information-theory-174

WHAT TO EXPECT: This work aims to provide a summary of information theory that achieves a balance, as indicated by the title—not overly concise yet not as exhaustive as a comprehensive textbook. My intention was to explore essential information theory content with more depth than a typical summary offers, yet not as extensively as a canonical textbook. While this isn't just a collection of equations, it also isn't a book or paper that aims to make a monumental contribution like MacKay (2002); Cover and Thomas (2006) have done. Ultimately, my goal is to present a practical approach to information theory that may be applicable for those studying artificial intelligence. In any case, I hope this proves useful to someone beyond myself. If you've read this far, thank you, and stay safe!

COLOR GUIDE: This document uses four colors to convey specific types of information:

- Color 1 - Indicates super important information, memorize it!
- Color 2 - Used exclusively for arrows, which signify important details, curiosities, or useful symbols and information.
- Color 3 - Marks important information designed to catch your attention.
- Color 4 - Reserved for citations, links, lines, and other objects.

Contents

1 Introduction to Information Theory

"While probability theory allows us to make uncertain statements and to reason in the presence of uncertainty, information theory enables us to quantify the amount of uncertainty in a probability distribution." Goodfellow et al. (2016).

Information theory plays a crucial role in various fields, including Machine Learning and neural networks. It provides a rigorous framework to measure and manage uncertainty, which is inherent in data and models. By quantifying information and uncertainty, information theory enables us to design robust algorithms, assess model performance, and optimize data representations. This capability is vital for enhancing the reliability and efficiency of Machine Learning systems across diverse applications.

In the realm of neural networks, information theory offers profound insights into understanding how information flows through network layers. Concepts such as information gain, compression, and coding efficiency are pivotal in designing architectures that can handle large-scale datasets effectively while minimizing computational overhead. Information-theoretic principles guide the selection of activation functions, regularization techniques, and model architectures, thereby improving the interpretability and generalization capabilities of neural networks. Integrating information theory into the study of artificial intelligence helps our understanding of learning processes in complex systems.

With that in mind, let's now review the fundamentals of probability theory :)

2 Basic Concepts in Probability Theory

⟹ The following subsections were written using the books Feller (1968); Goodfellow et al. (2016) and **specially** Blitzstein (2024) online lectures as references.

Mathematics is the logic of certainty; Probability is the logic of uncertainty. Probability provides procedures for principled problem-solving pitfalls and paradoxes.

2.0.1 Sample spaces

The mathematical framework for probability is built around sets, that is

Definition 2.1 *The Sample Space S of an experiment is the set of all possible outcomes of the experiment. An event A is a subset of the sample space S, and we say that A occured if the actual outcome is in A.*

We can visualize a sample space as shown in the figure 1. The sample space of an experiment can be finite, countably infinite, or uncountably infinite.

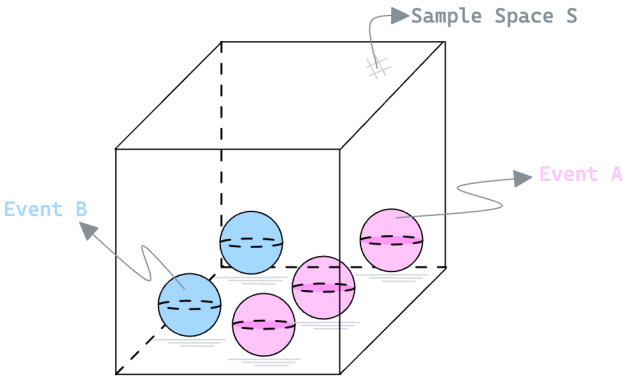

Figure 1: Illustration of the sample space as a pebble world, with two events A and B spotlighted

⟹When the sample space is finite, we can visualize it as a pebble world. Each pebble represents an outcome, and an event is a set of pebbles.

The table 1 gives a set of symbols useful when describing events and sets.

English	Sets
Events and occurrences	
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
something must happen	$s_{\text{actual}} \in S$
New events from old events	
A or B (inclusive)	$A \cup B$
A and B	$A \cap B$
not A	A^c
A or B, but not both	$(A \cap B^c) \cup (A^c \cap B)$
at least one of $A_1, \ldots, A_n$	$A_1 \cup \ldots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \ldots \cap A_n$
Relationships between events	
A implies B	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \emptyset$

Table 1: Sets Symbols Dictionary

2.1 Naive Definition of Probability

Definition 2.2 *Let A be an event for an experiment with a finite sample space S. The naive probability of A is*

$$P_{naive} = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to } A}{\text{Total number of outcomes in } S} \tag{1}$$

here $|\cdot|$ means the size of the sample space.

The naive definition is very restrictive in that it requires S to e finite, with equal mass for each pebble.

2.2 How to Count

Theorem 1 (Multiplication Rule) *Consider a compound experiment consisting of two sub-experiments A and B. Suppose that Exp A has a possible outcomes, and for each of these outcomes Exp B has b possible outcomes. Then the compound experiment has ab possible outcomes.*

We can see such relation by drawing a tree diagram as the figure 2 shows. As we can see we have three ramifications for Exp A and four for Exp B. With a total of 12 possible outcomes. Note, that it's often easier to think about the experiments as being in chronological order, but there is no requirements in the multiplication rule that experiment experiment A has to be performed before experiment B.

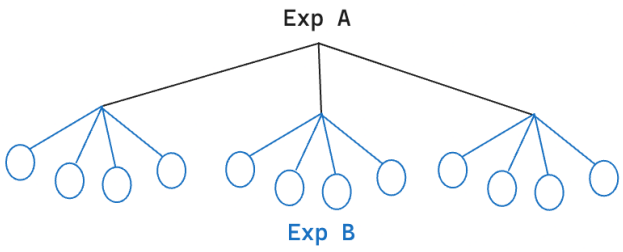

Figure 2: Tree Diagram.

Example: Suppose that 10 people are participating in a race. Assume that ties are not possible and that all 10 will complete the race, resulting in well-defined first, second, and third place winners. How many possible outcomes are there for these top three positions?

First, there are 10 possible choices for who finishes in first place. After the first place is determined, there remain 9 competitors for the second place, and once both the first and second places are fixed, there are 8 possibilities for the third place. Therefore, by the multiplication rule, the total number of possible outcomes is $10 \cdot 9 \cdot 8 = 720$.

$\Longrightarrow$ Note that we did not necessarily have to determine the first place winner first. We could have instead considered the third place winner initially, with 10 possibilities, followed by 9 possibilities for the second place, and finally 8 possibilities for the first place. This approach yields the same result, as expected from the multiplication rule.

Now, let's consider another more common example. Suppose we have a set of 10 unique books and we want to arrange three of them on a shelf. The order of the books matters. How many different arrangements are possible? To solve this, we again use the multiplication rule. We have 10 choices for the first position, 9 choices for the second position, and 8 choices for the third position. Therefore, the total number of different arrangements is $10 \cdot 9 \cdot 8 = 720$.

We can use the multiplication rule to derive formulas for sampling with and without replacement.

Theorem 2 (Sampling with replacement) *Consider n objects and making k choices from them, one at a time with replacement (i.e., choosing a certain object does not preclude it from being chosen again). Then there are n^k possible outcomes (where order matters, in the sense that, e.g., choosing object 3 and object 7 is counted as a different outcome than choosing object 7 and object 3).*

Example: Imagine a jar with n balls, labeled from 1 to n. We sample balls one at a time with replacement, meaning that each time a ball is chosen, it's returned to the jar.

By the multiplication rule there are n^k ways to obtain a sample of size k. We can visualize this in the figure 3.

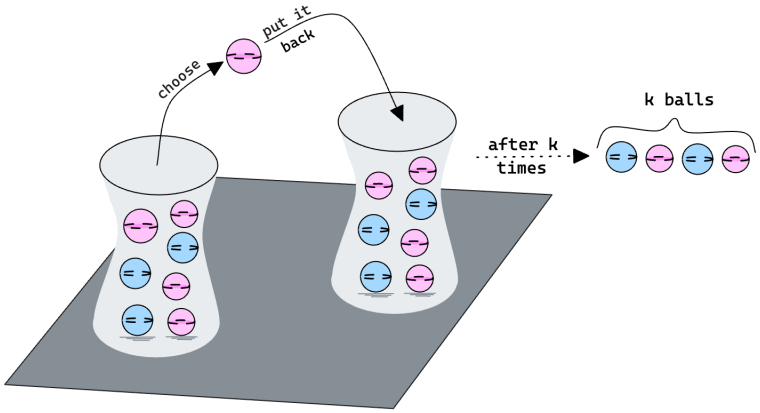

Figure 3: Illustration of sampling with replacement.

Theorem 3 (Sampling without replacement) *Consider n objects and making k choices from them, one at a time without replacement. (i.e., choosing a certain objects precludes it from being chosen again). The there are*

$$(n)_k = n(n-1)\cdots(n-k+1) \tag{2}$$

possible outcomes for $1 \leq k \leq n$, and 0 possibilities for $k > n$ (where order matters). By convention

$$n(n-1)\cdots(n-k+1) = n \text{ for } k = 1 \tag{3}$$

$\Longrightarrow$ Special case when $k = n$, $(n)_{k=n} = n!$

Example: There are k people in the room and their birthdays are independent. What is the probability that at least one pair of people in the group have the same birthday?

It's easier to calculate the probability that no two people share a birthday. This amounts to sampling 365 days of the year without replacement.

$$P(\text{no birthday match}) = \frac{365 \cdot 364 \cdots (365-k+1)}{365^k} = P_1$$

Then,

$$P(\text{at least 1 birthday match}) = 1 - P_1 = P_2$$

For $k = 23$ people, $P_2 > 50\%$ and for $k = 57$ people, $P_2 > 99\%$.

Example from Feller (1968): Let the population consist of the ten digits $0, 1, \cdots, 9$. Every succession of five digits represents a sample of size $k = 5$ and we assume that each such arrangement has a probability $p_1 = \frac{1}{10^5} = 10^{-5}$. Then, the *probability P that five consecutive random digits are all different* is

$$P = \frac{(10)_k}{10^k} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{10^5} = 0.3024$$

2.2.1 "Leibniz mistake"

$\Longrightarrow$ It's crucial to think of the objects or people in the population as named or labeled.

The "Leibniz mistake" in probability refers to an error made by the mathematician Gottfried Wilhelm Leibniz in the late 17th century regarding the probability of mutually exclusive events.

Leibniz incorrectly calculated the probability of drawing at least one ace from two draws in a deck of cards without replacement. He assumed that the probability of drawing an ace on the first or the second draw could be simply added together, which is only correct for mutually exclusive events. However, these events are not mutually exclusive because drawing an ace on the first draw affects the probability of drawing an ace on the second draw.

Leibniz's mistake was in not accounting for the dependency between the events. The correct calculation should involve conditional probability, which considers the changed probabilities after each draw. This error highlights the importance of understanding the conditions under which probabilities can be directly added, which is a fundamental concept in probability theory.

2.3 Adjusting for Overcounting

In many counting problems, it's not easy to directly count each possibility once and only once. However, if we're able to count each possibility exactly c times for some c, then we can adjust by dividing by c.

Definition 2.3 (Binomial Coefficient) *For any nonnegative integers k and n, the binomial coefficient $\binom{n}{k}$, read as "n choose k", is the number of subsets of size k from a set of size n.*

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!} \tag{4}$$

3 Conditional Probability And Stochastic Independence

$\Longrightarrow$ The following subsections were written using the book Feller (1968) and **specially** Venkatesh (2024) online lectures as references.

3.1 Conditional Probability

Suppose a population of N people includes N_A colorblind people and $N_{\mathcal{H}}$ females.

$$p(A) = \frac{N_A}{N} \rightarrow \text{probability of being colorblind} \quad p(\mathcal{H}) = \frac{N_{\mathcal{H}}}{N} \rightarrow \text{probability of being a female}$$

We may now restrict our attention to the subpopulation consisting of females. The probability that a person chosen at random from this subpopulation is colorblind equals

$$P(A|\mathcal{H}) = \frac{N_{A\mathcal{H}}}{N_{\mathcal{H}}} = \frac{P(A \cap \mathcal{H})}{P(\mathcal{H})}$$

where $P(A|\mathcal{H})$ is read as "the probability of the event A (colorblindness), assuming the event $\mathcal{H}$ (that the person chosen is female)."

Definition 3.1 *Let $\mathcal{H}$ be an event with positive probability. For an arbitrary event A, we shall write*

$$P(A|\mathcal{H}) = \frac{P(A \cap \mathcal{H})}{P(\mathcal{H})} \tag{5}$$

The quantity so defined will be called the conditional probability of A on the hypothesis $\mathcal{H}$ (or for given $\mathcal{H}$). When all sample points have equal probabilities, $P(A|\mathcal{H})$ is the ratio $\frac{N_{A\mathcal{H}}}{N_{\mathcal{H}}}$ of the number of sample points common to A and $\mathcal{H}$, to the number of points in $\mathcal{H}$.

Taking conditional probabilities of various events with respect to a particular hypothesis $\mathcal{H}$ amounts to choosing $\mathcal{H}$ as a new sample space with probabilities proportional to the original ones. All general theorems on probabilities are valid also for conditional probabilities with respect to any particular hypothesis $\mathcal{H}$, for instance

$$P(A \cup B|\mathcal{H}) = P(A|\mathcal{H}) + P(B|\mathcal{H}) - P(A \cap B|\mathcal{H}) \tag{6}$$

Theorem 4 (Compound probabilities)

$$P(A \cap \mathcal{H}) = P(A|\mathcal{H}) \cdot P(\mathcal{H}) \tag{7}$$

Example: Suppose we have a deck of 52 cards, and we want to find the probability that a randomly drawn card is a king, given that it is a face card (jack, queen, or king).

Let A be the event "the card is a king," and $\mathcal{H}$ be the event "the card is a face card." There are 12 face cards in total, and 4 of them are kings. The probability that a card is a face card is

$$P(\mathcal{H}) = \frac{12}{52} = \frac{3}{13}.$$

The probability that a card is a king and a face card is

$$P(A \cap \mathcal{H}) = \frac{4}{52} = \frac{1}{13}.$$

Therefore, the conditional probability that a card is a king given that it is a face card is

$$P(A|\mathcal{H}) = \frac{P(A \cap \mathcal{H})}{P(\mathcal{H})} = \frac{\frac{1}{13}}{\frac{3}{13}} = \frac{1}{3}.$$

Example: A family with two children is known to have (at least) one boy. What are the chances the other child is also a boy?

Let's consider the following notation:

- Let $\mathcal{H}$ be the event of at least one boy
- Let A be the event of the other child is also a boy
- Let denote our sample space as $\{bb, bg, gb, gg\}$ where g represents a girl and b represents a boy

We aim to calculate $P(A|\mathcal{H})$. Thus

$$P(bb) = P(bg) = P(gb) = P(gg) = \frac{1}{4}$$

$$P(A|\mathcal{H}) = P(b|b) = \frac{P(A \cap \mathcal{H})}{P(\mathcal{H})} = \frac{P(bb)}{P\{bb, gb, bg\}} = \frac{1/4}{3 \cdot 1/4} \implies P(A|\mathcal{H}) = \frac{1}{3}$$

$\Longrightarrow$ Note how our hypothesis space was initially $\{bb, bg, gb, gg\}$ but at the moment we received the information that the family already had at least one boy our hypothesis space changed to $\{bb, bg, gb\}$ nor more gg being considered. This example illustrates, how side information makes the sample space to change.

$\Longrightarrow$ It's interesting that most people expect the answer to be 1/2 . This is the correct answer to a different question, namely: a boy is chosen at random and found to come from a family with two children; what's the probability that the other child is a boy?

Definition 3.2 (Chain Rule for Conditional Probabilities) *The chain rule for conditional probabilities states that for any sequence of events $A_1, A_2, \ldots, A_n$, the joint probability of these events can be expressed as the product of conditional probabilities:*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \tag{8}$$

This rule allows us to break down the joint probability of a sequence of events into a product of simpler conditional probabilities.

Example: Suppose we want to find the joint probability of three events: A, B, and C. According to the chain rule for conditional probabilities, we can express this as:

$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B).$$

Let's consider a specific scenario: drawing three cards sequentially from a standard deck of 52 cards without replacement.

- Let A be the event that the first card drawn is an Ace.
- Let B be the event that the second card drawn is a King.
- Let C be the event that the third card drawn is a Queen.

We want to find the probability $P(A \cap B \cap C)$.

$$P(A) = \frac{4}{52} \quad \text{(since there are 4 Aces in a deck of 52 cards).}$$
$$P(B|A) = \frac{4}{51} \quad \text{(after drawing an Ace, there are 51 cards left, and 4 of them are Kings).}$$
$$P(C|A \cap B) = \frac{4}{50} \quad \text{(after drawing an Ace and a King, there are 50 cards left, and 4 of them are Queens).}$$

Using the chain rule:

$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B) = \frac{4}{52} \cdot \frac{4}{51} \cdot \frac{4}{50}.$$

Theorem 5 (Law of Total Probability) *Let $\{\mathcal{H}_j\}$ be a partition of the sample space, meaning that the events $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n$ are mutually exclusive and exhaustive (i.e., one of them must occur). Then, for any event A,*

$$P(A) = \sum_j P(A \cap \mathcal{H}_j) \tag{9}$$

Using the definition of conditional probability, we have:

$$P(A \cap \mathcal{H}_j) = P(A|\mathcal{H}_j) \cdot P(\mathcal{H}_j). \tag{10}$$

Substituting this into the equation above, we get:

$$P(A) = \sum_j P(A \cap \mathcal{H}_j) = \sum_j P(A|\mathcal{H}_j) \cdot P(\mathcal{H}_j). \tag{11}$$

Example: Suppose there are two bags, Bag 1 and Bag 2. Bag 1 contains 3 red balls and 7 blue balls, while Bag 2 contains 5 red balls and 5 blue balls. A bag is chosen at random with equal probability, and then a ball is drawn from the chosen bag. We want to find the probability of drawing a red ball.

Let:

- $\mathcal{H}_1$ be the event of choosing Bag 1.
- $\mathcal{H}_2$ be the event of choosing Bag 2.
- A be the event of drawing a red ball.

We know:

$$P(\mathcal{H}_1) = \frac{1}{2} \quad \text{and} \quad P(\mathcal{H}_2) = \frac{1}{2}.$$
$$P(A|\mathcal{H}_1) = \frac{3}{10} \quad \text{(probability of drawing a red ball from Bag 1).}$$
$$P(A|\mathcal{H}_2) = \frac{5}{10} \quad \text{(probability of drawing a red ball from Bag 2).}$$

Using the law of total probability:

$$P(A) = P(A|\mathcal{H}_1) \cdot P(\mathcal{H}_1) + P(A|\mathcal{H}_2) \cdot P(\mathcal{H}_2).$$

Substituting the values:

$$P(A) = \left(\frac{3}{10}\right) \cdot \left(\frac{1}{2}\right) + \left(\frac{5}{10}\right) \cdot \left(\frac{1}{2}\right) = \frac{3}{20} + \frac{5}{20} = \frac{8}{20} \implies P(A) = \frac{2}{5}.$$

So, the probability of drawing a red ball is $\frac{2}{5}$.

3.2 Bayes' Rule For Events

Definition 3.3 (Bayes' Rule) *Given a partition $\{A_j, j \geq 1\}$ of the sample space S, a priori probabilities $P(A_j)$, and a forward conditional probability $P(\mathcal{H}|A_j)$, Bayes' rule states that to determine the posterior probability $P(A_k|\mathcal{H})$, we have:*

$$P(A_k|\mathcal{H}) = \frac{P(\mathcal{H}|A_k) \cdot P(A_k)}{P(\mathcal{H})} \tag{12}$$

$$P(\mathcal{H}) = \sum_j P(\mathcal{H}|A_j) \cdot P(A_j) \tag{13}$$

$$P(A_k \cap \mathcal{H}) = \mathcal{H}|A_k) \cdot P(A_k) \tag{14}$$

Where $\{A_j\}$ are the events that partition S and

3.2.1 Intuitive Explanation of Bayes' Rule

Imagine you are trying to understand the likelihood of a specific event happening based on new information. Bayes' Rule helps you update your beliefs about this event. Here's an intuitive explanation:

- **Prior Belief**: Start with what you already know or believe about the event. This is your prior probability.
- **New Evidence**: You receive new information or evidence related to the event.
- **Likelihood**: Consider how likely this new evidence is if your initial belief (hypothesis) is true.
- **Update Belief**: Adjust your initial belief based on the new evidence. This gives you a new, updated probability, called the posterior probability.

In simpler terms, Bayes' Rule allows you to refine your initial assumptions (prior beliefs) by incorporating new data (evidence) to get a more accurate understanding (posterior belief) of the event's likelihood. It's like revising your opinion about something after considering additional information. There are two schematics that help me a lot in understanding how Bayes' Rule is computed. Please refer to Figure 4.

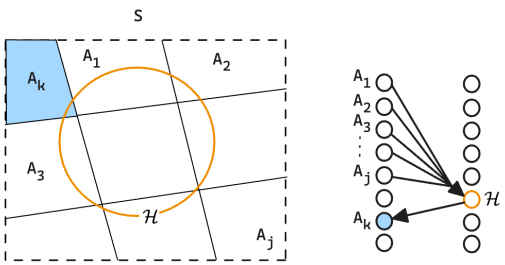

Figure 4: Illustration of how to understand Bayes' formula.

- **Left Diagram**: The sample space S is partitioned into several events $\{A_1, A_2, A_3, \ldots, A_j\}$. Each of these events represents a possible state or outcome. The event A_k is highlighted in blue, indicating the specific event for which we want to determine the posterior probability given the hypothesis $\mathcal{H}$, which is represented by the orange circle. The goal is to calculate the probability of A_k given $\mathcal{H}$, denoted as $P(A_k|\mathcal{H})$.

- **Right Diagram**: This diagram represents the flow of information between events and the hypothesis. Each event A_j (with $j \geq 1$) has a direct influence on the hypothesis $\mathcal{H}$. The highlighted event A_k (in blue) directly contributes to the hypothesis. The arrows indicate the direction of conditional dependencies. We start with the prior probabilities $P(A_j)$ and the conditional probabilities $P(\mathcal{H}|A_j)$, and use Bayes' rule to update our belief about A_k given the occurrence of $\mathcal{H}$.

  The key idea is to combine our prior belief about the events (left diagram) with new evidence (right diagram) to update our understanding of the specific event A_k in the context of the hypothesis $\mathcal{H}$.

3.2.2 Side information

One might believe that side information always increases the probability being calculated. However, conditioning provides information that can affect event probabilities in unexpected ways.

Example: Suppose we have a test for a rare disease that affects 1 in 1,000 people. The test is 99% accurate for those with the disease (true positive rate) and 99% accurate for those without the disease (true negative rate). If a person tests positive, what is the probability that they actually have the disease?

Let:

- D be the event that the person has the disease and D^c the event that the person doesn't have the disease.
- T^+ be the event that the person tests positive.

We know:

$$\begin{aligned} P(D) &= \frac{1}{1000} \quad \text{(prior probability of having the disease).} \\ P(T^+|D) &= 0.99 \quad \text{(probability of testing positive given the disease).} \\ P(T^+|D^c) &= 0.01 \quad \text{(probability of testing positive given no disease).} \\ P(D^c) &= 1 - P(D) = \frac{999}{1000} \quad \text{(prior probability of not having the disease).} \end{aligned}$$

We know from our conditional probability definition that

$$P(A|\mathcal{H}) = \frac{P(A \cap \mathcal{H})}{P(\mathcal{H})}$$

applying this to our problem, our hypothesis is that the person tests positive, represented as $P(\mathcal{H}) = P(T^+)$, and we want to calculate the probability that they actually have the disease given that their test result is positive. In other words, we aim to find $P(D|T^+)$, thus

$$P(D|T^+) = \frac{P(D) \cap P(T^+)}{P(T^+)} = \frac{P(T^+|D) \cdot P(D)}{P(T^+)}$$

Where

$$P(T^+) = P(T^+|D) \cdot P(D) + P(T^+|D^c) \cdot P(D^c).$$

Substituting the values:

$$P(T^+) = (0.99) \cdot \left(\frac{1}{1000}\right) + (0.01) \cdot \left(\frac{999}{1000}\right) = \frac{0.99}{1000} + \frac{9.99}{1000} = \frac{10.98}{1000} \implies P(T^+) = 0.01098.$$

Therefore,

$$P(D|T^+) = \frac{0.99 \cdot \frac{1}{1000}}{0.01098} = \frac{0.00099}{0.01098} \implies P(D|T^+) = 0.0902 = 9.02\%.$$

So, the probability that a person actually has the disease given a positive test result is approximately 9.02%, illustrating how conditioning on additional information can yield surprising results.

4 Introduction to Information Theory

$\Longrightarrow$ The following section was written using the book MacKay (2002) and MacKay (nd) online lectures as references.

4.1 How can we achieve perfect communication over an imperfect, noisy communication channel?

Example: Imagine you are playing a game of telephone, where one person whispers a message to the next person, and so on. Each person transmits the message correctly with a probability of $(1-p)$ and incorrectly with a probability of p.

Let's denote:

- p: The probability of transmitting the message incorrectly.
- $(1-p)$: The probability of transmitting the message correctly.
- n: The number of people the message is passed through.

The binary symmetric channel model can be used to analyze this scenario. The probability that the message is received incorrectly after being passed through n people can be calculated as follows:

$$P(\text{error}) = 1 - (1-p)^n \tag{15}$$

This formula assumes that each person makes an independent error with probability p. For example, if the message is passed through 3 people and each person has a 10% chance of making an error, the probability that the message is received incorrectly is:

$$P(\text{error}) = 1 - (1-0.1)^3 = 1 - 0.9^3 \approx 0.271 \tag{16}$$

Thus, there is approximately a 27.1% chance that the message will be incorrect after being passed through 3 people.

The figure 5 shows how the Binary Symmetric Channel (BSC) can be visualized. It can basically me understood as a communication channel that follows the Binomial Distribution.

$\Longrightarrow$ For a binomial distribution, the mean is given by Np and the variance is $Npq = \sigma^2$.

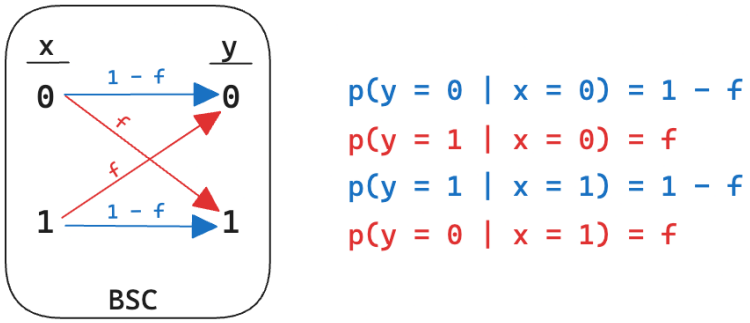

Figure 5: Binary Symmetric Channel (BSC)

4.1.1 What do we mean by *channel*?

Figure 6 illustrates the basic concept of transmitting a message. We start with a source message S that needs to be encoded before transmission. For example, when speaking to someone, we encode our ideas into a form that can be transmitted, and the listener must understand this encoding process to decode the message. For instance, if we are speaking in English, the English language serves as the encoding method for our source message. After encoding the message, we use our voice to transmit it, typically through the air channel in a normal conversation or through a call. The listener then needs to decode the message by understanding how to convert the sounds we make with our mouth into meaningful messages. This is, of course, a simplified view of the process of encoding, transmitting, and decoding a message.

However, during this process of communication, something might happen that introduces sufficient noise to the point where the other person cannot understand what we intended to say. With that in mind, we wonder: is there a way to correct these errors? And if there is, *what is the best error-correcting performance we can achieve?*

Let's see some practical examples of why correcting communication errors is very important.

Example: Imagine we want to buy a flash drive. However, the flash drive has a problem. Every bit you send to it to be saved will sometimes be flipped. Let's suppose that your large .mp3 file has a chunk of bits like $0011100\cdots$, and you transfer this .mp3 file to be saved on this flash drive. But the flash drive makes an error and saves it as $0011101\cdots$. The

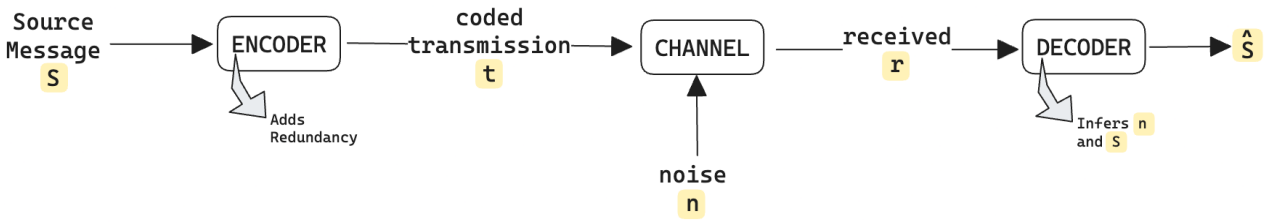

Figure 6: Process of the Encoding of a Message

manufacturer of the drive says that around 10% of the bits will be flipped when saving files. How worried should we be if we're buying this flash drive?

Let's imagine our .mp3 file has $N = 10,000$ bits which we need to store in the flash drive. Based on the manufacturer's information, the probability of error is given by $f = 0.1$. We can model this process of saving our file as a Binary Symmetric Channel, as shown in Figure 5. Thus, if we send a bit $x = 1$ to be saved on the flash drive, but a zero is saved $y = 0$, we can express this probability as $P(y = 0|x = 1) = f$. This reads as, *given that we have transmitted a 1, what's the probability we receive a zero*?

With that in mind, how can we calculate the total number of bits flipped if we try to save a file with $N = 10,000$ bits? As we've seen, the Binary Symmetric Channel (BSC) can be modeled as a binomial distribution, and we know that the mean or expected value of a binomial distribution is given by $E[X] = Np$. So, what is the expected number of bits to be flipped by this flash drive?

$$E[X] = 10,000 \cdot f = 1,000$$

This means that around 1,000 bits will be saved incorrectly. However, for a more precise estimate, we should also consider the standard deviation σ, which for a binomial distribution is calculated as $\sigma = \sqrt{Var[X]} = \sqrt{Npq}$.

$$Var[X] = 10,000 \cdot 0.1 \cdot 0.9 = 900 \implies \sigma = \sqrt{900} = 30$$

Finally, the number of bits flipped can be expressed as:

$$\text{Number of bits flipped} = Np \pm \sqrt{Npq} = 1,000 \pm 30$$

Therefore, we should be a bit worried :(

4.1.2 Adding Redundancy

Example: How can we help the flash drive company correct this saving malfunction?

The keyword here is redundancy. We want our encoders to add *redundancy*. This way, if our flash drive makes a mistake, we can try to check if an error occurred. Remember, sometimes we don't have access to the original transmitted file or message. We need to be clever in this approach.

Let's analyze two simple strategies we could employ to fix add redundancy to the encoding process.

1. **Parity Encoding**: Imagine we receive the bits 10101. We could append one extra bit at the end of every message to indicate whether we have an even or odd quantity of bits. For instance, 10101 would have a $parity = 1$ because there are three ones (odd). Similarly, 11110 would have a $parity = 0$ because there are four ones (even). This way, if we receive a message with an even number of ones but the parity bit is 1, we can be certain that an error occurred when saving this file. However, with parity coding, as you have probably noticed, we can only detect errors, not correct them :/

2. **Repetition Code**: Repetition codes are a simple form of error detection and correction. The idea is to repeat each bit of the message multiple times. For instance, if we want to transmit the bit '1', we might repeat it three times, sending '111'. Similarly, to transmit the bit '0', we would send '000'. When the receiver gets the message, they use a majority vote to determine the original bit. For example, if the receiver gets '110', they would interpret it as '1' because the majority of the bits are '1'. This method helps to correct errors as long as the number of errors is less than half the number of repetitions.

Let's imagine we receive a file with three bits, and we wish to use the repetition code to make it work. Thus, let's understand the table 2. It shows the received sequences r, their likelihood ratios $\frac{P(r|s=1)}{P(r|s=0)}$, and the decoded sequences $\hat{s}$. The decoding algorithm assumes the channel is a binary symmetric channel with $\gamma = \frac{(1-f)}{f}$, where f is the error probability.

Received sequence r	Likelihood ratio $\frac{P(r\|s=1)}{P(r\|s=0)}$	Decoded sequence $\hat{s}$
000	γ^{-3}	0
001	γ^{-1}	0
010	γ^{-1}	0
100	γ^{-1}	0
101	γ^{1}	1
110	γ^{1}	1
011	γ^{1}	1
111	γ^{3}	1

Table 2: Likelihood ratios and decoded sequences for a binary symmetric channel MacKay (2002).

To explain the optimal decoding decision (i.e., the decision that minimizes the probability of being wrong), we need to find the value of s that is most probable, given r. Consider the decoding of a single bit s, which was encoded as the sequence $r = r_1 r_2 r_3$. By Bayes' theorem, the posterior probability of s is:

$$P(s|r_1 r_2 r_3) = \frac{P(r_1 r_2 r_3|s)P(s)}{P(r_1 r_2 r_3)} \tag{17}$$

We can spell out the posterior probabilities of the two alternatives thus:

$$P(s = 1|r_1 r_2 r_3) = \frac{P(r_1 r_2 r_3|s = 1)P(s = 1)}{P(r_1 r_2 r_3)} \tag{18}$$

$$P(s = 0|r_1 r_2 r_3) = \frac{P(r_1 r_2 r_3|s = 0)P(s = 0)}{P(r_1 r_2 r_3)} \tag{19}$$

The posterior probability is determined by two factors: the prior probability $P(s)$ and the data-dependent term $P(r_1 r_2 r_3|s)$, which is called the likelihood of s. Assuming the prior probabilities are equal ($P(s = 0) = P(s = 1) = 0.5$), then maximizing the posterior probability $P(s|r)$ is equivalent to maximizing the likelihood $P(\boldsymbol{r}|s)$.

Assuming the channel is a binary symmetric channel with noise level $f < 0.5$, which in our case for the flash drive manufacturer is $f = 0.1$, the likelihood is:

$$P(\boldsymbol{r}|s) = P(\boldsymbol{r}|t(s)) = \prod_{n=1}^{N} P(r_n|t_n(s)) \tag{20}$$

where $N = 3$ is the number of transmitted bits in the block , and:

$$P(r_n|t_n) = \begin{cases} (1-f) & \text{if } r_n = t_n \\ f & \text{if } r_n \neq t_n \end{cases} \tag{21}$$

Thus, the likelihood ratio for the two hypotheses is:

$$\frac{P(r|s = 1)}{P(r|s = 0)} = \prod_{n=1}^{N} \frac{P(r_n|t_n(1))}{P(r_n|t_n(0))} \tag{22}$$

Each factor $\frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$ equals $\left(\frac{1-f}{f}\right)$ if $r_n = 1$ and $\left(\frac{f}{1-f}\right)$ if $r_n = 0$. The ratio

$$\gamma = \left(\frac{1-f}{f}\right)$$

implies that the winning hypothesis is the one with the most 'votes', each vote counting for a factor of γ in the likelihood ratio.

Thus, the majority-vote decoder shown in table 2 is the optimal decoder if we assume that the channel is a binary symmetric channel and that the two possible values of s are equally likely. For a more in depth explanation please refer to MacKay (2002).

$\Longrightarrow$ If we use repetition code to communicate data over a telephone line for example, it'll reduce the error frequency, but it'll also reduce our communication rate. We'll have to pay three times as much for each phone call.

5 Markov Chains and Information Source

$\Longrightarrow$ The following sections used the following books as reference MacKay (2002); Feller (1968)

Independent trials can be described as a set of possible outcomes where $E_1, E_2, \cdots$ (finite or infinite in number) is given, and with each there's associated a probability P_k; the probabilities of sample sequences are defined by the multiplicative property

$$P\{(E_{j0}, E_{j1}, \cdots, E_{jn})\} = P_{j0}P_{j1}\cdots P_{jn} \tag{23}$$

$\Longrightarrow$ In the theory of Markov chains, we consider the simplest generalization which consists in permitting the outcome of any trial to depend on the outcome of the directly preceding trial (and only on it).

The outcome E_k is no longer associated with a fixed probability P_k, but to every pair (E_j, E_k) there corresponds a conditional probability P_{jk}; given that E_j has occurred at some trial, the probability of E_k at the next trial is P_{jk}. In addition to P_{jk}, we must be given the probability a_k of the outcome E_k of the initial trial. For P_{jk} to have the meaning attributed to them, the probabilities of sample sequences corresponding to two trials are

$$P\{(E_j, E_k)\} = a_k P_{jk}$$

For three trials:

$$P\{(E_j, E_k, E_r)\} = a_k P_{jk} P_{kr}$$

For four trials:

$$P\{(E_j, E_k, E_r, E_s)\} = a_k P_{jk} P_{kr} P_{rs}$$

And generally we have:

$$P\{(E_{j0}, E_{j1}, \cdots, E_{jn})\} = a_{j0} P_{j0j1} P_{j1j2} \cdots P_{j_{n-2}j_{n-1}} P_{j_{n-1}j_n} \tag{24}$$

A sequence of trials with possible outcomes $E_1, E_2, \cdots$ is called a Markov chain if the probabilities of sample sequences are defined by the equation above in terms of a probability distribution $\{a_k\}$ for each E_k at the initial (or zero-th) trial and fixed conditional probabilities P_{jk} of E_k given that E_j has occurred at the preceding trial.

$\Longrightarrow$ P_{jk} is called the probability of a transition from E_j to E_k given by $P(X_{t+1} = j | X_t = k) = p_{jk}$.

We can compute a matrix of transition probabilities as shown below, which gives us the probability of changing each state.

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2k} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{j1} & p_{j2} & \cdots & p_{jk} & \cdots & p_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nk} & \cdots & p_{nn} \end{bmatrix}$$

Definition 5.1 *A stochastic process is said to be* ***stationary*** *if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index; that is,*

$$Pr\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} = Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \cdots, X_{n+l} = x_n\} \tag{25}$$

For every n and every shift l and for all $x_1, x_2, \cdots, x_n \in \mathcal{X}$.

5.0.1 Markov Information Source

This is a source of information that allows the probability distribution of the output at any point in time to be determined in a fixed manner from only the m outputs immediately before that point, regardless of the previous outputs. More formally, for random variables $X_1, X_2, \cdots X_n$, where $n \geq 3$, $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if

$$P(x_1, x_2, \cdots, x_n)P(x_2)P(x_3)\cdots P(x_{n-1}) = P(x_1, x_2)P(x_2, x_3)\cdots P(x_{n-1}, x_n) \tag{26}$$

For all $x_1, x_2, \cdots, x_n$, or equivalently

$$P(x_1, x_2, \cdots, x_n) = P(x_1, x_2)P(x_3|x_2)\cdots P(x_n|x_{n-1}) \text{ if } P(x_2), P(x_3), \cdots, P(x_{n-1}) > 0 \text{ and 0 otherwise.} \tag{27}$$

Or we could define it as:

$$P_{X_i|X_{i-1}\cdots X_{i-n}}(x_i|x_{i-1}, \cdots, x_{i-n}) = P_{X_i|X_{i-1},\cdots,X_{i-m}}(x_i|x_{i-1}, \cdots, x_{i-m}) \text{ for } n \geq m \tag{28}$$

5.0.2 Stationary Distributions

⟹ This subsection used the following tutorial as reference GSIS-Tohuku (2017)

*What's the probability of making a transition from state i to state j over **two** steps?*

$$P(X_2 = j|X_0 = i) = \sum_{k=1}^{N} P_i(X_2 = j|X_1 = k)P_i(X_1 = k) = \sum_{k=1}^{N} P(X_2 = j|X_1 = k, X_0 = i)P(X_1 = k|X_0 = i)$$

$$P(X_2 = j|X_0 = i) = \sum_{k=1}^{N} P(X_2 = j|X_1 = k)P(X_1 = k|X_0 = i)$$

$$P(X_2 = j|X_0 = i) = \sum_{k=1}^{N} p_{kj}p_{ik}$$

$$P(X_2 = j|X_0 = i) = \sum_{k=1}^{N} p_{ik}p_{kj} = (P^2)_{ij}, \text{ where } P \text{ is the transition matrix.}$$

Thus we can represent our possible states using the following matrix π,

$$\pi = \begin{bmatrix} P(X_0 = s_1) \\ P(X_0 = s_2) \\ \vdots \\ P(X_0 = s_N) \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{bmatrix}$$

With that in mind, let $\{X_n\}$ be a Markov chain on S with transition probability matrix P. A distribution π on S is called stationary (or invariant) if

$$\pi = \pi P \tag{29}$$

or equivalently if

$$\pi_j = \sum_{i \in S} \pi_i p_{ij}, \quad j \in S \tag{30}$$

Thus, in order to find a stationary distribution of a Markov chain with a transition probability matrix $P = [p_{ij}]$, we need to solve the linear system (6.1) (or equivalently (6.2)) together with the conditions:

$$\sum_i \pi_i = 1 \quad \text{and} \quad \pi_i \geq 0 \text{ for all } i \in S.$$

Example: **2-state Markov chain**. Consider the transition matrix:

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

Solving the equation $\pi P = \pi, \pi = [\pi_0\, \pi_1]$, which is equivalent to $\pi_0 p + \pi_1 q = 0$, together with $\pi_0 + \pi_1 = 1$, we obtain

$$\pi_0 = \frac{q}{p+q}, \quad \pi_1 = \frac{p}{p+q}$$

Thus we conclude:

(i) If $p + q > 0$, there exists a unique stationary distribution.

(ii) If $p = q = 0$, a stationary distribution is not uniquely determined. In fact, any distribution $\pi = [\pi_0, \pi_1]$ is stationary.

Example: A particle moves along the x-axis in such a way that its absolute speed remains constant but the direction of the motion can be reversed. The system is said to be in state E_1 if the particle moves in the positive direction, and in state E_2 if the motion is to the left. Let's denote

- $\alpha \rightarrow$ probability of reversal when the particle moves to the right
- $\beta \rightarrow$ probability of reversal when the particle moves to the left

Given the transition matrix shown below and the transition diagram in Figure 7, we can solve this problem by using the stationary equation $\pi = \pi P$, where $\pi = [\pi_0, \pi_1]$. That is

$$[\pi_0\, \pi_1] \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} = [\pi_0, \pi_1]$$

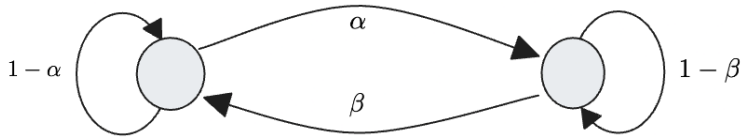

Figure 7: Two State Markov Chain

This implies the system of equations:

$$\pi_0(1-\alpha) + \pi_1\beta = \pi_0 \quad \text{and} \quad \pi_0\alpha + \pi_1(1-\beta) = \pi_1$$

and we know that $\sum_i \pi_i = 1 \implies \pi_0 + \pi_1 = 1$. Thus,

$$\pi_0(1-\alpha) + \beta(1-\pi_0) = \pi_0 \implies \pi_0 - \alpha\pi_0 + \beta - \beta\pi_0 = \pi_0 \implies \pi_0 = \frac{\beta}{\alpha+\beta}$$

Similarly,

$$(1-\pi_1)\alpha + \pi_1(1-\beta) = \pi_1 \implies \pi_1 = \frac{\alpha}{\alpha+\beta}$$

By setting $\alpha = 0.3$ and $\beta = 0.2$, we created an animation and code that can be executed using the following link. Additionally, we included Figure 8, which illustrates the particle movements at each step based on the stationary equations we derived for π_0 and π_1.

Example: A user moves among four different places: Home, Work, Park, and Bar. The system is said to be in state E_i if the user is at place i, where $i \in \{\text{Home, Work, Park, Bar}\}$. Let's denote:

- $p_{ij} \rightarrow$ probability of moving from place i to place j

Given the transition matrix shown below and the transition diagram, we can solve this problem by using the stationary equation $\pi = \pi P$, where $\pi = [\pi_0, \pi_1, \pi_2, \pi_3]$. That is,

$$\pi = [\pi_{\text{Home}}\ \pi_{\text{Work}}\ \pi_{\text{Park}}\ \pi_{\text{Bar}}]$$

and

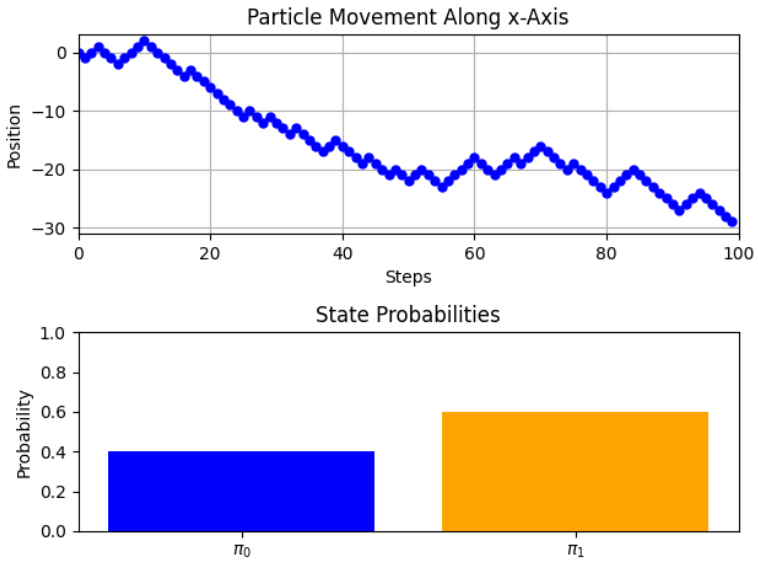

Figure 8: Particle movement with markov chains

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.1 & 0.1 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.3 & 0.1 \\ 0.2 & 0.3 & 0.2 & 0.3 \end{bmatrix}.$$

This implies the system of equations:

$$\begin{aligned}
\pi_{\text{Home}}(0.5) + \pi_{\text{Work}}(0.2) + \pi_{\text{Park}}(0.3) + \pi_{\text{Bar}}(0.2) &= \pi_{\text{Home}} \\
\pi_{\text{Home}}(0.3) + \pi_{\text{Work}}(0.5) + \pi_{\text{Park}}(0.3) + \pi_{\text{Bar}}(0.3) &= \pi_{\text{Work}} \\
\pi_{\text{Home}}(0.1) + \pi_{\text{Work}}(0.2) + \pi_{\text{Park}}(0.3) + \pi_{\text{Bar}}(0.2) &= \pi_{\text{Park}} \\
\pi_{\text{Home}}(0.1) + \pi_{\text{Work}}(0.1) + \pi_{\text{Park}}(0.1) + \pi_{\text{Bar}}(0.3) &= \pi_{\text{Bar}}
\end{aligned}$$

and we know that $\sum_i \pi_i = 1 \implies \pi_{\text{Home}} + \pi_{\text{Work}} + \pi_{\text{Park}} + \pi_{\text{Bar}} = 1$. Thus,

$$\begin{aligned}
\pi_{\text{Home}}(0.5) + \pi_{\text{Work}}(0.2) + \pi_{\text{Park}}(0.3) + \pi_{\text{Bar}}(0.2) &= \pi_{\text{Home}} \\
\pi_{\text{Home}}(0.3) + \pi_{\text{Work}}(0.5) + \pi_{\text{Park}}(0.3) + \pi_{\text{Bar}}(0.3) &= \pi_{\text{Work}} \\
\pi_{\text{Home}}(0.1) + \pi_{\text{Work}}(0.2) + \pi_{\text{Park}}(0.3) + \pi_{\text{Bar}}(0.2) &= \pi_{\text{Park}} \\
\pi_{\text{Home}}(0.1) + \pi_{\text{Work}}(0.1) + \pi_{\text{Park}}(0.1) + \pi_{\text{Bar}}(0.3) &= \pi_{\text{Bar}} \\
\pi_{\text{Home}} + \pi_{\text{Work}} + \pi_{\text{Park}} + \pi_{\text{Bar}} &= 1
\end{aligned}$$

Solving these equations, we find the stationary distribution π. Using a Python script available at this link, we can visualize the evolution of state probabilities over time, showing how the user's location probabilities change with each step in the Markov chain. Figure 9 illustrates the likelihood of the user being in each of the four states (Home, Work, Park, Bar) at any given step.

$\Longrightarrow$ I recommend watching the following video Sachdeva (2022) for an explanation and visual animations for Markov Chains.

6 Entropy, Relative Entropy, and Mutual Information

$\Longrightarrow$ The following section was written using mainly the book Cover and Thomas (2006) as references.

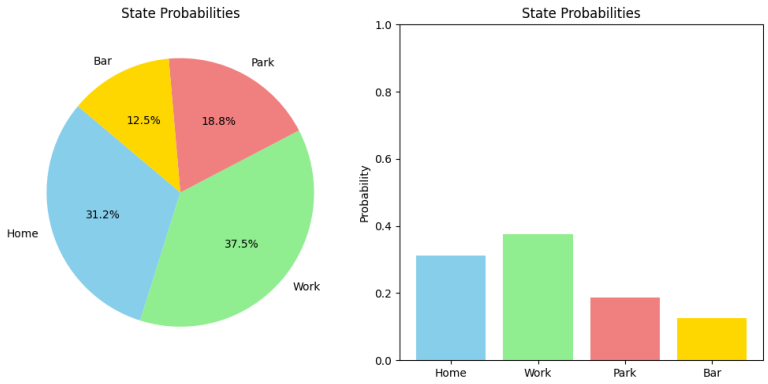

Figure 9: Likelihood of the user being in each of the four states (Home, Work, Park, Bar) at any given step.

6.1 Entropy

Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $P(x) = P\{X = x\}, x \in \mathcal{X}$. We denote the probability mass function by $p(x)$ rather than $P_X(x)$.

Definition 6.1 *The entropy $H(X)$ of a discrete random variable X is defined by*

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log(p(x)) \tag{31}$$

where entropy is expressed in bits.

$\Longrightarrow$ Note that entropy is a functional of the distribution of X. It doesn't depend on the actual values taken by the random variable X, but only on the probabilities.

We denote expectation by $\mathbb{E}$. Thus, if $X \sim p(x)$, the expected value of the random variable $g(x)$ is written as

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

6.2 Joint Entropy and Conditional Entropy

Definition 6.2 *The joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) is defined as*

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x,y)) \tag{32}$$

which can also be expressed as

$$H(X,Y) = -\mathbb{E}[\log(p(x,y))] \tag{33}$$

Definition 6.3 *If $(X,Y) \sim p(x,y)$, the conditional entropy $H(Y|X)$ is defined as*

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(y|x)\log(p(y|x)) \tag{34}$$

or equivalently,

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(y|x)) = -\mathbb{E}[\log(P(Y|X))] \tag{35}$$

Theorem 6 (Chain Rule)

$$H(X,Y) = H(X) + H(Y|X) \quad or \quad H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \tag{36}$$

6.3 Relative Entropy and Mutual Information

The relative entropy is a measure of the distance between two distributions.

Definition 6.4 *The relative entropy, or Kullback-Leibler divergence, between two probability mass functions $p(x)$ and $q(x)$ is defined as*

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) = \mathbb{E}_p\left[\log\left(\frac{p(x)}{q(x)}\right)\right] \tag{37}$$

where the relative entropy is nonnegative and is zero if and only if $p = q$.

We can also refer to relative entropy as the information divergence between two probability distributions p and q.

Definition 6.5 *Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:*

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = D(p(x,y)||p(x)p(y)) \tag{38}$$

or equivalently,

$$I(X;Y) = E_{p(x,y)}\left[\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right] \tag{39}$$

We can rewrite the definition of mutual information $I(X;Y)$ as

$$I(X;Y) = \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \sum_{x,y} p(x,y)\log\left(\frac{p(x|y)}{p(x)}\right)$$

$$I(X;Y) = \sum_{x,y} p(x,y)\log(p(x)) + \sum_{x,y} p(x,y)\log(p(x|y))$$

$$I(X;Y) = \sum_{x} p(x)\log(p(x)) - \left(-\sum_{x,y} p(x,y)\log(p(x|y))\right)$$

Finally,

$$I(X;Y) = H(X) - H(X|Y) \tag{40}$$

$\Longrightarrow$ The mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y.

Theorem 7 (Mutual Information and entropy)

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

we can see this relationship between entropy and mutual information by the Venn diagram shown in Figure 10.

6.4 Chain Rules For Entropy, Relative Entropy, and Mutual Information

Theorem 8 (Chain rule for entropy) *Let $X_1, X_2, \cdots, X_n$ be drawn according to $p(x_1, x_2, \cdots, x_n)$. Then*

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1) \tag{41}$$

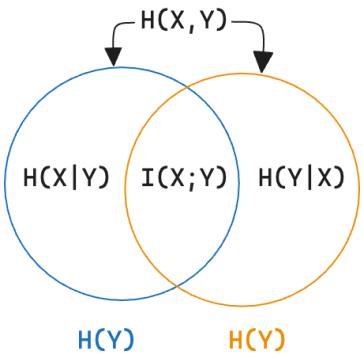

Figure 10: Relationship between entropy and mutual information.

Example: $H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2)$

We now define the conditional mutual information as the reduction in the uncertainty of X due to the knowledge of Y when Z is given.

Definition 6.6 (Conditional Mutual Information) *The conditional mutual information of random variables X and Y given Z is defined by*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \mathbb{E}_{p(x,y,z)} \log \left(\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \right) \tag{42}$$

6.5 Jensen's Inequality and its Consequences

A function $f(x)$ is said to be convex over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{43}$$

A function $f(x)$ is said to be strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$. The figure 11 shows an example of each kind of function.

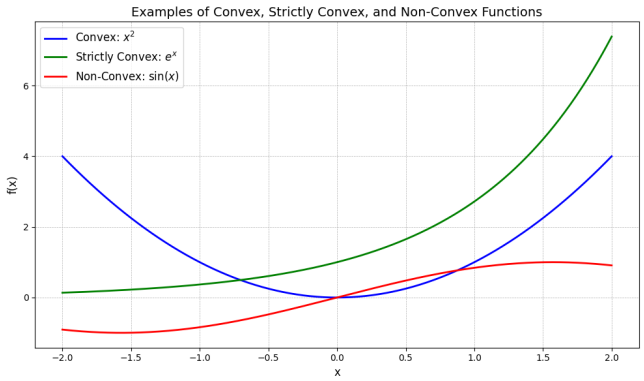

Figure 11: Illustration of a striclty convex, convex, and non convex functions

Theorem 9 *If the function f has a second derivative that is non-negative/positive over an interval, the function is convex/strictly convex over that interval.*

The figure 12 illustrates this theorem.

Theorem 10 (Jensen's Inequality) *If f is a convex function and X is a random variable*

$$\mathbb{E}f[(x)] \geq f(\mathbb{E}[X]) \tag{44}$$

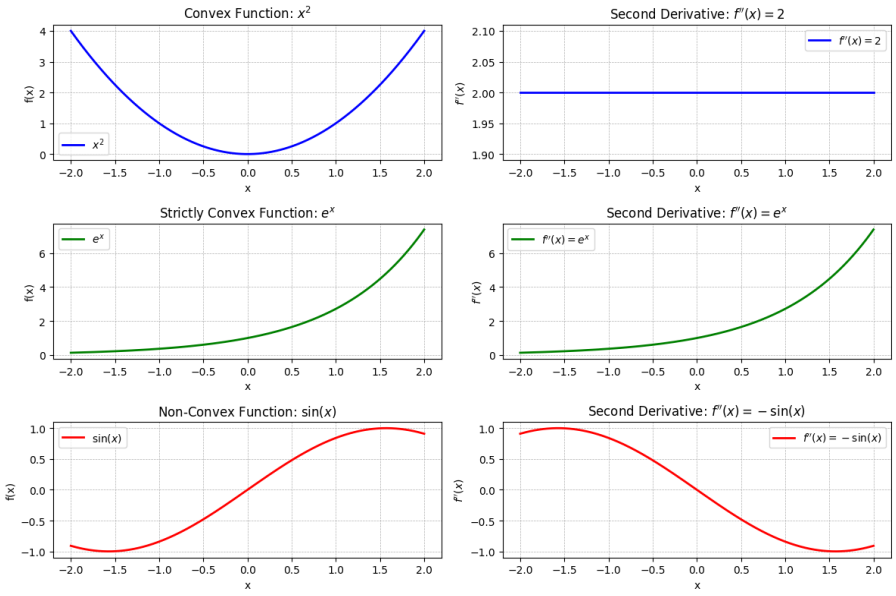

Figure 12: Illustration of the second derivative per type of function - Theorem 9.

Moreover, if f is a stricly convex, the equality above implies that $X = EX$ with probability 1 (i.e., X is a constant). Figure 13 illustrates some examples of Jensen's Inequality.

Theorem 11 (Information Inequality) *Let $p(x)$, $q(x)$, $x \in \mathcal{X}$ be two probability mass functions. Then*

$$D(p||q) \geq 0 \tag{45}$$

With equality if and only if $p(x) = q(x) \forall x$. This theorem Inequality is also called the Divergence Inequality .

Corollary 1 (Nonnegativity of Mutual Information) *For any two random variables X, Y*

$$I(X;Y) \geq 0 \tag{46}$$

with equality if and only if X and Y are independent.

Corollary 2

$$D(p(y|x)||q(y|x)) \geq 0 \tag{47}$$

with equality if and only if $p(y|x) = q(y|x)$ for all y and x such that $p(x) > 0$.

Corollary 3

$$I(X;Y|Z) \geq 0 \tag{48}$$

with equality if and only if X and Y are conditionally independent given Z that is $X \perp\!\!\!\perp Y \mid Z$.

Theorem 12

$$H(X) \leq \log(|H(\mathcal{X})|) \tag{49}$$

where $|\mathcal{X}|$ denotes the number of elements in the range of $\mathcal{X}$, with equality if and only if X has an uniform distribution over $\mathcal{X}$.

Theorem 13 (Conditioning reduces entropy)

$$H(X|Y) \leq H(X) \tag{50}$$

with equality if and only if $X \perp\!\!\!\perp Y$ (X and Y are independent).

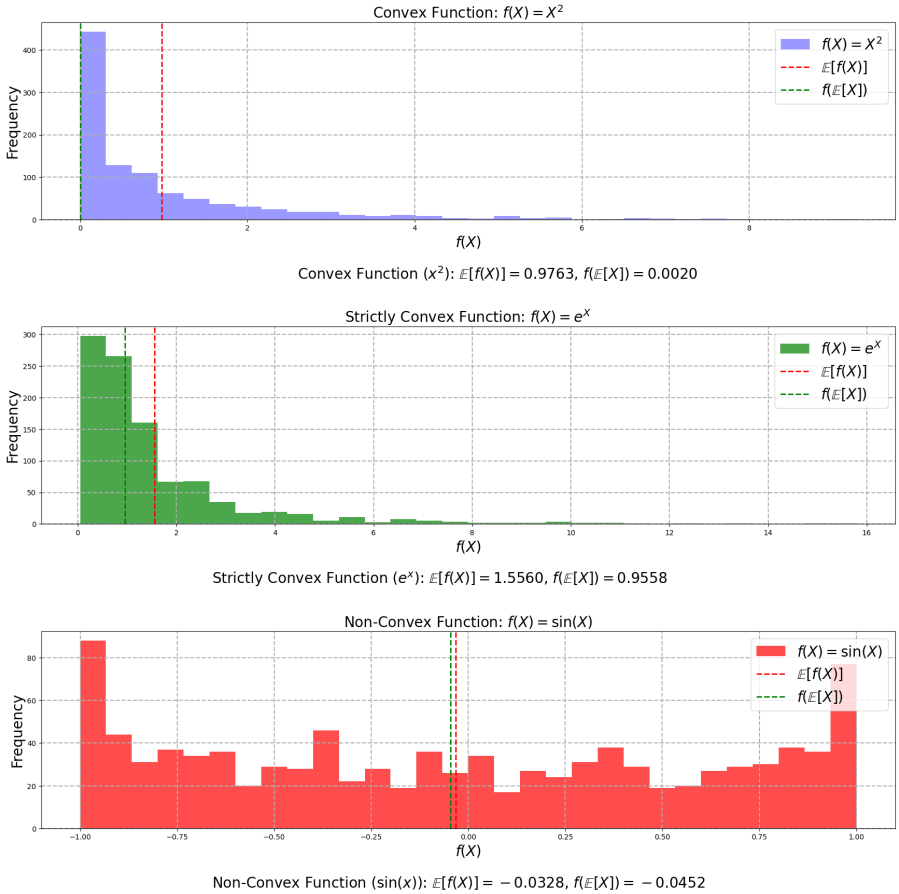

Figure 13: Illustration of Jensen's Inequality - Theorem 10.

Theorem 14 (Independence bound on entropy) *Let $X_1, X_2, \cdots, X_n$ be drawn according to $p(x_1, x_2, \cdots, x_n)$. Then*

$$H(X_1, X_2, \cdots, X_n) \leq \sum_{n}^{i=1} H(X_i) \tag{51}$$

with equality if and only if X_i are independent.

Figure 14 shows an example of the independence bound on entropy.

Lemma 1 (Fundamental Inequality) *For any $a > 0$*

$$\ln(a) \leq a - 1 \tag{52}$$

with equality if and only if $a = 1$.

Figure 14 illustrates the fundamental inequality.

Theorem 15 (Concavity of Entropy) *$H(P)$ is a concave function of P.*

Theorem 16 *Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.*

6.6 Data Processing Inequality

The data-processing inequality can be used to show that no clever manipulation of the data can improve the inferences that can be made from the data.

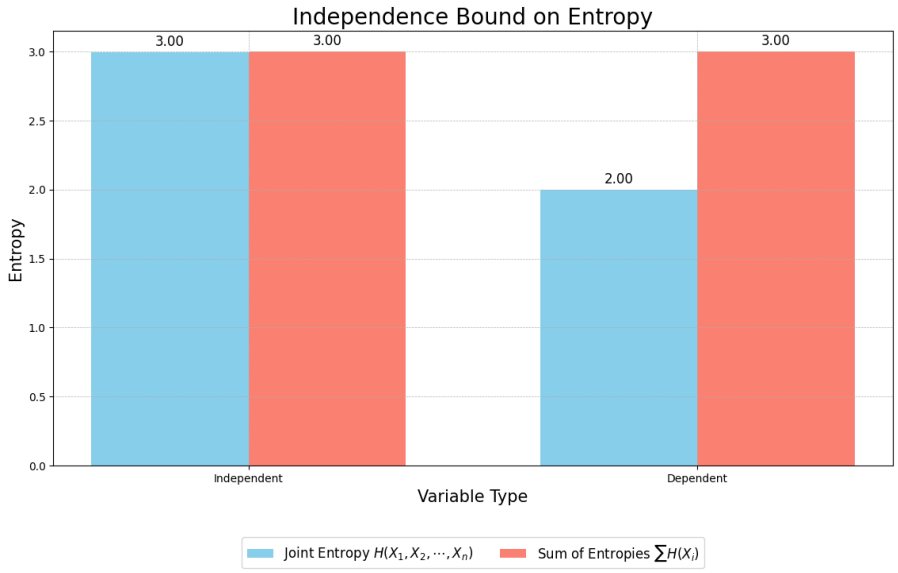

Figure 14: Example of the independence bound on entropy - Theorem 14.

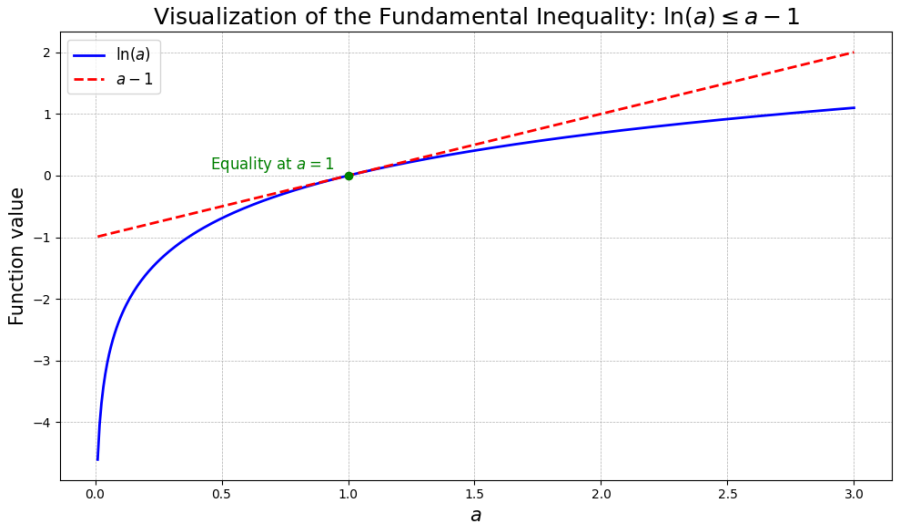

Figure 15: Example of the fundamental inequality - Lemma 1.

Definition 6.7 *Random variables X, Y, Z are said to form a Markov chain in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y, and Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written as*

$$p(x, y, z) = p(x)p(y|x)p(z|y) \tag{53}$$

Theorem 17 (Data Processing Inequality) *If $X \to Y \to Z$ forms a Markov chain, then*

$$I(X;Y) \geq I(X;Z) \text{ equivalently } I(X;Z) \geq I(X;Z) \tag{54}$$

It can also be stated that

$$I(X;Y,Z) \geq I(X;Z) \tag{55}$$

If the random variables are closer in the Markov Chain, they convey higher mutual information. The figure 16 illustrates such a theorem.

6.7 Sufficient Statistics

Definition 6.8 *A statistic $T(X)$ is said to be a sufficient statistic for θ if it contains all the information in X about θ.*

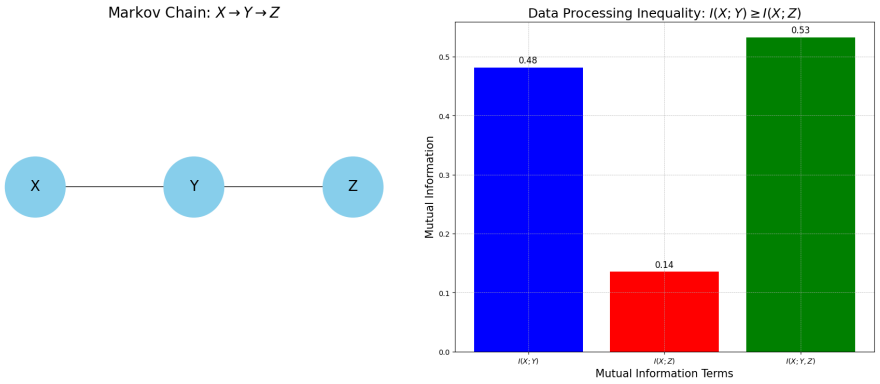

Figure 16: Example of the data processing inequality - Theorem 17.

Suppose that we have a family of probability mass functions $\{f(x|\theta)\}$ indexed by θ. Let X be a sample from a distribution in the family. Let $T(X)$ be our statistic (function of the sample). Like the sample mean or sample variance. Then $\Theta \to X \to T(X)$ we have:

$$I(\theta; X) = I(\theta; T(X)) \text{ for any distribution on } \theta$$

due to the data processing inequality we have:

$$I(\theta; T(X)) \leq I(\theta; X)$$

A statistic $T(X)$ is called sufficient for Θ if it contains all the information in X about Θ.

Definition 6.9 *A statistic $T(X)$ is a minimal sufficient statistic relative to $\{f_\theta(x)\}$ if it's a function of every other sufficient statistic U. Interpreting this in terms of the data-processing inequality, this implies that*

$$\Theta \to T(X) \to U(X) \to X$$

6.8 Fano's Inequality

Suppose that we know a random variable Y and we wish to guess the value of a correlated random variable X. Fano's Inequality relates the probability of error in guessing X to its conditional entropy $H(X|Y)$.

Theorem 18 (Fano's Inequality) *For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = Pr\{X \neq \hat{X}\}$*

$$H(P_e) + P_e \log(|\mathcal{X}|) \geq H(X|\hat{X}) \geq H(X|Y) \tag{56}$$

This inequality can be weakened to:

$$1 + P_e \log(|\mathcal{X}| \geq H(X|Y))$$

That is

$$P_e \geq \frac{H(X|Y) - 1}{\log |X|} \tag{57}$$

7 Data Compression

⟹ The following section was written using mainly the book Cover and Thomas (2006) as references.

Definition 7.1 *A source code C for a random variable X is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{D}^*$, the set of finite-length strings of symbols from a $\mathcal{D}$-any alphabet. Let $C(x)$ denote the codeword corresponding to x and let $l(x)$ denote the length of $C(x)$*

Definition 7.2 *The expected length $L(C)$ of a source code $C(x)$ for a random variable X with probability mass function $p(x)$ is given by*

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x) \tag{58}$$

Example: Suppose we have a random variable X with the following probability mass function:

$$p(X) = \begin{cases} 0.4 & \text{if } X = A \\ 0.3 & \text{if } X = B \\ 0.2 & \text{if } X = C \\ 0.1 & \text{if } X = D \end{cases}$$

Let's consider a source code C with the following codewords:

$$C(A) = 0, \quad C(B) = 10, \quad C(C) = 110, \quad C(D) = 111$$

The lengths of the codewords are:

$$l(A) = 1, \quad l(B) = 2, \quad l(C) = 3, \quad l(D) = 3$$

The entropy $H(X)$ of the random variable X is given by:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \lg p(x)$$

Substituting the given probabilities:

$$H(X) = -[0.4 \lg 0.4 + 0.3 \lg 0.3 + 0.2 \lg 0.2 + 0.1 \lg 0.1] \implies H(X) = 1.8464\text{bits}$$

The expected length $L(C)$ of the source code C is given by:

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)$$

Substituting the given probabilities and codeword lengths:

$$L(C) = 0.4 \cdot 1 + 0.3 \cdot 2 + 0.2 \cdot 3 + 0.1 \cdot 3 \implies L(C) = 1.9\text{bits}$$

7.1 Nonsingular Code

Definition 7.3 *A nonsingular code is a code in which different source symbols are mapped to distinct codewords. Formally, a code C is nonsingular if $C(x_i) \neq C(x_j)$ for all $x_i \neq x_j$ in the source alphabet $\mathcal{X}$.*

Example: Consider the source alphabet $\mathcal{X} = \{A, B, C\}$ and the codebook $C = \{00, 01, 10\}$. This code is nonsingular because each symbol maps to a unique codeword:

$$C(A) = 00, \quad C(B) = 01, \quad C(C) = 10$$

7.2 Uniquely Decodable Code

Definition 7.4 *A uniquely decodable code is a code in which every sequence of codewords maps to at most one sequence of source symbols. In other words, no two different sequences of source symbols result in the same sequence of codewords.*

Example: Consider the source alphabet $\mathcal{X} = \{A, B\}$ and the codebook $C = \{0, 10\}$. This code is uniquely decodable because any sequence of codewords can be uniquely parsed:

$$\text{Code sequence: } 010 \quad \text{decodes to: } AB$$

$$\text{Code sequence: } 1010 \quad \text{decodes to: } BB$$

7.3 Prefix Code

Definition 7.5 *A prefix code (or prefix-free code) is a code in which no codeword is a prefix of any other codeword. This property ensures that the code can be instantly decoded without the need for lookahead.*

Example: Consider the source alphabet $\mathcal{X} = \{A, B, C\}$ and the codebook $C = \{0, 10, 110\}$. This code is a prefix code because no codeword is a prefix of any other.

7.4 Instantaneous Code

Definition 7.6 *A instantaneous code is a code in which each codeword can be decoded immediately upon receipt of the final symbol in the codeword. All prefix codes are instantaneous codes.*

Example: Using the same example as the prefix code: consider the source alphabet $\mathcal{X} = \{A, B, C\}$ and the codebook $C = \{0, 10, 110\}$. This code is also an instantaneous code. The code is instantaneous because once we receive a codeword, we can immediately decode it without waiting for any additional symbols.

Example: Consider the source alphabet $\mathcal{X} = \{A, B, C\}$ and the codebook $C = \{0, 01, 011\}$. This code is non-instantaneous because the codeword for B is a prefix of the codeword for C.

7.5 Kraft Inequality

We wish to construct instaneous codes for minimum expected length to describe a given source.

Theorem 19 (Kraft Inequality) *For any instaneous code (prefix code) over an alphabet of size D, the codeword lengths $l_1, l_2, \cdots, l_m$ must satisfy the Inequality*

$$\sum_i D^{l_i} \leq 1 \tag{59}$$

Conversely, given a set of codewords lengths that satisfy this inequality there exists an instantneous code with these word lengths. The idea of Kraft Inequality can be understood through the table **??** shown in the online class of MacKay (nd).

References

Blitzstein, J. (2024). Harvard university - statistics 101. https://youtube.com/playlist?list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo&si=BQbsxrSDx5yjeIcp. Accessed: 2024-07-17.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA. http://www.deeplearningbook.org.

GSIS-Tohuku (2017). Stationary distributions. https://www.math.is.tohoku.ac.jp/~obata/student/graduate/file/2017-GSIS-ProbModel6-9.pdf. Accessed: 2024-07-19.

MacKay, D. (n.d.). Information theory, pattern recognition, and neural networks. https://youtube.com/playlist?list=PLruBu5BI5n4aFpG32iMbdWoRVAA-Vcso6&si=V1ajzYjkK3x_T1la. Produced by David MacKay, University of Cambridge.

MacKay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, USA.

Sachdeva, K. (2022). Markov chains - visually explained + history! https://www.youtube.com/watch?v=CIe869Rce2k. Accessed: 2024-07-20.

Venkatesh, S. S. (2024). University of pennsylvania - lectures on conditional probability. https://www.youtube.com/watch?v=vu_2qT-qB_Y&list=PLhCDzMM3Yov1DiPMQceoaj-QVy6ZCUiLc. Accessed: 2024-07-17.