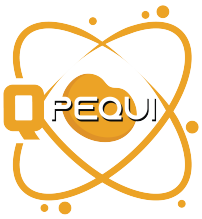


Federal University of Goiás

Institute of Physics

Quantum Pequi Group

Introduction to Classical and Quantum Information

Lecture Notes

Prof. Lucas Chibebe Céleri

Goiânia, 2026

Brazil

“*Lasciate ogne speranza, voi ch’ intrate.*”

D. A.

To Ana, who taught me the meaning of life.

Contents

Part I

Classical Information Theory

Chapter 1

Information measure

The mathematical theory of information was introduced by Claude Shannon in his seminal paper of 1948 *A Mathematical Theory of Communication* [1]. The theory aims at quantifying information in a precise and operational way. The central problem that motivated Shannon's work was the following: given a source that produces messages and a communication channel through which those messages are transmitted, what are the fundamental limits on the reliable compression and transmission of information?

Before addressing such questions, one must first clarify what is meant by the word information. In everyday language, information is often associated with meaning or knowledge. A sentence may be informative because it tells us something about the world, because it answers a question, or because it changes our understanding of a situation. Shannon's approach ignores this semantic aspect. Instead, it focuses exclusively on the statistical structure of the messages. In this way, Shannon was able to precisely define the information and develop a theory for its transformations.

The main idea is that information is related to uncertainty. Suppose that a system can produce several possible outcomes. Before we observe the system, we are uncertain about which outcome will occur. When the outcome is revealed, this uncertainty is reduced. The amount of information obtained from the observation should therefore be related to the degree of uncertainty that existed beforehand.

This immediately suggests that the information gained when an event occurs should depend on the probability of that event. If an event was almost certain, then learning that it occurred provides very little new information. However, if an event was extremely unlikely, then learning that it occurred provides a large amount of information. Rare events are therefore highly informative, while highly probable events carry little information. In other words, information is a measure of how surprised we are upon witnessing the occurrence of an event.

The task is therefore to construct a mathematical function that assigns a numerical quantity to the occurrence of an event of probability p. This quantity will be the

information content of the event. Let us denote by $\mathrm{I}(p_x)$ the amount of information obtained when we learn that the event x, whose probability of occurrence is p_x, has occurred. The goal is to determine the form of this function from basic principles.

The first requirement is that the information associated with an event should depend only on its probability. Two events with the same probability should carry the same amount of information when they occur. This reflects the fact that Shannon's notion of information is purely statistical: it depends only on the likelihood of events, not on their semantic meaning. In this way, the theory is independent of the physical properties of any system, being thus valid for all systems.

The second requirement is that rarer events should be more informative. If an event becomes less probable, the information associated with its occurrence should increase. In other words, the function $\mathrm{I}(p_x)$ must decrease as the probability p increases. A particularly important consequence of this principle concerns events that are certain to occur. If $p_x = 1$, the event was completely predictable and learning that it occurred provides no new information. Therefore, the information associated with a certain event must vanish. This is a fairly intuitive feature of any candidate for being a measure of information.

The third requirement concerns continuity. If the probability of an event changes slightly, the information associated with it should also change slightly. A reasonable measure of information should therefore be a continuous function of the probability.

The last requirement concerns independent events. Suppose that two independent events x and y occur, with probabilities p_x and q_y. The probability that both events occur is then $p_x q_y$. If we first learn that the first event occurred and then learn that the second event occurred, the total information gained should be the sum of the information obtained in the two events. However, learning directly that the joint event occurred should produce the same total information. This must be so since the events are independent, and, theretofore, learning one of them tells us nothing about the other. This property, called additivity, is, despite being natural, the most restrictive one. Thus, for independent events, the function I must satisfy the condition

$$\mathrm{I}(p_x q_y) = \mathrm{I}(p_x) + \mathrm{I}(q_y).$$

These simple requirements are sufficient to uniquely determine the form of the information measure as being

$$\mathrm{I}(p_x) = \ln \frac{1}{p_x} = -\ln p_x. \tag{1.1}$$

Proof. Let us first examine some immediate consequences of the additivity property. If we set $p_x = 1$, the relation becomes

$$\mathrm{I}(p_x q_y) = \mathrm{I}(1 q_y) = \mathrm{I}(1) + \mathrm{I}(q_y) = \mathrm{I}(q_y),$$

which implies that $\mathrm{I}(1) = 0$. This confirms that an event that was certain to occur does not contain information.

Next, consider the probability p_x^n, where n is a positive integer. So, we are considering n independent occurrences of the same event whose probability is p_x. Repeated application of the additivity property gives

$$\begin{aligned} \mathrm{I}(p_x^n) &= \mathrm{I}(p_x p_x^{n-1}) = \mathrm{I}(p_x) + \mathrm{I}(p_x^{n-1}) = \mathrm{I}(p_X) + \mathrm{I}(p_x p_x^{n-2}) = 2\mathrm{I}(p_x) + \mathrm{I}(p_x^{n-2}) \\ &= n\,\mathrm{I}(p_x). \end{aligned}$$

Thus, the information associated with n independent occurrences of an event is simply n times the information associated with a single occurrence.

Now, the structure of the additivity property strongly suggests that the function I should be related to the logarithm, which transforms multiplication into addition. This observation can be made precise by introducing a new variable $p_x = e^x$. By defining the function $F(x) = \mathrm{I}(e^x)$ we obtain

$$F(x+y) = \mathrm{I}(e^{x+y}) = \mathrm{I}(e^x e^y) = \mathrm{I}(e^x) + \mathrm{I}(e^y) = F(x) + F(y).$$

Thus, the function $F(x)$ satisfies the functional equation $F(x+y) = F(x) + F(y)$. This equation is known as Cauchy's functional equation. A fundamental result in analysis states that every continuous solution of this equation must be linear [2][1]. And we want a continuous function for the measure of information! Therefore, there exists a constant c such that $F(x) = cx$. Returning to the original variable p_x, we obtain

$$\mathrm{I}(p_X) = c \ln p_x.$$

However, the information associated with the rarer events must be larger, which means that the function must decrease as p_x increases. Since $\ln p_X$ increases with p_x, the constant c must be negative. Writing $c = -k$ with $k > 0$, we obtain

$$\mathrm{I}(p) = -k \ln p_x = k \ln \frac{1}{p_x}.$$

[1]The additivity condition already implies linearity in the rational numbers. Continuity extends this result from the rationals to all real numbers since $\mathbb{Q}$ is dense in $\mathbb{R}$, thus implying that for every real x there exists a sequence of rational numbers (r_n) such that $r_n \to x$.

Thus, the information associated with the occurrence of an event of probability p_x must be proportional to the logarithm of the inverse probability. □

The constant k merely sets the unit in which the information is measured. If logarithms are taken to the base 2, the unit of information is the *bit*. If we choose the natural logarithm, information is measured in *nats*. Since the choice of the base for the logarithm changes only a multiplicative constant, I can be defined in terms of any basis. We chose here the base 2. In that case, the information content of an event is

$$\mathrm{I}(p_x) = -\log p_x. \tag{1.2}$$

This captures in a precise mathematical form the intuitive relation between information and surprise. $\mathrm{I}(p_x)$ can be understood in two different ways. First, we can think of it as a measure of the information we obtain when we learn about the event. On the other hand, we can also think about it as the uncertainty we have about the event before its occurrence. It is clear that both views are equivalent.

As a final remark, let us reflect on the meaning of the information bit. Consider the simple example of tossing a fair coin. Before the outcome is observed, the uncertainty associated with the experiment is one bit. Once the outcome is revealed, this uncertainty is completely removed, and one bit of information is gained. The coin itself is the physical system that carries or encodes the outcome, whereas the bit introduced by Shannon is an abstract logical quantity that measures the information conveyed by that outcome. Shannon's theory deliberately separates the physical realization of information from its mathematical description, allowing the same logical bit to be encoded in many different physical systems, such as coins, electrical circuits, magnetic media, or quantum states.

1.1 Entropy

An essential concept in Shannon's theory is that of an information source. An information source is any physical process capable of producing a sequence of outcomes that can be observed and recorded. These outcomes do not need to be deterministic. In fact, Shannon's theory is primarily concerned with situations in which there is uncertainty about which outcome will occur. Because of this uncertainty, the behaviour of an information source is naturally described using the language of probability theory.

Mathematically, an information source is represented by a random variable. This does not mean that the physical system itself is random, but rather that the outcome of the experiment or observation is modelled as a random variable taking values

in a set of possible outcomes, called the alphabet of the source. Each element of the alphabet is assigned a probability that reflects our knowledge, or lack thereof, about the occurrence of that outcome. The random variable therefore provides a mathematical description of the information produced by the source, independently of the physical mechanism responsible for generating it.

A simple example is the toss of a fair coin. The physical source is the coin-tossing experiment, while the associated random variable X takes values in the alphabet $\mathcal{X} = \{\text{Heads}, \text{Tails}\}$ with probabilities $p_X(\text{Heads}) = p_X(\text{Tails}) = 1/2$. Before the outcome is observed, there is uncertainty regarding which value of X will occur. Once the outcome is revealed, this uncertainty is eliminated and one bit of information is gained.

As another example, consider rolling a fair six-sided die. The physical source is the rolling experiment, and the corresponding random variable takes values in the alphabet $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, with each outcome occurring with probability $p_X(x) = 1/6$ for all x. More generally, a weather station may report whether the next day will be sunny, cloudy, or rainy, a communication channel that transmits the letters of the English alphabet, or a sensor that measures the temperature of a physical system. In each case, the information source is the physical process that generates the observations, while the random variable provides a probabilistic description of its possible outcomes.

This distinction between the physical source and its mathematical representation is one of the fundamental ideas underlying Shannon's theory. The same logical information can be encoded in many different physical systems, and conversely, very different physical processes may give rise to the same probability measure[2]. Shannon's theory is therefore concerned not with the physical nature of the source but with the statistical properties of the symbols it produces. This abstraction is precisely what gives the theory its remarkable generality and allows it to be applied to such diverse fields as communication, computation, statistics, biology, and physics.

Therefore, we assume that any information source is described by the random variable $X = \{\mathcal{X}, p_X(x)\}$, with the outcome $x \in \mathcal{X}$ occurring with probability $p_X(x)$. When the outcome of a single event is revealed, the information obtained is $\mathrm{I}(x) = -\log p_x$. Since outcome x occurs with probability $p_X(x)$, the average information produced by the source is obtained by taking the expectation value of $\mathrm{I}[p_X(x)]$. This average information is called the Shannon entropy of the random

[2]We use probability measure to mean both the continuous or the discrete cases. In the continuous case we have a probability distribution density, while a probability mass function applies to the discrete case.

variable and is defined as

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x). \tag{1.3}$$

Entropy therefore represents the average amount of information obtained when the value of the random variable is revealed. Equivalently, it measures the average uncertainty associated with the variable before the observation is made. This is a fundamental quantity in Shannon theory, where it acquires operational meaning as the maximum rate of data compression.

Before stating some of its properties, let us see some examples. The entropy associated with the flipping of a fair coin is given by

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1,$$

i.e., a fair coin has one bit of information. The information associated with the coin can answer one binary question. A fair die has entropy $H(X) = \log 6 \approx 2.58$. This means that it takes an average of 2.585 yes/no questions to determine the exact face of the die in a single roll. Since we cannot ask a fractional number of questions, we will need 2 or 3. For example, we can ask the following questions: Question 1: "Is it 1, 2, or 3?", Question 2: "Is it 1?", and Question 3: "Is it 2?".

Now that the intuitive meaning of the entropy is clear, let us see some properties of the Shannon entropy (1.3). The most important ones can be stated as follows.

Theorem 1.1 (Non-negativity)**.** *For every random variable, $H(X) \geq 0$. Moreover, $H(X) = 0$ if and only if one outcome occurs with probability one.*

Proof. Since $0 \leq p_X(x) \leq 1$, we have $\ln p_X(x) \leq 0$. Therefore, $-p_X(x)\ln p_X(x) \geq 0$ for every x. Summing over all outcomes gives $H(X) =\geq 0$. The equality holds only if each term vanishes. Since $-p_X(x)\ln p_X(x) = 0$ only when $p_X(x) = 0$ or $p_X(x) = 1$, the entropy vanishes precisely when one outcome has probability one and all the others have probability zero. □

Theorem 1.2 (Maximum Entropy)**.** *Let X have $|\mathcal{X}|$ possible outcomes. Then $H(X) \leq \log N$, with equality if and only if $p_X(x) = 1/|\mathcal{X}|$ for all x.*

Proof. We seek the probability measure that maximises the entropy $H(X)$ subject to the normalisation condition $\sum_{x=1}^{|\mathcal{X}|} p_X(x) = 1$. Introducing a Lagrange multiplier α, we define the auxiliary function

$$\Phi = -\sum_{x=1}^{|\mathcal{X}|} p_X(x)\ln p_X(x) - \alpha\left(\sum_{x=1}^{|\mathcal{X}|} p_X(x) - 1\right).$$

The extremum condition is

$$\frac{\partial \Phi}{\partial p_X(x)} = 0,$$

which gives $\ln p_i = -1 - \alpha$. Since the right-hand side is independent of x, all probabilities must be equal:

$$p_X(x) = e^{-1-\alpha}.$$

The normalisation condition then implies $\sum_{x=1}^{|\mathcal{X}|} p_X(x) = \mathcal{X} e^{-1-\alpha} = 1$. Therefore, $e^{-1-\alpha} = p_X(x) = 1/|\mathcal{X}|$ for all x. Substituting into the entropy gives the upper bound stated in the theorem.

To verify that this stationary point is a maximum, note that

$$\frac{\partial^2 H}{\partial p_X^2(x)} = -\frac{1}{p_X(x)},$$

which is strictly negative for every $p_X(x) > 0$. Hence, the entropy is a concave function of the probabilities, and the stationary point found above is necessarily the unique global maximum.

The concavity of the entropy is one of its most important mathematical properties. It expresses the idea that uncertainty increases when probability measures are mixed. In particular, if one is uncertain about which probability measure correctly describes a system, the resulting entropy is greater than or equal to the average entropy of the individual distributions. This can be expressed as follows. Let $p_X = \{p_X(x)\}$ and $q_X = \{q_X(x)\}$ be two probability measures, and let $0 \leq \lambda \leq 1$. Then

$$H(\lambda p_X + (1-\lambda) q_X) \geq \lambda H(p_X) + (1-\lambda) H(q_X).$$

□

Consider two sources that are described by the random variables X and Y, associated with the alphabets $\mathcal{X}$ and $\mathcal{Y}$, and whose results $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ occur with probabilities $p_X(x)$ and $p_X(y)$, respectively. The joint probability measure of the occurrence of both outcomes x and y is written as $p_{X,Y}(x,y)$. From this we can naturally extend the entropy function (1.3) by defining the joint entropy

$$H(X,Y) = -\sum_{x,y} p_{X,Y}(x,y) \log p_{X,Y}(x,y). \tag{1.4}$$

This is the joint information of both sources. An important property of this quantity is given by the following theorem.

Theorem 1.3 (Additivity for Independent Variables)**.** *If X and Y are independent random variables, then $H(X,Y) = H(X) + H(Y)$.*

Proof. Independence implies $p_{X,Y}(x,y)(x,y) = p_X(x)p_Y(y)$. Therefore

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p_{X,Y}(x,y)\log p_{X,Y}(x,y) = -\sum_{x,y} p_X(x)p_Y(y)\log\big[p_X(x)p_Y(y)\big] \\ &= -\sum_{x,y} p_X(x)p_Y(y)\log p_X(x) - \sum_{x,y} p_X(x)p_Y(y)\log p_Y(y) \\ &= -\sum_{x} p_X(x)\log p_X(x) - \sum_{y} p_Y(y)\log p_Y(y) \\ &= H(X) + H(Y), \end{aligned}$$

where we used the fact that the probabilities are normalised. □

A different question we can ask is how much we learn about X when Y is observed. The conditional entropy quantifies precisely this. If the value of Y is known, our uncertainty about X is no longer characterised by the probability measure $p_X(x)$, but rather by the conditional distribution

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

For each possible value of Y, the corresponding uncertainty about X is therefore measured by the entropy of the conditional distribution,

$$H(X|Y=y) = -\sum_{x} p_{X|Y}(x|y)\log p_{X|Y}(x|y).$$

Since the value of Y itself is random, the overall uncertainty about X after observing Y is obtained by averaging over all possible values of Y. This leads to the definition of the conditional entropy,

$$H(X|Y) = -\sum_{x,y} p_{X,Y}(x,y)\log p_{X|Y}(x|y). \tag{1.5}$$

Conditional entropy measures the average uncertainty that remains about X once the value of Y is known. If Y completely determines X, then there is no remaining uncertainty and $H(X|Y) = 0$. On the other hand, if X and Y are statistically independent, knowledge of Y does not provide information about X, and the conditional entropy reduces to the ordinary entropy, $H(X|Y) = H(X)$. Thus, conditional entropy provides a quantitative measure of how much uncertainty survives after part of the available information has been revealed. The relation between the joint and the conditional entropies is given in the following theorem.

Theorem 1.4 (Chain Rule). *For any pair of random variables X and Y,*

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

Proof. Using $p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x)$, we obtain $\log p_{X,Y}(x,y) = \log p_X(x) + \log p_{Y|X}(y|x)$. Substituting into the definition of the joint entropy yields

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p_{X,Y}(x,y) \log p_X(x) - \sum_{x,y} p_{X,Y}(x,y) \log p_{Y|X}(y,x) \\ &= -\sum_{x} \left[\sum_{y} p_{X,Y}(x,y) \right] \log p_X(x) - \sum_{x,y} p_{X,Y}(x,y) \log p_{Y|X}(y,x) \\ &= H(X) + H(Y|X), \end{aligned}$$

where we used the marginal probability measure $p_X(x) = \sum_y p_{X,Y}(x,y)$. □

Our knowledge of X changes when we learn about Y only when the variables are not independent. Intuitively, they must be correlated, thus raising the question of how much correlated they are. Mutual information quantifies the amount of information that two random variables share. It measures how much the uncertainty about one variable is reduced when the value of the other variable is known. If the variables are completely independent, knowledge of one provides no information about the other, and mutual information should vanish. On the other hand, if one variable completely determines the other, then learning the value of one removes all uncertainty about the second, and the mutual information is maximal.

This intuition can be naturally expressed in terms of entropy. Suppose that we are interested in a random variable X. Before any observation is made, our uncertainty about X is measured by its entropy $H(X)$. If we are subsequently informed of the value of a second random variable Y, the remaining uncertainty about X is measured by conditional entropy $H(X|Y)$. The difference between these two quantities therefore measures the average reduction in uncertainty produced by knowledge of Y. This motivates the definition of the mutual information as

$$I(X:Y) = H(X) - H(X|Y). \tag{1.6}$$

The mutual information is symmetric with respect to the exchange of X and Y, as it should. To see this, recall the chain rule for entropy,

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

Using these relations, we may rewrite the mutual information as

$$I(X:Y) = H(X) + H(Y) - H(X,Y). \tag{1.7}$$

This expression makes the symmetry between X and Y explicit and is often taken as the fundamental definition of mutual information. The above equation admits a simple geometric interpretation. The sum $H(X)+H(Y)$ represents the total uncertainty associated with the two variables if they are considered separately. However, when the variables are correlated, part of this uncertainty is counted twice, since some information is shared by both variables. The joint entropy $H(X,Y)$ measures the total uncertainty of the pair without double counting. The difference on the right side of Eq. (1.7) therefore quantifies precisely the information that is common to both variables. For this reason, mutual information is interpreted as a measure of correlation. A diagrammatic representation of the information quantities introduced here is shown in Fig. 1.1.

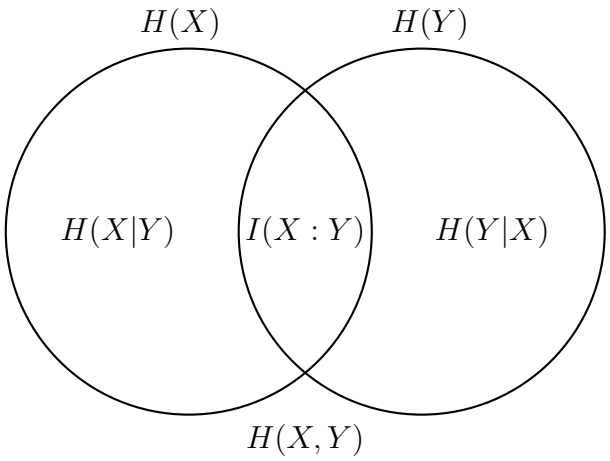


Figure 1.1: Venn diagram representation of the relation between entropy, conditional entropy, and mutual information. The left circle represents $H(X)$ and the right circle represents $H(Y)$, which quantify the uncertainties associated with the random variables X and Y, respectively. Their overlap corresponds to the mutual information $I(X:Y)$, which measures the information shared by the two variables and therefore quantifies the correlations between them. The regions outside the overlap correspond to the conditional entropies $H(X|Y)$ and $H(Y|X)$. These quantify the uncertainty that remains about one variable after the value of the other has been specified, that is, after the information contained in their correlations has been taken into account. From this diagram, we can see that $H(X,Y) = H(X) + H(Y) - I(X:Y)$ quantifies the total uncertainty in both variables.

Several important special cases follow immediately from the definition. If X

and Y are statistically independent, then $H(X|Y) = H(X)$, and therefore $I(X : Y) = 0$. If the value of Y completely determines the value of X, then $H(X|Y) = 0$, and the mutual information reduces to $I(X : Y) = H(X)$. In this case, all the uncertainty about X can be removed by observing Y.

Mutual information thus provides a quantitative measure of statistical dependence. Unlike ordinary correlation coefficients, it detects arbitrary forms of dependence between random variables and plays a central role in information theory, statistics, machine learning, and communication theory.

Mutual information can also be written directly in terms of joint and marginal probability measures. Let $p_{X,Y}(x, y)$ be the joint distribution of X and Y, and let $p_X(x) = \sum_y p_{X,Y}(x, y)$ and $p_Y(y) = \sum_x p_{X,Y}(x, y)$ be the corresponding marginal distributions. Starting from Eq. (1.7) and using the explicit form of the entropy given in Eq. (1.3) and the joint entropy shown in Eq. (1.4) we obtain

$$\begin{aligned} I(X : Y) &= -\sum_x p_X(x) \log p_X(x) - \sum_y p_Y(y) \log p_Y(y) \\ &+ \sum_{x,y} p_{X,Y}(x, y) \log p_{X,Y}(x, y). \end{aligned}$$

Now, we can write $-\sum_x p_X(x) \log p_X(x) = -\sum_{x,y} p_{X,Y}(x, y) \log p_X(x)$ and similarly for the sum over y. Hence, mutual information can be written as

$$I(X : Y) = \sum_{x,y} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)}.$$

This form makes explicit that mutual information compares the true joint distribution $p_{X,Y}(x, y)$ with the product distribution $p_X(x)p_Y(y)$. If X and Y are independent, then $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ and $I(X : Y) = 0$. Thus, mutual information measures how far the joint distribution is from the distribution that would be obtained if the variables were statistically independent.

A very intuitive and important property of the conditional entropy can be derived from its relation to the mutual information, and it is expressed in the following theorem.

Theorem 1.5 (Conditioning Reduces Entropy)**.** *For random variables X and Y, $H(X|Y) \leq H(X)$, with equality holding if and only if X and Y are independent.*

Proof. For each value y of Y, let $p_{X|Y}(x|y)$ denote the conditional probability measure of X given $Y = y$. Then $H(X|Y) = \sum_y p_Y(y)H(X|Y = y)$. The marginal distribution of X can be obtained by averaging the conditional distributions: $p_X(x) =$

$\sum_y p_Y(y)p_{X|Y}(x|y)$. Since Shannon entropy is concave as a function of the probability measure, we have

$$H\left(\sum_y p_Y(y)p_{X|Y}(x|y)\right) \geq \sum_y p_Y(y)H\left(p_{X|Y}(x|y)\right).$$

The left side is precisely $H(X)$, while the right side is $H(X|Y)$. Therefore, $H(X) \geq H(X|Y)$.

The equality holds if and only if all conditional distributions $p_{X|Y}(x,y)$ are the same for every y such that $p_Y(y) > 0$. Thus, $p_{X|Y}(x|y) = p_X(x)$ for all such y. This is exactly the condition that X and Y are independent, since it implies $p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x|y) = p_Y(y)p_X(x)$. Conversely, if X and Y are independent, then $p_{X|Y}(x,y) = p_X(x)$ and therefore $H(X|Y) = H(X)$. □

1.2 Relative Entropy

So far, we have focused on quantities that characterise uncertainty and correlations associated with probability measures. In many situations, however, the relevant problem is not quantifying this, but rather comparing different probability measures for the same system. For example, one may wish to compare a theoretical model with experimental data, determine how distinguishable two statistical descriptions are, or quantify the cost of using an incorrect probabilistic model.

The fundamental quantity used for this purpose is the relative entropy, also known as the Kullback-Leibler divergence. Given two probability measures $p_X = \{p_X(x)\}$ and $q_X = \{q_X(x)\}$ defined in the same sample space, the relative entropy of p with respect to q is defined by

$$D(p_X\|q_X) = \sum_x p_X(x)\log\frac{p_X(x)}{q_X(x)}. \tag{1.8}$$

Relative entropy measures the distinguishability between two probability measures. It is important to note that it is not a distance in the mathematical sense. In general, the relative entropy is not symmetric under exchange of its arguments. Moreover, it does not satisfy the triangle inequality. Nevertheless, it plays a central role in information theory, statistics, machine learning, and statistical physics.

The asymmetry of the relative entropy reflects the different roles played by the distributions p_X and q_X. The distribution p_X determines the weighting of the average, while q_X serves as a reference distribution. For this reason, relative entropy

should not be interpreted as a geometric distance, but rather as a measure of distinguishability between probabilistic descriptions.

A useful interpretation follows from rewriting the definition as

$$\begin{aligned} D(p_X\|q_X) &= -\sum_x p_X(x)\log q_X(x) + \sum_x p_X(x)\log p_X(x) \\ &= -\sum_x p_X(x)\log q_X(x) - H(p). \end{aligned}$$

Suppose that an information source is characterised by the probability measure q_X. Thus, the content of information in the appearance of x is $-\log p_X(x)$. Now, let us assume that we mistakenly think that the source is characterised by the probability measure p_X. The first term is the above equation is the information we associate to the source. The second term corresponds to the real information of the source. The relative entropy therefore quantifies the cost of confusing the probability measures.

We now establish some of the most important properties of the relative entropy, its positivity.

Theorem 1.6 (Gibbs' Inequality)**.** *For any two probability measures p_X and q_X,*

$$D(p_X\|q_X) \geq 0. \tag{1.9}$$

Furthermore, $D(p_X\|q_X) = 0$ if and only if $p_X(x) = q_X(x)$ for all outcomes x.

Proof. We begin with the elementary inequality $\log t \leq t - 1$, which holds for all positive numbers t, with equality if and only if $t = 1$. Let $t = q_X(x)/p_X(x)$. Then

$$\log\frac{q_X(x)}{p_X(x)} \leq \frac{q_X(x)}{p_X(x)} - 1 \quad \Rightarrow \quad p_X(x)\log\frac{p_X(x)}{q_X(x)} \geq p_X(x) - q_X(x).$$

Summing over x, we obtain

$$D(p_X\|q_X) = \sum_x p_X(x)\log\frac{p_X(x)}{q_X(x)} \geq \sum_x (p_x - q_x) = 0.$$

The last equality follows from the normalisation of the probability measures.

It is straightforward to see that the equality holds if and only if $p_X(x) = q_X(x)$ for all x. □

The non-negativity of relative entropy is one of the most important inequalities in information theory. Several results in the theory can be proved using this result. An example is the positivity of mutual information.

One of the deepest interpretations of relative entropy comes from statistical hypothesis testing. Suppose that a sequence of independent observations is generated

according to one of two competing hypotheses. Under the first hypothesis, the observations are distributed according to p_X, whereas under the second hypothesis they are distributed according to q_X. The goal is to determine, from the observed data, which hypothesis is correct.

If only a single observation is available, it can be difficult to distinguish the two hypotheses. However, as the number of observations increases, the statistical evidence accumulates. Intuitively, the more different the distributions are, the easier it should be to identify the correct hypothesis.

This intuition is made precise by Stein's lemma [3, 4].

Theorem 1.7 (Stein's Lemma)**.** *Let p_X and q_X be two probability measures. In the limit of a large number n of independent observations, the optimal error probability for mistaking p_X for q_X decreases exponentially as*

$$P_{\text{error}} = e^{-nD(p_X \| q_X)}.$$

The relative entropy therefore determines the asymptotic rate at which the two hypotheses become distinguishable. A rigorous proof of this Lemma requires the theory of typical sequences and large deviations and will not be presented at this point. However, the result can be understood heuristically from the behaviour of the likelihood ratio. Suppose that we observe a sequence $x_1, x_2, \ldots, x_n$ of independent samples. Under the hypothesis that the data are generated according to p_X, the probability of this sequence is

$$P_p(x_1, \ldots, x_n) = \prod_{k=1}^{n} p_X(x_k).$$

Under the hypothesis that the data are generated according to q_X, the probability of the same sequence is

$$P_q(x_1, \ldots, x_n) = \prod_{k=1}^{n} q_X(x_k).$$

To compare the two hypotheses, one considers the likelihood ratio

$$\frac{P_p(x_1, \ldots, x_n)}{P_q(x_1, \ldots, x_n)} = \prod_{k=1}^{n} \frac{p_X(x_k)}{q_X(x_k)}.$$

Taking the logarithm gives

$$\log \frac{P_p(x_1, \ldots, x_n)}{P_q(x_1, \ldots, x_n)} = \sum_{k=1}^{n} \log \frac{p_X(x_k)}{q_X(x_k)}.$$

Now assume that the sequence was actually generated according to p_X. Let N_x be the number of times the symbol x appears in the sequence. Then

$$\sum_{k=1}^{n} \log \frac{p_X(x_k)}{q_X(x_k)} = \sum_x N_x \log \frac{p_X(x)}{q_X(x)}.$$

This implies, for large n, that

$$\frac{1}{n} \sum_{k=1}^{n} \log \frac{p_X(x_k)}{q_X(x_k)} \to \sum_x p_X(x) \log \frac{p_X(x)}{q_X(x)},$$

since the empirical frequency of each outcome x is approximately $p_X(x)$, $N_x/n \to p_X(x)$ when n is very large. The right-hand side of this equation is precisely the relative entropy $D(p_X \| q_X)$. Hence,

$$\frac{P_p(x_1, \ldots, x_n)}{P_q(x_1, \ldots, x_n)} \approx e^{nD(p_X \| q_X)}.$$

This means that a typical sequence generated by p_X is exponentially more likely under hypothesis p_X than under hypothesis q_X, with exponential rate $D(p_X \| q_X)$. Thus, if the true distribution is p_X, the probability that the data look compatible with q_X is exponentially suppressed. The exponent controlling this suppression is the relative entropy. This is the heuristic content of Stein's lemma.

The relative entropy therefore determines the asymptotic rate at which the error probability decreases. If $D(p_X \| q_X)$ is large, only a modest amount of data is required to distinguish the two hypotheses. If $D(p_X \| q_X)$ is small, a much larger number of observations is necessary. Relative entropy therefore quantifies the statistical distinguishability of competing probabilistic models.

The connection between relative entropy and mutual information is particularly important. Starting from the expression of mutual information in terms of the probabilities, we immediately recognise the structure of a relative entropy:

$$I(X : Y) = D\left(p_{X,Y}(x, y) \middle\| p_X(x) p_Y(y)\right).$$

Mutual information may therefore be interpreted as the relative entropy between the true joint probability measure and the product of its marginals. Since the product corresponds to statistical independence, mutual information measures how distinguishable the actual correlations are from the absence of correlations.

As a direct consequence of Gibbs' inequality is the positivity of the mutual information

$$I(X : Y) \geq 0,$$

with equality if and only if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$.

The last property we discuss in this section is the convexity of the relative entropy.

Theorem 1.8 (Convexity of Relative Entropy). *Let p_X^1, p_X^2, q_X^1, and q_X^2 be probability measures, and let $0 \leq \lambda \leq 1$. Define*

$$p_X = \lambda p_X^1 + (1-\lambda)p_X^2 \quad \textit{and} \quad q_X = \lambda q_X^1 + (1-\lambda)q_X^2.$$

Then

$$D(p_X \| q_X) \leq \lambda D(p_X^1 \| q_X^1) + (1-\lambda) D(p_X^2 \| q_X^2).$$

Proof. For each outcome x, apply the log-sum inequality (see Appendix **??**) to the two pairs $[a_1 = \lambda p_X^1 \,,\, a_2 = (1-\lambda)p_X^2]$, and $[b_1 = \lambda q_X^1 \,,\, b_2 = (1-\lambda)q_X^2]$. The log-sum inequality gives

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2}.$$

Substituting the above definitions, we obtain

$$\lambda p_X^1 \log \frac{p_X^1}{q_X^1} + (1-\lambda) p_X^2 \log \frac{p_X^2}{q_X^2} \geq p_X \log \frac{p_X}{q_X}.$$

Summing over x results

$$\lambda \sum_x p_X^1 \log \frac{p_X^1}{q_X^1} + (1-\lambda) \sum_x p_X^2 \log \frac{p_X^2}{q_X^2} \geq \sum_x p_X \log \frac{p_X}{q_X}.$$

Therefore,

$$\lambda D(p_X^1 \| q_X^1) + (1-\lambda) D(p_X^2 \| q_X^2) \geq D(p_X \| q_X),$$

which proves the joint convexity of the relative entropy. □

1.3 The Data Processing Inequality

One of the most important principles in information theory is that information cannot be increased by processing data. Intuitively, if two probability measures become more difficult to distinguish after a transformation, no subsequent analysis of the transformed data can recover the lost distinguishability. This idea is formalised by the data processing inequality.

Before we introduce this important inequality, we need to define the class of transformations to which it applies. Let $T(y|x)$ be a Markov transformation, that

is, a conditional probability measure that satisfies $T(y|x) \geq 0$ and $\sum_y T(y|x) = 1$. $T(y|x)$ is the probability that the system evolves from the state x to the state y in a single step. Given an input distribution p_X, the transformed distribution is

$$p'_Y(y) = \sum_x T(y|x)p_X(x).$$

This is also called a stochastic map.

Theorem 1.9 (Data Processing Inequality)**.** *Let p_X and q_X be two probability measures. Let $T(y|x)$ be a stochastic map. Define the transformed distributions $p'_Y(y) = \sum_x T(y|x)p_X(x)$ and $q'_Y(y) = \sum_x T(y|x)q_X(x)$. Then*

$$D(p'_Y \| q'_Y) \leq D(p_X \| q_X).$$

Proof. Starting from the definition of relative entropy and using the definitions of p'_Y and q'_Y, we have

$$D(p'_Y \| q'_Y) = \sum_y \left(\sum_x T(y|x)p_X(x) \right) \log \frac{\sum_x T(y|x)p_X(x)}{\sum_x T(y|x)q_X(x)}.$$

Applying the log-sum inequality for each value of y gives

$$\left(\sum_x T(y|x)p_X(x) \right) \log \frac{\sum_x T(y|x)p_X(x)}{\sum_x T(y|x)q_X(x)} \leq \sum_x T(y|x)p_X(x) \log \frac{p_X(x)}{q_X(x)}.$$

Summing over y yields

$$D(p'_Y \| q'_Y) \leq \sum_{x,y} T(y|x)p_X(x) \log \frac{p_X(x)}{q_X(x)}.$$

Since $\sum_y T(y|x) = 1$, we obtain

$$D(p'_Y \| q'_Y) \leq \sum_x p_X(x) \log \frac{p_X(x)}{q_X(x)} = D(p_X \| q_X).$$

□

A particularly important application arises when three random variables form a Markov chain $X \to Y \to Z$. A Markov chain is a mathematical model for describing the evolution of a system that moves through a sequence of possible states in discrete steps. Its defining characteristic is the memoryless property: the probability of the next state depends only on the current state of the system and not

on the sequence of states that preceded it. In other words, the present state contains all the information needed to predict the future evolution of the system. This makes Markov chains a fundamental tool for modelling stochastic processes in a wide range of fields, including physics, information theory, biology, and finance, where the dynamics can be approximated as a succession of random transitions between states.

In this situation, all information that Z possesses about X must pass through Y. The data processing inequality therefore implies

$$I(X : Z) \leq I(X : Y).$$

Thus, every processing step can only decrease, or at best preserve, the information available about the original source.

1.4 Fano's Inequality

The quantities introduced so far quantify uncertainty and information. A natural question is how these quantities relate to the probability of making mistakes when attempting to infer an unknown variable from observed data.

Suppose that X is a random variable taking values in an alphabet of size $|\mathcal{X}|$, and let $\hat{X}$ denote an estimate of X obtained from another observation. Define the probability of error

$$p_e = \Pr\Big(X \neq \hat{X}\Big).$$

Fano's inequality provides an upper bound on the conditional entropy in terms of the error probability [**<empty citation>**].

Theorem 1.10 (Fano's Inequality)**.** *Let X be a random variable with $|\mathcal{X}|$ possible outcomes and let $\hat{X}$ be an estimator of X. Then*

$$H(X|\hat{X}) \leq H_b(p_e) + p_e \log(|\mathcal{X}| - 1),$$

where

$$H_b(p) = -p \log p - (1 - p) \log(1 - p)$$

is the binary entropy function.

Proof. Introduce the error indicator

$$E = \begin{cases} 0, & X = \hat{X}, \\ 1, & X \neq \hat{X}. \end{cases}$$

By definition, $\Pr(E = 1) = p_e$. Using the chain rule for the conditional entropy, we obtain

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}).$$

Since E is completely determined by X and $\hat{X}$, $H(E|X, \hat{X}) = 0$. Therefore, it follows that $H(X|\hat{X}) = H(E, X|\hat{X})$.

Applying the chain rule in the opposite order gives

$$H(X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}).$$

Now, since conditioning cannot increase entropy,

$$H(E|\hat{X}) \leq H(E) = H_b(p_e).$$

We now bound the second term. If $E = 0$, then $X = \hat{X}$, so that $H(X|E = 0, \hat{X}) = 0$. If $E = 1$, then X can take at most $|\mathcal{X}| - 1$ values different from $\hat{X}$. Hence $H(X|E = 1, \hat{X}) \leq \log(|\mathcal{X}| - 1)$. Combining these observations,

$$H(X|E, \hat{X}) \leq p_e \log(|\mathcal{X}| - 1).$$

Therefore,

$$H(X|\hat{X}) \leq H_b(p_e) + p_e \log(|\mathcal{X}| - 1).$$

which proves Fano's inequality. □

Fano's inequality establishes a fundamental connection between information and inference. If the conditional entropy $H(X|\hat{X})$ is small, then the probability of error must also be small. In contrast, a high probability of error implies a large residual uncertainty about the value of X after the observation has been made.

Combining Fano's inequality with the identity

$$I(X : \hat{X}) = H(X) - H(X|\hat{X}),$$

one obtains lower bounds on the probability of error in terms of mutual information. Such bounds play a central role in communication theory, statistical learning, and information-theoretic limits on inference.

Chapter 2

Information theory and dynamical systems

As we saw in the previous chapter, Information Theory is concerned with the mathematical description of information sources. An information source is characterized by an alphabet, that is, a finite or countable set of possible symbols, together with probability laws that specify the likelihood of individual symbols and finite sequences of symbols. The central object of the theory is therefore not the physical mechanism responsible for generating the messages, but rather the statistical structure of the emitted symbolic sequences. This abstraction makes Shannon's theory remarkably general, allowing it to be applied far beyond the context of communication systems for which it was originally developed.

Among several interesting questions, one of particular interest in the context of classical mechanics emerges from this observation: can a deterministic physical system be regarded as an information source? At first sight, the answer appears to be negative. A dynamical system evolves continuously in time according to deterministic equations of motion, whereas an information source emits discrete symbols according to a probability distribution. The two descriptions seem to belong to entirely different mathematical frameworks. Nevertheless, by introducing an appropriate symbolic representation of the dynamics, it becomes possible to establish a precise correspondence between them. Once this correspondence is constructed, the set of tools of Information Theory can be employed to characterize several features of deterministic dynamical systems, providing both, new predictions and new lenses to consider old problems.

The objective of this chapter is to develop this correspondence. We begin by introducing the notion of a dynamical system and the probability measures in the state space. We then show how a partition of the state space naturally induces a symbolic representation of the dynamics, transforming trajectories into sequences of symbols

that can be interpreted as messages emitted by an information source. This symbolic description allows us to assign probabilities to finite trajectories and to quantify their uncertainty using Shannon entropy. Based on these ideas, we introduce the refinement of partitions and the Kolmogorov-Sinai entropy, which measures the asymptotic rate at which information is generated by the dynamics. Finally, we discuss the physical significance of the Kolmogorov-Sinai entropy, emphasizing its interpretation as a measure of dynamical complexity and its connection with nonequilibrium thermodynamics and entropy production.

2.1 Dynamical systems

A dynamical system describes how the state of a physical system evolves in time. It is specified by two fundamental ingredients: a state space containing all possible configurations of the system and a dynamical law that determines how these configurations evolve. We denote the state space by Ω and point $s \in \Omega$ specifies the complete state of the system at a given instant.

The precise nature of the state space depends on the physical problem under consideration. In classical Hamiltonian mechanics, for example, Ω is naturally identified with the phase space Γ, whose points are pairs $s = (q, p)$ consisting of generalized coordinates and their conjugate momenta. In other contexts, the state space may consist of functions, symbolic sequences, probability distributions, or more abstract mathematical objects. The construction developed throughout this chapter is independent of the particular realization of Ω. The only assumption is that the dynamics acts on its elements in a well-defined manner.

The evolution of the system is described by a deterministic transformation

$$T : \Omega \longrightarrow \Omega,$$

which assigns to every state its state after one unit of time. Determinism means that every initial condition possesses a unique image under the transformation. Consequently, once the initial state is specified, the future evolution of the system is completely determined.

Repeated applications of the transformation generate the trajectory, or orbit, associated with an initial condition $s_0 \in \Omega$,

$$\gamma(s_0) = \left\{ s_0, s_1 = T(s_0), s_2 = T^2(s_0), \ldots \right\},$$

where

$$T^n = \underbrace{T \circ T \circ \cdots \circ T}_{n \text{ times}}$$

denotes the n-fold composition of the transformation. The orbit therefore represents the complete history of the system generated by the initial condition s_0. Figure 2.1 provides an illustration of this orbit.

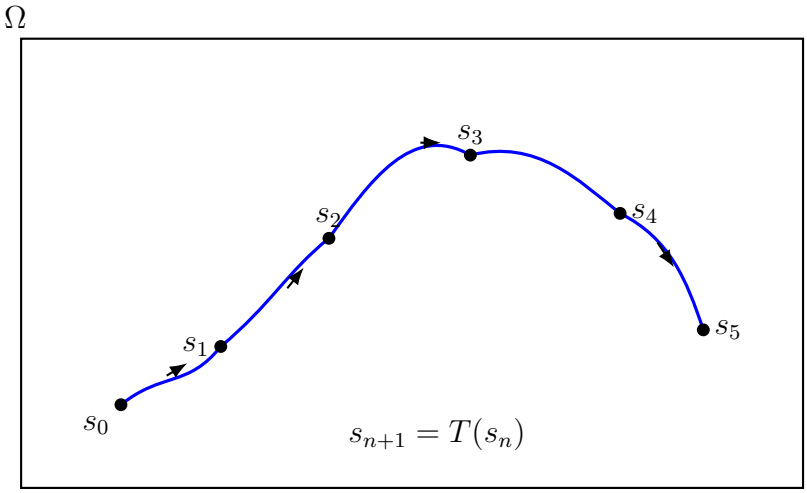


Figure 2.1: A deterministic dynamical system generates an orbit by successive applications of the transformation T. Starting from the initial condition $s_0 \in \Omega$, each iterate is uniquely determined by the previous one according to $s_{n+1} = T(s_n)$. The resulting orbit represents the complete history of the system generated from the initial state.

Throughout this chapter we formulate the theory in discrete time, so that the evolution is indexed by the integer $n \in \mathbb{N}$. This assumption considerably simplifies the notation while preserving all the essential ideas underlying symbolic dynamics and Kolmogorov-Sinai entropy. Moreover, many systems of physical interest, including iterated maps, cellular automata and periodically driven systems, are naturally described in discrete time.

However, continuous-time dynamical systems fit naturally into the same framework. If Φ_t denotes the continuous flow generated by the dynamics, choosing a fixed sampling interval $\tau > 0$ defines the discrete transformation $T = \Phi_\tau$, whose iterates reproduce the continuous evolution at the discrete times $\{0, \tau, 2\tau, \ldots\}$. Consequently, the construction developed here applies equally well to continuous- and discrete-time dynamical systems.

The transformation T need not be invertible. Non-invertible maps arise naturally in many physical situations, particularly in dissipative systems, where distinct initial conditions may evolve to the same final state. When the transformation is invertible, however, one may also define the inverse map T^{-1} and reconstruct the

past evolution of the system. Although invertibility is not essential for the concepts introduced in this section, inverse images will play a fundamental role in the symbolic description of the dynamics developed later in this chapter.

Now, we need to introduce the notion of probability, which provides the bridge between deterministic dynamics and information theory.

2.2 Probability measures

The deterministic dynamics introduced in the previous section completely specifies the evolution of every individual trajectory. Once the initial condition s_0 is known, the entire orbit of the system is uniquely determined by successive applications of the transformation T. In many physical situations, however, one is interested not only in following a single trajectory but also in describing the statistical behaviour of an ensemble of systems prepared under identical macroscopic conditions. Such a statistical description requires assigning probabilities to the possible initial conditions.

To formulate this probabilistic description, the state space Ω is endowed with a σ-algebra[1] $\mathcal{B}$ of subsets of Ω. The elements of $\mathcal{B}$ are called *measurable sets* and represent precisely those subsets of the state space to which probabilities may be assigned. The pair $(\Omega, \mathcal{B})$ is called a *measurable space*.

A probability measure is a mapping

$$p : \mathcal{B} \longrightarrow [0, 1],$$

satisfying the normalization condition $p(\Omega) = 1$, along with countable additivity

$$p\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} p(A_j),$$

whenever the measurable sets $A_1, A_2, \ldots$, with $A_i \subset \Omega$, are pairwise disjoint. The quantity $p(A)$ is interpreted as the probability that the initial condition of the system belongs to the measurable subset $A \subset \Omega$. It is important to emphasize that the probability measure does not introduce randomness into the dynamics itself. The evolution generated by the transformation T remains completely deterministic. Rather, the probability measure is introduced to allow us to formulate a statistical distribution of the system when complete knowledge is unavailable.

[1] A σ-algebra on a set Ω is a collection $\mathcal{F}$ of subsets of Ω satisfying the following properties: *i*) $\Omega \in \mathcal{F}$; *ii*) if $A \in \mathcal{F}$, then its complement $A^c = \Omega \setminus A$ also belongs to $\mathcal{F}$; *iii*) if $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ is a countable collection of sets, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$. Consequently, $\mathcal{F}$ is also closed under countable intersections and contains the empty set.

To combine probability with the deterministic dynamics, the transformation T is assumed to be measurable. This means that, for every measurable set $A \in \mathcal{B}$, its inverse image

$$T^{-1}(A) = \{s \in \Omega : T(s) \in A\}$$

also belongs to $\mathcal{B}$. In other words, it maps measurable spaces into measurable spaces.

The inverse image plays a central role because probabilities are assigned to the initial conditions rather than to the trajectories themselves. The set $T^{-1}(A)$ consists of all initial conditions whose evolution reaches the region A after one application of the transformation. Consequently, if the initial conditions are distributed according to the probability measure p, the probability that the system is found inside the region A after one iteration of the dynamics is simply

$$p(T^{-1}(A)).$$

More generally, the probability that the system occupies the measurable set A after n iterations is

$$p(T^{-n}(A)).$$

This expression is obtained by considering all initial conditions whose trajectories arrive in A after n applications of the transformation. It is important to observe here that, even in the case of a non-invertible transformation, T^{-1} is still well defined as the pre-image of A. See Fig. 2.2 for an illustration of this idea.

The pair consisting of the deterministic map T and the probability measure p therefore provides a complete statistical description of the dynamics. The map determines how every individual trajectory evolves, while the probability measure specifies the likelihood of the different initial conditions. Together they induce a probability distribution over the ensemble of trajectories generated by the dynamical system.

The probabilistic description developed in this section forms the basis for the information-theoretic construction that follows. Shannon's theory is formulated in terms of sequences of symbols emitted by a source, whereas the trajectories generated by a dynamical system are sequences of states in the space Ω. The next step is therefore to construct a symbolic representation of these trajectories. Once such an encoding has been introduced, the probability measure p naturally induces probability distributions over the corresponding symbolic sequences.

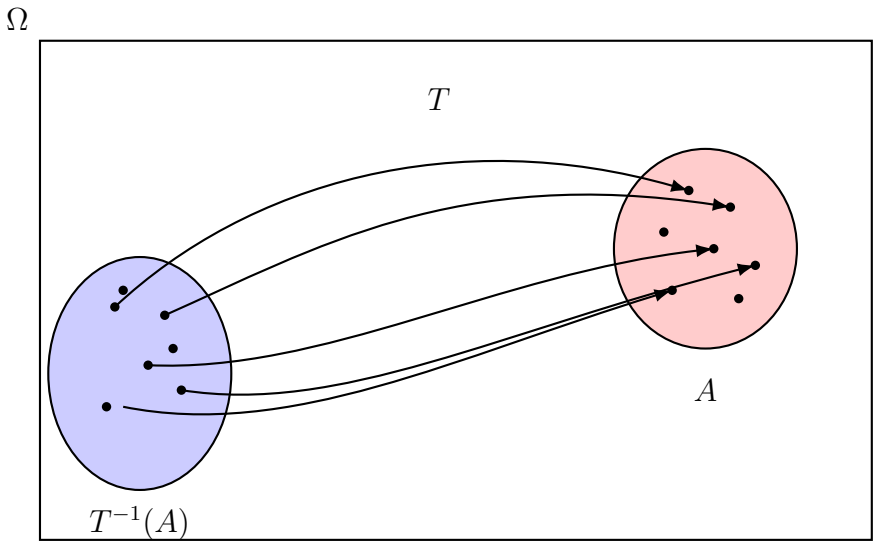


Figure 2.2: Illustration of the inverse image of a measurable set. The shaded blue region on the left represents the set $T^{-1}(A)$ of all initial conditions whose evolution under the transformation T reaches the measurable set A (red shaded region on the right) after one iteration. Since probabilities are assigned to the initial conditions, the probability that the system is found in A after one time step is obtained by measuring its preimage, namely $p(T^{-1}(A))$.

2.3 Symbolic representation of the dynamics

The probabilistic description introduced in the previous section assigns probabilities to trajectories evolving in the state space Ω. Shannon's information theory, however, is formulated in terms of sequences of symbols drawn from a finite alphabet rather than sequences of points in a continuous state space. To establish a connection between deterministic dynamical systems and information theory, it is therefore necessary to convert the trajectories of the dynamical system into symbolic sequences, thus producing a symbolic dynamics.

The first step consists in partitioning the state space into a finite collection of mutually disjoint measurable sets,

$$\mathcal{P} = \{A_{x_1}, A_{x_2}, \ldots, A_{x_m}\},$$

called a *partition* of Ω. These sets satisfy

$$A_{x_i} \cap A_{x_j} = \varnothing, \qquad i \neq j,$$

and

$$\bigcup_{i=1}^{m} A_{x_i} = \Omega.$$

This technique allows us to associate each element of the partition with a symbol belonging to the finite alphabet

$$\mathcal{X} = \{x_1, x_2, \ldots, x_m\}\,.$$

The partition therefore establishes a correspondence between regions of the state space and symbols of the alphabet. Note that we are not mapping states into symbols, we are mapping sets of states into distinguishable symbols. The partition $\mathcal{P}$ thus defines an encoding map

$$\pi_{\mathcal{P}} : \Omega \longrightarrow \mathcal{X},$$

given by

$$\pi_{\mathcal{P}}(s) = x_i, \qquad \text{whenever } s \in A_{x_i}.$$

In other words, the encoding map replaces every point of the state space by the symbol associated with the partition element containing that point. Therefore, this map provides the translation of the trajectories in the state space to symbolic trajectories, which are sequences of the symbols of the alphabet $\mathcal{X}$.

Combining the deterministic evolution with the encoding map naturally generates a random variable X_n at every instant n, defined by

$$X_n(s) = \pi_{\mathcal{P}}(T^n(s))\,.$$

The random variable X_n specifies the symbol associated with the region occupied by the trajectory after n iterations of the dynamics.

Starting from an initial condition $s_0 \in \Omega$, the orbit $\gamma(s_0)$ is therefore transformed into the symbolic sequence

$$\gamma(s_0) \to X_0(s_0),\, X_1(s_0),\, X_2(s_0),\, \ldots,$$

Every trajectory of the dynamical system thus generates a unique infinite sequence of symbols. This sequence may be regarded as a message emitted by the information source describing the dynamical system. This idea is illustrated in Fig. 2.3.

It is important to emphasize that the symbolic sequence does not represent an approximation of the dynamics. Rather, it provides an alternative description in which the detailed position of the system inside each partition element is ignored, while the order in which the partition elements are visited is preserved. Different partitions therefore produce different symbolic descriptions of the same underlying dynamics.

Since the initial condition is distributed according to the probability measure p, the symbolic sequences generated by the encoding map inherit a probability distribution from the underlying dynamics. Consequently, the deterministic dynamical system together with the partition defines a stochastic process over the finite alphabet $\mathcal{X}$. This process constitutes the information source from which Shannon's information-theoretic quantities will be constructed.

The next step is then to determine the probability associated with individual symbols and symbolic words. As we shall see, these probabilities are obtained directly from the probability measure on the state space through the action of the dynamical map.

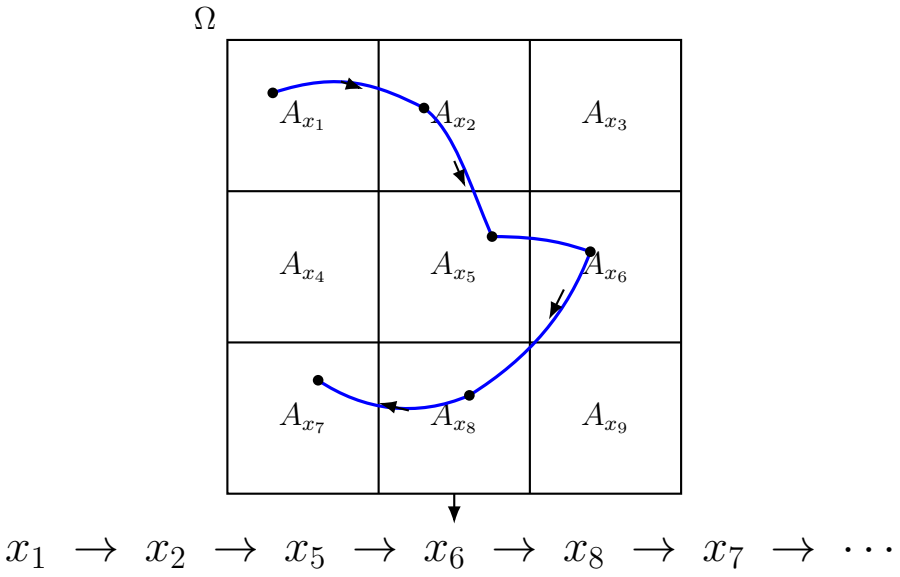


Figure 2.3: A partition of the state space Ω defines a symbolic representation of the dynamics. Each partition element A_{x_i} is associated with a symbol x_i. As the trajectory evolves through the state space, the successive regions it visits generate a symbolic sequence over the finite alphabet $\mathcal{X} = \{x_1, x_2, \ldots, x_9\}$.

2.4 Probabilities of symbolic trajectories

The symbolic representation introduced in the previous section associates a sequence of symbols with every trajectory of the dynamical system. Since the initial condition is distributed according to the probability measure p, the symbolic sequences generated by the dynamics inherit a probability distribution. Our objective in this section is to determine these probabilities explicitly.

We begin by considering the probability that the symbol observed after n iterations is equal to a given symbol x_i. By definition, $X_n(s) = \pi_{\mathcal{P}}(T^n(s))$, so that the

event $X_n = x_i$ occurs whenever the state of the system after n iterations belongs to the partition element A_{x_i}. Consequently,

$$X_n = x_i \quad \Longleftrightarrow \quad T^n(s) \in A_{x_i}.$$

From the dynamics of the system, the probability of observing the symbol x_i after n iterations is therefore

$$P(X_n = x_i) = p(T^{-n}(A_{x_i})).$$

In other words, one must determine all initial conditions whose trajectories reach the partition element A_{x_i} after n iterations and then evaluate the probability assigned to this set. This is equivalent to compute $p_n(T^n(s_0))$, where p_n is the evolved probability distribution to time n. So, instead of evolving the probability distribution we simply ask for all the points at the initial time reaches the element A_{x_i} after n iterations of the dynamical map.

The same reasoning extends naturally to several consecutive symbols. Consider, for example, the probability of observing the two-symbol word (x_i, x_j). This event occurs if and only if the initial condition belongs to the partition element A_{x_i} and, after one iteration of the dynamics, reaches the partition element A_{x_j}. The corresponding set of initial conditions is therefore

$$A_{x_i} \cap T^{-1}(A_{x_j}),$$

and the probability of the word becomes

$$P(X_0 = x_i, X_1 = x_j) = p\big(A_{x_i} \cap T^{-1}(A_{x_j})\big)\,.$$

This construction immediately generalizes to symbolic words of arbitrary length. Consider the finite word

$$(x_0, x_1, \ldots, x_{n-1}).$$

This word is produced whenever the trajectory begins inside the partition element A_{x_0}, enters A_{x_1} after one iteration, enters A_{x_2} after two iterations, and so forth until reaching $A_{x_{n-1}}$ after $n-1$ iterations. The corresponding set of initial conditions is therefore

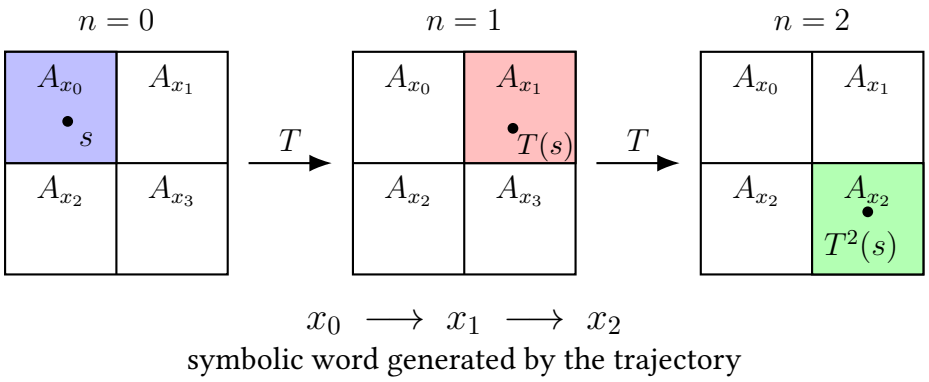
$$A_{x_0} \cap T^{-1}(A_{x_1}) \cap T^{-2}(A_{x_2}) \cap \cdots \cap T^{-(n-1)}(A_{x_{n-1}}),$$

whose probability is

$$P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) = p\left(A_{x_0} \cap T^{-1}(A_{x_1}) \cap \cdots \cap T^{-(n-1)}(A_{x_{n-1}})\right).$$

This expression constitutes the fundamental link between deterministic dynamics and information theory. It shows that the probability of every symbolic word is determined entirely by the probability measure on the state space together with the deterministic evolution generated by the map T. No stochastic dynamics has been introduced. The apparent randomness of the symbolic process arises solely from the uncertainty in the initial condition. See Fig. 2.4 for a graphic illustration of this concept.

The collection of all finite symbolic words together with their associated probabilities defines a stochastic process over the finite alphabet $\mathcal{X}$. It is this stochastic process, rather than the original deterministic dynamics, that will be analysed using the tools of information theory. In particular, the probabilities derived above provide the ingredients required to quantify the uncertainty associated with symbolic trajectories through the Shannon entropy of finite words.



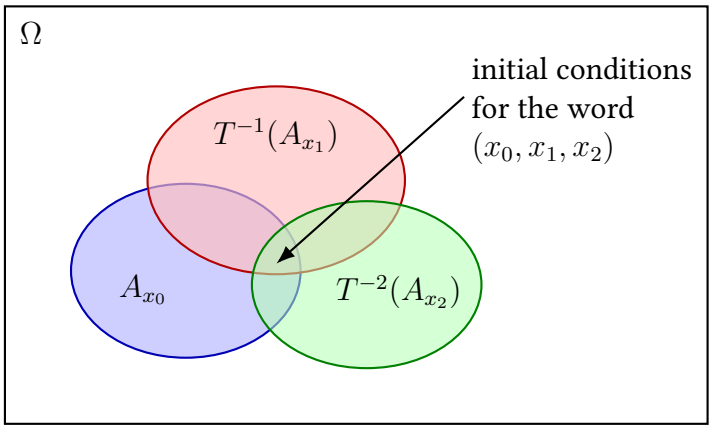


Figure 2.4: Geometric interpretation of the probability of a symbolic word. In the upper row, an initial condition $s \in A_{x_0}$ evolves into $T(s) \in A_{x_1}$ and subsequently into $T^2(s) \in A_{x_2}$, thereby generating the word (x_0, x_1, x_2). The lower panel represents the set of all initial conditions producing this word. Such initial conditions must belong simultaneously to A_{x_0}, $T^{-1}(A_{x_1})$, and $T^{-2}(A_{x_2})$. The probability of the word is therefore the probability measure of their intersection.

2.5 Shannon entropy of symbolic words

The symbolic dynamics developed in the previous sections transforms a deterministic dynamical system into a stochastic process over the finite alphabet $\mathcal{X}$. Every finite symbolic word is assigned a probability through the probability measure on the state space, thereby providing all the ingredients required to apply information theory. We may therefore quantify the uncertainty associated with symbolic trajectories by means of the Shannon entropy.

Consider the symbolic sequence of random variables

$$(X_0, X_1, \ldots, X_{n-1}),$$

whose probability distribution was derived in the previous section. The Shannon entropy of this sequence is

$$H(X_0, \ldots, X_{n-1}) = -\sum_{x_0,\ldots,x_{n-1}} P(x_0, \ldots, x_{n-1}) \log P(x_0, \ldots, x_{n-1}),$$

where the sum extends over all possible symbolic messages of length n and

$$P(x_0, \ldots, x_{n-1}) \equiv P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}).$$

The Shannon entropy measures the average uncertainty associated with observing a message of length n. Equivalently, it quantifies the average amount of information gained when the message produced by the dynamical system becomes known. If only a few symbolic messages possess significant probability, the uncertainty is small and the entropy assumes a relatively low value. Conversely, if many different messages occur with comparable probabilities, the uncertainty increases, leading to a larger entropy.

The entropy generally depends on the length of the message. As the number of symbols increases, more information about the trajectory becomes available and additional correlations among successive symbols are revealed. Consequently, the Shannon entropy of the symbolic process is naturally viewed as a function of the message length.

The increase of the entropy with the length reflects the growth of the information required to specify progressively longer segments of the symbolic trajectory. If the symbols were statistically independent, the entropy would simply be proportional to the number of symbols. In deterministic dynamical systems, however, successive symbols are generally correlated because they originate from the same underlying trajectory. These correlations reduce the uncertainty of future symbols

once the previous ones are known and therefore play a fundamental role in the information-theoretic characterization of the dynamics.

Our objective is not merely to quantify the uncertainty associated with symbolic messages of a fixed length, but rather to determine the average amount of new information generated by the dynamics per iteration. To accomplish this, we must understand how the Shannon entropy changes as the messages become progressively longer. The key to this analysis is the observation that messages of length n possess a natural geometric interpretation in terms of progressively finer descriptions of the state space. This geometric viewpoint, developed in the next section, leads directly to the definition of the Kolmogorov-Sinai entropy.

2.6 Refining the symbolic description

In the previous section we showed that the uncertainty associated with the dynamics is quantified by the Shannon entropy of symbolic messages. As the length of these messages increases, progressively longer segments of the trajectory become available and the symbolic description contains more information about the evolution of the system. It is natural to ask whether this increase in information possesses a simple geometric interpretation in the state space. The answer is affirmative. Longer symbolic messages correspond to progressively finer partitions of the state space.

To understand this correspondence, let us first consider a message of length one. Observing the symbol x_0 merely tells us that the initial condition belongs to the partition element A_{x_0}. At this stage, the precise location of the system inside A_{x_0} remains completely unknown.

Suppose now that the first two symbols, (x_0, x_1), are observed. The first symbol implies that the initial condition belongs to the set A_{x_0}, whereas the second symbol requires that, after one iteration of the dynamics, the trajectory reaches the partition element A_{x_1}. Consequently, the initial condition must simultaneously satisfy the two conditions $s \in A_{x_0}$ and $T(s) \in A_{x_1}$, or equivalently, $s \in A_{x_0} \cap T^{-1}(A_{x_1})$. The knowledge of two symbols therefore localizes the initial condition to a smaller region of the state space than either symbol alone. See Fig. 2.4.

The same reasoning immediately extends to messages of arbitrary length. If the trajectory generates the word $(x_0, x_1, \ldots, x_{n-1})$, then the initial condition must satisfy

$$s \in A_{x_0} \cap T^{-1}(A_{x_1}) \cap T^{-2}(A_{x_2}) \cap \cdots \cap T^{-(n-1)}(A_{x_{n-1}}).$$

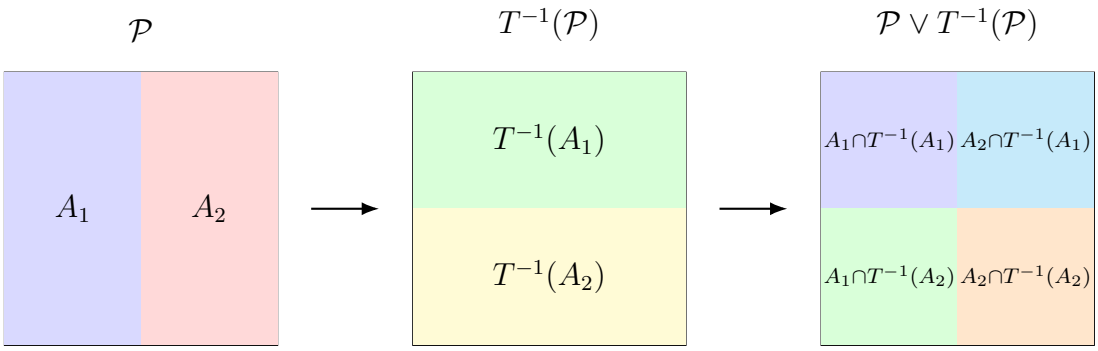


Figure 2.5: Construction of the refinement of a partition. The original partition $\mathcal{P}$ and its inverse image $T^{-1}(\mathcal{P})$ provide two different descriptions of the state space. Their common refinement, $\mathcal{P} \vee T^{-1}(\mathcal{P})$, consists of all non-empty intersections between elements of the two partitions. Each element of the refined partition corresponds to one symbolic word of length two. Successive refinements are obtained by intersecting additional inverse images, producing symbolic messages of increasing length.

Every message therefore determines a unique measurable subset of the state space consisting of all initial conditions that generate exactly the same symbolic sequence during the first n iterations.

This observation naturally leads to the notion of a refinement of a partition. Given two partitions,

$$\mathcal{P} = \{A_1, \ldots, A_r\} \qquad \text{and} \qquad \mathcal{Q} = \{B_1, \ldots, B_s\},$$

their common refinement is defined as

$$\mathcal{P} \vee \mathcal{Q} = \{A_i \cap B_j : A_i \cap B_j \neq \varnothing\}.$$

The refined partition consists of all non-empty intersections of elements of the original partitions and therefore provides a finer description of the state space.

Applying this construction to the successive inverse images of the original partition in the dynamical system yields the refinement

$$\mathcal{P}^{(n)} = \mathcal{P} \vee T^{-1}(\mathcal{P}) \vee T^{-2}(\mathcal{P}) \vee \cdots \vee T^{-(n-1)}(\mathcal{P}).$$

Each element of $\mathcal{P}^{(n)}$ is obtained by intersecting one element of the partition with one element of each of its successive inverse images. Consequently, every element of the refined partition is uniquely associated with a symbolic message of length n, while every symbolic message determines one element of the refined partition. There is therefore a one-to-one correspondence between symbolic messages and

the elements of $\mathcal{P}^{(n)}$. Such construction is represented in Fig. 2.5

This correspondence provides the geometric interpretation of symbolic dynamics. Observing progressively longer symbolic messages does not merely increase the amount of information available about the trajectory. It simultaneously refines our knowledge of the initial condition by restricting it to progressively smaller subsets of the state space. The refinement of the partition therefore represents the geometric counterpart of increasing the length of the symbolic message.

This result establishes a direct correspondence between the combinatorial description of the dynamics in terms of symbolic messages and the geometric structure of the state space. The Shannon entropy of messages may therefore be interpreted equivalently as the entropy associated with the refined partitions. This observation forms the basis for the definition of the Kolmogorov-Sinai entropy.

2.7 Kolmogorov-Sinai entropy

The previous sections established a direct correspondence between the dynamics of the system and a stochastic process over a finite alphabet. The uncertainty associated with finite symbolic trajectories is quantified by the Shannon entropy

$$H(X_0, \ldots, X_{n-1}),$$

while the refinement of the partition provides the geometric interpretation of progressively longer symbolic messages. We are now in a position to introduce one of a central quantities in the theory of dynamical systems: the Kolmogorov-Sinai entropy.

The Shannon entropy of symbolic messages generally increases with their length, since longer messages contain more information about the trajectory. Nevertheless, the quantity of greatest interest is not the total uncertainty associated with a message of length n, but rather the average amount of new information generated by the dynamics at each iteration. This motivates the introduction of the entropy per symbol

$$\frac{1}{n} H(X_0, \ldots, X_{n-1}),$$

which measures the average information required to specify one symbol of a sufficiently long trajectory.

For a broad class of dynamical systems, this quantity approaches a well-defined limit as the message length tends to infinity. The resulting entropy rate is called the

Kolmogorov-Sinai entropy associated with the partition $\mathcal{P}$,

$$h_p(T, \mathcal{P}) = \lim_{n\to\infty} \frac{1}{n} H(X_0, \ldots, X_{n-1}),$$

provided the limit exists. The subindex p is introduced to remember that the entropy is defined with respect to the measure Lebesgue p.

The interpretation of this quantity is the following. If the symbolic trajectory is highly predictable, observing additional symbols contributes relatively little new information, and the entropy rate is small. Conversely, if successive observations continue to reveal genuinely new information about the evolution of the system, the entropy rate is larger. The Kolmogorov-Sinai entropy therefore measures the asymptotic rate at which information is produced by the dynamics.

The geometric interpretation developed in the previous section allows this definition to be expressed in an equivalent form. Since every symbolic message of length n is in one-to-one correspondence with an element of the refined partition $\mathcal{P}^{(n)}$, the Shannon entropy of symbolic messages is precisely the Shannon entropy of the refined partition,

$$H(X_0, \ldots, X_{n-1}) = H(\mathcal{P}^{(n)}),$$

where

$$H(\mathcal{P}^{(n)}) = -\sum_{A\in\mathcal{P}^{(n)}} p(A) \log p(A).$$

Consequently, the entropy rate may be written as

$$h_p(T, \mathcal{P}) = \lim_{n\to\infty} \frac{1}{n} H(\mathcal{P}^{(n)}).$$

The value of $h_p(T, \mathcal{P})$ generally depends on the partition used to construct the symbolic dynamics. Different partitions provide different symbolic descriptions of the same deterministic trajectory and may therefore reveal different amounts of information about the dynamics. Some partitions fail to capture important dynamical features, whereas others provide a much more faithful representation of the evolution.

To obtain a quantity that depends only on the dynamical system itself and not on the particular symbolic description, one considers the supremum over all finite measurable partitions,

$$h_p(T) = \sup_{\mathcal{P}} h(T, \mathcal{P}).$$

This quantity is called the Kolmogorov-Sinai entropy of the dynamical system. It

is an invariant of measure-preserving transformations and constitutes one of the fundamental quantities used to characterize deterministic dynamics.

From an information-theoretic perspective, the Kolmogorov-Sinai entropy represents the maximum asymptotic rate at which information can be extracted from the evolution of the system by means of symbolic observations. Equivalently, it quantifies the maximum average amount of information generated per iteration among all possible symbolic descriptions of the dynamics. The introduction of the supremum guarantees that this quantity is an intrinsic property of the dynamical system rather than an artifact of a particular partition.

The Kolmogorov-Sinai entropy establishes a profound bridge between dynamical systems and information theory. Starting from a deterministic evolution, we introduced a symbolic description through a partition of the state space, assigned probabilities to symbolic trajectories using a probability measure on the state space, quantified their uncertainty through Shannon entropy, and finally obtained the asymptotic information production rate of the dynamics. This construction shows that deterministic systems may be characterized not only by their trajectories in phase space but also by the rate at which they generate information. Based on this construction, in the following section we provide a thermodynamic interpretation of the this dynamical entropy.

2.8 Thermodynamic interpretation

The construction developed in the previous sections shows that the Kolmogorov-Sinai entropy is the asymptotic rate at which information is generated by the symbolic description of a deterministic dynamical system. Although this definition is mathematically precise, it does not immediately reveal why the Kolmogorov-Sinai entropy is physically important. One may naturally ask whether this quantity merely characterizes the complexity of symbolic trajectories or whether it also possesses a direct thermodynamic interpretation.

Remarkably, for driven classical Hamiltonian systems, the randomness generated by the dynamics, quantified by the Kolmogorov-Sinai entropy, imposes a lower bound on the energetic cost required to drive the system out of equilibrium [5]. This result reveals that dynamical complexity and thermodynamic irreversibility are fundamentally connected.

To understand the origin of this connection, consider a Hamiltonian system initially prepared in thermal equilibrium with a heat reservoir at inverse temperature β. The system is subsequently isolated from the reservoir while an external control parameter is varied according to a prescribed protocol. During this process, work

is performed on the system and, in general, part of this work is irreversibly dissipated. Since the dynamics remains Hamiltonian throughout the driving protocol, Liouville's theorem guarantees the conservation of phase-space volume, allowing the symbolic description introduced in the previous sections to remain valid.

The phase space is partitioned into a finite collection of cells, and every trajectory is represented by a symbolic sequence indicating the successive partition elements visited during the evolution. The probability of each symbolic trajectory is therefore completely determined by the probability measure over the phase space. Consequently, the Kolmogorov-Sinai entropy measures the asymptotic rate at which new information is generated by the coarse-grained dynamics.

The central idea is to compare two different descriptions of the evolving system. The first corresponds to the exact probability density in phase space, whereas the second is obtained after coarse graining the dynamics according to the symbolic partition. The difference between these two descriptions quantifies the information that is discarded by the coarse-graining procedure. Since coarse graining necessarily increases the Shannon entropy, a lower bound relating the thermodynamic entropy production to the amount of information generated by the dynamics can be obtained

$$\beta\langle W_{\mathrm{d}}\rangle \geq \beta\left(\langle H\rangle - F\right) - I_t(\mathcal{P}),$$

where $\langle W_{\mathrm{d}}\rangle$ denotes the average dissipated work, F is the equilibrium free energy associated with the external control parameter, and

$$I_t(\mathcal{P}) = h_p(T) - c_t(\mathcal{P}) - d_t(\mathcal{P})$$

contains the Kolmogorov-Sinai entropy together with correction terms associated with the finite coarse-graining procedure. In the limit of increasingly refined partitions, the dominant contribution is provided by the Kolmogorov-Sinai entropy itself.

This inequality possesses a remarkably interpretation. The Kolmogorov-Sinai entropy quantifies the average amount of new information produced by the dynamics per iteration, while the dissipated work measures the degree of thermodynamic irreversibility of the driven process. The inequality therefore shows that these two quantities cannot vary independently. A dynamical system that generates information at a higher rate necessarily possesses a larger minimum energetic cost associated with nonequilibrium driving.

Throughout this chapter, the Kolmogorov-Sinai entropy emerged as a purely informational quantity constructed from symbolic trajectories and Shannon entropy. The above inequality demonstrates that this informational measure is not

merely an abstract mathematical invariant but also possesses direct physical consequences. The information generated by the dynamics is ultimately constrained by energetic considerations, revealing that information production and thermodynamic irreversibility are two complementary manifestations of the same underlying physical process.

This connection also provides a deeper interpretation of chaos. Traditionally, positive Kolmogorov-Sinai entropy is regarded as a signature of chaotic dynamics because nearby trajectories become increasingly difficult to distinguish through finite-resolution observations. The thermodynamic interpretation extends this viewpoint by showing that the unpredictability generated by chaotic dynamics is accompanied by an unavoidable energetic cost when the system is driven out of equilibrium. In this sense, the Kolmogorov-Sinai entropy measures not only the rate at which deterministic dynamics produces information but also the minimum thermodynamic resources required to sustain such informational complexity.

The relationship between information production and dissipation illustrates one of the central themes of modern statistical physics: information is a physical quantity. Quantities originally introduced to characterize uncertainty in communication theory naturally acquire energetic significance when applied to physical systems. The Kolmogorov-Sinai entropy therefore occupies a central position at the interface between dynamical systems, information theory, and thermodynamics, providing a unified framework for understanding deterministic chaos from both informational and physical perspectives.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication", Bell Syst. Tech. J. **27**, 379 (1948).

[2] M. Kuczma, *An introduction to the theory of functional equations and inequalities: cauchy's equation and jensen's inequality*, edited by A. Gilányi, 2nd ed. (Birkhäuser, 2009).

[3] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. (Wiley, 2006).

[4] C. Stein, "Information and efficiency of tests", Ann. Math. Stat. **27**, 272 (1956).

[5] K. Micadei, R. M. Serra, and L. C. Céleri, "Thermodynamic cost of acquiring information", Phys. Rev. E **88**, 062123 (2013).