Chapter 2
Review of Classical Information Theory

Abstract This chapter presents a review of the classical information theory which plays a crucial role in this thesis. We introduce the various types of informational measures such as Shannon entropy, the relative entropy, the mutual information and the transfer entropy. We also briefly discuss the noisy-channel coding theorem which represents the meaning of informational measures in the artificial information transmission over a communication channel.

Keywords Information theory · Noisy-channel coding theorem

We review the classical information theory in this chapter. The classical information theory had been well established by Shannon in his historical paper entitled "A mathematical theory of communication" [1]. Shannon discussed the relationship between the entropy and the accuracy of the information transmission through a noisy communication channel with an artificial coding device, which is well known as the noisy channel coding theory. In this chapter, we introduce various types of the entropy (i.e., the Shannon entropy, the relative entropy, the mutual information and the transfer entropy [2]) as measures of information, and the noisy channel coding theorem [1, 3, 4].

2.1 Entropy

First of all, we briefly introduce various types of the entropy, which quantify measures of information [1, 3].

2.1.1 Shannon Entropy

We first introduce the Shannon entropy, which characterizes the uncertainty of random variables. Let $p(x)$ be the probability distribution of a discrete random variable

© Springer Science+Business Media Singapore 2016

S. Ito, *Information Thermodynamics on Causal Networks and its Application to Biochemical Signal Transduction*, Springer Theses,
DOI 10.1007/978-981-10-1664-6_2

x. The probability distribution $p(x)$ satisfies the normalization of the probability and the nonnegativity (i.e., $\sum_x p(x) = 1, 0 \leq p(x) \leq 1$). The Shannon entropy $S(x)$ is defined as

$$S(x) := -\sum_x p(x) \ln p(x). \tag{2.1}$$

In the case of a continuous random variable x with probability density function $p(x)$ which satisfies the normalization and the nonnegativity (i.e., $\int dx p(x) = 1$, $0 \leq p(x) \leq 1$), the Shannon entropy (or differential entropy) is defined as

$$S(x) := -\int dx p(x) \ln p(x). \tag{2.2}$$

In this thesis, the logarithm (ln) denotes the natural logarithm. To discuss the discrete and continuous cases in parallel, we introduce the ensemble average $\langle f(x) \rangle$ for any function $f(x)$ as

$$\begin{aligned} \langle f(x) \rangle &= \langle f(x) \rangle_p \\ &:= \sum_x p(x) f(x) \end{aligned} \tag{2.3}$$

for a discrete random variable x and

$$\begin{aligned} \langle f(x) \rangle &= \langle f(x) \rangle_p \\ &:= \int dx p(x) f(x) \end{aligned} \tag{2.4}$$

for a continuous random variable x. From the definition of ensemble average Eqs. (2.3) and (2.4), the two definitions of the Shannon entropy Eqs. (2.1) and (2.2) are rewritten as

$$\begin{aligned} S(x) &= \langle -\ln p(x) \rangle \\ &= \langle s(x) \rangle, \end{aligned} \tag{2.5}$$

where we here say $s(x) := -\ln p(x)$ is a stochastic Shannon entropy. The Shannon entropy $S(X)$ of a set of random variables $X = \{x_1, \ldots, x_N\}$ with a joint probability distribution $p(X)$ is also defined as

$$\begin{aligned} S(X) &:= \langle -\ln p(X) \rangle \\ &= \langle s(X) \rangle. \end{aligned} \tag{2.6}$$

Let the conditional probability distribution of X under the condition Y be $p(X|Y) := p(X, Y)/p(Y)$. The conditional Shannon entropy $S(X|Y)$ with a joint probability

$p(X)$ is defined as

$$\begin{aligned} S(X|Y) &:= \langle -\ln p(X|Y) \rangle \\ &= \langle s(X|Y) \rangle, \end{aligned} \tag{2.7}$$

where $s(X|Y) := -\ln p(X|Y)$ is a stochastic conditional Shannon entropy. We note that its ensemble takes an integral over a joint distribution $p(X)$. From $0 \leq p \leq 1$, the Shannon entropy S satisfies the nonnegativity $S \geq 0$.

By the definition of the conditional probability distribution $p(X|Y) := p(X, Y)/p(Y)$, we have the chain rule in probability theory. The chain rule in probability theory produces the product of conditional probabilities:

$$p(X) = p(x_1) \prod_{k=2}^{N} p(x_k|x_{k-1}, \ldots, x_1). \tag{2.8}$$

From this chain rule Eq. (2.8) and the definitions of the Shannon entropy Eqs. (2.6) and (2.7), we obtain the chain rule for (stochastic) Shannon entropy:

$$s(X) = s(x_1) + \sum_{k=2}^{N} s(x_k|x_{k-1}, \ldots, x_1). \tag{2.9}$$

$$S(X) = S(x_1) + \sum_{k=2}^{N} S(x_k|x_{k-1}, \ldots, x_1). \tag{2.10}$$

The chain rule indicates that the (stochastic) joint Shannon entropy is always rewritten by a sum of the (stochastic) conditional Shannon entropy.

2.1.2 Relative Entropy

We next introduce the relative entropy (or the Kullback–Leibler divergence), which is an asymmetric measure of the difference between two probability distributions. The thermodynamic relationships (e.g., the second law of thermodynamics) and several theorems in information theory can be derived from the nonnegativity of the relative entropy. The relative entropy or the Kullback–Leibler divergence between two probability distributions $p(x)$ and $q(x)$ is defined as

$$D_{\mathrm{KL}}(p(x)||q(x)) = \langle \ln p(x) - \ln q(x) \rangle_p. \tag{2.11}$$

We will show that the relative entropy is always nonnegative and is 0 if and only if $p = q$.

To show this fact, we introduce Jensen's inequality [3]. Let $\phi(f(x))$ be a convex function, which satisfies $\phi(\lambda a+(1-\lambda)b) \leq \lambda\phi(a)+(1-\lambda)\phi(b)$ with $\forall a, b \in f(x)$ and $\forall\lambda \in [0, 1]$. Jensen's inequality states

$$\phi(\langle f(x)\rangle) \leq \langle\phi(f(x))\rangle. \tag{2.12}$$

The equality holds if and only if $f(x)$ is constant or ϕ is linear.

We notice that $-\ln(f(x))$ is a convex nonlinear function. By applying Jensen's inequality (2.12), we can derive the nonnegativity of the relative entropy,

$$\begin{aligned} D_{\mathrm{KL}}(p(x)||q(x)) &= \langle -\ln[q(x)/p(x)]\rangle_p \\ &\geq -\ln\langle q(x)/p(x)\rangle_p \\ &= -\ln 1 \\ &= 0, \end{aligned} \tag{2.13}$$

where we used the normalization of the distribution q, $\langle q(x)/p(x)\rangle_p = \int dx q(x) = 1$. The equality holds if and only if $q(x)/p(x) = c$, where c is a constant. Because p and q satisfy the normalizations $\int dx p(x) = 1$ and $\int dx q(x) = 1$, a constant c should be $c = 1$, and we can show that the relative entropy $D_{\mathrm{KL}}(p(x)||q(x))$ is 0 if and only if $p = q$.

From the nonnegativity of the relative entropy, we can easily show that $S(x) \leq \ln|x|$ where $|x|$ denotes the number of elements of a discrete random variable x with the equality satisfied if and only if x is uniformly distributed. Let $p_u(x) = 1/|x|$ be a uniform function over x. The relative entropy $D_{\mathrm{KL}}(p(x)||p_u(x))$ is calculated as $D_{\mathrm{KL}}(p(x)||p_u(x)) = \ln|x| - S(x)$, and its negativity gives $S(x) \leq \ln|x|$.

The joint relative entropy $D_{\mathrm{KL}}(p(X)||q(X))$ is defined as

$$D_{\mathrm{KL}}(p(X)||q(X)) = \langle\ln p(X) - \ln q(X)\rangle_p \tag{2.14}$$

and the conditional relative entropy $D_{\mathrm{KL}}(p(X|Y)||q(X|Y))$ is defined as

$$D_{\mathrm{KL}}(p(X|Y)||q(X|Y)) = \langle\ln p(X|Y) - \ln q(X|Y)\rangle_p. \tag{2.15}$$

The joint and conditional relative entropy satisfy $D_{\mathrm{KL}} \geq 0$ with the equality satisfied if and only if $p = q$. The chain rule in probability theory Eq. (2.8) and the definition of the relative entropy Eqs. (2.15) and (2.14) give the chain rule for relative entropy as

$$\begin{aligned} &D_{\mathrm{KL}}(p(X)||q(X)) \\ &\quad = D_{\mathrm{KL}}(p(x_1)||q(x_1)) + \sum_{k=2}^{N} D_{\mathrm{KL}}(p(x_k|x_{k-1},\ldots,x_1)||q(x_k|x_{k-1},\ldots,x_1)). \end{aligned} \tag{2.16}$$

2.1.3 Mutual Information

We introduce the mutual information I, which characterizes the correlation between random variables. The mutual information between X and Y is given by the relative entropy between the joint distribution $p(X, Y)$ and the product distribution $p(X)p(Y)$:

$$\begin{aligned} I(X:Y) &:= D_{\mathrm{KL}}(p(X,Y)||p(X)p(Y)) \\ &= \langle \ln p(X,Y) - \ln p(X) - \ln p(Y) \rangle \\ &= \langle s(X) + s(Y) - s(X,Y) \rangle \\ &= S(X) + S(Y) - S(X,Y) \\ &= S(X) - S(X|Y) \\ &= S(Y) - S(Y|X). \end{aligned} \tag{2.17}$$

The mutual information quantifies the amount of information in X about Y (or information in Y about X). From the nonnegativity of the relative entropy $D_{\mathrm{KL}} \geq 0$, the mutual information is nonnegative $I(X:Y) \geq 0$ with the equality satisfied if and only if X and Y are independent $p(X, Y) = p(X)p(Y)$. This nonnegativity implies the fact that conditioning reduces the Shannon entropy (i.e., $S(X|Y) \leq S(X)$). From the nonnegativity of the Shannon entropy $S \geq 0$, the mutual information is bounded by the Shannon entropy of each variable X or Y ($I(X:Y) \leq S(X)$ and $I(X:Y) \leq S(Y)$). To summarize the nature of the mutual information, the following Venn's diagram is useful (see Fig. 2.1).

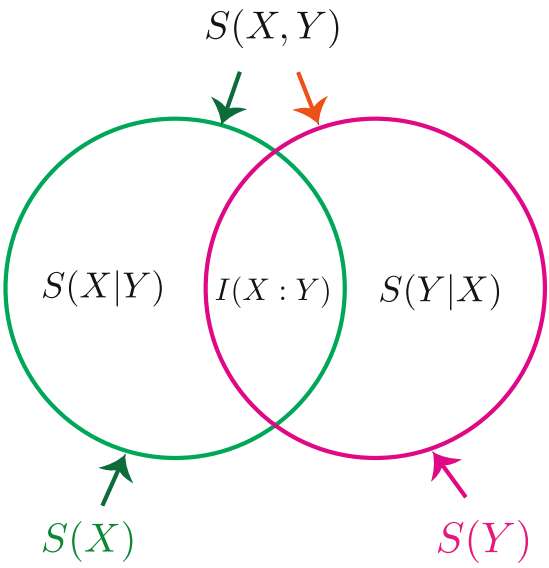


Fig. 2.1 The Venn's diagram which represents the nature of the mutual information (i.e., Eq. 2.17). The mutual information between X and Y represents the correlation between them in the sense of the Shannon entropy

The conditional mutual information between X and Y under the condition Z is also defined as

$$\begin{aligned} I(X:Y|Z) &:= D_{\mathrm{KL}}(p(X,Y|Z)||p(X|Z)p(Y|Z)) \\ &= \langle \ln p(X,Y|Z) - \ln p(X|Z) - \ln p(Y|Z) \rangle \\ &= \langle s(X|Z) + s(Y|Z) - s(X,Y|Z) \rangle \\ &= S(X|Z) + S(Y|Z) - S(X,Y|Z) \\ &= S(X|Z) - S(X|Y,Z) \\ &= S(Y|Z) - S(Y|X,Z). \end{aligned} \tag{2.18}$$

For Z independent of X and Y (i.e., $p(X,Y|Z) = p(X,Y)$), we have $I(X:Y|Z) = I(X:Y)$. The conditional mutual information is also nonnegative $I(X:Y|Z) \geq 0$ with the equality satisfied if and only if X and Y are independent under the condition of Z: $p(X,Y|Z) = p(X|Z)p(Y|Z)$ (or $p(X|Y,Z) = p(X|Z)$). We also define the stochastic mutual information $i(X:Y)$ and the stochastic conditional mutual information $i(X:Y|Z)$ as

$$i(X:Y) := s(X|Z) + s(Y|Z) - s(X,Y|Z) \tag{2.19}$$

$$i(X:Y|Z) := s(X|Z) + s(Y|Z) - s(X,Y|Z) \tag{2.20}$$

From the chain rule for (stochastic) Shannon entropy Eq. (2.10) and the definition of the mutual information Eqs. (2.17) and (2.18), we have the chain rule for (stochastic) mutual information

$$i(X:Y) := i(x_1:Y) + \sum_{k=2}^{N} i(x_k:Y|x_{k-1},\ldots,x_1), \tag{2.21}$$

$$I(X:Y) := I(x_1:Y) + \sum_{k=2}^{N} I(x_k:Y|x_{k-1},\ldots,x_1). \tag{2.22}$$

2.1.4 Transfer Entropy

Here, we introduce the transfer entropy, which characterizes the directed information flow between two systems in evolving time $X = \{x_k|k = 1,\ldots,N\}$ and $Y = \{y_k|k = 1,\ldots,N\}$. The transfer entropy was ordinarily introduced by Schreiber in 2000 [2] as a measure of the causal relationship between two random time series. The transfer entropy from X to Y at time k is defined as the conditional mutual information:

$$\begin{aligned} \mathcal{T}_{X\to Y} &:= I(y_{k+1}:\{x_k,\ldots,x_{k-l}\}|y_k,\ldots,y_{k-l'}) \\ &= \langle s(y_{k+1}|x_k,\ldots,x_{k-l},y_k,\ldots,y_{k-l'}) - s(y_{k+1}|y_k,\ldots,y_{k-l'}) \rangle \end{aligned}$$

$$= \langle \ln p(y_{k+1}|y_k, \ldots, y_{k-l'}) - \ln p(y_{k+1}|x_k, \ldots, x_{k-l}, y_k, \ldots, y_{k-l'}) \rangle. \tag{2.23}$$

The indexes l and l' denote the lengths of two causal time sequences $\{x_k, \ldots, x_{k-l}\}$ and $\{y_{k+1}, y_k, \ldots, y_{k-l'}\}$. Because of the nonnegativity of the mutual information $I(y_{k+1} : \{x_k, \ldots, x_{k-l}\}|y_k, \ldots, y_{k-l'})$, the transfer entropy is always nonnegative and is 0 if and only if the time evolution of Y system at time k does not depend on the history of X system,

$$p(y_{k+1}|y_k, \ldots, y_{k-l'}) = p(y_{k+1}|x_k, \ldots, x_{k-l}, y_k, \ldots, y_{k-l'}). \tag{2.24}$$

Thus, the transfer entropy quantifies the causal dependence between them at time k. If the dynamics of X and Y is Markovian (i.e., $p(y_{k+1}|x_k, \ldots, x_{k-l}, y_k, \ldots, y_{k-l'}) = p(y_{k+1}|x_k, y_k)$), the most natural choices of l and l' becomes $l = l' = 0$ in the sense of the causal dependence.

Here, we compare other entropic quantities which represent the direction exchange of information. Such conditional mutual informations have been discussed in the context of the causal coding with feedback [5]. Massey defined the sum of the conditional mutual information

$$I^{\mathrm{DI}}(X \to Y) := \sum_{k=1}^{N} I(y_k : \{x_1, \ldots, x_k\}|y_{k-1}, \ldots, y_1), \tag{2.25}$$

called the directed information. It can be interpreted as a slight modification of the sum of the transfer entropy over time. Several authors [6, 7] have also introduced the mutual information with time delay to investigate spatiotemporal chaos.

In recent years, the transfer entropy has been investigated in several contexts. For a Gaussian process, the transfer entropy is equivalent to the Granger causality test [8], which is an economic statistical hypothesis test for detecting whether one time series is useful in forecasting another [9, 10]. Using a technique of symbolization, a fast and robust calculation method of the transfer entropy has been proposed [11]. In a study of the order-disorder phase transition, the usage of the transfer entropy has been also proposed to predict an imminent transition [12]. In relation to our study, thermodynamic interpretations of the transfer entropy [13, 14] and a generalization of the transfer entropy for causal networks [15] have been proposed.

2.2 Noisy-Channel Coding Theorem

In this section, we show the fact that the mutual information between input and output is related to the accuracy of signal transmission. This fact is well known as the noisy-channel coding theorem (or Shannon's theorem) [1, 3]. The noisy-channel coding theorem was proved by Shannon in his original paper in 1948 [1]. In the case

of Gaussian channel, a similar discussion of information transmission was also given by Hartley previously [4].

2.2.1 Communication Channel

We consider the noisy communication channel. Let x be the input of the signal and y be the output of the signal. The mutual information $I(x : y)$ represents the correlation between input and output, which quantifies the ability of information transmission through the noisy communication channel. Here, we introduce two simple examples of the mutual information of the communication channel.

2.2.1.1 Example 1: Binary Symmetric Channel

Suppose that the input and output are binary states $x = 0, 1$ and $y = 0, 1$. The noise in the communication channel is represented by the conditional probability $p(y|x)$. The binary symmetric channel is given by the following conditional probability,

$$p(y = 1|x = 1) = p(y = 0|x = 0) = 1 - e, \tag{2.26}$$

$$p(y = 1|x = 0) = p(y = 1|x = 1) = e, \tag{2.27}$$

where e denotes the error rate of the communication. We assume that the distribution of the input signal is given by $p(x = 1) = 1 - r$ and $p(x = 0) = r$. The mutual information between input and output is calculated as

$$I(x : y) = (1 - e)\ln(1 - e) + e\ln e - (1 - e')\ln(1 - e') - e'\ln e', \tag{2.28}$$

where $e' := (1 - e)r + e(1 - r)$. This mutual information represents the amount of information transmitted through the noisy communication channel. In the case of $e = 1/2$, we have $I(x : y) = 0$, which means that we cannot infer the input signal x from reading the output y. The mutual information depends on the bias of the input signal r. To discuss the nature of the communication channel, the supremum value of the mutual information between input and output with respect to the input distribution. Let the channel capacity for the discrete input be

$$C := \sup_{p(x)} I(x : y). \tag{2.29}$$

For a binary symmetric channel, the mutual information has a supremum value with $r = 1/2$, and the channel capacity C is given as

$$C = \ln 2 + e\ln e + (1 - e)\ln(1 - e). \tag{2.30}$$

2.2.1.2 Example 2: Gaussian Channel

Suppose that the input and output have continuous values: $x \in [-\infty, \infty]$ and $y \in [-\infty, \infty]$. The Gaussian channel is given by the Gaussian distribution:

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(x-y)^2}{2\sigma_N^2}\right], \tag{2.31}$$

where σ_N^2 denotes the intensity of the noise in the communication channel. We assume that the initial distribution is also Gaussian:

$$p_P(x) = \frac{1}{\sqrt{2\pi\sigma_P^2}} \exp\left[-\frac{x^2}{2\sigma_P^2}\right], \tag{2.32}$$

where $\sigma_P^2 = \langle x^2 \rangle$ means the power of input signal. The mutual information $I(x : y)$ is calculated as

$$I(x : y) = \frac{1}{2} \ln\left(1 + \frac{\sigma_P^2}{\sigma_N^2}\right). \tag{2.33}$$

In the limit $\sigma_N^2 \to \infty$, we have $I(x : y) \to 0$, which indicates that any information of input signal x cannot be obtained from output y if the noise in communication channel is extremely large. We have $I(x : y) \to \infty$ in the limit $\sigma_P^2 \to \infty$, which means that the power of input is needed to send much information.

In the continuous case, the definition of the channel capacity is modified with the power constraint:

$$C = \sup_{\langle x^2 \rangle \leq \sigma_P^2} I(x : y). \tag{2.34}$$

The channel capacity C is given by the mutual information with the initial Gaussian distribution Eq. (2.33),

$$C = \frac{1}{2} \ln\left(1 + \frac{\sigma_P^2}{\sigma_N^2}\right). \tag{2.35}$$

To show this fact, we prove that the mutual information $I_q(x : y) = \langle \ln p(y|x) - \ln \int dx p(y|x) q(x) \rangle$ for any initial distribution $q(x)$ with $\langle x^2 \rangle = \sigma_{P'}^2 \leq \sigma_P^2$ is always lower than the mutual information for a Gaussian initial distribution Eq. (2.33).

$$\begin{aligned} I(x : y) - I_q(x : y) &= -\langle \ln p_P(y) \rangle_{p_P} + \langle \ln q(y) \rangle_q \\ &\geq -\langle \ln p_{P'}(y) \rangle_{p_{P'}} + \langle \ln q(y) \rangle_q \\ &= -\langle \ln p_{P'}(y) \rangle_q + \langle \ln q(y) \rangle_q \end{aligned}$$

$$= D_{\mathrm{KL}}(q(y)||p_{P'}(y))$$
$$\geq 0, \tag{2.36}$$

where $p_P(x, y) := p(y|x)p_P(x)$, $q(x, y) := p(y|x)q(x)$, and we use $\langle -\ln p_P(y)\rangle_{p_P} = 2^{-1}\ln[2\pi(\sigma_P^2 + \sigma_N^2)] + 2^{-1}$.

2.2.2 Noisy-Channel Coding Theorem

We next review the noisy-channel coding theorem, which is the basic theorem of information theory stated by Shannon in his original paper [1].

Here, we consider the situation of information transmission through a noisy communication channel (see also Fig. 2.2). First, the input message is encoded to generate a sequence of the signal (e.g., 0010101010). Second, this signal is transmitted through a noisy communication channel. Finally, the output signal (e.g., 0011101011) is decoded to generate the output message.

To archive the exact information transmission through a noisy communication channel, the length of encoding should be sufficiently large to correct the error in output signal. The noisy-channel coding theorem states the relationship between the length of encoding (i.e., the archivable rate) and the noise in the communication channel (i.e., the channel capacity). Strictly speaking, the noisy-channel coding theorem contains two statements, the noisy-channel coding theorem and the converse to the noisy-channel coding theorem. The former states the existence of coding, and the latter states an upper bound of the coding length. In this section, we introduce the noisy-channel coding theorem for a simple setting.

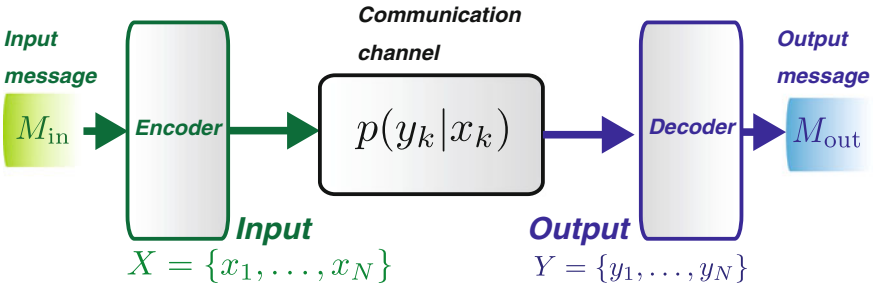


Fig. 2.2 Schematic of the communication system. To send an input message M_{in} through a noisy artificial communication channel, the input message should be encoded in a redundant sequence of bits by a channel coding protocol, and the encoded bits sequence X is transmitted through a noisy communication channel $p(y_k|x_k)$ (e.g., a Gaussian channel). The output sequence Y does not necessary coincide with the input sequence X, because of the noise in the communication channel. However, if the redundancy N of the encoded bit sequence is sufficiently large, one can recover the original input message $M_{\mathrm{out}} = M_{\mathrm{in}}$ correctly from the output sequence Y. This is a sketch of the channel coding

Let the input message be $M_{\text{in}} \in \{1, \ldots, M\}$, where M denotes the number of types of the message. The input message is assumed to be uniformly distributed: $p(M_{\text{in}}) = 1/M$ $(M_{\text{in}} = 1, \ldots, M)$. By the encoder, the input message is encoded as a discrete sequence $X(M_{\text{in}}) := \{x_1(M_{\text{in}}), \ldots, x_N(M_{\text{in}})\}$. Through a noisy communication channel defined as the conditional probability $p(y_k|x_k)$ (e.g., the binary symmetric channel), the output signal $Y = \{y_1, \ldots, y_N\}$ is stochastically obtained from the input signal X:

$$p(Y|X(M_{\text{in}})) = \prod_{k=1}^{N} p(y_k|x_k(M_{\text{in}})), \tag{2.37}$$

which represents a discrete memoryless channel. The output message $M_{\text{out}}(Y)$ is a function of the output signal Y. We define the rate as

$$R := \frac{\ln M}{N}, \tag{2.38}$$

which represents the encoding length N to describe the number of the input messages M. A code (M, N) indicates (i) an index set $\{1, \ldots, M\}$, (ii) an encoding function $X(M_{\text{in}})$ and (iii) a decoding function $M_{\text{out}}(Y)$. A code (e^{NR}, N) means $(\lceil e^{NR} \rceil, N)$, where $\lceil \ldots \rceil$ denotes the ceiling function. Let the arithmetic average probability of error P_e for a code (e^{NR}, N) be

$$P_e := \frac{1}{M} \sum_{j}^{M} p(M_{\text{out}}(Y) \neq j | X(M_{\text{in}} = j)). \tag{2.39}$$

In this setting, we have the noisy-channel coding theorem.

Theorem (Noisy-channel coding theorem) *(i) For every rate $R < C$, there exists a code (e^{NR}, N) with $P_e \to 0$.*

(ii) Conversely, any code (e^{NR}, N) with $P_e \to 0$ must have $R \leq C$.

The converse theorem (ii) can be easily proved using the nonnegativity of the relative entropy. Here, we show the proof of the converse theorem. From the initial distribution $p(M_{\text{in}})$, we have

$$\begin{aligned} NR &= S(M_{\text{in}}) \\ &= I(M_{\text{in}} : M_{\text{out}}) + S(M_{\text{in}}|M_{\text{out}}). \end{aligned} \tag{2.40}$$

We introduce a binary state E: $E := 0$ for $M_{\text{in}} = M_{\text{out}}$ and $E := 1$ for $M_{\text{in}} \neq M_{\text{out}}$. From $S(E|M_{\text{in}}, M_{\text{out}}) = 0$, we have

$$\begin{aligned} S(M_{\rm in}|M_{\rm out}) &= S(M_{\rm in}|M_{\rm out}) + S(E|M_{\rm in}, M_{\rm out}) \\ &= S(E|M_{\rm out}) + S(M_{\rm in}|E, M_{\rm out}) \\ &\leq \ln 2 + P_e N R, \end{aligned} \tag{2.41}$$

where we use $S(M_{\rm in}|E, M_{\rm out}) = P_e S(M_{\rm in}|E = 0, M_{\rm out}) \leq P_e S(M_{\rm in})$, and $S(E|M_{\rm out}) \leq S(E) \leq \ln 2$. This inequality (2.41) is well known as Fano's inequality.

The Markov property $p(M_{\rm in}, X, Y, M_{\rm out}) = p(M_{\rm in})p(X|M_{\rm in})p(Y|X)p(M_{\rm out}|Y)$ is satisfied in this setting. From the Markov property, we have $I(M_{\rm in} : M_{\rm out}|Y) = 0$, $I(M_{\rm in} : Y|X) = 0$, and

$$\begin{aligned} I(M_{\rm in} : M_{\rm out}) &\leq I(M_{\rm in} : Y) + I(M_{\rm in} : M_{\rm out}|Y) \\ &\leq I(X : Y) + I(M_{\rm in} : Y|X) \\ &= I(X : Y) \\ &= \langle \ln p(Y) \rangle - \sum_k \langle \ln p(x_k|y_k) \rangle \\ &\leq \sum_k I(x_k : y_k) \\ &\leq NC. \end{aligned} \tag{2.42}$$

The inequality using the Markov property (e.g., Eq. (2.42)) is well known as the data processing inequality. From Eqs. (2.40)–(2.42), we have

$$R \leq C + \frac{\ln 2}{N} + P_e R. \tag{2.43}$$

For sufficiently large N, we have $\ln 2/N \to 0$. Thus we have proved the converse to the noisy-channel coding theorem, which indicates that the channel capacity C gives a bound of the archivable rate R with $P_e \to 0$. We add that the mutual information $I(M_{\rm in} : M_{\rm out})$ (or $I(X : Y)$) also becomes a tighter bound of the rate R with $P_e \to 0$ from Eq. (2.42).

References

1. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
2. T. Schreiber, Measuring information transfer. Phys. Rev. Lett. **85**, 461 (2000)
3. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991)
4. R.V.L. Hartley, Transmission of information. Bell Syst. Tech. J. **7**, 535–563 (1928)
5. J. Massey, Causality, feedback and directed information. in Proceedings of International Symposium on Information Theory and Its Applications, pp. 303–305 (1990)
6. K. Kaneko, Lyapunov analysis and information flow in coupled map lattices. Physica D **23**, 436 (1986)

7. J.A. Vanstano, H.L. Swinney, Information transport in spatiotemporal systems. Phys. Rev. Lett. **60**, 1733 (1988)
8. L. Barnett, A.B. Barrett, A.K. Seth, Granger causality and transfer entropy are equivalent for Gaussian variables. Phys. Rev. Lett. **103**, 238701 (2009)
9. C.W.J. Granger, Investigating causal relations by econometric and cross-spectral methods. Econometrica **37**, 424438 (1969)
10. A. Hlavackova-Schindler, M. Palus, M. Vejmelka, Causality detection based on information-theoretic approaches in time series analysis. Phys. Rep. **441**, 1–46 (2007)
11. M. Staniek, K. Lehnertz, Symbolic transfer entropy. Phys. Rev. Lett. **100**, 158101 (2008)
12. L. Barnett, J.T. Lizier, M. Harre, A.K. Seth, T. Bossomaier, Information flow in a kinetic Ising model peaks in the disordered phase. Phys. Rev. Lett. **111**, 177203 (2013)
13. A.E. Allahverdyan, J. Dominik, M. Guenter, Thermodynamic efficiency of information and heat flow. J. Stat. Mech. P09011 (2009)
14. M. Prokopenko, J.T. Lizier, D.C. Price, On thermodynamic interpretation of transfer entropy. Entropy **15**, 524 (2013)
15. N. Ay, D. Polani, Information flows in causal networks. Adv. Complex Syst. **11**, 17 (2008)



http://www.springer.com/978-981-10-1662-2

Information Thermodynamics on Causal Networks and its Application to Biochemical Signal Transduction

Ito, S.

2016, XIII, 133 p. 32 illus., 28 illus. in color., Hardcover
ISBN: 978-981-10-1662-2