Quantities of Information: Definitions 2

This lecture and the next one introduce definitions and key properties of the most important quantities of information, including entropy, relative entropy, and mutual information. These quantities of information provide basic tools for the study of engineering problems in the subsequent lectures, and meanwhile they also constitute a coherent mathematical unity in their own right.

2.1 Discrete Probability Theory

We provide a sketchy review of basic probablity theory for discrete random variables. Our treatment of information theory will be confined within the discrete case, until Lectures 10 and 11, where continuous-valued random variables will be considered.

A probability space consists of three parts:

- A sample space, Ω, which is the set of all outcomes of the random experiment in question. For the discrete case, Ω is a finite or countably infinite set, i.e., $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, with n possibly being ∞.
- A set of events, $\mathcal{F}$, each element of which is a subset of Ω, satisfying:
  - at least Ω (certain event) and $\emptyset$ (impossible event) are contained in $\mathcal{F}$;
  - if $A \in \mathcal{F}$ then its complementary $A^c \in \mathcal{F}$;
  - for any finite collection of elements of $\mathcal{F}$, $A_1, A_2, \dots, A_n$, $\bigcup_{i=1}^n A_i \in \mathcal{F}$, and for any sequence of elements of $\mathcal{F}$, $A_1, A_2, \dots$, $\bigcup_{i=1}^\infty A_i \in \mathcal{F}$.
- A probability measure, P, which is a function assigning to each element of $\mathcal{F}$ a real number in $[0, 1]$, satisfying:
  - $P(\Omega) = 1$;
  - $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$, $\forall A_1, A_2, \dots, A_n \in \mathcal{F}$, s.t. $\forall i \neq j, A_i \cap A_j = \emptyset$;
  - $P(\bigcup_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$, $\forall A_1, A_2, \dots \in \mathcal{F}$, s.t. $\forall i \neq j$, $A_i \cap A_j = \emptyset$.

Definition 2.1 Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable is a mapping from the sample space Ω into a specified finite or countably infinite set.

Example 2.1 Let $\Omega = \{\omega_1, \omega_2\}$. We may define the following random variables.

- Bernoulli: $X(\omega_1) = 0$, $X(\omega_2) = 1$.
- Toss a coin: $Y(\omega_1) =$ head, $Y(\omega_2) =$ tail.
- Fruits: $Z(\omega_1) =$ apple, $Z(\omega_2) =$ banana.

So we may understand a probability space as an abstract structure, and endow any specific interpretation to it via defining a suitable random variable. A probability space may induce different random variables, and their underlying probability measures are identical.

We can then define a probability distribution with respect to a random variable and its underlying probability measure.

Definition 2.2 Given a random variable X on a probability space $(\Omega, \mathscr{F}, P)$, its probability distribution is

$$P_X(x) = P(\{\omega : X(\omega) = x\}), \tag{2.1}$$

where x is taken over $X(\Omega)$, the range of X. Since we consider discrete probability spaces, P_X is also called the probability mass function (pmf) of X. In later lectures, sometimes we also denote $X(\Omega)$ by $\mathscr{X}$, and call it the alphabet of X.

It is immediate that P_X satisfies

1. $P_X(x) \geq 0, \forall x \in X(\Omega)$;
2. $\sum_{x \in X(\Omega)} P_X(x) = 1$.

When a collection of random variables on a common probability space is considered, it is of interest to study their joint behavior.

Definition 2.3 Given a collection of random variables $X_1, X_2, \ldots, X_n$ on a probability space $(\Omega, \mathscr{F}, P)$, their joint probability distribution is defined by

$$\begin{aligned} & P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \\ = \; & P(\{\omega : X_1(\omega) = x_1\} \cap \{\omega : X_2(\omega) = x_2\} \cap \ldots \\ & \qquad \ldots \cap \{\omega : X_n(\omega) = x_n\}), \end{aligned} \tag{2.2}$$

where $(x_1, x_2, \ldots x_n)$ is taken over $X_1(\Omega) \times X_2(\Omega) \times \ldots \times X_n(\Omega)$.

The joint probability distribution satisfies

1. $P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \geq 0, \forall (x_1, x_2, \ldots x_n) \in X_1(\Omega) \times X_2(\Omega) \times \ldots \times X_n(\Omega)$;
2. $$\sum_{(x_1,x_2,\ldots x_n) \in X_1(\Omega)\times X_2(\Omega)\times\ldots\times X_n(\Omega)} P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = 1;$$

3. (Marginalization) for any $1 \leq i \leq n$,

$$\begin{aligned}&\sum_{x_i \in X_i(\Omega)} P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \\ = \; & P_{X_1,X_2,\ldots,X_{i-1},X_{i+1},\ldots X_n}(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).\end{aligned}$$

For a collection of random variables on a common probability space, it is also of interest to study the behavior of some of them, with the remaining fixed. This is described by the conditional probability distribution.

Definition 2.4 Given random variables X and Y on a probability space $(\Omega, \mathcal{F}, P)$, for $x \in X(\Omega)$ with $P_X(x) > 0$, the conditional probability distribution of Y conditioned upon $\{\omega : X(\omega) = x\}$ is defined by

$$P_{Y|X}(y|x) = \frac{P_{X,Y}(x, y)}{P_X(x)}, \tag{2.3}$$

and for $x \in X(\Omega)$ with $P_X(x) = 0$, $P_{Y|X}(y|x)$ is undefined.

From the definition of conditional probability distribution, the celebrated Bayes' rule is an immediate consequence.

Theorem 2.1 Considering random variables X and Y on a probability space $(\Omega, \mathcal{F}, P)$, the Bayes' rule states that for any $y \in Y(\Omega)$ with $P_Y(y) > 0$,

$$P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x) P_X(x)}{P_Y(y)}. \tag{2.4}$$

Independence is a special type of joint behavior among random variables, and is defined as follows.

Definition 2.5 Random variables $X_1, X_2, \ldots, X_n$ on a probability space $(\Omega, \mathcal{F}, P)$ are said to be mutually independent if

$$P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P_{X_1}(x_1) P_{X_2}(x_2) \ldots P_{X_n}(x_n), \tag{2.5}$$

for any $(x_1, x_2, \ldots x_n) \in X_1(\Omega) \times X_2(\Omega) \times \ldots \times X_n(\Omega)$; and are said to be pairwise independent if for any pair of $i \neq j$,

$$P_{X_i,X_j}(x_i, x_j) = P_{X_i}(x_i) P_{X_j}(x_j), \tag{2.6}$$

for any $(x_i, x_j) \in X_i(\Omega) \times X_j(\Omega)$.

Note that only when $n = 2$, pairwise independence and mutual independence are equivalent. As a simple example, let X and Z be mutually independent Bernoulli random variables, with $P_X(0) = P_X(1) = \frac{1}{2}$, $P_Z(0) = P_Z(1) = \frac{1}{2}$, and let Y be the modulo-two sum of X and Z, $Y = X \oplus Z$. It can be readily verified that X, Y, Z are pairwise independent, but not mutually independent.

An important case is where $X_1, X_2, \ldots, X_n$ are mutually independent, and $P_{X_1}, P_{X_2}, \ldots, P_{X_n}$ are identical. We call these n random variables independent and identically distributed (i.i.d.).

Conditional independence is a crucial concept with important applications in information theory.

Definition 2.6 For random variables X, Y and Z on a probability space $(\Omega, \mathscr{F}, P)$, if for any $(x, y, z) \in X(\Omega) \times Y(\Omega) \times Z(\Omega)$,

$$P_{X,Z|Y}(x, z|y) = P_{X|Y}(x|y)P_{Z|Y}(z|y), \tag{2.7}$$

then X and Z are said to be conditionally independent given Y. This relationship can be represented as $X \leftrightarrow Y \leftrightarrow Z$, and is called a Markov chain.

Next we proceed to introduce the expectation of random variables.

Definition 2.7 For a random variable X on a probability space $(\Omega, \mathscr{F}, P)$, and a function $F : X(\Omega) \mapsto \mathbb{R}$, the expectation of $F(X)$ is defined as

$$\mathbf{E}[F(X)] = \sum_{x \in X(\Omega)} F(x)P_X(x). \tag{2.8}$$

Similarly, for a collection of random variables $X_1, X_2, \ldots, X_n$ on a probability space $(\Omega, \mathscr{F}, P)$, and a function

$$F : X_1(\Omega) \times X_2(\Omega) \times \ldots \times X_n(\Omega) \mapsto \mathbb{R}, \tag{2.9}$$

the expectation of $F(X_1, X_2, \ldots, X_n)$ is defined as

$$\mathbf{E}[F(X_1, X_2, \ldots, X_n)] = \sum_{(x_1,x_2,\ldots,x_n) \in X_1(\Omega)\times X_2(\Omega)\times\ldots\times X_n(\Omega)} F(x_1, x_2, \ldots, x_n)P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n). \tag{2.10}$$

The following basic properties of expectation will be useful.

Theorem 2.2 For random variables X and Y on a probability space $(\Omega, \mathscr{F}, P)$, and functions $F : X(\Omega) \mapsto \mathbb{R}$ and $G : Y(\Omega) \mapsto \mathbb{R}$, we have

- Linearity:

$$\mathbf{E}[F(X) + G(X)] = \mathbf{E}[F(X)] + \mathbf{E}[G(X)]. \tag{2.11}$$

- Scaling: for any $c \in \mathbb{R}$,

$$\mathbf{E}[cF(X)] = c\mathbf{E}[F(X)]. \tag{2.12}$$

- If X and Y are independent,

$$\mathbf{E}[F(X)G(Y)] = \mathbf{E}[F(X)]\mathbf{E}[G(Y)]. \tag{2.13}$$

Regarding the third property (2.13), if $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$, then we say that X and Y are uncorrelated. Note that independence implies uncorrelatedness, but the converse is not true in general.

Indicator function is usually convenient for derivations. Consider a probability space $(\Omega, \mathcal{F}, P)$. For any event $A \in \mathcal{F}$, define a function E_A over Ω as $E_A(\omega) = 1$ if $\omega \in A$ and 0 otherwise. This can also be written as $E_A(\omega) = \mathbf{1}_{\{\omega \in A\}}$. Then E_A is a random variable indicating whether the event A occurs, and we have $\mathbf{E}[E_A] = P(A)$.

We can further define conditional expectation as follows.

Definition 2.8 For random variables X and Y on a probability space $(\Omega, \mathcal{F}, P)$, and a function $F : X(\Omega) \mapsto \mathbb{R}$, the conditional expectation of $F(X)$ conditioned upon the event $\{\omega : Y(\omega) = y\}$, $y \in Y(\Omega)$, is defined as

$$\mathbf{E}[F(X)|y] = \sum_{x \in X(\Omega)} F(x) P_{X|Y}(x|y). \tag{2.14}$$

It should be kept in mind that $\mathbf{E}[F(X)|Y]$ itself is a random variable, induced by the random variable Y.

The following property of conditional expectation, called the law of total expectation, is useful.

Theorem 2.3 For random variables X and Y on a probability space $(\Omega, \mathcal{F}, P)$, we have

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]]. \tag{2.15}$$

At this point, we introduce the convergence of random variables. There are several different notions of convergence for random variables, but for most of our purposes in the lecture notes we only need a weak form of convergence, as follows.

Definition 2.9 For a sequence of random variables, $X_1, X_2, \ldots$, if there exists a random variable X such that for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - X| < \epsilon) = 1, \tag{2.16}$$

then $X_1, X_2, \dots$ converge in probability to X, denoted as "$X_n \xrightarrow{P} X$" or "$X_n \to X$ in probability".

When $X_1, X_2, \dots$ are i.i.d. random variables, the following weak law of large numbers (WLLN) holds.

Theorem 2.4 For a sequence of i.i.d. random variables, $X_1, X_2, \dots$, define $Y_n = (X_1 + X_2 + \dots + X_n)/n$, $n = 1, 2, \dots$. Then we have $Y_n \to \mathbf{E}[X]$ in probability.

The WLLN only suggests that the probability of Y_n being close to $\mathbf{E}[X]$ is arbitrarily close to one for any sufficiently large n, but it does not exclude the possibility that some particular trajectory of Y_n, $n = 1, 2, \dots$, eventually deviates from $\mathbf{E}[X]$. This possibility is excluded by the strong law of large numbers (SLLN), which states that the trajectory Y_n, $n = 1, 2, \dots$, has limit $\mathbf{E}[X]$ with probability one. The SLLN is beyond the scope of our lecture notes.

2.2 Entropy, Joint Entropy, and Conditional Entropy

Consider a random variable X on a probability space $(\Omega, \mathcal{F}, P)$. We observe $X = x$ with probability $P_X(x) = P(\{\omega : X(\omega) = x\})$, and call the information or "surprise" of this observation as

$$i(x) = \log \frac{1}{P_X(x)}, \quad \forall x \in X(\Omega). \tag{2.17}$$

We define the entropy of X as the expectation of $i(X)$.

Definition 2.10 For a random variable X on a probability space $(\Omega, \mathcal{F}, P)$, its entropy is defined as

$$H(X) = \mathbf{E}[i(X)] = -\sum_{x \in X(\Omega)} P_X(x) \log P_X(x). \tag{2.18}$$

The concept of entropy in information theory has close connection with its homonymous counterpart in statistical physics, but as will be shown in next lectures, it has its own engineering interpretation, beyond its statistical physics counterpart.

The base of the logarithm in the definition of entropy can be arbitrary. Commonly used bases include 2 and e. When the base is 2, the unit of entropy is bit; when the base is e, the unit of entropy is nat. It can be easily verified that 1 nat is equal to $\log_2 e \approx 1.443$ bits. Another base, rarely used, is 10, and the corresponding unit of entropy is hartley. Engineers may prefer the usage of bits for practical purposes, because in digital circuits binary states are universal, but for information theoretic study it is usually more convenient to work with nats, thanks to the good analytical property of the ln function.

In the definition of entropy, the summation can skip over x with $P_X(x) = 0$, justified by the limit $\lim_{p \to 0^+} p \log p = 0$. In other words, we may adopt the convention of $0 \log 0 = 0$.

Example 2.2 For Bernoulli random variable X, $X(\Omega) = \{0, 1\}$, $P_X(0) = 1 - \epsilon$ and $P_X(1) = \epsilon$, we have

$$\begin{aligned} H(X) &= -\sum_{x \in \{0,1\}} P_X(x) \log P_X(x) \\ &= -(1-\epsilon)\log(1-\epsilon) - \epsilon \log \epsilon. \end{aligned} \tag{2.19}$$

We usually denote this entropy by $h_2(\epsilon)$, where the subscript 2 means that the random variable is binary. We plot in Figure 2.1 $h_2(\epsilon)$ in bits. Note that $h_2(\epsilon)$ is symmetric with respect to $\epsilon = 0.5$, i.e., $h_2(\epsilon) = h_2(1-\epsilon)$. The following special points on the graph can be useful: $h_2(0) = h_2(1) = 0$, $h_2(0.5) = 1$, and $h_2(0.11) = h_2(0.89) \approx 0.5$ (all in bits). Expanding $h_2(\epsilon)$ near $\epsilon = 0$, we have $h_2(\epsilon) = \epsilon \ln \frac{1}{\epsilon} + \epsilon + o(\epsilon)$ (in nats), whose first term is dominant. Graphically, this implies that the slope of $h_2(\epsilon)$ at 0 is infinite.

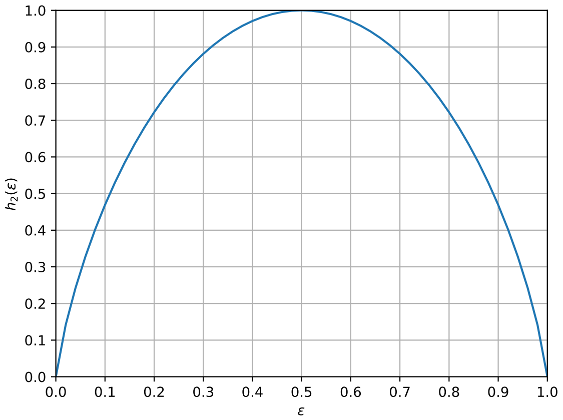

Figure 2.1: Entropy (in bits) of Bernoulli random variable.

Example 2.3 Consider a geometric random variable X, $X(\Omega) = \{1, 2, \ldots\}$, with $P_X(x) = \epsilon(1-\epsilon)^{x-1}$. Such a random variable can be interpreted as the number of trials when the first "1" occurring in a sequence of i.i.d. Bernoulli(ϵ) random variables. It can be verified that the expectation of X is $\frac{1}{\epsilon}$. The entropy of X can then be calculated as

$$\begin{aligned} H(X) &= -\mathbf{E}\log P_X(X) \\ &= -\mathbf{E}\log \epsilon(1-\epsilon)^{X-1} \\ &= -\log\epsilon - \log(1-\epsilon)\cdot \mathbf{E}[X-1] \\ &= -\log\epsilon - \log(1-\epsilon)\cdot\left(\frac{1}{\epsilon}-1\right) \\ &= \frac{-\epsilon\log\epsilon-(1-\epsilon)\log(1-\epsilon)}{\epsilon} = \frac{h_2(\epsilon)}{\epsilon}. \qquad (2.20) \end{aligned}$$

We can verify that $H(X)$ is monotonically decreasing in $\epsilon \in (0,1]$, and that $\lim_{\epsilon\to 0^+} H(X) = \infty$, as displayed in Figure 2.2.

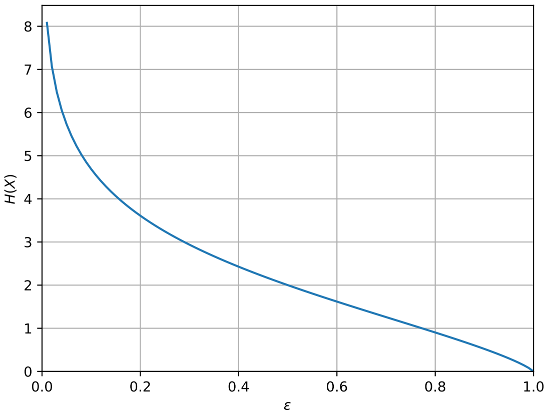

Figure 2.2: Entropy (in bits) of geometric random variable.

When there are multiple random variables, we may treat them together and define their joint entropy as follows.

Definition 2.11 For random variables X and Y on a probability space $(\Omega, \mathcal{F}, P)$, their joint entropy is defined as

$$\begin{aligned} H(X,Y) &= \mathbf{E}[i(X,Y)] \\ &= -\sum_{(x,y)\in X(\Omega)\times Y(\Omega)} P_{X,Y}(x,y)\log P_{X,Y}(x,y). \qquad (2.21) \end{aligned}$$

For more than two random variables, their joint entropy can be defined in an analogous way.

Now we consider the entropy of a random variable conditioned upon another random variable.

Definition 2.12 For random variables X and Y on a probability space $(\Omega, \mathcal{F}, P)$, the entropy of Y conditioned on the occurrence of $\{\omega : X(\omega) = x\}$ is

$$H(Y|X=x) = -\sum_{y\in Y(\Omega)} P_{Y|X}(y|x)\log P_{Y|X}(y|x). \qquad (2.22)$$

The conditional entropy of Y given X is then defined as

$$\begin{aligned} H(Y|X) &= \sum_{x\in X(\Omega)} P_X(x)H(Y|X=x) \\ &= -\sum_{(x,y)\in X(\Omega)\times Y(\Omega)} P_{X,Y}(x,y)\log P_{Y|X}(y|x) \\ &= -\mathbf{E}[\log P_{Y|X}(Y|X)], \qquad (2.23) \end{aligned}$$

where the expectation is with respect to the joint probability distribution $P_{X,Y}(x,y)$.

Example 2.4 Consider independent Bernoulli random variables X and Z, each taking values 1 and 0 with equal probability 0.5, and let $Y = X \cdot Z$, i.e., the product between X and Z. So we can calculate

$$\begin{aligned} H(X,Y) &= \sum_{(x,y)\in X(\Omega)\times Y(\Omega)} P_{X,Y}(x,y)\log\frac{1}{P_{X,Y}(x,y)} \\ &= \frac{1}{2}\log 2 + 0 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 \\ &= \frac{3}{2}\log 2, \qquad (2.24) \\ H(Y|X=0) &= h_2(0) = 0, \qquad (2.25) \\ H(Y|X=1) &= h_2(0.5) = \log 2, \qquad (2.26) \\ H(Y|X) &= \frac{1}{2}H(Y|X=0) + \frac{1}{2}H(Y|X=1) \\ &= \frac{1}{2}\log 2. \qquad (2.27) \end{aligned}$$

2.3 Relative Entropy and Mutual Information

Given two probability distributions, relative entropy (also called Kullback-Leibler distance or I-divergence) is a widely used quantity for describing the discrepancy between them.

Definition 2.13 The relative entropy between two probability distributions $P(x)$ and $Q(x)$ is defined as

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}, \qquad (2.28)$$

where $\mathcal{X}$ denotes the alphabet of x.

In general it is possible that for certain $x \in \mathcal{X}$, $P(x)$ or $Q(x)$ is zero, and we adopt the following convention when calculating $D(P\|Q)$:

- $0\log\frac{0}{0} = 0$;
- $0\log\frac{0}{q} = 0$ for $q > 0$;

- $p \log \frac{p}{0} = \infty$ for $p > 0$.

Example 2.5 Consider two Bernoulli distributions P and Q, with $P_X(0) = 1 - \epsilon$, $P_X(1) = \epsilon$, $Q_X(0) = 1 - \delta$, $Q_X(1) = \delta$. We have

$$D(P\|Q) = (1-\epsilon)\log\frac{1-\epsilon}{1-\delta} + \epsilon\log\frac{\epsilon}{\delta}, \qquad (2.29)$$

$$D(Q\|P) = (1-\delta)\log\frac{1-\delta}{1-\epsilon} + \delta\log\frac{\delta}{\epsilon}. \qquad (2.30)$$

It is interesting to note that whenever $\epsilon > 0$ and $\delta = 0$, $D(P\|Q) = \infty$. This example also clearly shows that in general, $D(P\|Q) \neq D(Q\|P)$; that is, the relative entropy is asymmetric with respect to its two probability distributions.

Example 2.6 Consider three Bernoulli distributions P, Q, and Q', with $P_X(0) = 1$, $P_X(1) = 0$, $Q_X(0) = \frac{1}{2}$, $Q_X(1) = \frac{1}{2}$, $Q'_X(0) = \frac{1}{4}$, $Q'_X(1) = \frac{3}{4}$. Then we have $D(P\|Q) = \log 2$, $D(P\|Q') = 2\log 2$, and $D(Q\|Q') = \log 2 - \frac{1}{2}\log 3$. Therefore, $D(P\|Q) + D(Q\|Q') - D(P\|Q') = -\frac{1}{2}\log 3 < 0$. This example hence shows that the relative entropy does not satisfy the triangle inequality in general.

Now we are ready to define the mutual information between two random variables with a joint probability distribution, as a specialized relative entropy.

Definition 2.14 For random variables X and Y on a probability space $(\Omega, \mathcal{F}, P)$, their mutual information is defined as

$$I(X;Y) = D(P_{X,Y}\|P_X P_Y) \qquad (2.31)$$

$$= \sum_{(x,y)\in X(\Omega)\times Y(\Omega)} P_{X,Y}(x,y)\log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}. \qquad (2.32)$$

For any $(x, y) \in X(\Omega) \times Y(\Omega)$, we define its associated information density as*

$$i(x;y) = \log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}. \qquad (2.33)$$

We see that the mutual information $I(X;Y)$ is the expecation of $i(X;Y)$.

When there is an additional random variable Z, the conditional mutual information between X and Y, given Z, is defined as follows.

* Note that we have reused the symbol i for both information or "surprise" in (2.17) and information density here.

Definition 2.15 The conditional mutual information between X and Y given Z is

$$I(X;Y|Z) = \sum_{(x,y,z)\in X(\Omega)\times Y(\Omega)\times Z(\Omega)} P_{X,Y,Z}(x,y,z) \log \frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}. \quad (2.34)$$

2.4 Entropy Rate

In our lecture notes, let us be content with viewing a discrete-time stochastic process as a sequence of random variables, on a certain probability space. Starting from time index 1, a stochastic process **X** is denoted as $X_1, X_2, \ldots$. The first n random variables in **X** has joint entropy $H(X_1, X_2, \ldots, X_n)$, and we examine the trend of $H(X_1, X_2, \ldots, X_n)$ as n grows without bound. This leads to the definition of entropy rate of a stochastic process.

Definition 2.16 For a stochastic process **X**, its entropy rate is defined by

$$H(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n), \quad (2.35)$$

when the limit exists.

Analogously, for two stochastic processes **X** and **Y**, we can define their mutual information rate by

$$I(\mathbf{X};\mathbf{Y}) = \lim_{n\to\infty} \frac{1}{n} I(X_1, \ldots, X_n; Y_1, \ldots, Y_n), \quad (2.36)$$

when the limit exists.

Properties and examples of entropy rate will be provided in the next lecture.

Notes

Both entropy and mutual information first appeared in Shannon's landmark paper [1], whereas therein the term "mutual information" was not used. Instead, Shannon called conditional entropy "equivocation" and worked with the difference between entropy $H(X)$ and equivocation $H(X|Y)$, which is exactly the mutual information $I(X;Y)$, as will be shown in our next lecture. The name "mutual information" appeared a few years later after the publication of [1] (see, e.g., [4]). Rumor has it that John von Neumann suggested to use the name "entropy", but Shannon in an interview clarified that this was not the case. The unit “bit” is “binary digit” for short, a name suggested by John W. Tukey, one of the inventors of the fast Fourier transform.

Before Shannon, Harry Nyquist around 1924 argued that the transmission rate of a communication system should be proportional to the logarithm of the number of signal levels in a unit duration, and Ralph Hartley around 1928 proposed to quantitatively measure information of a variable as the logarithm of the size of its alphabet. For example, for a jar containing balls of two colors (black and white), the amount of information in drawing a ball from the jar is $\log 2$. These thoughts may be viewed as a precursor of entropy, but they were certainly inadequate as they did not take the probabilistic nature into consideration.

There are a number of generalizations of entropy. As an example, the Rényi entropy of order α, named after Hungarian mathematician Alfréd Rényi, where $\alpha \geq 0$ and $\alpha \neq 1$, for a random variable X, is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{x \in X(\Omega)} P_X(x)^\alpha \right). \tag{2.37}$$

As $\alpha \longrightarrow 1$, the Rényi entropy converges to the entropy we have introduced in Definition 2.10.

The information density $i(x; y)$ was first used by the Russian mathematician Mark S. Pinsker in his study of quantities of information [5], and this, over the years, has led to the information spectrum approach to information theory (see, e.g., [6]).

The relative entropy was first introduced by American mathematicians Solomon Kullback and Richard Leibler [7], and has found extensive applications in information theory, statistics and machine learning.

Exercises

1. For a probability space $(\Omega, \mathcal{F}, P)$, prove the following properties:
   a) $P(\emptyset) = 0$.
   b) For any $A, B \in \mathcal{F}$, if $A \subseteq B$ then $P(A) \leq P(B)$.
   c) For any $A, B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
2. For discrete random variables X and Y over a probability space $(\Omega, \mathcal{F}, P)$,
   a) Prove the law of total expectation,

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]].$$

b) Prove the law of total variance,

$$\mathrm{var}X = \mathbf{E}[\mathrm{var}[X|Y]] + \mathrm{var}\mathbf{E}[X|Y].$$

3. Let X_1, X_2, X_3, X_4 be random variables such that $X_1 \leftrightarrow (X_2, X_3) \leftrightarrow X_4$ and $X_1 \leftrightarrow (X_2, X_4) \leftrightarrow X_3$ simultaneously hold.

   a) If $P_{X_1,X_2,X_3,X_4}(x_1, x_2, x_3, x_4) > 0$, for any $(x_1, x_2, x_3, x_4) \in X_1(\Omega)\times X_2(\Omega)\times X_3(\Omega)\times X_4(\Omega)$, prove that $X_1 \leftrightarrow X_2 \leftrightarrow (X_3, X_4)$ holds.

   b) Can you give an example, where for some (x_1, x_2, x_3, x_4), $P_{X_1,X_2,X_3,X_4}(x_1, x_2, x_3, x_4) = 0$, and $X_1 \leftrightarrow X_2 \leftrightarrow (X_3, X_4)$ does not hold? This illustrates the delicacy of probability distributions with strictly zero probability [8, Prop. 2.12].

4. Prove the following basic inequalities:

   a) Markov's inequality: for a nonnegative random variable X with finite expectation, and any $a > 0$,

   $$P(X \geq a) \leq \frac{\mathbf{E}[X]}{a}. \tag{2.38}$$

   b) Chebyshev's inequality: for a random variable X with finite expectation and variance, and any $a > 0$,

   $$P(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathrm{var}X}{a^2}. \tag{2.39}$$

   c) Chernoff's inequality: for a random variable X and any a,

   $$P(X \geq a) \leq \min_{\lambda \geq 0} e^{-\lambda a}\mathbf{E}[e^{\lambda X}]. \tag{2.40}$$

5. If we model a pair of random variables X and Y with weak dependence as $P_{X,Y}(x, y) = P_X(x)P_Y(y)(1 + \epsilon(x, y))$, such that there exists $\delta < 1$ satisfying $|\epsilon(x, y)| \leq \delta$, $\forall(x, y) \in X(\Omega) \times Y(\Omega)$, can you provide an upper bound on the difference between $H(X, Y)$ and $H(X) + H(Y)$?

6. Cross entropy is an important concept in machine learning, usually used as objective function when training neural networks for classification tasks. For two probability distributions $P(x)$ and $Q(x)$ with domain $\mathcal{X}$, the cross entropy of $Q(x)$ relative to $P(x)$ is defined as

   $$H_c(P, Q) = -\sum_{x\in\mathcal{X}} P(x) \log Q(x).$$

   Calculate the cross entropy $H_c(P, Q)$ when $P(x)$ and $Q(x)$ are geometric distributions with parameters ϵ_P and ϵ_Q, respectively.

7. For a random variable whose range has size m, we may denote its pmf as a vector of elements $\{p_i\}_{i=1}^m$, and denote its entropy as $H(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^m p_i \log p_i$. Verify the following properties of entropy:

    a) expansibility: $H(p_1, p_2, \ldots, p_m, 0) = H(p_1, p_2, \ldots, p_m)$.

    b) additivity:

    $$\begin{aligned} & H(p_1, p_2, \ldots, p_m) + H(q_1, q_2, \ldots, q_n) \\ = \; & H(p_1 q_1, \ldots, p_1 q_n, p_2 q_1, \ldots, p_m q_1, \ldots, p_m q_n). \end{aligned}$$

    c) grouping:

    $$\begin{aligned} H(p_1, p_2, \ldots, p_m, q_1, q_2, \ldots, q_n) &= H\left(\sum_{i=1}^m p_i, \sum_{j=1}^n q_j\right) \\ &+ \left(\sum_{i=1}^m p_i\right) H\left(\frac{p_1}{\sum_{i=1}^m p_i}, \frac{p_2}{\sum_{i=1}^m p_i}, \ldots, \frac{p_m}{\sum_{i=1}^m p_i}\right) \\ &+ \left(\sum_{j=1}^n q_j\right) H\left(\frac{q_1}{\sum_{j=1}^n q_j}, \frac{q_2}{\sum_{j=1}^n q_j}, \ldots, \frac{q_n}{\sum_{j=1}^n q_j}\right). \end{aligned}$$

    These are among a variety of "axiomatic" properties. It can be proved that from a suitable collection of such axiomatic properties, the definition of the entropy function is unique except for a scaling factor; see [9] for a summary of such type of results.

8. Consider independent random variables X and Y, each uniformly distributed over $\{1, 2, \ldots, n\}$.

    a) Use computer to numerically study $H(X + Y)$ and plot its growth with n.

    b) Use computer to numerically study $H(X \cdot Y)$ and plot its growth with n.

Quantities of Information: Properties 3

This lecture continues our exploration of quantities of information, developing their key properties. When random variables are given, their entropy and mutual information can be calculated according to their definitions introduced in Lecture 2. As will be shown in this lecture, however, there are a number of useful properties of entropy and mutual information, which not only greatly facilitate the calculation, but also play key roles when studying information theoretic problems in our subsequent lectures.

3.1 Chain Rules

Chain rules provide a useful tool for decomposing entropy or mutual information involving multiple random variables into multiple terms each of which invovles a smaller number of random variables. Mathematically, this decomposition is no surprise, as a direct consequence of the property of logarithmic function; that is, the logarithm of the product of several variables is equal to the sum of the logarithms of individual variables.

Theorem 3.1 The chain rule of entropy:

- (Basic form) For two random variables X and Y, we have

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \qquad (3.1)$$

- (Conditional form) For three random variables X, Y and Z, we have

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z). \qquad (3.2)$$

- (General form) For n random variables $X_1, X_2, \dots, X_n$, we have

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1), \qquad (3.3)$$

  where X_0 is understood as a degenerated random variable, say, a constant, and hence $H(X_1|X_0) = H(X_1)$.

Proof: We only prove the basic form. The other two forms can be proved analogously and are left as exercise. From the definition of joint entropy (Definition 2.11), we have

$$\begin{aligned} H(X,Y) &\overset{(a)}{=} \mathbf{E}\left[\log \frac{1}{P_{X,Y}(X,Y)}\right] \\ &= \mathbf{E}\left[\log \frac{1}{P_X(X)P_{Y|X}(Y|X)}\right] \\ &= \mathbf{E}\left[\log \frac{1}{P_X(X)} + \log \frac{1}{P_{Y|X}(Y|X)}\right] \\ &\overset{(b)}{=} \mathbf{E}\left[\log \frac{1}{P_X(X)}\right] + \mathbf{E}\left[\log \frac{1}{P_{Y|X}(Y|X)}\right] \\ &\overset{(c)}{=} H(X) + H(Y|X), \qquad (3.4) \end{aligned}$$

where (a) is by noting that $i(x,y) = \log \frac{1}{P_{X,Y}(x,y)}$, (b) is from the linearity property of expectation (Theorem 2.2), and in (c) we change the probability distribution in the first expectation from $P_{X,Y}$ to P_X. This shows that $H(X,Y) = H(X) + H(Y|X)$; the other identity, $H(X,Y) = H(Y) + H(X|Y)$, can be proved in the same way by exchanging the roles of X and Y. □

Now inspect mutual information. The following properties of mutual information are basic.

Theorem 3.2 Mutual information satisfies the following basic properties:

- (Symmetry) $I(X;Y) = I(Y;X)$.
- (Self-information) $I(X;X) = H(X)$.
- (Decomposition)

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && (3.5) \\ &= H(Y) - H(Y|X) && (3.6) \\ &= H(X) + H(Y) - H(X,Y). && (3.7) \end{aligned}$$

- (Decomposition with conditioning)

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) && (3.8) \\ &= H(Y|Z) - H(Y|X,Z) && (3.9) \\ &= H(X|Z) + H(Y|Z) - H(X,Y|Z). && (3.10) \end{aligned}$$

Proof: All these basic properties are immediate consequences of the definition of mutual information (Definition 2.14) and are therefore left as exercise. □

The chain rule of mutual information is given by the following theorem.

Theorem 3.3 For random variables $X_1, X_2, \ldots, X_n, Y$, we have

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1). \tag{3.11}$$

Proof: We start with the decomposition property in Theorem 3.2 to get

$$\begin{aligned} & I(X_1, X_2, \ldots, X_n; Y) \\ = \; & H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y). \end{aligned} \tag{3.12}$$

Then, from the chain rule of entropy (Theorem 3.1), we have

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1), \tag{3.13}$$

$$H(X_1, X_2, \ldots, X_n | Y) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1, Y). \tag{3.14}$$

So upon combining these two decompositions, we obtain

$$\begin{aligned} & I(X_1, X_2, \ldots, X_n; Y) \\ = \; & \sum_{i=1}^{n} [H(X_i | X_{i-1}, \ldots, X_1) - H(X_i | X_{i-1}, \ldots, X_1, Y)] \\ = \; & \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1). \end{aligned} \tag{3.15}$$

This proves the chain rule of mutual information. □

3.2 Non-negativity Properties

The most fundamental non-negativity property in information theory is that relative entropy is always non-negative.

Theorem 3.4 The relative entropy $D(P\|Q)$ is always non-negative, and is equal to zero if and only if $P(x) = Q(x), \forall x \in \mathcal{X}$.

Proof: It loses no generality to use the natural logarithm, and by its definition (Definition 2.13) the relative entropy between P and Q can be written as

$$\begin{aligned} D(P\|Q) &= \sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)} \\ &= \sum_{x \in \mathcal{X}'} P(x) \ln \frac{P(x)}{Q(x)}, \end{aligned} \tag{3.16}$$

where $\mathcal{X}' = \mathcal{X} \setminus \{x : P(x) = 0\}$.

Applying the inequality $\ln t \leq t - 1, \forall t \geq 0$, with equality if and only if $t = 1$, we have

$$\begin{aligned} D(P\|Q) &= -\sum_{x\in\mathcal{X}'} P(x)\ln\frac{Q(x)}{P(x)} \\ &\geq -\sum_{x\in\mathcal{X}'} P(x)\left[\frac{Q(x)}{P(x)} - 1\right] \\ &= -\sum_{x\in\mathcal{X}'} [Q(x) - P(x)] \\ &= -\sum_{x\in\mathcal{X}'} Q(x) + \sum_{x\in\mathcal{X}'} P(x) \\ &\geq -\sum_{x\in\mathcal{X}} Q(x) + \sum_{x\in\mathcal{X}} P(x) \\ &= -1 + 1 = 0. \end{aligned} \tag{3.17}$$

Inspecting the two inequalities in the steps above, the first one becomes equality if and only if $P(x) = Q(x)$ for any $x \in \mathcal{X}'$, and the second one becomes equality if and only if $Q(x) = 0$ whenever $P(x) = 0$. So in summary we have $D(P\|Q) \geq 0$, with equality if and only if $P(x) = Q(x), \forall x \in \mathcal{X}$. □

The non-negativity of relative entropy has a number of important and useful corollaries for entropy and mutual information. First, we provide general lower and upper bounds of entropy as follows.

Corollary 3.1 The entropy $H(X)$ is bounded as $0 \leq H(X) \leq \log|X(\Omega)|$. Furthermore, $H(X) = 0$ holds if and only if X is a deterministic constant, and $H(X) = \log|X(\Omega)|$ holds if and only if X obeys the uniform distribution over $X(\Omega)$.

Proof: The lower bound zero immediately follows by noting $H(X) = \mathbf{E}[i(X)]$ and $i(x) = \log\frac{1}{P_X(x)} \geq 0, \forall x \in X(\Omega)$. The lower bound zero is achieved if and only if either $i(x) = 0$ or $P_X(x) = 0$ is satisfied, $\forall x \in X(\Omega)$, leading to the requirement that X is a deterministic constant.

To prove the upper bound $\log|X(\Omega)|$, we examine the relative entropy between P_X and the uniform distribution over $X(\Omega)$, $P_{X,\mathrm{u}}$:

$$\begin{aligned} D(P_X\|P_{X,\mathrm{u}}) &= \sum_{x\in X(\Omega)} P_X(x)\log\frac{P_X(x)}{P_{X,\mathrm{u}}(x)} \\ &= \sum_{x\in X(\Omega)} P_X(x)\log\left(|X(\Omega)|P_X(x)\right) \\ &= -H(X) + \log|X(\Omega)|. \end{aligned} \tag{3.18}$$

According to Theorem 3.4, $D(P_X\|P_{X,\mathrm{u}}) \geq 0$ holds and equality is achieved if and only if P_X is also the uniform distribution over

$X(\Omega)$. So we have $H(X) \leq \log|X(\Omega)|$, with equality if and only if X obeys the uniform distribution over $X(\Omega)$. □

Another corollary of the non-negativity property of relative entropy is the following maximum entropy result.

Corollary 3.2 For a random variable X over $X(\Omega) = \{1, 2, \ldots\}$ satisfying $\mathbf{E}X = A > 1$, the geometric distribution with pmf $P_{X,g}(x) = \frac{1}{A}\left(1 - \frac{1}{A}\right)^{x-1}$ maximizes the entropy $H(X)$.

Proof: For a geometric random variable with expectation A, as calculated in Example 2.3 in Lecture 2, its entropy is

$$\begin{aligned} H(X) &= \frac{h_2(1/A)}{1/A} \\ &= A \log A - (A-1)\log(A-1). \end{aligned} \quad (3.19)$$

For an arbitrary random variable X over $X(\Omega) = \{1, 2, \ldots\}$ satisfying $\mathbf{E}X = A > 1$, let us examine the relative entropy between its probability distribution P_X and that of the geometric distribution, $P_{X,g}$:

$$\begin{aligned} D(P_X \| P_{X,g}) &= \sum_{x=1}^{\infty} P_X(x) \log \frac{P_X(x)}{P_{X,g}(x)} \\ &= -H(X) - \sum_{x=1}^{\infty} P_X(x) \log \left[\frac{1}{A}\left(1 - \frac{1}{A}\right)^{x-1} \right] \\ &= -H(X) + \log A - \log \frac{A-1}{A} \mathbf{E}[X-1] \\ &= -H(X) + A \log A - (A-1)\log(A-1).(3.20) \end{aligned}$$

Hence $H(X)$ is always upper bounded by the entropy of the geometric distribution, due to the non-negativity of $D(P_X \| P_{X,g})$, and this completes the proof. □

Since mutual information is a specialized relative divergence, the non-negativity property also holds for mutual information.

Corollary 3.3 For random variables X and Y, the mutual information satisfies $I(X;Y) \geq 0$, with equality holding if and only if X and Y are independent.

Proof: Recalling that $I(X;Y) = D(P_{X,Y} \| P_X P_Y)$, the non-negativity of $I(X;Y)$ immediately follows. The necessary and sufficient condition of $D(P_{X,Y} \| P_X P_Y) = 0$, $P_{X,Y}(x,y) = P_X(x)P_Y(y)$, $\forall (x,y) \in X(\Omega) \times Y(\Omega)$, is exactly that X and Y are independent. □

Conditional mutual information also satisfies the non-negativity property.

Corollary 3.4 For random variables X, Y and Z, the conditional mutual information $I(X;Y|Z) \geq 0$, with equality holding if and only if X and Y are conditionally independent given Z; that is, the Markov chain $X \leftrightarrow Z \leftrightarrow Y$ holds.

Proof: Recall that $I(X;Y|Z)$ is defined in Definition 2.15 in Lecture 2 as

$$I(X;Y|Z) = \sum_{(x,y,z)\in X(\Omega)\times Y(\Omega)\times Z(\Omega)} P_{X,Y,Z}(x,y,z) \log \frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}, \tag{3.21}$$

which can be rewritten as

$$I(X;Y|Z) = \sum_{z\in Z(\Omega)} P_Z(z) \sum_{(x,y)\in X(\Omega)\times Y(\Omega)} P_{X,Y|Z}(x,y|z) \log \frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}. \tag{3.22}$$

The inner summation is exactly the relative entropy between $P_{X,Y|Z}$ and $P_{X|Z}P_{Y|Z}$, for each of $z \in Z(\Omega)$. So the non-negativity of $I(X;Y|Z)$ immediately follows from the non-negativity of relative entropy. Furthermore, in order to have $I(X;Y|Z) = 0$, it is necessary and sufficient that the relative entropy between $P_{X,Y|Z}$ and $P_{X|Z}P_{Y|Z}$ is zero, for each of $z \in Z(\Omega)$ unless $P_Z(z) = 0$. This is equivalent to that X and Y are conditionally independent given Z (Definition 2.6). □

Corollary 3.4 is particularly useful in combination with the chain rule of mutual information, when a Markov chain can be identified among several random variables.

An important consequence of the non-negativity of conditional mutual information is the data processing inequality (DPI), which asserts that mutual information decreases along a Markov chain.

Theorem 3.5 For a triple of random variables X, Y and Z satisfying the Markov chain $X \leftrightarrow Y \leftrightarrow Z$, we have $I(X;Y) \geq I(X;Z)$, with equality holding if and only if $X \leftrightarrow Z \leftrightarrow Y$ forms a Markov chain as well.

Proof: Let us expand the mutual information $I(X;Y,Z)$ using the chain rule in two ways:

$$\begin{aligned} I(X;Y,Z) &= I(X;Y) + I(X;Z|Y) && (3.23)\\ &= I(X;Z) + I(X;Y|Z). && (3.24) \end{aligned}$$

Since $X \leftrightarrow Y \leftrightarrow Z$, we have $I(X;Z|Y) = 0$ according to Corollary 3.4. This immediately leads to $I(X;Y) \geq I(X;Z)$. In order to have equality hold, we need $I(X;Y|Z) = 0$, which, by Corollary 3.4, is equivalent to the condition that $X \leftrightarrow Z \leftrightarrow Y$ forms a Markov chain as well. □

As a special case, if we process Y via a mapping f to obtain $f(Y)$, then it always holds that $I(X;f(Y)) \leq I(X;Y)$. Intuitively, processing the raw data (i.e., Y) reduces our ability of extracting its information content (i.e., X).

If equality holds in the DPI, Z is called a sufficient statistic of Y, a concept that plays a key role in statistics.

3.3 Effect of Conditioning

From Corollaries 3.3 and 3.4 we can obtain the following relationship between entropy and conditional entropy, usually called "conditioning reduces entropy."

Theorem 3.6 For random variables X, Y and Z,

$$H(X) \geq H(X|Y), \tag{3.25}$$

with equality holding if and only if X and Y are independent, and

$$H(X|Z) \geq H(X|Y,Z), \tag{3.26}$$

with equality holding if and only if X and Y are conditionally independent given Z.

Proof: Since

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y), & (3.27)\\ I(X;Y|Z) &= H(X|Z) - H(X|Y,Z), & (3.28) \end{aligned}$$

the inequalities (3.25) and (3.26) immediately follow from the nonnegativity of mutual information (Corollary 3.3) and conditional mutual information (Corollary 3.4), respectively. □

The following corollary is then apparent from Theorem 3.6 and the chain rule of entropy (Theorem 3.1).

Corollary 3.5 For random variables $X_1, X_2, \ldots, X_n$, we have

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i), \tag{3.29}$$

with equality holding if and only if $X_1, X_2, \ldots, X_n$ are mutually independent.

As an application of the non-negativity of entropy and the "conditioning reduces entropy" property, we have the following useful result regarding the relationship between a random variable and its mappings.

Corollary 3.6 Consider a random variable X,

- There exists a random variable Y such that $H(Y|X) = 0$ holds, if and only if Y is deterministic given X, i.e., there exists a mapping f such that $Y = f(X)$.
- For any mapping f, $H(f(X)) \leq H(X)$ holds, with equality holding if and only if f is a bijection, i.e., it possesses an inverse mapping f^{-1} such that $f^{-1}(f(X)) = X$.

Proof: According to the definition of conditional entropy (Definition 2.12), we have

$$H(Y|X) = \sum_{x \in X(\Omega)} P_X(x) H(Y|X = x). \tag{3.30}$$

So from the non-negativity of entropy (Corollary 3.1), $H(Y|X) = 0$ is equivalent to $H(Y|X = x) = 0$ for each $x \in X(\Omega)$ (except for those with $P_X(x) = 0$). Furthermore, from Corollary 3.1, $H(Y|X = x) = 0$ if and only if Y is a deterministic constant given $X = x$. This means that Y is determined by X; that is, there exists a mapping f from $X(\Omega)$ to $Y(\Omega)$.

In order to prove the second claim, let us start with $H(X, f(X))$, and expand it using the chain rule of entropy (Theorem 3.1) in two ways:

$$\begin{aligned} H(X, f(X)) &= H(X) + H(f(X)|X) && (3.31) \\ &= H(f(X)) + H(X|f(X)). && (3.32) \end{aligned}$$

Since we have just shown that $H(f(X)|X) = 0$, we immediately have $H(f(X)) \leq H(X)$ by noting that $H(X|f(X)) \geq 0$. When equality holds, we need $H(X|f(X)) = 0$, and this leads to the requirement that f is a bijection. □

Example 3.1 In a secrecy system, there are three parties: a sender, an intended receiver, and an eavesdropper, as illustrated in Figure 3.1. The sender and the intended receiver agree upon a secret key in advance, which is unknown to the eavesdropper. The sender uses the secret key to encipher his plain text, which he wants to share with the intended receiver. The inteded receiver, upon receiving the enciphered codeword, uses the secret key to decipher the plain text. The eavesdropper observes the enciphered codeword and attempts to decipher, without the secret key.

Let the plain text be a random variable X, and the secret key be a random variable Z independent of X. There is a mapping f that

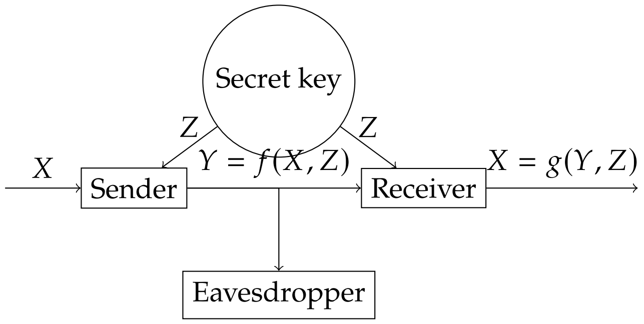

Figure 3.1: A secrecy system.

enciphers X into an enciphered codeword Y, i.e., $Y = f(X, Z)$; and there is another mapping g that deciphers Y into the plain text X, i.e., $X = g(Y, Z)$. A concrete example is the so-called "one-time pad": both X and Z are binary strings and Y is the modulo-2 sum of X and Z, $Y = X \oplus Z$. The intended receiver simply takes modulo-2 sum of Y and Z to reproduce X, $X = Y \oplus Z$.

A perfect secrey system requires that the eavesdropper cannot do any better than pure guessing, and this requirement boils down to the condition of $I(X; Y) = 0$, i.e., X and Y being independent, in light of Corollary 3.3.

For this setup, since $I(X; Y) = 0$, we have

$$\begin{aligned} H(X) &\overset{(a)}{=} H(X|Y) \\ &\overset{(b)}{\leq} H(X, Z|Y) \\ &\overset{(c)}{=} H(Z|Y) + H(X|Y, Z) \\ &\overset{(d)}{=} H(Z|Y) \\ &\overset{(e)}{\leq} H(Z), \end{aligned} \tag{3.33}$$

where, (a) is due to $I(X; Y) = H(X) - H(X|Y)$ (Theorem 3.2), (b) is by expanding $H(X, Z|Y) = H(X|Y) + H(Z|X, Y)$ via the chain rule of entropy (Theorem 3.1) and using the non-negativity of entropy (Corollary 3.1), (c) is also via the chain rule of entropy, (d) is by applying Corollary 3.6 to $X = g(Y, Z)$, and (e) is due to the "conditioning reduces entropy" property (Theorem 3.6).

At this point, it suffices to heuristically interpret the relationship $H(X) \leq H(Z)$ as the fact that, for a perfect secrecy system, the amount of information in the secret key must be no less than the amount of information in the plain text.

At the end of this section, we discuss the convexity and concavity of mutual information, and provide their proofs based on the basic properties we have developed thus far. It is helpful to recall that our definition of mutual information $I(X; Y)$ is given with respect to a pair of random variables X and Y, characterized by their joint probability distribution $P_{X,Y} = P_X P_{Y|X}$.

Theorem 3.7 The mutual information $I(X;Y)$ is concave with respect to P_X, for any fixed $P_{Y|X}$, and is convex with respect to $P_{Y|X}$, for any fixed P_X.

Proof:

(a) The meaning of the concavity of $I(X;Y)$ with respect to P_X is as follows: Take two arbitrary probability distributions on X, $P_X^{(i)}$, $i \in \{0,1\}$, and denote the mutual information for joint probability distribution $P_X^{(i)} P_{Y|X}$ as $I^{(i)}(X;Y)$. For an arbitrary $\lambda \in [0,1]$, define $P_X^{(\lambda)} = (1-\lambda)P_X^{(0)} + \lambda P_X^{(1)}$, and denote the mutual information for joint probability distribution $P_X^{(\lambda)} P_{Y|X}$ as $I^{(\lambda)}(X;Y)$. Then the concavity of $I(X;Y)$ means that

$$(1-\lambda)I^{(0)}(X;Y) + \lambda I^{(1)}(X;Y) \leq I^{(\lambda)}(X;Y). \tag{3.34}$$

In order to prove (3.34), we construct a Markov chain as follows. Let Q be a Bernoulli random variable taking 0 with probability $1-\lambda$ and 1 with probability λ, X have probaility distribution $P_X^{(Q)}$ given Q, and Y have conditional probability distribution $P_{Y|X}$ given X. So $Q \leftrightarrow X \leftrightarrow Y$ forms a Markov chain, and the joint probability distribution of (X,Y) is exactly $P_X^{(\lambda)} P_{Y|X}$. Now, expand the mutual information $I(X,Q;Y)$ using the chain rule of mutual information in two ways:

$$\begin{aligned} I(X,Q;Y) &= I(X;Y) + I(Q;Y|X) && (3.35)\\ &= I(Q;Y) + I(X;Y|Q). && (3.36) \end{aligned}$$

By our preceding convention, $I(X;Y) = I^{(\lambda)}(X;Y)$; applying Corollary 3.4 to the Markov chain $Q \leftrightarrow X \leftrightarrow Y$, $I(Q;Y|X) = 0$; and from the definition of conditional mutual information (Definition 2.15), we can evaluate $I(X;Y|Q)$ according to

$$\begin{aligned} I(X;Y|Q) &= P_Q(0)I(X;Y|Q=0) + P_Q(1)I(X;Y|Q=1) \\ &= (1-\lambda)I^{(0)}(X;Y) + \lambda I^{(1)}(X;Y). && (3.37) \end{aligned}$$

It hence holds, by putting together (3.35) and (3.36) that

$$I^{(\lambda)}(X;Y) \geq (1-\lambda)I^{(0)}(X;Y) + \lambda I^{(1)}(X;Y), \tag{3.38}$$

which is exactly (3.34).

(b) Similar to (a), the meaning of the convexity of $I(X;Y)$ with respect to $P_{Y|X}$ is as follows: Take two arbitrary conditional probability distributions on Y given X, $P_{Y|X}^{(i)}$, $i \in \{0,1\}$, and denote the mutual information for joint probability distribution $P_X P_{Y|X}^{(i)}$ as $I^{(i)}(X;Y)$. For an arbitrary $\lambda \in [0,1]$, define

$P^{(\lambda)}_{Y|X} = (1-\lambda)P^{(0)}_{Y|X} + \lambda P^{(1)}_{Y|X}$, and denote the mutual information for joint probability distribution $P_X P^{(\lambda)}_{Y|X}$ as $I^{(\lambda)}(X;Y)$. Then the convexity of $I(X;Y)$ means that

$$(1-\lambda)I^{(0)}(X;Y) + \lambda I^{(1)}(X;Y) \geq I^{(\lambda)}(X;Y). \tag{3.39}$$

In order to prove (3.39), we introduce a Bernoulli random variable Q taking 0 with probability $1-\lambda$ and 1 with probability λ, independent of X. Let Y have conditional probability $P^{(Q)}_{Y|X}$ given Q. Clearly, the joint probability distribution of (X, Y) is exactly $P_X P^{(\lambda)}_{Y|X}$. Now, expand the mutual information $I(X, Q; Y)$ using the chain rule in two ways:

$$\begin{aligned} I(X;Q,Y) &= I(X;Y) + I(X;Q|Y) \qquad (3.40) \\ &= I(X;Q) + I(X;Y|Q). \qquad (3.41) \end{aligned}$$

By definition, $I(X;Y) = I^{(\lambda)}(X;Y)$; by the independence between Q and X, $I(X;Q) = 0$ (see Corollary 3.3); and we can evaluate $I(X;Y|Q)$ according to

$$\begin{aligned} I(X;Y|Q) &= P_Q(0)I(X;Y|Q=0) + P_Q(1)I(X;Y|Q=1) \\ &= (1-\lambda)I^{(0)}(X;Y) + \lambda I^{(1)}(X;Y). \qquad (3.42) \end{aligned}$$

It hence holds, by putting together (3.40) and (3.41) that

$$I^{(\lambda)}(X;Y) \leq (1-\lambda)I^{(0)}(X;Y) + \lambda I^{(1)}(X;Y), \tag{3.43}$$

which is exactly (3.39). □

3.4 Fano's Inequality

For a pair of random variables X and Y with joint probability distribution $P_{X,Y}$, if we can only observe Y and wish to decide the value of X, how well can we do?

Mathematically, our decision, denoted by $\hat{X}$, is a random variable induced by Y according to some conditional probability distribution $P_{\hat{X}|Y}$. So there exists a Markov chain $X \leftrightarrow Y \leftrightarrow \hat{X}$.

In order to prceed, we need to be specific about how to assess the quality of a decision. Let us use the probability that $\hat{X}$ is not equal to X as the performance measure, which is usually called the error probability, $P_e = P(\hat{X} \neq X)$.

The decision that minimizes P_e is given by the following theorem.

Theorem 3.8 The following maximum a posteriori (MAP) decision minimizes $P_e = P(\hat{X} \neq X)$: for an observation $Y = y \in Y(\Omega)$,

$$\hat{X} = \arg \max_{x \in X(\Omega)} P_{X|Y}(x|y). \tag{3.44}$$

Proof: We expand $P_e = P(\hat{X} \neq X)$ as follows:

$$\begin{aligned}
& P(\hat{X} \neq X) \\
= \; & \sum_{x \in X(\Omega)} P_X(x) P(\hat{X} \neq x | X = x) \\
= \; & \sum_{x \in X(\Omega)} P_X(x) \sum_{y \in Y(\Omega)} P(\hat{X} \neq x, Y = y | X = x) \\
= \; & \sum_{x \in X(\Omega)} P_X(x) \sum_{y \in Y(\Omega)} P(\hat{X} \neq x | Y = y, X = x) P_{Y|X}(y|x) \\
= \; & \sum_{x \in X(\Omega)} P_X(x) \sum_{y \in Y(\Omega)} [1 - P(\hat{X} = x | Y = y, X = x)] P_{Y|X}(y|x) \\
\overset{(a)}{=} \; & \sum_{x \in X(\Omega)} P_X(x) \sum_{y \in Y(\Omega)} [1 - P_{\hat{X}|Y}(x|y)] P_{Y|X}(y|x) \\
= \; & \sum_{(x,y) \in X(\Omega) \times Y(\Omega)} P_X(x) P_{Y|X}(y|x) [1 - P_{\hat{X}|Y}(x|y)] \\
= \; & 1 - \sum_{(x,y) \in X(\Omega) \times Y(\Omega)} P_{X,Y}(x, y) P_{\hat{X}|Y}(x|y),
\end{aligned} \tag{3.45}$$

in which (a) is due to the Markov chain $X \leftrightarrow Y \leftrightarrow \hat{X}$.

So we turn to maximizing $\sum_{(x,y) \in X(\Omega) \times Y(\Omega)} P_{X,Y}(x, y) P_{\hat{X}|Y}(x|y)$, for which we have

$$\begin{aligned}
& \sum_{(x,y) \in X(\Omega) \times Y(\Omega)} P_{X,Y}(x, y) P_{\hat{X}|Y}(x|y) \\
= \; & \sum_{y \in Y(\Omega)} P_Y(y) \sum_{x \in X(\Omega)} P_{X|Y}(x|y) P_{\hat{X}|Y}(x|y).
\end{aligned} \tag{3.46}$$

Therefore, for each $y \in Y(\Omega)$, we need to maximize the inner summation, i.e., $\sum_{x \in X(\Omega)} P_{X|Y}(x|y) P_{\hat{X}|Y}(x|y)$. Since this is the sum of $|X(\Omega)|$ non-negative terms, under the constraint

$$\sum_{x \in X(\Omega)} P_{\hat{X}|Y}(x|y) = 1, \tag{3.47}$$

the optimal solution is clearly to let $P_{\hat{X}|Y}(x|y) = 1$ for $x \in X(\Omega)$ that achieves the maximum $P_{X|Y}(x|y)$, and let the remaining $P_{\hat{X}|Y}(x|y) = 0$. When there are more than one $x \in X(\Omega)$ maximizing $P_{X|Y}(x|y)$, we may pick any of them arbitrarily. This is exactly the MAP decision (3.44). □

Using the Bayes' rule, the MAP decision (3.44) can be rewritten as

$$\hat{X} = \arg \max_{x \in X(\Omega)} P_X(x) P_{Y|X}(y|x). \tag{3.48}$$

Note that the denominator $P_Y(y)$ has been removed in (3.48) because the maximization is over $X(\Omega)$ and does not depend on the value of $P_Y(y)$. Furthermore, if X is uniform over $X(\Omega)$, the decision becomes $\hat{X} = \arg\max_{x \in X(\Omega)} P_{Y|X}(y|x)$, which is called the maximum likelihood (ML) decision.

Although we have started with a fairly general requirement on $\hat{X}$ as a random variable induced by Y satisfying $X \leftrightarrow Y \leftrightarrow \hat{X}$, the resulting MAP decision (3.44) is a deterministic rule based on Y. When considering criteria other than P_e, this may not be true and randomized decision may be necessary.

The MAP decision is fundamental for many statistical inference problems. Its performance, however, is usually difficult to obtain in closed form, or even numerically, — if the dimension of a problem grows large. Furthermore, there surely exist other kinds of decision that may be used for certain reasons, and hence we want to assess their error probability as well. Fano's inequality, built upon basic properties of entropy, provides a lower bound on P_e, for any (not necessarily MAP) decision $\hat{X}$.

Theorem 3.9 For any decision $\hat{X}$ such that $X \leftrightarrow Y \leftrightarrow \hat{X}$, $P_e = P(\hat{X} \neq X)$ satisfies

$$H(X|Y) \leq H(X|\hat{X}) \leq h_2(P_e) + P_e \log(|X(\Omega)| - 1). \tag{3.49}$$

Proof: The first inequality $H(X|Y) \leq H(X|\hat{X})$ directly follows from the DPI, and subsequently we prove the second inequality.

Define an indicator random variable E as follows: $E = 1$ if $\hat{X} \neq X$, and $E = 0$ otherwise. It is clear that E is a Bernoulli random variable taking value 1 with probability P_e. Applying the chain rule of entropy (Theorem 3.1), we obtain

$$\begin{aligned} H(X, E|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) && (3.50) \\ &= H(E|\hat{X}) + H(X|\hat{X}, E). && (3.51) \end{aligned}$$

Since E is determined by $(X, \hat{X})$, $H(E|X, \hat{X}) = 0$ according to Corollary 3.6; since conditioning reduces entropy (Theorem 3.6), $H(E|\hat{X}) \leq H(E) = h_2(P_e)$; for $H(X|\hat{X}, E)$, we can expand it as

$$\begin{aligned} H(X|\hat{X}, E) &= P_E(0) H(X|\hat{X}, E = 0) + P_E(1) H(X|\hat{X}, E = 1) \\ &\leq (1 - P_e) \cdot 0 + P_e \cdot \log(|X(\Omega)| - 1), && (3.52) \end{aligned}$$

where

- $H(X|\hat{X}, E = 0) = 0$ since when $E = 0$, $\hat{X} = X$ holds;
- $H(X|\hat{X}, E = 1) \leq \log(|X(\Omega)|-1)$ since when $E = 1$, X cannot be equal to $\hat{X}$ and hence can only be in the set of $X(\Omega) \setminus \{\hat{X}\}$.

So we have, from (3.50) and (3.51),

$$\begin{aligned} H(X|\hat{X}) + H(E|X, \hat{X}) &= H(E|\hat{X}) + H(X|\hat{X}, E) \\ H(X|\hat{X}) + 0 &\leq h_2(P_e) + (1 - P_e) \cdot 0 \\ &\qquad + P_e \cdot \log(|X(\Omega)| - 1) \\ H(X|\hat{X}) &\leq h_2(P_e) + P_e \log(|X(\Omega)| - 1), \qquad (3.53) \end{aligned}$$

and this completes the proof. □

Since Fano's inequality holds for any decision $\hat{X}$, it will be found useful for proving impossibility results, i.e., that the error probability cannot be lower than a certain level no matter what decision is used. We will use it to prove the converse part of Shannon's fundamental theorem for channel coding in Lecture 6. Inspecting (3.49), we see that when $|X(\Omega)| < \infty$, if $P_e = 0$, i.e., error-free decision, then it is necessary to have $H(X|Y) = 0$, i.e., X being deterministic given Y, according to Corollary 3.6.

3.5 Entropy Rate for Stationary Processes

In Lecture 2 we have defined entropy rate for stochastic processes. In this lecture we study its properties.

Let us first examine some examples.

Example 3.2 Consider a stochastic process $\mathbf{X} : X_1, X_2, \ldots$ in which all elements are i.i.d. random variables. Such a stochastic process is said to be "memoryless" in the sense that X_i does not rely upon its "history" $\{X_{i-1}, \ldots, X_1\}$, for any i.

Then, since

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &\overset{(a)}{=} \sum_{i=1}^{n} H(X_i) \\ &= nH(X_1), \qquad (3.54) \end{aligned}$$

where (a) is due to Corollary 3.5, we immediately obtain the entropy rate of $\mathbf{X}$ as

$$\begin{aligned} H(\mathbf{X}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \\ &= \lim_{n\to\infty} \frac{1}{n} nH(X_1) \\ &= H(X_1). \qquad (3.55) \end{aligned}$$

Example 3.3 Consider a stochastic process $\mathbf{X} : X_1, X_2, \ldots$ in which all elements are mutually independent random variables, and X_i obeys geometric distribution with parameter 2^{-i}, $\forall i$.

Then, we have

$$\begin{aligned}
\frac{1}{n}H(X_1, X_2, \ldots, X_n) &\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n} H(X_i) \\
&\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n} \frac{h_2(2^{-i})}{2^{-i}} \\
&\overset{(c)}{>} \frac{1}{n}\sum_{i=1}^{n} 2^i 2^{-i} \log 2^i \\
&= \frac{\log 2}{n}\sum_{i=1}^{n} i \\
&= \frac{\log 2}{2}(n+1) \\
&\longrightarrow \infty \qquad (3.56)
\end{aligned}$$

as $n \longrightarrow \infty$, where (a) is due to Corollary 3.5, (b) is from Example 2.3 in Lecture 2, and (c) is obtained by expanding $h_2(\epsilon)$ and only retaining the term $\epsilon \log \frac{1}{\epsilon}$. So the entropy rate of this stochastic process, the limit of $\frac{1}{n}H(X_1, X_2, \ldots, X_n)$ as $n \longrightarrow \infty$, does not exist.

Example 3.4 Consider a stochastic process $\mathbf{X} : X_1, X_2, \ldots$ generated as follows: using a Bernoulli random variable Z with parameter $1/2$ as switch, when $Z = 1$, $X_1, X_2, \ldots$ are i.i.d. random variables each obeying Bernoulli distribution with parameter $1/2$, and when $Z = 0$, $X_1, X_2, \ldots$ are constant zero.

For this stochastic process, we have

$$\begin{aligned}
& P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \\
= \;& \frac{1}{2}P_{X_1,\ldots,X_n|Z}(x_1, \ldots, x_n|Z = 1) + \frac{1}{2}P_{X_1,\ldots,X_n|Z}(x_1, \ldots, x_n|Z = 0) \\
= \;& 2^{-(n+1)} + \frac{1}{2}\mathbf{1}_{\{(x_1,\ldots,x_n)=(0,\ldots,0)\}}. \qquad (3.57)
\end{aligned}$$

So we can obtain, with some calculation,

$$H(X_1, \ldots, X_n) = 1 + \frac{1 - 2^{-n}}{2}n - \frac{1 + 2^{-n}}{2}\log_2(1 + 2^{-n}) \quad (3.58)$$

in bits, and hence

$$\begin{aligned}
H(\mathbf{X}) &= \lim_{n\to\infty} \frac{1}{n}H(X_1, \ldots, X_n) \\
&= \frac{1}{2} \text{ (bits).} \qquad (3.59)
\end{aligned}$$

This result may be intuitively understood as follows: the stochastic process $\mathbf{X}$ is a mixture of two component stochastic processes, one being memoryless Bernoulli with entropy rate one bit, and the other being deterministic with entropy rate zero, so the overall entropy rate is simply the average of the entropy rates of these two component stochastic processes.

Subsequently, we focus on stationary stochastic processes.

Definition 3.1 A discrete-time stochastic process $\mathbf{X}$ is stationary if the joint probability distribution of any subset of $\mathbf{X}$ is invariant with respect to shifts in time; that is,

$$\begin{aligned} & P(X_{i_1} = x_1, X_{i_2} = x_2, \ldots, X_{i_n} = x_n) \\ = \ & P(X_{i_1+l} = x_1, X_{i_2+l} = x_2, \ldots, X_{i_n+l} = x_n), \qquad (3.60) \end{aligned}$$

for any n, any subscript indices $i_1, i_2, \ldots, i_n$, any time shift l, and any collection of $x_1, x_2, \ldots, x_n \in X(\Omega)$.

For a stationary stochastic process $\mathbf{X}$, its entropy rate is given by the following theorem.

Theorem 3.10 For a stationary stochastic process $\mathbf{X}$ satisfying $H(X_1) < \infty$, its entropy rate exists, and satisfies

$$H(\mathbf{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_1). \qquad (3.61)$$

Proof: Let us begin with applying the chain rule of entropy (Theorem 3.1) to $H(X_1, \ldots, X_n)$:

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \qquad (3.62)$$

The terms in the summation constitute a monotonically non-increasing sequence, because

$$\begin{aligned} H(X_{i+1}|X_i, \ldots, X_1) & \overset{(a)}{\leq} H(X_{i+1}|X_i, \ldots, X_2) \\ & \overset{(b)}{=} H(X_i|X_{i-1}, \ldots, X_1), \qquad (3.63) \end{aligned}$$

where (a) is due to the property that conditioning reduces entropy (Theorem 3.6), and (b) is due to the stationarity of $\mathbf{X}$. So the sequence $\{H(X_i|X_{i-1}, \ldots, X_1)\}_{i=1}^{\infty}$, being non-negative monotonically non-increasing, possesses a limit, which can be denoted by

$$H'(\mathbf{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_1). \qquad (3.64)$$

Hence, for any $\epsilon > 0$, there exists an integer N_ϵ such that for any $i > N_\epsilon$, $|H(X_i|X_{i-1}, \ldots, X_1) - H'(\mathbf{X})| < \epsilon$.

Now returning to (3.62), let us examine the gap

$$\left|\frac{1}{n}H(X_1,\ldots,X_n) - H'(\mathbf{X})\right|$$

as follows:

$$\begin{aligned}
&\left|\frac{1}{n}H(X_1,\ldots,X_n) - H'(\mathbf{X})\right| \\
\overset{(a)}{=}\ & \left|\frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1) - H'(\mathbf{X})\right| \\
\overset{(b)}{\leq}\ & \frac{1}{n}\sum_{i=1}^{n} |H(X_i|X_{i-1},\ldots,X_1) - H'(\mathbf{X})| \\
=\ & \frac{1}{n}\sum_{i=1}^{N_\epsilon} |H(X_i|X_{i-1},\ldots,X_1) - H'(\mathbf{X})| \\
& + \frac{1}{n}\sum_{i=N_\epsilon+1}^{n} |H(X_i|X_{i-1},\ldots,X_1) - H'(\mathbf{X})| \\
<\ & \frac{1}{n}\sum_{i=1}^{N_\epsilon} |H(X_i|X_{i-1},\ldots,X_1) - H'(\mathbf{X})| + \frac{1}{n}(n - N_\epsilon)\epsilon \\
<\ & \frac{1}{n}\sum_{i=1}^{N_\epsilon} |H(X_i|X_{i-1},\ldots,X_1) - H'(\mathbf{X})| + \epsilon, \qquad (3.65)
\end{aligned}$$

where (a) is via the chain rule of entropy (Theorem 3.1), (b) is the triangle inequality for absolute values. The right hand side of (3.65) can then be upper bounded by 2ϵ for all sufficiently large n. So in summary, the limit $\lim_{n\to\infty} \frac{1}{n}H(X_1,\ldots,X_n)$ exists and is equal to $H'(\mathbf{X})$. □

According to Theorem 3.10, the entropy rate of a stationary stochastic process is the conditional entropy of the "current" given "all the past".

Another special type of stochastic process is Markov chain. A stochastic process $\mathbf{X} : X_1, X_2, \ldots$ is a Markov chain, If

$$(X_1,\ldots,X_{n-1}) \leftrightarrow X_n \leftrightarrow X_{n+1} \qquad (3.66)$$

holds for any n. A Markov chain is time invariant if $P_{X_{n+1}|X_n}$ does not depend on n. Therefore, a time invariant Markov chain is described by the conditional probability distribution $P_{X_2|X_1}(b|a)$, $\forall a, b \in X_1(\Omega)$.

For a time invariant Markov chain, if there exists a probability distribution on X_1, P_{X_1}, such that $\forall b \in X_1(\Omega)$,

$$P_{X_1}(b) = \sum_{a\in X_1(\Omega)} P_{X_1}(a)P_{X_2|X_1}(b|a), \qquad (3.67)$$

holds, then the resulting Markov chain is also stationary, and P_{X_1}

is called the stationary distribution of the Markov chain.

If a Markov chain is both stationary and time invariant, its entropy rate is given by the following corollary of Theorem 3.10.

Corollary 3.7 The entropy rate of a stationary time invariant Markov chain **X** satisfies

$$H(\mathbf{X}) = H(X_2|X_1). \tag{3.68}$$

Proof: According to Theorem 3.10,

$$H(\mathbf{X}) = \lim_{n\to\infty} H(X_n|X_{n-1},\ldots,X_1), \tag{3.69}$$

so we apply the Markov chain relationship $(X_1,\ldots,X_{n-2}) \leftrightarrow X_{n-1} \leftrightarrow X_n$ to obtain

$$\begin{aligned} H(\mathbf{X}) &= \lim_{n\to\infty} H(X_n|X_{n-1},\ldots,X_1) \\ &= \lim_{n\to\infty} H(X_n|X_{n-1}). \end{aligned} \tag{3.70}$$

Since **X** is also stationary, we have

$$H(\mathbf{X}) = H(X_2|X_1), \tag{3.71}$$

thereby completing the proof. □

Example 3.5 Consider a time invariant two-state Markov chain with $X_1(\Omega) = \{0,1\}$, and transition probability distribution given by $P_{X_2|X_1}(0|0) = 1-\alpha$, $P_{X_2|X_1}(1|0) = \alpha$, $P_{X_2|X_1}(0|1) = \beta$ and $P_{X_2|X_1}(1|1) = 1-\beta$, as illustrated in Figure 3.2. The stationary distribution P_{X_1} can be solved via (3.67) as

$$\begin{aligned} P_{X_1}(0) &= P_{X_1}(0)P_{X_2|X_1}(0|0) + P_{X_1}(1)P_{X_2|X_1}(0|1) \\ &= (1-\alpha)P_{X_1}(0) + \beta P_{X_1}(1), \\ P_{X_1}(1) &= P_{X_1}(0)P_{X_2|X_1}(1|0) + P_{X_1}(1)P_{X_2|X_1}(1|1) \\ &= \alpha P_{X_1}(0) + (1-\beta)P_{X_1}(1), \\ P_{X_1}(0) + P_{X_1}(1) &= 1, \end{aligned} \tag{3.72}$$

and is given by $P_{X_1}(0) = \frac{\beta}{\alpha+\beta}$ and $P_{X_1}(1) = \frac{\alpha}{\alpha+\beta}$. So from Corollary 3.7 the entropy rate of this stochastic process is

$$H(\mathbf{X}) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta}h_2(\alpha) + \frac{\alpha}{\alpha+\beta}h_2(\beta). \tag{3.73}$$

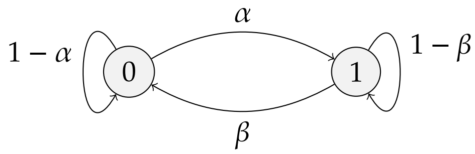

Figure 3.2: A time invariant two-state Markov chain.

Notes

Most properties in this lecture can be found in standard textbooks of information theory. The non-negativity property of relative entropy serves as the foudation of other non-negativity properties; in many textbooks the exposition of properties follows a different approach, starting with Jensen's inequality of convex functions.

Fano's inequality is due to Robert Fano, a pioneer in the forming of information theory, who developed the first course in early 1950s and wrote one of the first comprehensive textbooks [10] on this subject at MIT.

For a more in-depth and systematic treatment of quantities of information, refer to [8, Chapter 6], where a one-to-one correspondence between quantities of information and set operations is developed. The inequalities encountered in this lecture belong to the so-called Shannon-type inequalities, but there also exist non-Shannon-type inequalities; see [8, Chapters 13 and 14].

The perfect secrecy system of Example 3.1 was treated by Shannon in his foundational article "Communication Theory of Secrecy Systems" [11]. The conclusion that a perfect secrecy system requires the entropy of secret key to be no smaller than that of the plain text had to some extent discouraged the development of information theoretic cryptography until late 1970s; see [12].

Historically, the ML decision was proposed viewing X as a deterministic parameter, without considering any probabilistic structure of X. When X is a random variable with a uniform probability distribution over $X(\Omega)$, the resulting MAP decision exhibits the same form as the ML decision.

Exercises

1. For random variables X and Y, prove that $H(X+Y) \leq H(X) + H(Y)$ holds.
2. For random variables $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$, when does

$$\begin{aligned} & H(X_1, X_2, \ldots, X_n | Y_1, Y_2, \ldots, Y_n) \\ = \; & H(X_1|Y_1) + H(X_2|Y_2) + \ldots + H(X_n|Y_n) \end{aligned}$$

hold?

3. We know from Theorem 3.6 that conditioning reduces entropy. For mutual information $I(X;Y)$ and conditional mutual information $I(X;Y|Z)$, does an analogous property hold?
4. Using the non-negativity of relative divergence, prove the log-sum inequality: for non-negative numbers $\{a_i\}_{i=1,\ldots,n}$ and $\{b_i\}_{i=1,\ldots,n}$,

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

   where $a = \sum_{i=1}^{n} a_i$ and $b = \sum_{i=1}^{n} b_i$, with equality holding if and only if there exists c such that $a_i = cb_i$ for all i.
5. In this exercise we apply Corollary 3.2 to a guessing problem due to James L. Massey [13]. Suppose that we want to guess the value of a random variable X over $X(\Omega) = \{1, 2, \ldots\}$. How many times do we need to guess, on average? Without loss of generality, we can always relabel the random variable so that $P_X(1) \geq P_X(2) \geq \ldots$ holds. Prove that, on average, we need to guess no less than $e^{H(X)-1}$ times, where the unit of entropy is nat.
6. Consider a random variable X over $X(\Omega) = \{1, 2, \ldots\}$.

   a) Prove that if $\mathbf{E}X$ is finite, then $H(X)$ is also finite.
   b) Prove that if $\mathbf{E}\log X$ is finite, then $H(X)$ is also finite.
   c) Prove that if $H(X)$ is finite and $P_X(x)$ is monotonically non-increasing with x, then $\mathbf{E}\log X$ is finite.
   d) Give an example to illustrate that the monotonically non-increasing condition of $P_X(x)$ in the previous statement is necessary.

7. Consider a uniform random variable X over $\{0, 1, \ldots, m-1\}$, and its observation Y is drawn uniformly from $\{(X-1) \bmod m, X, (X+1) \bmod m\}$. Define $P_e = P(Y \neq X)$.

   a) Give a lower bound of P_e using the Fano inequality.
   b) Find the gap between the lower bound and the exact value of P_e of the MAP decision.
   c) Can you resolve the gap by inspecting the proof of the Fano inequality and improving it?

8. Construct an example where equality holds in the Fano inequality.
9. If the estimate $\hat{X}$ is a size-L subset of $X(\Omega)$, and define the error event to be $\{X \notin \hat{X}\}$, establish an extension of the Fano inequality.

10. Prove the Csiszár identity:

$$\begin{aligned}
& \sum_{i=1}^{n} I(X_{i+1}, \ldots, X_n; Y_i | Y_1, \ldots, Y_{i-1}) \\
= & \sum_{i=1}^{n} I(Y_1, \ldots, Y_{i-1}; X_i | X_{i+1}, \ldots, X_n),
\end{aligned}$$

where X_{n+1} and Y_0 are understood as degenerated.

11. In this exercise, we provide an information-theoretic proof of the well known number-theoretic result that there are infinitely many prime numbers. For this, consider an arbitrary integer n, and denote the number of primes no greater than n by $\pi(n)$. Take a random variable N uniformly distributed over $\{1, 2, \ldots, n\}$, and write it in its unique prime factorization, $N = p_1^{X_1} p_2^{X_2} \ldots p_{\pi(n)}^{X_{\pi(n)}}$, where $\{p_1, p_2, \ldots, p_{\pi(n)}\}$ are primes no greater than n, and each X_i is the largest power $k \geq 0$ such that p_i^k divides N. By inspecting $H(N)$, prove that $\pi(n) \longrightarrow \infty$ as $n \longrightarrow \infty$. For further reading, refer to [14].

12. For integer set $[n] := \{1, 2, \ldots, n\}$, drawing each of its elements independently with probability p leads to a random subset of $[n]$. For two such subsets, A and B, generated independently, calculate $H(A)$ and $H(A \cup B)$, and show that $H(A \cup B) > H(A)$ when $p \leq \frac{3-\sqrt{5}}{2}$.
This is related to the so-called union-closed sets conjecture, for which the first constant lower bound was established using an information-theoretic argument; for further reading, refer to [15].

13. Consider a random variable X generated as follows: conditioned upon a random variable Z taking values in $\{1, 2, \ldots\}$, let X be a geometric random variable (see Example 2.3) with parameter 2^{-Z}.

    a) Show that if $\mathbf{E}[Z] = \infty$ then $H(X) = \infty$.
    b) Define a random variable Y as follows: $Y = 0$ with probability $1 - \epsilon$ and $Y = X$ with probability ϵ. Let $\hat{Y} = 0$ with probability one. Show that if $H(X) = \infty$ then $H(Y|\hat{Y})$ does not tend to zero, no matter how small the decision error probability $P_e = P(Y \neq \hat{Y}) = \epsilon > 0$ is. This example illustrates the delicacy when applying Fano's inequality when the alphabet is infinite ([8, Example 2.49]).

14. Prove the submodularity property of entropy: for any two sets of random variables $\mathbf{S}_1$ and $\mathbf{S}_2$, $H(\mathbf{S}_1 \cup \mathbf{S}_2) + H(\mathbf{S}_1 \cap \mathbf{S}_2) \leq H(\mathbf{S}_1) + H(\mathbf{S}_2)$.

15. For random variables X and Y and a mapping f, under what condition does $H(X|f(Y)) = H(X|Y)$ hold?

16. Suppose that $\Theta \in (0, 1)$ is a random variable over the unit

interval, and conditioned upon Θ, $\mathbf{X} = (X_1, \ldots, X_n)$ consists of n i.i.d. random variables $X_i \sim \text{Bernoulli}(\Theta)$. Define $T = \sum_{i=1}^{n} X_i$. Is T a sufficient statistic for Θ?

17. For the two-state Markov chain in Example 3.5, if we undersample it to obtain a new stochastic process $X_1, X_3, X_5, \ldots$, is it still a Markov chain? Under stationarity, evaluate its entropy rate and compare with that of the original Markov chain $X_1, X_2, X_3, \ldots$.
18. Define an "almost Markov" relationship for three random variables (X, Y, Z) if they satisfy

$$p(z|x, y) = p(z|y)(1 + \epsilon(x, y, z)),$$

where $|\epsilon(x, y, z)| \leq \delta$ for any (x, y, z) tuple. Prove that for such an "almost Markov" relationship, we have the following "δ-approximate DPI" hold:

$$I(X; Z) \leq I(X; Y) + \delta^2.$$

19. For random variables $V, W_1, W_2, \ldots, W_n$, prove that

$$H(V) \geq \sum_{i=1}^{n} I(V; W_i),$$

when $W_1, W_2, \ldots, W_n$ are mutually independent.