Chapter 2

Entropy, Relative Entropy, and Mutual Information

Peng-Hua Wang

Graduate Institute of Communication Engineering

National Taipei University

Chapter Outline

Chap. 2 Entropy, Relative Entropy, and Mutual Information

- **2.1** Entropy
- **2.2** Joint entropy and conditional entropy
- **2.3** Relative entropy and mutual information
- **2.4** Relationship between entropy and mutual information
- **2.5** Chain Rules for Entropy, Relative Entropy, and Mutual Information

Chapter Outline

Chap. 2 Entropy, Relative Entropy, and Mutual Information

- **2.6** Jensen's inequality and its consequences
- **2.7** Log sum inequality and its applications
- **2.8** Data processing inequality
- **2.9** Sufficient Statistics
- **2.10** Fano's Inequality

2.1 Entropy

Entropy

Definition 1 (Entropy) *The entropy $H(X)$ of a discrete random variable X is defined by*

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

Entropy

- X be a discrete random variable with alphabet $\mathcal{X}$ and pmf $p(x) = \Pr[X = x]$, $x \in \mathcal{X}$.
- $\log_2 p(x)$, the entropy is expressed in bits.
- If the base is e, i.e., $\ln p(x)$, the entropy is expressed in nats.
- If the base is b, we denote the entropy as $H_b(X)$.
- $0 \log 0 \triangleq \lim_{t \to 0^+} t \log t = 0$.
- $H(X) = E[\log \frac{1}{p(X)}] = -E \log p(X)$
- $H(X)$ may not exist.

Properties of entropy

Lemma 1

$$H(X) \geq 0$$

Lemma 2

$$H_b(X) = \log_b(a) H_a(X)$$

Meaning of entropy

- The amount of information (code length) required on the average to describe the random variable.
- The minimum expected number of binary questions required to determine X lies between $H(X)$ and $H(X)+1$.
- The amount of "information" provided by an observation of a random variable.
  - If an event is less probable, we receive more information when it occurs.
  - A certain event provides no information.
- "Uncertainty" about a random variable.
- "Randomness" of a random variable.

Example 1.1.1

Consider a random variable that has a uniform distribution over 32 outcomes. To identify an outcome, we need a label that takes on 32 different values.

(1) How may bit is sufficient as label?
(2) Compute the entropy of the random variable.

Example 1.1.2

Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $(1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64)$. Suppose that we wish to send a message indicating which horse won the race.

(1) How may bit is sufficient for labeling the horse?
(2) Compute the entropy $H(X)$.
(3) Can we label the horse in average $H(X)$ bits?

Example 2.1.1

Let $\Pr[X=1]=p$ and $\Pr[X=0]=1-p$. The entropy $H(X) \triangleq H(p) = -p\log p - (1-p)\log(1-p)$.

- $H(p)$ is a concave function of the distribution.
- $H(p)=0$ if $p=0$ or 1.
- $H(p)=1$ is maximum if $p=1/2$.

Example 2.1.2

Let

$$X = \begin{cases} a, & \text{with probability } \frac{1}{2}, \\ b, & \text{with probability } \frac{1}{4}, \\ c, & \text{with probability } \frac{1}{8}, \\ d, & \text{with probability } \frac{1}{8}. \end{cases}$$

Compute $H(X)$.

- We wish to determine the value of X with the "Yes/No" questions.
- The minimum number of binary questions lies between $H(X)$ and $H(X)+1$.

2.2 Joint entropy and conditional entropy

Joint entropy

Definition 2 (Joint Entropy) *Let (X,Y) be a pair of discrete random variables with a joint distribution $p(x,y)$. The joint entropy $H(X,Y)$ is defined as*

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(X,Y) \\ &= -E\log p(x,y) \end{aligned}$$

Conditional entropy

Definition 3 (Conditional Entropy) *The conditional entropy* $H(Y|X)$ *is defined as*

$$\begin{aligned} H(X|Y) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\ &= -E\log p(X|Y) \end{aligned}$$

Example 2.2.1

Let (X, Y) have the following joint distribution:

	$X=1$	$X=2$	$X=3$	$X=4$
$Y=1$	$1/8$	$1/16$	$1/32$	$1/32$
$Y=2$	$1/16$	$1/8$	$1/32$	$1/32$
$Y=3$	$1/16$	$1/16$	$1/16$	$1/16$
$Y=4$	$1/4$	0	0	0

Compute $H(X)$, $H(Y)$, $H(X,Y)$, $H(Y|X)$, $H(X|Y)$.

Properties of conditional entropy

Theorem 1 (Chain rule)

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

Proof. *Take logarithm and expectation on*

$$[p(x,y)]^{-1} = [p(x)]^{-1}[p(y|x)]^{-1}. \quad \square$$

Properties of conditional entropy

Corollary 1

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

Proof. *Take logarithm and expectation on*

$$[p(x,y|z)]^{-1} = [p(x|z)]^{-1}[p(y|x,z)]^{-1}. \quad \square$$

- $H(Y|X) \neq H(X|Y)$.
- $H(X) - H(X|Y) = H(Y) - H(Y|X)$

2.3 Relative entropy and mutual information

Relative entropy

Definition 4 (Relative Entropy) *The relative entropy between two distributions $p(x)$ and $q(x)$ is defined as*

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$$

$$= E_p \log\frac{p(X)}{q(X)}$$

- $D(p||q)$ is also called the **Kullback–Leibler Distance**
- We will use $0\log\frac{0}{0} = 0$ and $p\log\frac{p}{0} = \infty$

Meaning of Relative entropy

- $D(p||q)$ is a measure of the distance between two distributions.
- $D(p||q)$ is a measure of the inefficiency of assuming that the distribution is $q(x)$ when the true distribution is $p(x)$.

Meaning of Relative entropy

- If we know the true distribution $p(x)$, we could construct a code with average description length

$$\sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} = H(p).$$

If, instead, we used the distribution $q(x)$ to construct the code (wrong code), the average code length is

$$L = \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{q(x)}.$$

The difference is

$$L - H(p) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = D(p||q)$$

Mutual information

Definition 5 (Mutual Information) *The mutual information* $I(X;Y)$ *is defined as*

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)||p(x)p(y)) \\ &= E_{p(x,y)}\log\frac{p(X,Y)}{p(X)p(Y)} \end{aligned}$$

- The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$.

Example 2.3.1

Consider two distributions p and q on $\mathcal{X} = \{0, 1\}$. Let $p(0) = 1 - r$, $p(1) = r$, and let $q(0) = 1 - s$, $q(1) = s$. Compute $D(p||q)$ and $D(q||p)$.

2.4 Relationship between entropy and mutual information

Mutual information and entropy

Theorem 2 (Mutual information and entropy)

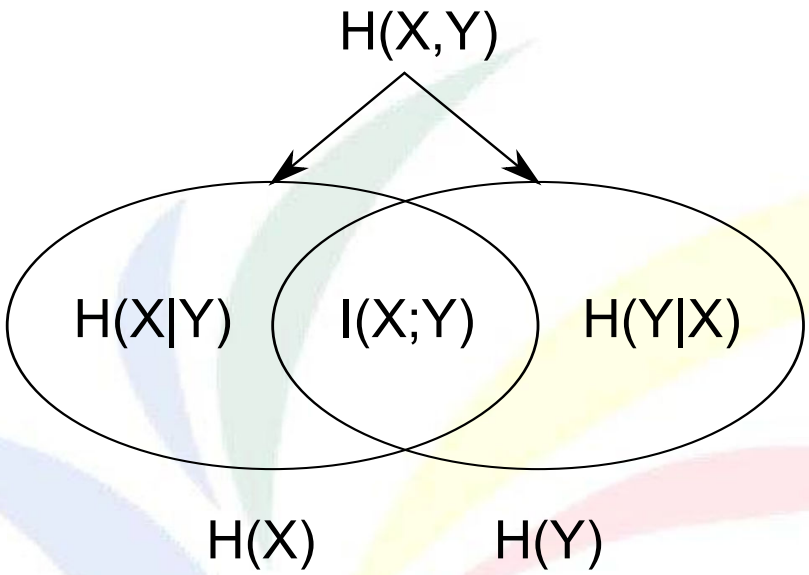
$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
I(X;Y) &= H(Y) - H(Y|X) \\
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
I(X;Y) &= I(Y;X) \\
I(X;X) &= H(X)
\end{aligned}
$$

Proof. 1. Take logarithm and expectation on

$$[p(x,y)/p(x)p(y)]^{-1} = [p(x)]^{-1} \div [p(x|y)]^{-1}. \quad \square$$

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

Mutual information and entropy



Relationships between mutual information and entropy

2.5 Chain Rules for Entropy, Relative Entropy, and Mutual Information

Chain rules

Theorem 3 (Chain rule for entropy)

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

Proof. Take logarithm and expectation on

$$\begin{aligned}&[p(x_1, x_2, \ldots, x_n)]^{-1} \\ =&[p(x_1)]^{-1}[p(x_2|x_1)]^{-1}[p(x_3|x_1, x_2)]^{-1}\cdots. \quad \square\end{aligned}$$

Theorem 4 (Chain rule for information)

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

Chain rules

Theorem 5 (Chain rule for relative entropy)

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Proof. Take logarithm and expectation on

$$\frac{p(x,y)}{q(x,y)} = \frac{p(x)}{q(x)}\frac{p(y|x)}{q(y|x)}. \quad \square$$

2.6 Jensen's inequality and its consequences

Convex function

Definition 6 (Convex Function) *A function $f(x)$ is said to be convex over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $\alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_1 + \alpha_2 = 1$, we have*

$$f(\alpha_1 x_1 + \alpha_2 x_2) \leq \alpha_1 f(x_1) + \alpha_2 f(x_2).$$

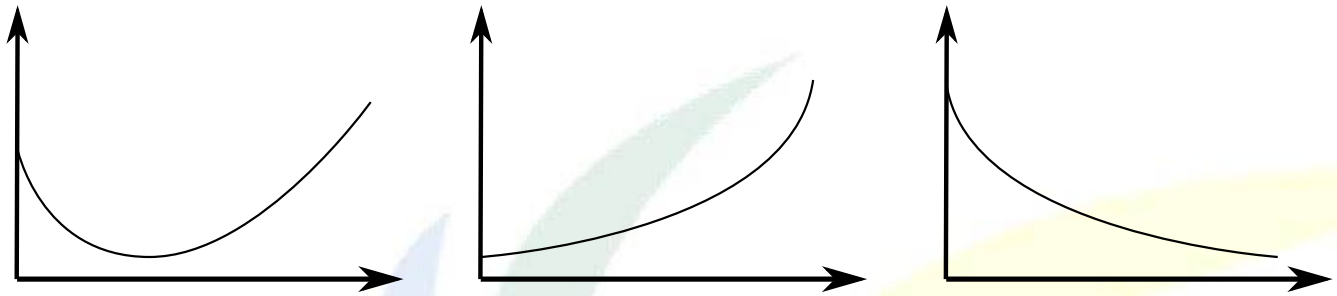
- A function f is said to be **strictly convex** if $\leq$ is replaced by $<$.
- A function is convex if it always lies below any chord.
- A function f is **concave** if $-f$ is convex.
- f is convex (strictly convex) $\Leftrightarrow f'' \geq 0$ ($f'' > 0$).

Convex function

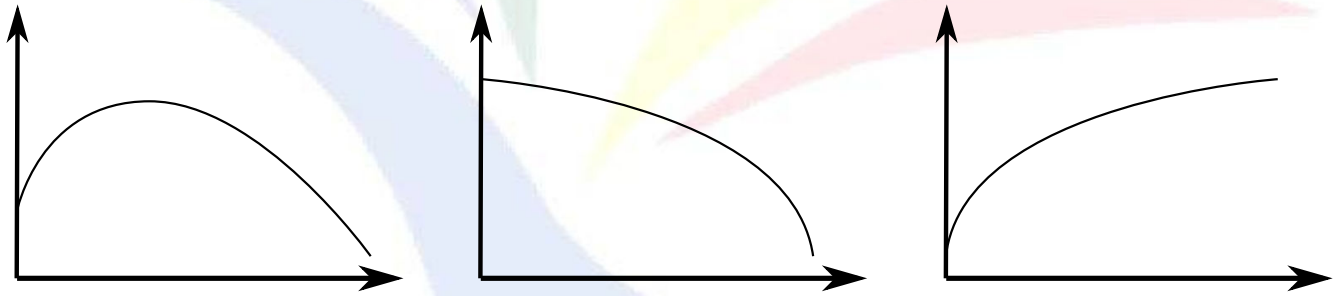
- It can be extended to linear combination of n values. For example,

$$\begin{aligned} f(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3) &= f\left(\alpha_1 x_1 + (\alpha_2 + \alpha_3)(\alpha_2' x_2 + \alpha_3' x_3)\right) \\ &\geq \alpha_1 f(x_1) + (\alpha_2 + \alpha_3) f(\alpha_2' x_2 + \alpha_3' x_3) \\ &\geq \alpha_1 f(x_1) + (\alpha_2 + \alpha_3)\left(\alpha_2' f(x_2) + \alpha_3' f(x_3)\right) \\ &= \alpha_1 f(x_1) + \alpha_2 f(x_2) + \alpha_3 f(x_3) \end{aligned}$$

Examples of convex and concave functions



(a) Convex functions



(b) Concave functions

Jensen's inequality

Theorem 6 (Jensen's inequality) *If f is a convex function and X is a random variable,*

$$Ef(x) \geq f(EX).$$

Moreover, if f is strictly convex, the equality implies that $X = EX$ with probability 1 (i.e., X is a constant).

Information inequality

Theorem 7 (Information inequality) *Let $p(x)$ and $q(x)$ be two pmf's. Then*

$$D(p||q) \geq 0.$$

with equality if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Information inequality

Corollary 2 (Nonnegative of mutual information)

$$I(X;Y) \geq 0$$

with equality iff X and Y are independent.

Corollary 3

$$D(p(y|x)||q(y|x)) \geq 0$$

with equality iff $p(y|x) = q(y|x)$ for all y and x.

Corollary 4

$$I(X;Y|Z) \geq 0$$

with equality iff X and Y are conditionally independent given Z.

Upper bound of entropy

Theorem 8 (Upper bound of entropy)

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

with equality if and only X has a uniform distribution over $\mathcal{X}$.

Proof.

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &\leq \log \sum_{x \in \mathcal{X}} p(x) \cdot \frac{1}{p(x)} \\ &= \log \sum_{x \in \mathcal{X}} 1 \\ &= \log |\mathcal{X}|. \quad \square \end{aligned}$$

Conditioning reduces entropy

Theorem 9 (Conditioning reduces entropy)

$$H(X|Y) \leq H(X)$$

with equality iff X and Y are independent.

Proof. $H(X) - H(X|Y) = I(X;Y) \geq 0$. $\square$

Example 2.6.1

Let (X, Y) have the following joint distribution:

	$X=1$	$X=2$
$Y=1$	0	$3/4$
$Y=2$	$1/8$	$1/8$

Compute $H(X)$, $H(X|Y)$.

Independence bound on entropy

Theorem 10 (Independence bound on entropy)

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

with equality if and only if the X_i are independent.

2.7 Log sum inequality and its applications

Log sum inequality

Theorem 11 (Log sum inequality) *For nonnegative numbers,* $a_1, a_2, \ldots, a_n$ *and* $b_1, b_2, \ldots, b_n$,

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

with equality iff a_i / b_i *is constant.*

Log sum inequality

Proof. By concavity of logarithm, suppose that $A > 0, \alpha_i > 0, \sum_{i=1}^{n} \alpha_i = 1$, we have

$$\sum_{i=1}^{n} \alpha_i \log \frac{b_i}{A\alpha_i} \leq \log \sum_{i=1}^{n} \alpha_i \cdot \frac{b_i}{A\alpha_i} = \log \frac{\sum_{i=1}^{n} b_i}{A}$$

Now, let $a_i = A\alpha_i$, we have $\sum_{i=1}^{n} a_i = A$, and

$$\sum_{i=1}^{n} \frac{a_i}{A} \log \frac{b_i}{a_i} \leq \log \frac{\sum_{i=1}^{n} b_i}{A}$$

$$\Rightarrow \sum_{i=1}^{n} a_i \log \frac{b_i}{a_i} \leq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} a_i}$$

$$\Rightarrow \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}. \quad \square$$

Convexity of relative entropy

Theorem 12 (Convexity of relative entropy) *$D(p||q)$ is convex in the pair (p,q). That is, if $s+t=1, s>0, t>0$, we have*

$$D(sp_1+tp_2||sq_1+tq_2) \leq sD(p_1||q_1)+tD(p_2||q_2).$$

Proof. By log-sum inequality, we have

$$sp_1(x)\log\frac{sp_1(x)}{sq_1(x)}+tp_2(x)\log\frac{tp_2(x)}{tq_2(x)} \geq (sp_1(x)+tp_2(x))\log\frac{sp_1(x)+tp_2(x)}{sq_1(x)+tq_2(x)}$$

Summing this over all $x\in\mathcal{X}$, we obtain the desired property. $\square$

Concavity of entropy

Theorem 13 (Concavity of entropy) *$H(p)$ is a concave function of p. That is, if $s+t=1, s>0, t>0$, we have*

$$H(sp_1+tp_2) \geq sH(p_1)+tH(p_2).$$

Concavity of entropy

Proof. By log-sum inequality, we have

$$sp_1(x)\log\frac{sp_1(x)}{s}+tp_2(x)\log\frac{tp_2(x)}{t}$$
$$\geq (sp_1(x)+tp_2(x))\log\frac{sp_1(x)+tp_2(x)}{s+t}$$
$$\Rightarrow -sp_1(x)\log p_1(x)+tp_2(x)\log p_2(x)$$
$$\leq -[sp_1(x)+tp_2(x)]\log[sp_1(x)+tp_2(x)]$$

Summing this over all $x\in\mathcal{X}$, we obtain the desired property. $\square$

Concavity of mutual information

Theorem 14 (Concavity of mutual information) *The mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.*

2.8 Data processing inequality

Definition of Markov chain

Definition 7 (Markov chain) *X, Y, and Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written as*

$$p(x,y,z) = p(x)p(y|x)p(z|y)$$

- $X \to Y \to Z$ iff X and Z are conditionally independent given Y.
- $X \to Y \to Z$ implies $Z \to Y \to X$. We can write $X \leftrightarrow Y \leftrightarrow Z$.
- If $Z = f(Y)$, then $X \to Y \to Z$.

Data-processing inequality

Theorem 15 (Data-processing inequality) *If* $X \to Y \to Z$, *then* $I(X;Y) \geq I(X,Z)$.

Proof.

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0}. \end{aligned}$$ □

Corollary 5 Corollary *If* $Z = g(Y)$, *we have* $I(X;Y) \geq I(X;g(Y))$.

Corollary 6 *If* $X \to Y \to Z$, *then* $I(X;Y|Z) \leq I(X;Y)$.