Elements of Information Theory

Lecture 2
Entropy and Mutual Information

Instructor: Yichen Wang

Ph.D./Professor

School of Information and Communications Engineering
Division of Electronics and Information Engineering
Xi'an Jiaotong University

Outlines

- Entropy, Joint Entropy, and Conditional Entropy
- Relative Entropy and Mutual Information
- Convexity Analysis for Entropy and Mutual Information
- Entropy and Mutual Information in Communications Systems

Entropy

Entropy —— 熵

Entropy in *Thermodynamics（热力学）*

- **It was first developed in the early 1850s by *Rudolf Clausius (French Physicist)*.**

- **System is composed of a very large number of constituents (atoms, molecule…).**

- **It is a measure of the number of the microscopic configurations that corresponds to a thermodynamic system in a state specified by certain macroscopic variables.**

- **It can be understood as a measure of molecular disorder within a macroscopic system.**

Entropy

Entropy in *Statistical Mechanics（统计力学）*

- The statistical definition was developed by *Ludwig Boltzmann* in the 1870s by analyzing the statistical behavior of the microscopic components of the system.

- *Boltzmann* showed that this definition of entropy was equivalent to the thermodynamic entropy to within a constant number which has since been known as Boltzmann's constant.

- Entropy is associated with the number of the microstates of a system.

How to define ENTROPY in information theory?

Entropy

Definition

The entropy H(X) of a discrete random variable X is defined by

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

- ***p(x)*** **is the probability mass function（概率质量函数）which can be written as**

$$p(x) = \Pr\{X = x\}, \ x \in \mathcal{X}$$

- **The base of the logarithm is *2* and the unit is *bits*.**
- **If the base of the logarithm is *e*, then the unit is *nats*.**

Entropy

Remark 1

What does the entropy measure?

----- It is a measure of the uncertainty of a random variable.

Remark 2

Must random variables with different sample spaces have different entropy?

----- It is only related with the distribution of the random variable. It does not depend on the actual values taken by the random variable, but only on the probabilities.

Remark 3

If the base of the logarithm is b, we denote the entropy as $H_b(X)$. Moreover, we have

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

Entropy

Another Explanation of Entropy

Question:
How to define information?

In information theory, information is associate with the uncertainty.

$$\begin{array}{ccc} \textit{Information} & \longrightarrow & \textit{Uncertainty} \\ \uparrow & & \downarrow \\ \textit{Probability} & \longleftarrow & \textit{Random} \end{array}$$

We use probabilistic model to describe information

- ***High probability → not so surprise → less information***
- ***Low probability → great surprise → more information***

Entropy

Another Explanation of Entropy

Example

- ***32 teams are in the FIFA World Cup 2002***
- ***Brazil, England, France, Germany, …, China***
- ***Brazil is the champion → not so surprise → less information***
- ***China wins the champion → great surprise → more information***

1. Probability reflects the prior knowledge

2. Information is defined as a function of probability

Entropy

Another Explanation of Entropy

We can define the Self-Information Function, which should satisfy the following requirements:

1. ***It should be the function of the probability that the event happens.***
2. ***It should be the decreasing function of probability that the event happens.***
3. ***If the event happens with probability ONE, the self-information should equal to ZERO.***
4. ***If the probability that the event happens is ZERO, the self-information should be INFINITE.***
5. ***The joint information of two independent events should be the SUM of the information of each event.***

Entropy

- The ***Self-Information*** of the event $X = x$ can be written as

$$I(x) = \log\left(\frac{1}{\Pr\{X = x\}}\right)$$

- If the base of the logarithm is ***2***, the unit is ***bits***.
- If the base of the logarithm is ***e***, the unit is ***nats***.

What does Self-Information imply?

1. ***Before the event occurs – The uncertainty of the event occurring;***
2. ***After the event occurs – The amount of information provided by the event.***

Entropy

Relationship Between Entropy and Self-Information

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

$$= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{p(x)}\right) \overset{p(x)=\Pr\{X=x\}}{=} \sum_{x\in\mathcal{X}} p(x)I(x) = \mathbb{E}\Big\{I(X)\Big\}$$

The above relationship tells us:

1. ***From the mathematical view – The entropy of random variable X is the expected value of the random variable*** $\log\left(1/p(X)\right)$ ***;***
2. ***From the information theory's view – The entropy of random variable X is the average self-information of X.***

Entropy

Example:

Let the random variable X equals 1 with probability p and equals 0 with probability 1-p, i.e.,

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } (1-p). \end{cases}$$

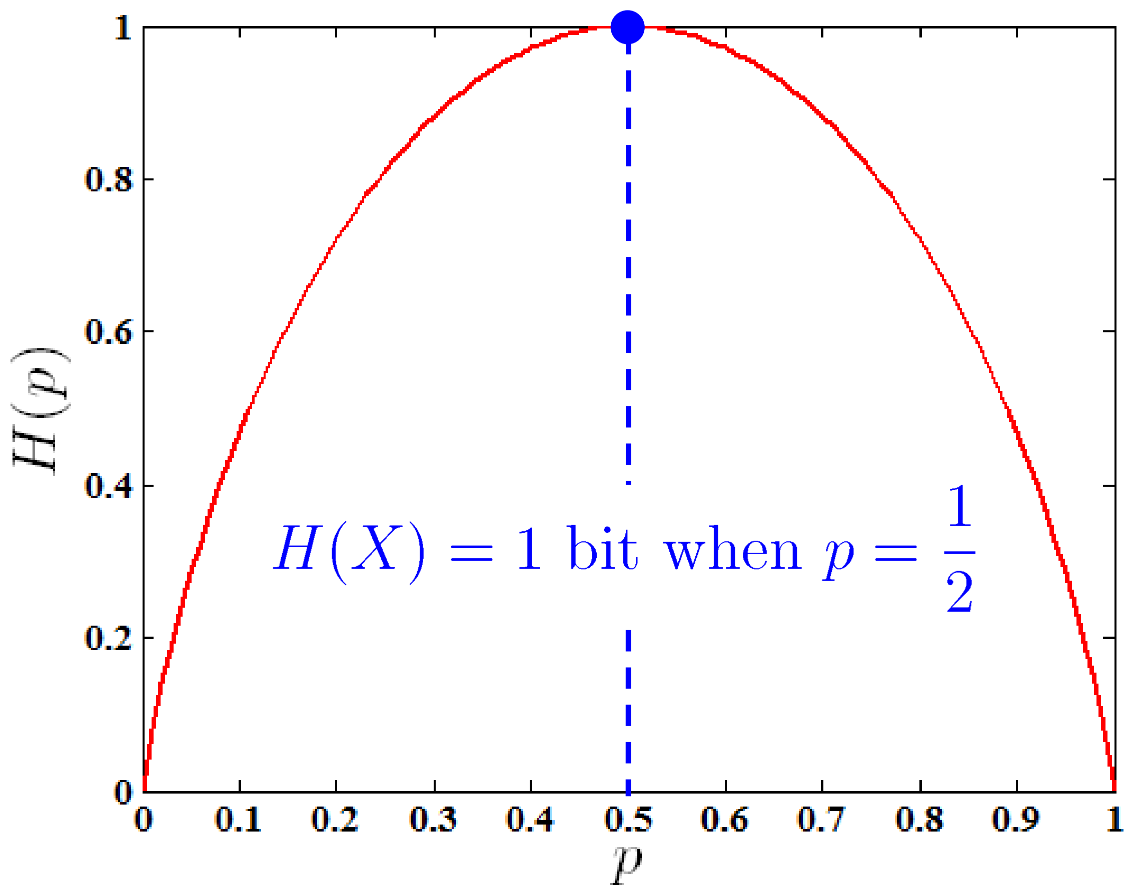
Please calculate the entropy H(X).

$$H(X) = -p \cdot \log(p) - (1-p) \cdot \log(1-p) \triangleq H(p)$$

It's easy. However, what can we obtain from this simple example?

Entropy

$$H(X) = -p \cdot \log(p) - (1-p) \cdot \log(1-p) \triangleq H(p)$$



1. ***The entropy is a concave function.***
2. ***Why does the entropy equal to zero when the value of p is 0 or 1?***
3. ***When does the entropy achieve its maximum?***

Entropy

Example:

Let the random variable X takes the value according to the following policy

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

Please calculate the entropy H(X).

Entropy

The entropy *H*(*X*) is

$$H(X) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{8}\log\left(\frac{1}{8}\right) - \frac{1}{8}\log\left(\frac{1}{8}\right) = \frac{7}{4}\ \text{bits}$$

How to determine the value of X with the minimum average number of binary questions?

1. ***First question – Is X = a? -- Splitting the probability in half***
2. ***Second question – Is X = b?***
3. ***Third question – Is X = c?***

The expected number of binary questions required is 1.75.

The minimum expected number of binary questions required to determine X lies between H(X) and H(X)+1

Entropy

Some Discussions

1. ***Before observation***

   ---- The average uncertainty of the random variable

2. ***After observation***

   ---- The average amount of information provided by each observation

3. ***Why does larger value of entropy imply higher uncertainty?***

   ---- Entropy is associated with the number of microstates of a system. Larger value of entropy means more microstates.

4. ***Continuity***

   ---- Changing the values of the probabilities by a very small amount should only change the entropy by a small amount.

Joint Entropy

- We have already defined the entropy of a **single** random variable
- Extend the definition to a **pair** of random variables – *Joint Entropy（联合熵）*

Definition

The Joint Entropy H(X,Y) of a pair of discrete random variables (X,Y) with a joint distribution p(x,y) is defined by

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

Joint Entropy

Some Discussions

1. ***In information theory, the joint entropy is a measure of the UNCERTAINTY associated with a set of random variables.***

2. ***Similar with the single variable case, the joint entropy can also be understood as***

$$H(X,Y) = -\mathbb{E}\left\{\log\, p(X,Y)\right\} = \mathbb{E}\left\{\log\frac{1}{p(X,Y)}\right\}$$

3. ***In this definition, we treat the two random variables (X,Y) as a single vector-valued random variable.***

4. ***Joint entropy in more general N random variables case***

$$H(X_1,\cdots X_N) = -\sum_{x_1\in\mathcal{X}_1}\cdots\sum_{x_N\in\mathcal{X}_N} p(x_1,\cdots,x_N)\log p(x_1,\cdots,x_N)$$

Conditional Entropy

- **Joint entropy is used for characterizing the uncertainty of a set of random variables.**
- **Observing one thing may help us predict another thing.**
- **Can we measure the uncertainty of one random variable while observing another one?**
- **The answer is YES – *Conditional Entropy（条件熵）***

Definition

The Conditional Entropy of a random variable given another random variable is defined as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable.

Conditional Entropy

Based on the above definition, if (*X*,*Y*) ~ *p*(*x*,*y*), conditional entropy *H*(*Y*|*X*) can be mathematically written as

$$H(Y|X) = \mathbb{E}_{X\sim p(x)}\left\{H\left(Y|X=x\right)\right\}$$

The average of the entropy of Y given X over all possible values of X

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

Conditional Entropy

Some Discussions

1. ***The conditional entropy H(Y|X) is a measure of what X does NOT say about Y, i.e., the amount of uncertainty remaining about Y after X is known.***

2. ***The larger the value of H(Y|X) is, the less we can predict the state of Y, knowing the state of X.***

3. ***Two extreme cases***

   Case 1: $H(Y|X) = 0 \Longleftrightarrow$ ***Y is completely determined by X***

   Case 2: $H(Y|X) = H(Y) \Longleftrightarrow$ ***X and Y are independent***

4. ***H(Y|X) = H(X|Y)?***

Conditional Entropy

We have already known:

- ***H(X) – the uncertainty of X***
- ***H(X,Y) – the uncertainty of (X,Y)***
- ***H(Y|X) – the uncertainty of Y while knowing X***

Question: Is there any relationship among the above three items?

Theorem (Chain rule)

The entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other, which can be mathematically written as

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

Conditional Entropy

Proof:

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

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

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

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

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

Is there other way to prove this theorem?

Conditional Entropy

Example

Let (X,Y) have the following joint distribution

Y \ X	1	2	3	4
1	1/8	1/16	1/32	1/32
2	1/16	1/8	1/32	1/32
3	1/16	1/16	1/16	1/16
4	1/4	0	0	0

Please calculate $H(X)$, $H(Y)$, $H(X|Y)$, $H(Y|X)$, $H(X,Y)$

Conditional Entropy

Based on the definition of entropy, we have

$$H(X) = -\sum_{i=1}^{4} p(X=i)\log p(X=i)$$

Thus, we need to derive the marginal distribution of *X*

$$p(X=i) = \sum_{j=1}^{4} p(X=i, Y=j) \Rightarrow \begin{bmatrix} X \\ p(X=i) \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{bmatrix}$$

Consequently, we have

$$H(X) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{8}\log\left(\frac{1}{8}\right) - \frac{1}{8}\log\left(\frac{1}{8}\right) = \frac{7}{4}\text{ bits}$$

Similarly, we can calculate *H*(*Y*)

$$\begin{bmatrix} Y \\ p(Y=j) \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \Rightarrow H(Y) = 2\text{ bits}$$

Conditional Entropy

Based on the definition of conditional entropy, we have

$$\begin{aligned} H(X|Y) &= \sum_{j=1}^{4} p(Y=j) H(X|Y=j) \\ &= -\sum_{j=1}^{4} p(Y=j) \sum_{i=1}^{4} p(X=i|Y=j) \log p(X=i|Y=j) \end{aligned}$$

where $p(X=i|Y=j) = \dfrac{p(X=i, Y=j)}{p(Y=j)}$

Then, we can obtain

$$\begin{aligned} H(X|Y) &= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) \\ &\quad + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\ &= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 = \frac{11}{8} \text{ bits} \end{aligned}$$

Conditional Entropy

$$
\begin{aligned}
H(Y|X) &= \sum_{i=1}^{4} p(X=i) H(Y|X=i) \\
&= -\sum_{i=1}^{4} p(X=i) \sum_{j=1}^{4} p(Y=j|X=i) \log p(Y=j|X=i)
\end{aligned}
$$

where $p(Y=j|X=i) = \dfrac{p(X=i, Y=j)}{p(X=i)}$

Then, we can obtain

$$
\begin{aligned}
H(Y|X) &= \frac{1}{2} H\left(\frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{2}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0\right) \\
&\quad + \frac{1}{8} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0\right) + \frac{1}{8} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0\right) \\
&= \frac{1}{2} \times \frac{7}{4} + \frac{1}{4} \times \frac{3}{2} + \frac{1}{8} \times \frac{3}{2} + \frac{1}{8} \times \frac{3}{2} = \frac{13}{8} \text{ bits}
\end{aligned}
$$

$$
H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) = \frac{27}{8} \text{ bits}
$$

Conditional Entropy

Example

Suppose probability distribution of random variable X are given as

X	a_1	a_2	a_3
$p(x)$	11/36	4/9	1/4

and the conditional probability $P(a_j|a_i)$ ***are given as***

		a_j		
a_i		a_1	a_2	a_3
	a_1	9/11	2/11	0
	a_2	1/8	3/4	1/8
	a_3	0	2/9	7/9

Please calculate $H(X^2)$

Conditional Entropy

Based on the definition of joint entropy, we have

$$H\left(X^2\right) = -\sum_{i=1}^{3}\sum_{j=1}^{3} p\left(a_i, a_j\right)\log p\left(a_i, a_j\right)$$

Thus, we need to calculate the joint probability $p\left(a_i, a_j\right)$

$$p\left(a_i, a_j\right) = p\left(a_i\right)p\left(a_j \middle| a_i\right) \begin{cases} p(a_1, a_1) = p(a_1)p(a_1|a_1) = \dfrac{11}{36} \times \dfrac{9}{11} = \dfrac{1}{4} \\ p(a_1, a_2) = p(a_1)p(a_2|a_1) = \dfrac{11}{36} \times \dfrac{2}{11} = \dfrac{1}{18} \\ \vdots \\ p(a_3, a_3) = p(a_3)p(a_3|a_3) = \dfrac{1}{4} \times \dfrac{7}{9} = \dfrac{7}{36} \end{cases}$$

$$\Downarrow$$

$$H\left(X^2\right) = 2.412 \text{ bits}$$

Some Properties

Properties of Entropy, Joint Entropy, and Conditional Entropy

1. Nonnegativity of entropy

$$H(X) \geq 0$$

2. Symmetry

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ p_1 & p_2 & \cdots & p_N \end{bmatrix} \longrightarrow H(X) = H(p_1, p_2, \cdots, p_N)$$

$$\Downarrow$$

$$H(p_1, p_2, \cdots, p_N) = H(p_2, p_3, \cdots, p_N, p_1) = \cdots = H(p_N, p_1, \cdots, p_{N-1})$$

Some Properties

3. Maximum

Suppose random variable X follows the following distribution

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ p(x_1) & p(x_2) & \cdots & p(x_N) \end{bmatrix}$$

Then, we have the following inequality

$$H(X) \leq \log N$$

with the equality if and only if X has a uniform distribution, i.e.,

$$p(x_1) = p(x_2) = \cdots = p(x_N) = \frac{1}{N}$$

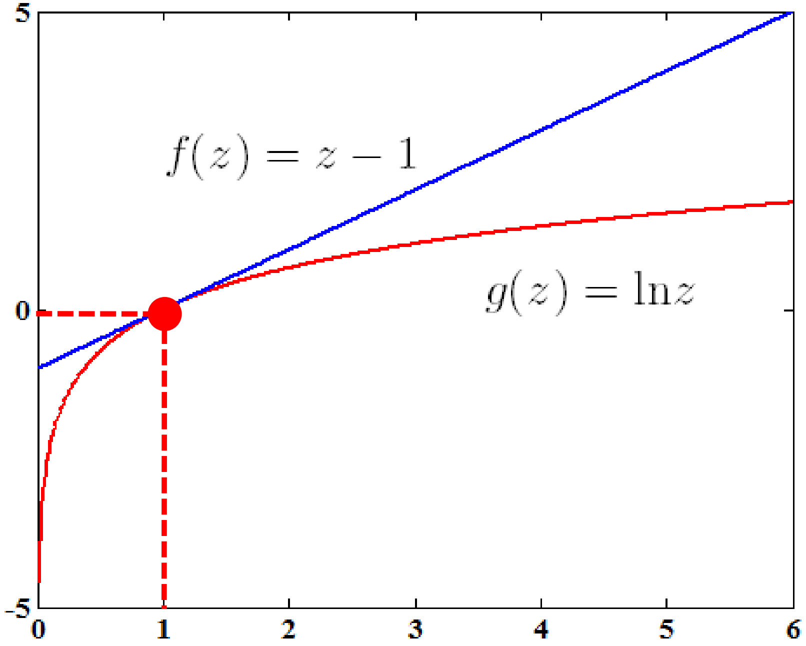
Some Properties

Proof

To prove the "Maximum" property, we need to use the following inequality:

$$\ln z \le z - 1,\ \ z \ge 0$$

with the equality if and only is z =1.



$$t(z) = \ln z - (z-1)$$

$$t'(z) = \frac{1}{z} - 1 \qquad t''(z) = -\frac{1}{z^2} \le 0$$

The difference has a negative second derivative and a stationary point at z=1

Some Properties

We now show that $H(X) - \log N \leq 0$

$$
\begin{aligned}
H(X) - \log N &= \sum_{i=1}^{N} p(x_i) \log \frac{1}{p(x_i)} - \sum_{i=1}^{N} p(x_i) \log N \\
&= (\log e) \sum_{i=1}^{N} p(x_i) \ln \frac{1}{p(x_i) \cdot N}
\end{aligned}
$$

By applying the abovementioned inequality, we can obtain

$$
\begin{aligned}
H(X) - \log N &\leq (\log e) \sum_{i=1}^{N} p(x_i) \left[\frac{1}{p(x_i) \cdot N} - 1 \right] \\
&= (\log e) \left[\sum_{i=1}^{N} \frac{1}{N} - \sum_{i=1}^{N} p(x_i) \right] = 0
\end{aligned}
$$

Some Properties

4. Adding or removing an event with probability zero does not contribute to the entropy

$$H_{N+1}(p_1, \cdots, p_N, 0) = H_N(p_1, \cdots, p_N)$$

5. Chain rule

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

If X and Y are independent, we have

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

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

General case: $(X_1, X_2, \cdots, X_N) \sim p(x_1, x_2, \cdots, x_N)$

$$H(X_1, X_2, \cdots, X_N) = \sum_{i=1}^{N} H(X_i|X_{i-1}, \cdots, X_1)$$

Some Properties

Another explanation for chain rule

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ p_1 & p_2 & \cdots & p_N \end{bmatrix} \qquad \begin{bmatrix} Y \\ q(y) \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \cdots & y_M \\ q_1 & q_2 & \cdots & q_M \end{bmatrix}$$

$$Q_{mn} = \Pr\{Y = y_m | X = x_n\},\ m = 1, \cdots, M\ ;\ n = 1, \cdots, N$$

$$\boxed{\begin{aligned} &H\left(p_1 Q_{11}, \cdots, p_1 Q_{M1}, p_2 Q_{12}, \cdots, p_2 Q_{M2}, \cdots, p_N Q_{1N}, \cdots, p_N Q_{MN}\right) \\ &\qquad = H\left(p_1, p_2, \cdots, p_N\right) + \sum_{n=1}^{N} p_n H\left(Q_{1n}, Q_{2n}, \cdots, Q_{Mn}\right) \end{aligned}}$$

What can we obtain from the above equality?

Some Properties

6. Conditioning reduces entropy

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

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

Corollary:

$$H(X,Y) \le H(X) + H(Y)$$

$$H\left(X_1, X_2, \cdots, X_N\right) \le \sum_{i=1}^{N} H\left(X_i\right)$$

Independence bound on entropy

Outlines

- Entropy, Joint Entropy, and Conditional Entropy
- Relative Entropy and Mutual Information
- Convexity Analysis for Entropy and Mutual Information
- Entropy and Mutual Information in Communications Systems

Relative Entropy

Definition (Relative Entropy)

The Relative Entropy（相对熵）or Kullback-Leibler Divergence（K-L散度）between two probability mass functions p(x) and q(x) is defined as

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

- ***We use the conventions that $0 \cdot \log(0/0) = 0$, $0 \cdot \log(0/q) = 0$, and $p \cdot \log(p/0) = \infty$.***
- ***Does symmetry hold for relative entropy, i.e. $D(p||q) = D(q||p)$?***

How to understand relative entropy?

Mutual Information

Definition (Mutual Information)

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

$$I(X;Y) = D\left(p(x,y)||p(x)p(y)\right)$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = \mathbb{E}_{X,Y}\left\{\log\frac{p(X,Y)}{p(X)p(Y)}\right\}$$

What does the mutual information imply?

Mutual Information

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x|y)}{p(x)} \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x|y) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \left(-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x|y)\right) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

The mutual information is the <u>reduction</u> in the uncertainty of X due to the knowledge of Y.

Mutual Information

Some Discussions

1. ***Mutual information measures the amount of uncertainty of X removed by knowing Y. In other words, this is the amount of information obtained about X by knowing Y.***

2. ***Symmetry*** $I(X;Y) = I(Y;X)$

3. ***Relationship between mutual information and entropy***

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

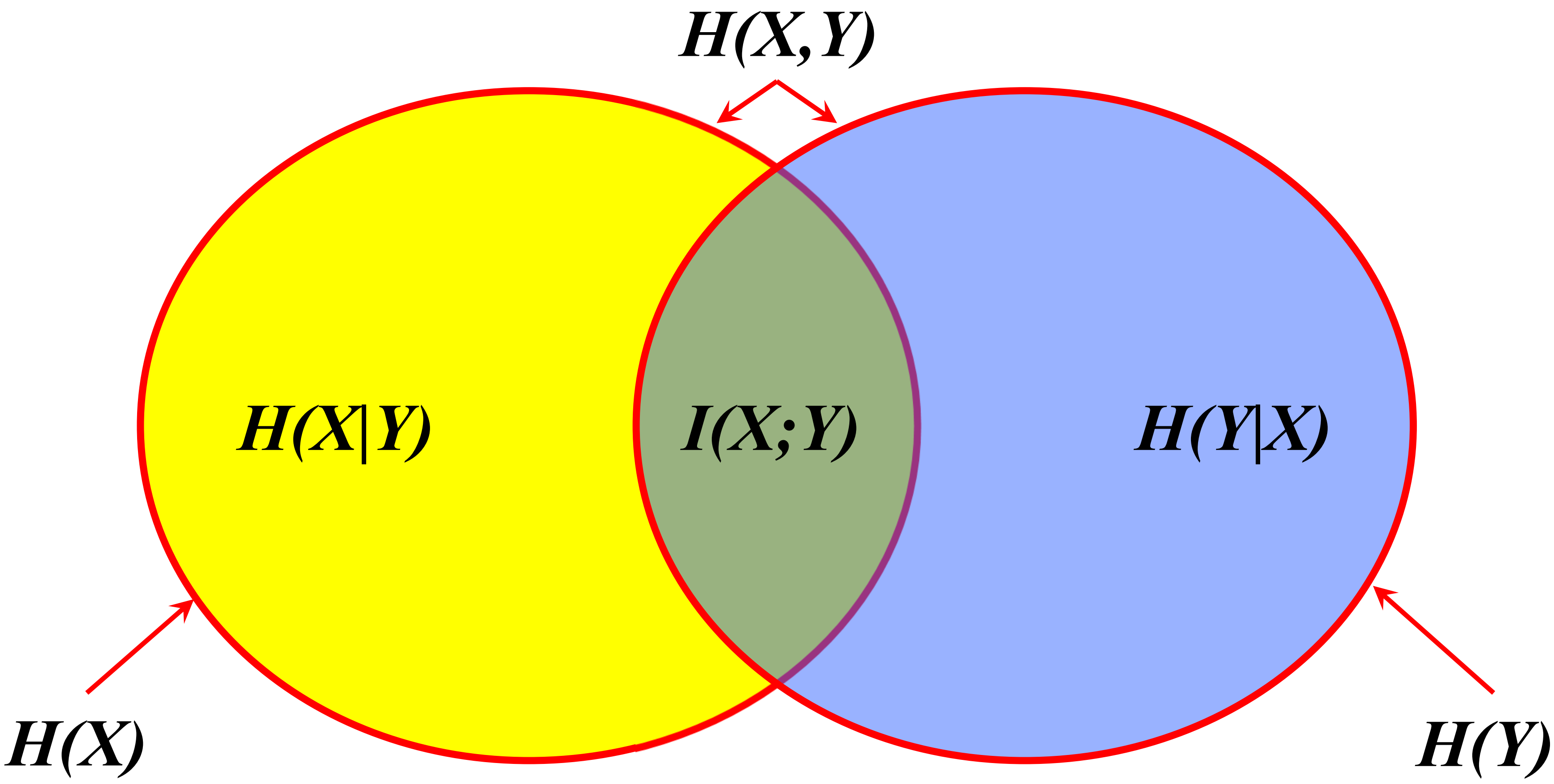
$$I(X;X) = H(X)$$

4. ***Chain rule for information***

$$I\left(X_1, X_2, \cdots, X_n; Y\right) = \sum_{i=1}^{n} I\left(X_i; Y | X_{i-1}, X_{i-2}, \cdots, X_1\right)$$

Mutual Information

Relationship among mutual information, entropy, joint entropy, and conditional entropy



Mutual Information

Definition (Markov Chain)

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

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

- ***$X \to Y \to Z$ implies that $Z \to Y \to X$; If $Z = f(Y)$, then $X \to Y \to Z$.***

Theorem (Data Processing Inequality)

If $X \to Y \to Z$, then $I(X; Y) \geq I(X; Z)$. In particular, if $Z = g(Y)$, we have $I(X; Y) \geq I(X; g(Y))$.

Mutual Information

- ***Suppose that we wish to estimate a random variable X with a distribution p(x).***

- ***We observe a random variable Y that is related to X by the conditional distribution p(y|x).***

- ***From Y, we calculate a function*** $g(Y)=\hat{X}$ ***, where*** $\hat{X}$ ***is an estimate of X and takes on values in*** $\hat{\mathcal{X}}$ ***.***

$$\Downarrow$$

$$\boxed{X \longrightarrow Y \longrightarrow \hat{X} \ \textit{forms a Markov Chain}}$$

- ***Define the probability of error***

$$P_e = \Pr\left\{\hat{X} \neq X\right\}$$

Mutual Information

Theorem (Fano's Inequality)

For any estimator $\hat{X}$ such that

$$X \longrightarrow Y \longrightarrow \hat{X}$$

with $P_e = \Pr\{\hat{X} \neq X\}$, we have

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

This inequality can be weakened to

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

or

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

Mutual Information

Proof

Define an error random variable

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

Then, we have

$$\begin{aligned} H\left(E, X|\hat{X}\right) &= H\left(X|\hat{X}\right) + \boxed{H\left(E|X, \hat{X}\right)} \quad \mathbf{= 0} \\ &= \boxed{H\left(E|\hat{X}\right)} + \underline{H\left(X|E, \hat{X}\right)} \quad \mathbf{= ?} \end{aligned}$$

Conditioning reduces entropy

$$H\left(E|\hat{X}\right) \leq H(E) = H\left(P_e\right)$$

Mutual Information

Proof

Based on the definition of conditional entropy, we have

$$H\left(X|E,\hat{X}\right) = \underline{\Pr\left\{E=0\right\}H\left(X|\hat{X},E=0\right)} = (1-P_e)0$$

$$\underline{+\Pr\left\{E=1\right\}H\left(X|\hat{X},E=1\right)} \le P_e \log\left|\mathcal{X}\right|$$

Then, we have

$$H(P_e) + P_e \log\left|\mathcal{X}\right| \ge H\left(X|\hat{X}\right)$$

By applying the data-processing inequality, we can obtain

$$H\left(X|\hat{X}\right) \ge H(X|Y)$$

Outlines

- Entropy, Joint Entropy, and Conditional Entropy
- Mutual Information
- Convexity Analysis for Entropy and Mutual Information
- Entropy and Mutual Information in Communications Systems

Convexity Analysis

Convex function (Cup)

- ***A function f(x) is said to be convex over an interval (a, b) if for every*** $x_1, x_2 \in (a, b)$ ***and 0 ≤ λ ≤ 1, the following inequality holds***

$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

- ***A function f(x) is said to be strictly convex if the equality holds only if λ = 0 or λ = 1.***

Concave function (Cap)

- ***f(x) is concave over (a, b) if for every*** $x_1, x_2 \in (a, b)$ ***, we have***

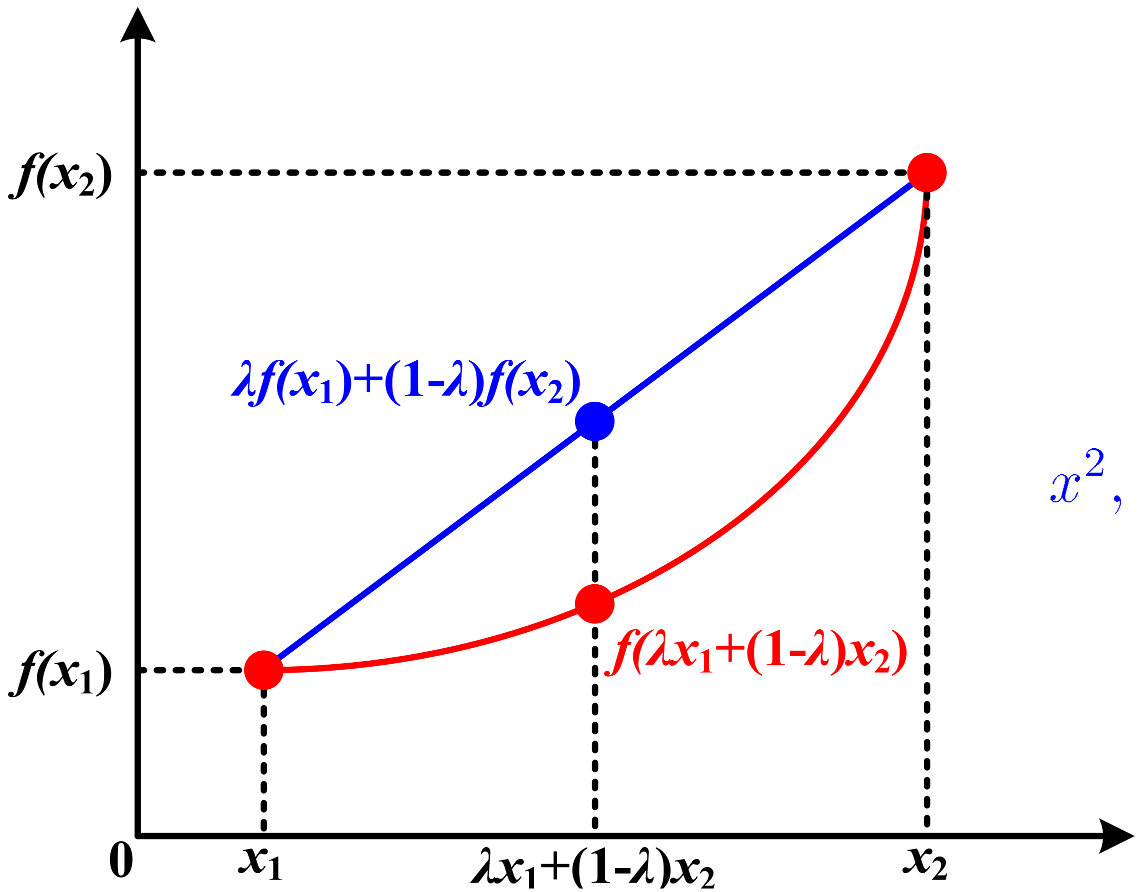
$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \ge \lambda f(x_1) + (1-\lambda) f(x_2), 0 \le \lambda \le 1$$

- ***f(x) is strictly concave if the equality holds only if λ = 0 or λ = 1.***

Function f(x) is convex, then we have –f(x) is concave.

Convexity Analysis

➢ Illustration for Convex Function

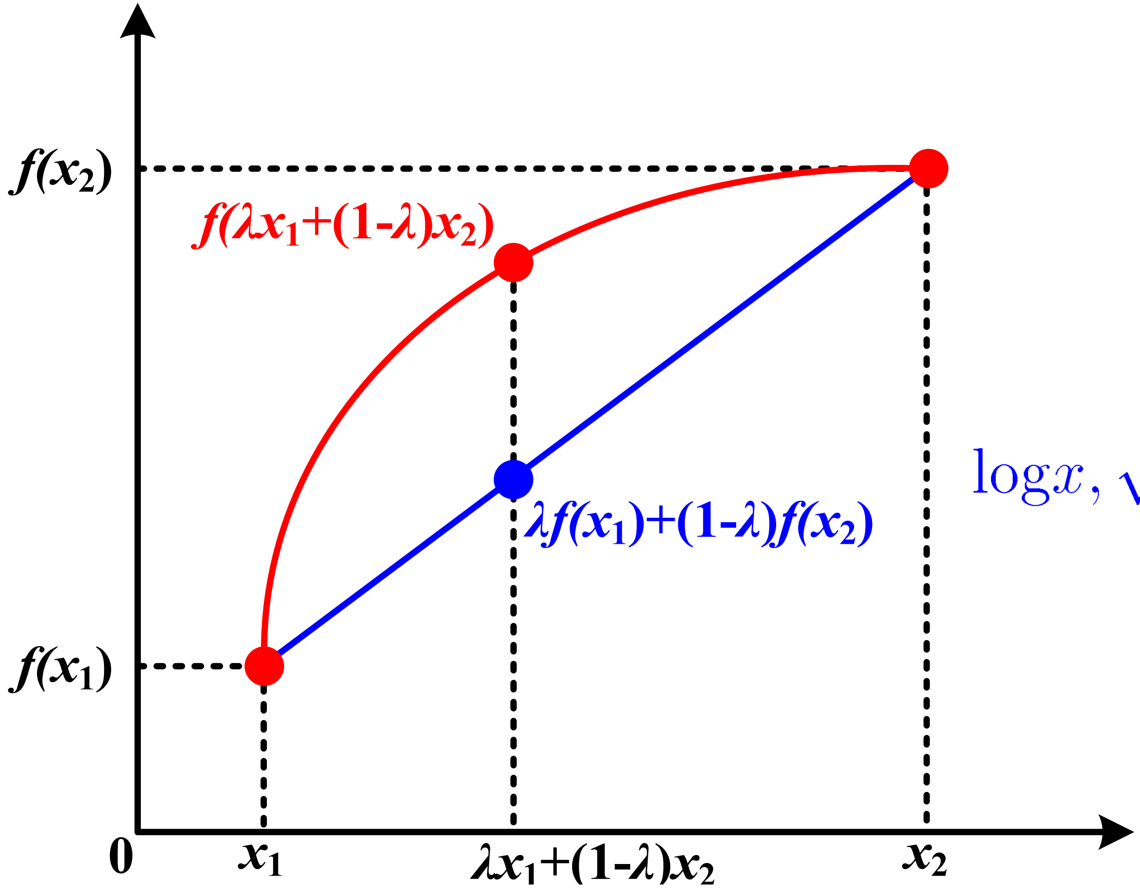


$x^2, |x|, e^x \cdots$

Convex (cup): Function always lies below any chord

Convexity Analysis

➢ **Illustration for Concave Function**



$\log x, \sqrt{x}(x \geq 0) \cdots$

Concave (cap): Function always lies above any chord

Convexity Analysis

Is there other approach to determine the convexity of a function?

Theorem

If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval. Mathematically,

- ***If*** $d^2 f(x)/dx^2 \geq 0$ ***holds, then f(x) is convex;***
- ***If*** $d^2 f(x)/dx^2 > 0$ ***holds, then f(x) is strictly convex.***

How can we extend the above theorem to a more general case $f(x_1, x_2, \ldots, x_N)$***?***

Convexity Analysis

Theorem (Jensen's inequality)

If f is a convex function and X is a random variable, then we have

$$\mathbb{E}\left\{f(X)\right\} \geq f\left(\mathbb{E}\{X\}\right)$$

Moreover, if f is strictly convex, the above equality implies that $X = \mathbb{E}\{X\}$ ***with probability 1 (i.e., X is a constant).***

- *Here we only consider the discrete random variable case*
- *We can employ Mathematical Induction to prove the above theorem*

Convexity Analysis

Proof

For a two-mass-point distribution

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ p_1 & p_2 \end{bmatrix}$$

the inequality becomes

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

The above inequality apparently holds as *f* is a convex function.

Suppose that the theorem is true for distributions with (*k*-1) mass points. Then, we prove it is true for *k*-mass-point distributions.

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \\ p_1 & p_2 & \cdots & p_k \end{bmatrix} \Rightarrow p_i' = \frac{p_i}{1-p_k}, i = 1, \cdots, k-1$$

Convexity Analysis

Then, we have

$$\sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1-p_k)\sum_{i=1}^{k-1} p_i' f(x_i)$$

$$\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right)$$

$$\geq f\left(p_k x_k + (1-p_k)\sum_{i=1}^{k-1} p_i' x_i\right) = f\left(\sum_{i=1}^{k} p_i x_i\right)$$

Question:

1. When does the equality hold if X is not a constant?

2. Why can we obtain the conclusion, i.e., the strict convexity of function f implies X is a constant?

Convexity Analysis

- *The expectation of a convex function (cup) of a random variable is no smaller than the convex function (cup) of the expectation of the random variable.*

- *The expectation of a concave function (cap) of a random variable is no larger than the concave function (cap) of the expectation of the random variable.*

Famous Puzzle:

A man says, "I am the average height and average weight of the population. Thus, I am an average man." However, he is still considered to be a little overweight. Why?

Convexity Analysis

Recalled that we have discussed the *"Maximum" property* of entropy. Now, Let's discuss it again.

Theorem (Uniform maximizes entropy)

$H(X) \le \log|\mathcal{X}|$ ***, where*** $|\mathcal{X}|$ ***denotes the number of elements in the range of X, with equality if and only X has a uniform distribution over*** $\mathcal{X}$**.**

Let $u(x) = 1/|\mathcal{X}|$ ***be the uniform probability mass function over*** $\mathcal{X}$ ***, and let p(x) be the probability mass function for X. Then, we have***

$$
\begin{aligned}
H(X) - \log|\mathcal{X}| &= \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} + \sum_{x\in\mathcal{X}} p(x)\log u(x) = \sum_{x\in\mathcal{X}} p(x)\log\frac{u(x)}{p(x)} \\
&\le \log\left(\sum_{x\in\mathcal{X}} p(x)\frac{u(x)}{p(x)}\right) = \log\left(\sum_{x\in\mathcal{X}} u(x)\right) = 0
\end{aligned}
$$

Convexity Analysis

We have known that the entropy is nonnegative, i.e., $H(X) \geq 0$

How about the mutual information?

Theorem (Nonnegative of mutual information)

For any two random variables X and Y, we have

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

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

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \\ &\geq -\log \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \frac{p(x)p(y)}{p(x,y)} \right) = -\log \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x)p(y) \right) = 0 \end{aligned}$$

If X and Y are independent, we have p(x,y)=p(x)p(y)

Convexity Analysis

Based on the theory of convex optimization, we can obtain that *the sum of convex (concave) functions is also a convex (concave) function*.

Theorem (Concavity of entropy)

The entropy of a random variable is a concave (cap) function.

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ p(x_1) & p(x_2) & \cdots & p(x_N) \end{bmatrix} \longrightarrow H(X) = -\sum_{i=1}^{N} p(x_i)\log p(x_i)$$

$$\downarrow$$

f(p) is concave over p $\longleftarrow f''(p) = -\log(e)\frac{1}{p} < 0 \longleftarrow f(p) = -p\log p$

$$\downarrow$$

H(X) is the sum of f(p) with different values of p. Thus, H(X) is concave.

Convexity Analysis

Theorem

Let (X,Y) ~ p(x,y)=p(x)p(y|x). Then, we can obtain that the mutual information I(X;Y) is a concave function of p(x) for fixed p(y|x).

Proof

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y|x)\log\left(\frac{p(y|x)}{\sum_{x\in\mathcal{X}} p(x)p(y|x)}\right)$$

The mutual information I(X;Y) is the function of p(x) $\longrightarrow I(X;Y) \triangleq I\{p(x)\}$

$$I\{\lambda p_1(x) + (1-\lambda)p_2(x)\} \geq \lambda I\{p_1(x)\} + (1-\lambda)I\{p_2(x)\}?$$

Convexity Analysis

For different distributions $p_1(x)$ and $p_2(x)$, we have

$$p_1(x, y) = p_1(x)p(y|x)$$

$$p_2(x, y) = p_2(x)p(y|x)$$

$$p_1(y) = \sum_{x \in \mathcal{X}} p_1(x, y) = \sum_{x \in \mathcal{X}} p_1(x)p(y|x)$$

$$p_2(y) = \sum_{x \in \mathcal{X}} p_2(x, y) = \sum_{x \in \mathcal{X}} p_2(x)p(y|x)$$

If we denote $p(x) = \lambda_1 p_1(x) + \lambda_2 p_2(x)$ ***, where*** $\lambda_1 + \lambda_2 = 1$ ***,***
we can obtain

$$\begin{aligned} p(x, y) &= p(x)p(y|x) \\ &= \Big[\lambda_1 p_1(x) + \lambda_2 p_2(x)\Big] p(y|x) = \lambda_1 p_1(x, y) + \lambda_2 p_2(x, y) \end{aligned}$$

Convexity Analysis

$$I\{p(x)\} - \lambda_1 I\{p_1(x)\} - \lambda_2 I\{p_2(x)\}$$

$$= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} - \sum_{x,y}\lambda_1 p_1(x,y)\log\frac{p_1(x,y)}{p_1(x)p_1(y)} - \sum_{x,y}\lambda_2 p_2(x,y)\log\frac{p_2(x,y)}{p_2(x)p_2(y)}$$

$$= \sum_{x,y}\left[\lambda_1 p_1(x,y) + \lambda_2 p_2(x,y)\right]\log\frac{p(y|x)}{p(y)}$$

$$-\sum_{x,y}\lambda_1 p_1(x,y)\log\frac{p(y|x)}{p_1(y)} - \sum_{x,y}\lambda_2 p_2(x,y)\log\frac{p(y|x)}{p_2(y)}$$

$$= \sum_{x,y}\lambda_1 p_1(x,y)\log\frac{p_1(y)}{p(y)} + \sum_{x,y}\lambda_2 p_2(x,y)\log\frac{p_2(y)}{p(y)}$$

$$= -\sum_{y}\lambda_1 p_1(y)\log\frac{p(y)}{p_1(y)} - \sum_{y}\lambda_2 p_2(y)\log\frac{p(y)}{p_2(y)}$$

$$\geq -\lambda_1\log\left(\sum_{y} p_1(y)\frac{p(y)}{p_1(y)}\right) - \lambda_2\log\left(\sum_{y} p_2(y)\frac{p(y)}{p_2(y)}\right) = 0$$

Convexity Analysis

Using definition for proving is sometimes quite complicated. Thus, we here provide another simple way.

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

- ***As p(y|x) is fixed, p(y) is a linear function of p(x)***

- ***H(Y) is the concave function of p(y). Thus, it is also the concave function of p(x)***

- ***H(Y|X) is a linear function of p(x)***

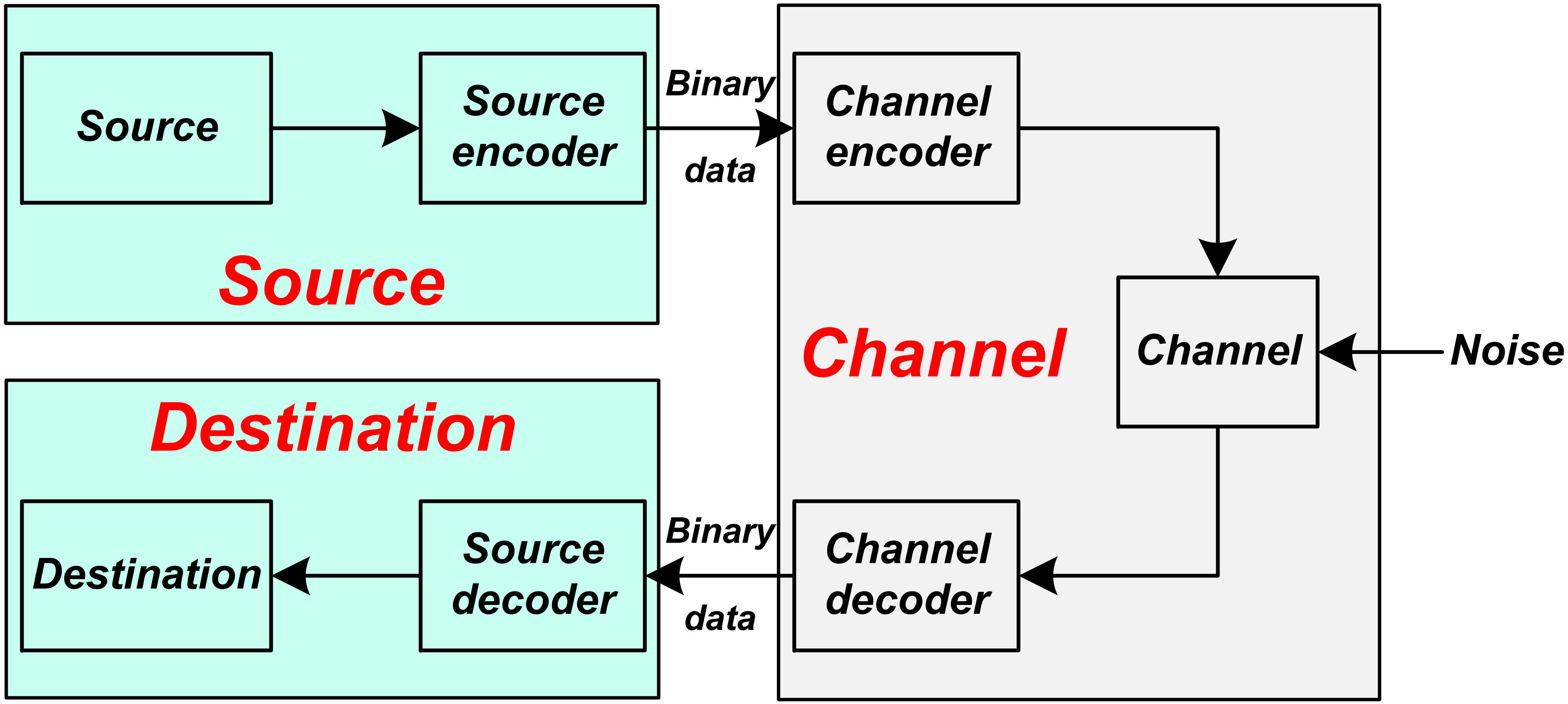
- ***Consequently, I(X;Y) is the concave function of p(x)***

Outlines

- Entropy, Joint Entropy, and Conditional Entropy
- Mutual Information
- Convexity Analysis for Entropy and Mutual Information
- Entropy and Mutual Information in Communications Systems

Explanation in Communications

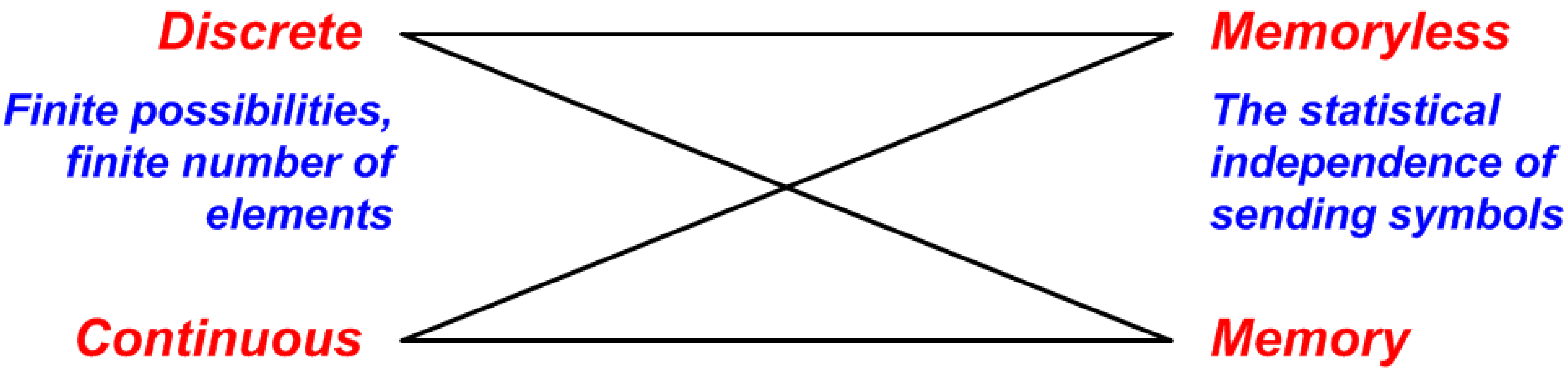
Block Diagram of Communication System



Explanation in Communications

Source

- The *Source* is the source of information.
- How to categorize "*Source*"?
  - *Discrete Source (The output is a sequence of symbols from a known discrete alphabet, e.g., English letters, Chinese characters.) and Continuous Source (Analog Waveform Source, the output is an analog real waveform, i.e., speech, image, video)*
  - *Memoryless (The outputs of source are statistically independent.) and Memory (The outputs are dependent.)*



Explanation in Communications

Source

- The ***Source*** is the source of information.
- How to categorize "***Source***"?
  - ***Discrete Source*** *(The output is a sequence of symbols from a known discrete alphabet, e.g., English letters, Chinese characters.)* ***and Continuous Source*** *(**Analog Waveform Source**, the output is an analog real waveform, i.e., speech, image, video)*
  - ***Memoryless*** *(The outputs of source are statistically independent.)* *and **Memory** (The outputs are dependent.)*

Example

10 black balls and 10 white balls in a bag

✓ Take a ball and put it back -- Memoryless

✓ Take a ball, but do not put it back -- Memory

Explanation in Communications

Source

K-order memory: If the currently transmitted symbol correlates with previously transmitted K symbols, the source is K-order discrete memory source.

1-order memory: Currently transmitted symbol only correlates with previously transmitted one symbol.

Question:

What will the memory result?

Explanation in Communications

Example

Suppose probability distribution of random variable X are given as

X	a_1	a_2	a_3
$p(x)$	11/36	4/9	1/4

and the conditional probability $P\left(a_j|a_i\right)$ ***are given as***

a_i	a_j			
		a_1	a_2	a_3
	a_1	9/11	2/11	0
	a_2	1/8	3/4	1/8
	a_3	0	2/9	7/9

Please calculate $H(X^2)$

Explanation in Communications

$$H\left(X^2\right) = -\sum_{i=1}^{3}\sum_{j=1}^{3} p\left(a_i, a_j\right)\log p\left(a_i, a_j\right) = 2.412 \text{ bits}$$

$$H\left(X\right) = -\sum_{i=1}^{3} p\left(a_i\right)\log p\left(a_i\right) = 1.542 \text{ bits}$$

$$H\left(X|X\right) = -\sum_{i=1}^{3}\sum_{j=1}^{3} p\left(a_i, a_j\right)\log p\left(a_i|a_j\right) = 0.870 \text{ bits}$$

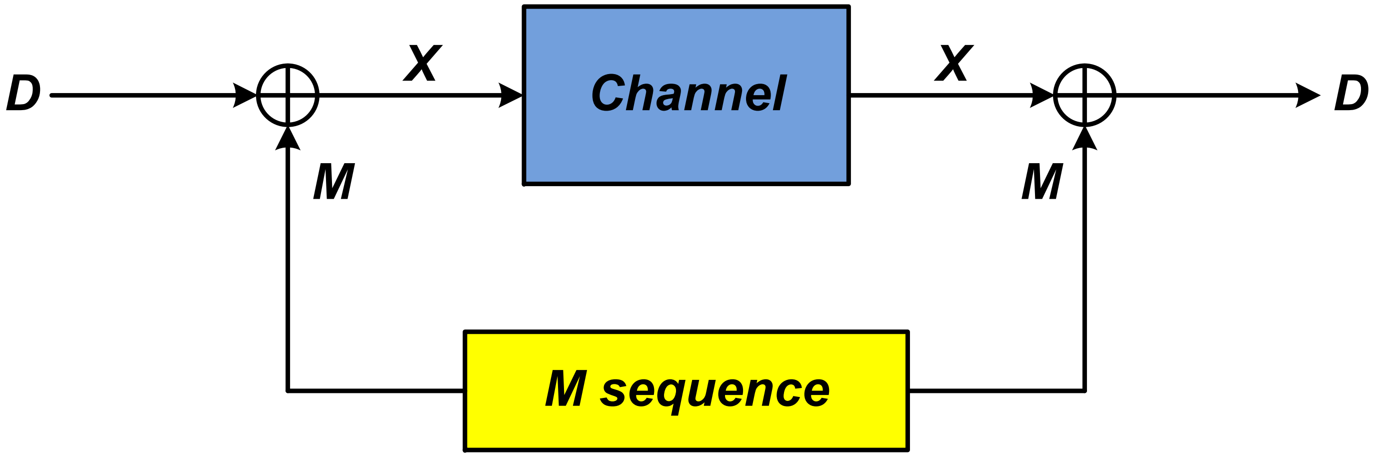
$$\Downarrow$$

$$H\left(X^2\right) = H\left(X\right) + H\left(X|X\right) < 2H\left(X\right)$$

Memory will reduce the amount of information of the source

Explanation in Communications

In realistic communication system, memory source can be transformed to memoryless source by scrambling.



$$D \oplus M \oplus M = D \oplus (M \oplus M) = D \oplus 0 = D$$

$$P(X=1) = P(D=0, M=1) + P(D=1, M=0) = \frac{1}{2}P(D=0) + \frac{1}{2}P(D=1) = \frac{1}{2}$$

$$P(X=0) = P(D=0, M=0) + P(D=1, M=1) = \frac{1}{2}P(D=0) + \frac{1}{2}P(D=1) = \frac{1}{2}$$

Explanation in Communications

Source Encoder -- Source Coding

- **Why should we use *Source Coding*?**
  - *Represent the source output by a sequence of binary digits*
  - *Data compression or bit-rate reduction*
- **Examples**
  - *Text – ASCII (128 symbols, 7 bits), GB2312 (6763 characters, at least 13 bits, actually 14 bits)*
  - *Voice – CD, MP3*
  - *Image – JEPG*
  - *Video – MPEG-1, MPEG-2, MPEG-4, RMVB*

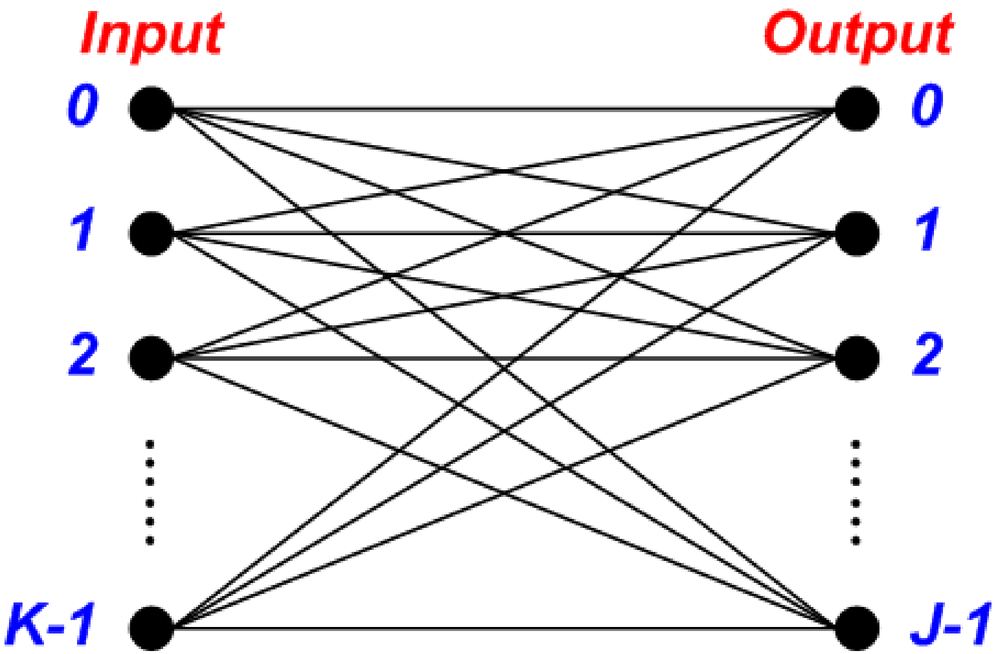
Explanation in Communications

Communication Channel

- ***Channel*** **is viewed as the part of the communication system between source and destination that is given and not under the control of designer.**
- **The *Channel* can be specified in terms of the set of *inputs* available at the input terminal, the set of *outputs* available at the output terminal, and for each input the *probability measure* on the output events conditional on that input**
  - ***Discrete memoryless channel***
  - ***Continuous amplitude, discrete-time memoryless channel***
  - ***Continuous time channel in which the input and output are waveforms***
  - ***Discrete channel with memory***

Explanation in Communications

Discrete Memoryless Channel (DMC)

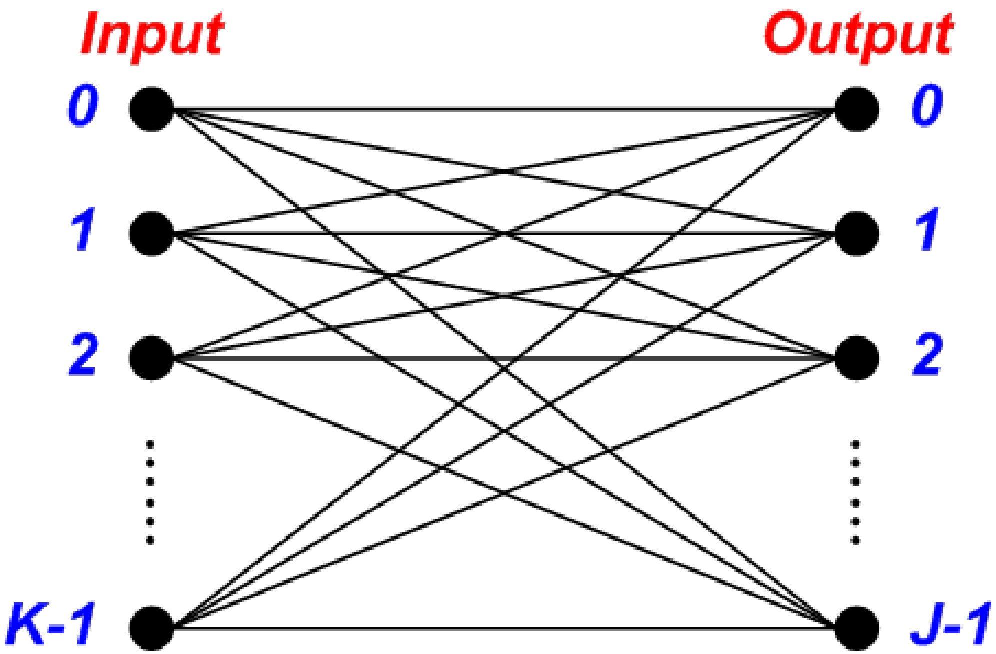


- ***Input alphabet X consists of K integers 0, 1, … , K-1***
- ***Output alphabet Y consists of J integers 0 , 1, … , J-1***

The channel is specified by transition probability P(j|k): The probability of receiving integers j given that integer k is the channel input.

Explanation in Communications

<u>Discrete Memoryless Channel (DMC)</u>



- ***A sequence of N input:***

$$\mathbf{x} = (x_1, \cdots, x_n, \cdots, x_N)$$

- ***The sequence of output:***

$$\mathbf{y} = (y_1, \cdots, y_n, \cdots, y_N)$$

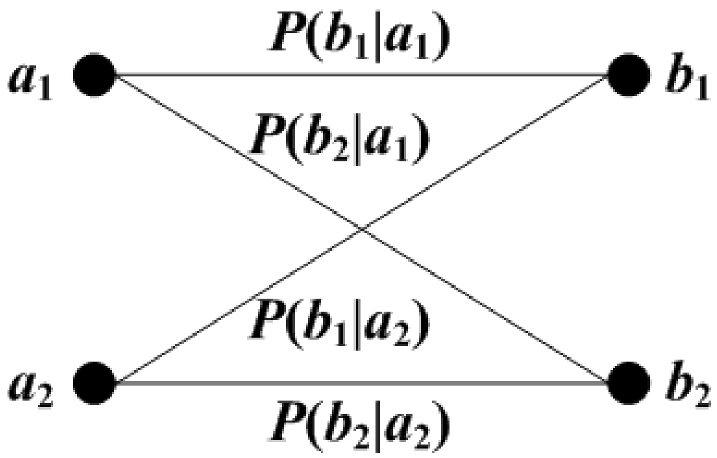
$$\boxed{P_N(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} P(y_n|x_n)}$$

More formally, a channel is memoryless if there is a transition probability assignment, P(j|k), such that the above equality is satisfied for all N, all y = (y_1, … , y_N) and all x = (x_1, … , x_N).

Explanation in Communications

Example 1:

Binary Discrete Memoryless Channel (BDMC)



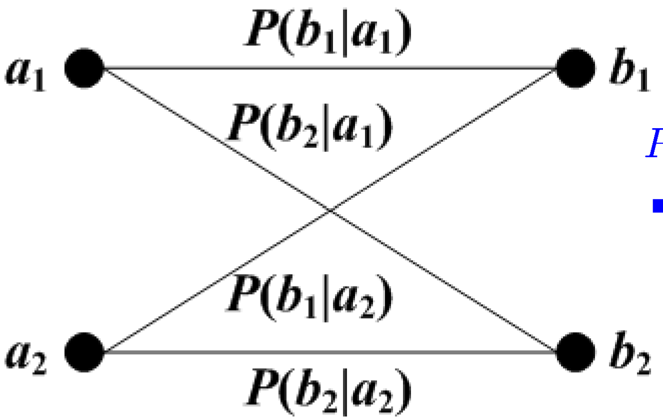
$P(b_1|a_1)$: The probabiliry of receiving b_1 on the condition of sending a_1

$P(b_2|a_1)$: The probabiliry of receiving b_2 on the condition of sending a_1

$P(b_1|a_2)$: The probabiliry of receiving b_1 on the condition of sending a_2

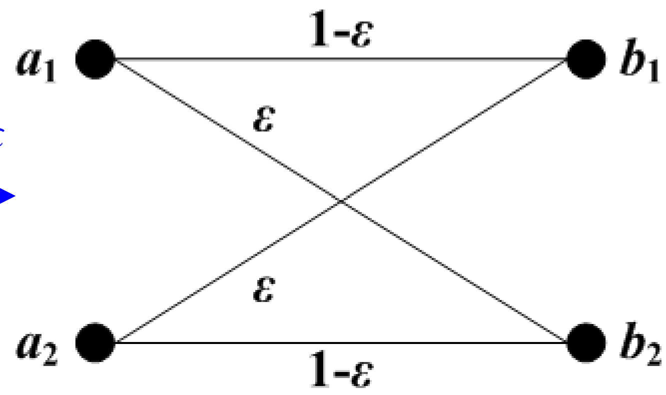
$P(b_2|a_2)$: The probabiliry of receiving b_2 on the condition of sending a_2

Explanation in Communications



$P(b_1|a_1) = P(b_2|a_2) = 1 - \varepsilon$

$P(b_2|a_1) = P(b_1|a_2) = \varepsilon$



Binary Symmetric Channel (BSC)

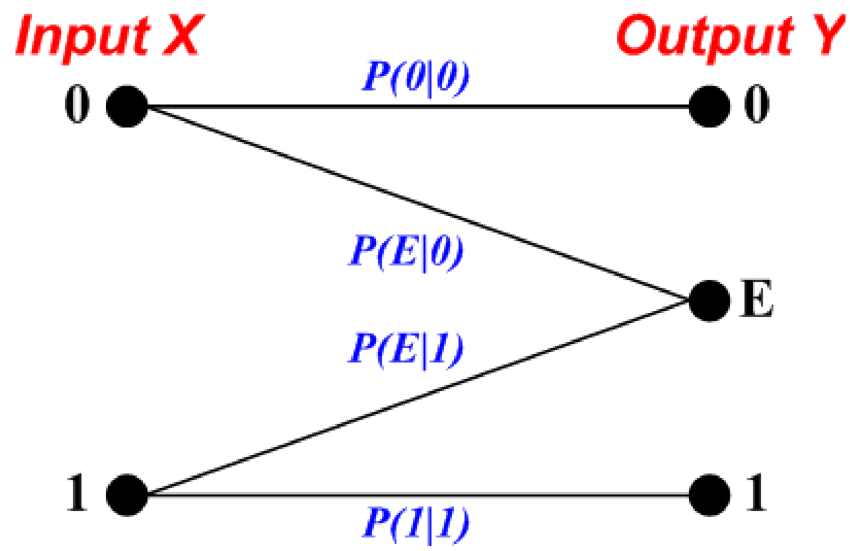
- ***When ε = 1/2, the input is independent with the output – <u>Completely-noisy-channel (CNC)</u> – cannot transmit information***
- ***When ε = 0, we have the <u>noiseless channel</u>***

Explanation in Communications

Example 2: Binary Erasure Channel (BEC)

The Binary Erasure Channel can transmit only one of two symbols (usually called 0 and 1).

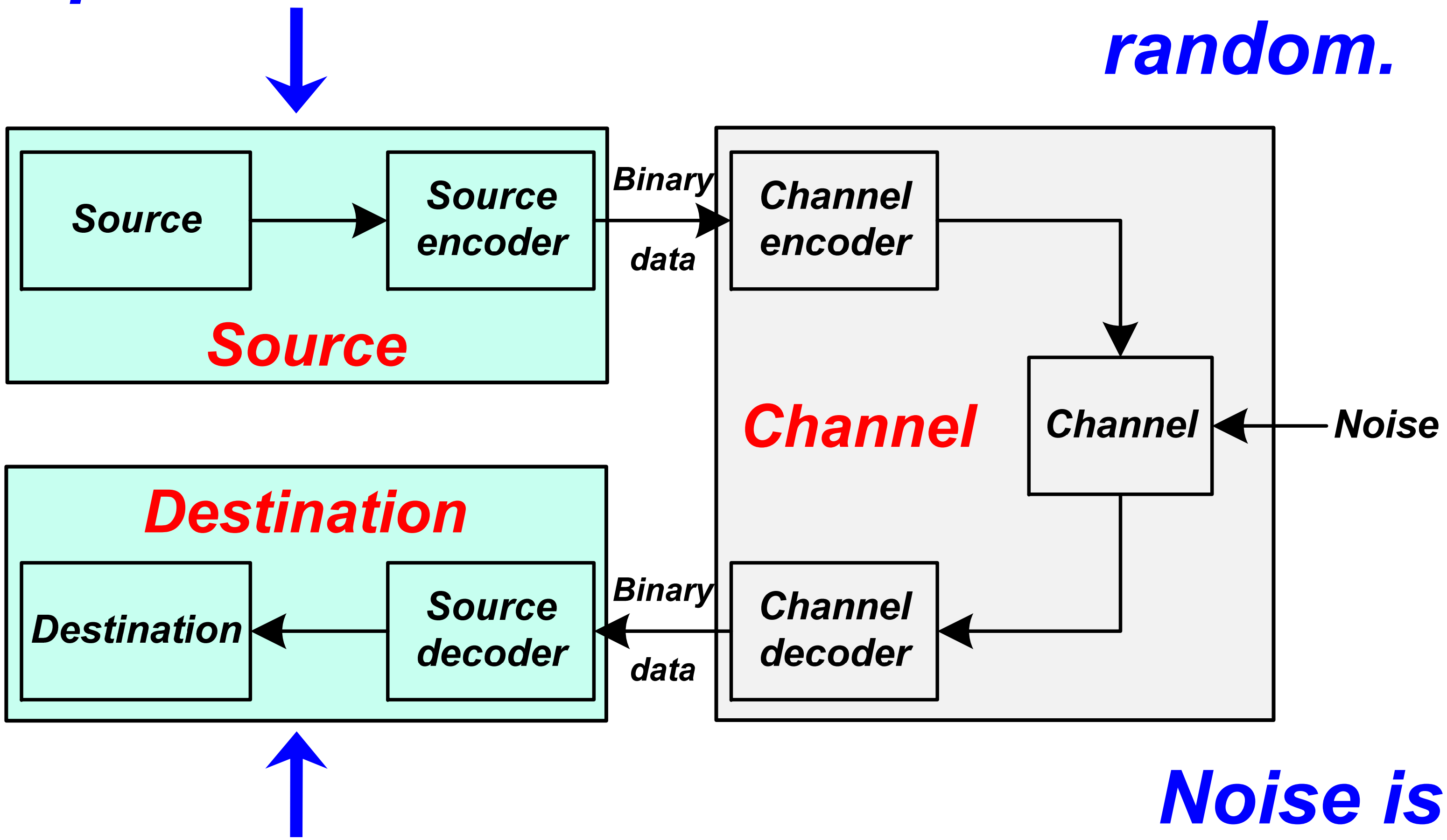
The channel is not perfect and sometimes the bit gets "erased" -- the receiver has no idea what the bit was.



Explanation in Communications

Input X -- Random Variable

Channel is also random.



Output Y -- Random Variable

Noise is also random.

Explanation in Communications

Entropy H(X)

The average uncertainty of the source X

Conditional Entropy H(X|Y)

The average remaining uncertainty of the source X after the observation of the output Y

Mutual Information I(X;Y)

The average amount of uncertainty in the source X resolved by the observation of the output Y.

Let's further discuss how to explain mutual information in communications systems

Explanation in Communications

- **Let the channel input (source) *X* is**

$$X \in \{a_1, a_2, \cdots, a_K\}$$

- **Let the channel output *Y* is**

$$Y \in \{b_1, b_2, \cdots, b_J\}$$

- **We denote the joint probability as $P(a_k, b_j)$, then we have the following results:**
  - **Input:** $P(a_k) = \sum_{j=1}^{J} P(a_k, b_j)$
  - **Output:** $P(b_j) = \sum_{k=1}^{K} P(a_k, b_j)$
  - **Forward transition:** $P(b_j|a_k) = P(a_k, b_j)/P(a_k)$
  - **Backward transition:** $P(a_k|b_j) = P(a_k, b_j)/P(b_j)$

Explanation in Communications

Recall the *self-information*, then we have

- **If the channel input is a_k, the information before the transmission is**

$$I(a_k) = \log \frac{1}{P(a_k)}$$

- **If the channel output is b_j, the information after the transmission about a_k is**

$$I(a_k|b_j) = \log \frac{1}{P(a_k|b_j)}$$

- **The transmission changes the probability of** $x = a_k$

$$P(a_k) \longrightarrow P(a_k|b_j)$$

Explanation in Communications

The information about the event $x = a_k$ provided by the occurrence of the event $y = b_j$ is

$$I(a_k;b_j) = I(a_k) - I(a_k|b_j) = \log\frac{P(a_k|b_j)}{P(a_k)}$$

The mutual information between events x=a_k and y=b_j

Questions:

1. ***The relationship between*** $I(a_k;b_j)$ ***and*** $I(b_j;a_k)$
2. ***The mutual information*** $I(a_k;b_j)$ ***is random or deterministic?***

Explanation in Communications

The mutual information between input X and output Y can be written as

$$I(X;Y) = \sum_{k=1}^{K} \sum_{j=1}^{J} P(a_k, b_j) \log \frac{P(a_k|b_j)}{P(a_k)}$$

In abbreviated notation, this is

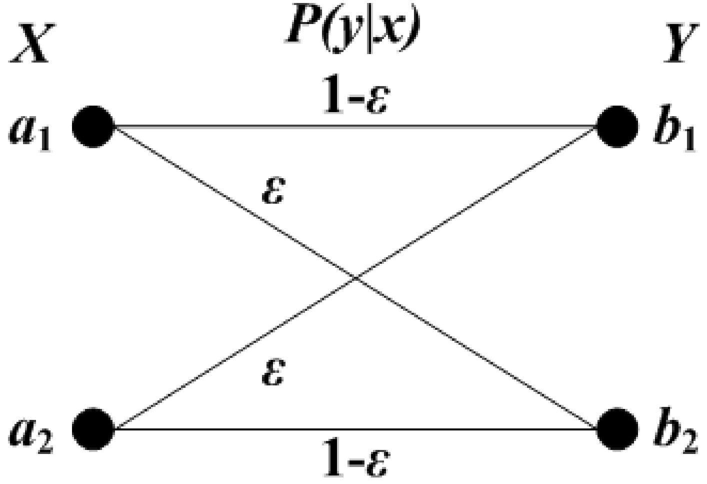
$$I(X;Y) = \sum_{x} \sum_{y} P(x, y) \log \frac{P(x|y)}{P(x)}$$

Similar approach can also be employed for analyzing the entropy, joint entropy, and conditional entropy

Explanation in Communications

Example

Consider a binary symmetric channel (BSC). Denote the probabilities of sending a_1 and a_2 as P and $1-P$, respectively.



(1) $H(X)$ and $H(X|Y)$

(2) $I(a_i;b_j)$ where $i = 1, 2$ and $j = 1, 2$

(3) $I(X;Y)$

Explanation in Communications

$$
\begin{aligned}
H(X) &= -P(a_1)\log P(a_1) - P(a_2)\log P(a_2) \\
&= -P\log P - (1-P)\log(1-P)
\end{aligned}
$$

If we denote $\Omega(z) = -z\log z - (1-z)\log(1-z)$ **, then** $H(X) = \Omega(P)$

$$H(X|Y) = \Omega(P) + \Omega(\varepsilon) - \Omega(P+\varepsilon-2P\varepsilon)$$

$$I(a_1;b_1) = \log\frac{1-\varepsilon}{P+\varepsilon-2P\varepsilon} \qquad I(a_2;b_2) = \log\frac{1-\varepsilon}{1-P-\varepsilon+2P\varepsilon}$$

$$I(a_1;b_2) = \log\frac{\varepsilon}{1-P-\varepsilon+2P\varepsilon} \qquad I(a_2;b_1) = \log\frac{\varepsilon}{P+\varepsilon-2P\varepsilon}$$

$$I(X;Y) = \Omega(P+\varepsilon-2P\varepsilon) - \Omega(\varepsilon)$$

What can we obtain from this example?

Explanation in Communications

Some Discussions

1. ***When does the source uncertainty achieves its maximum?***

2. ***How does the value of parameter ε impact the channel?***
   - ***When $\varepsilon = 0$, what can we obtain?***

     ------ Noiseless Channel

   - ***When $\varepsilon = ½$, what can we obtain?***

     ------ Completely Noisy Channel

3. ***The mutual information of two events can be negative, but the mutual information of two random variables cannot.***

Summary

- **Entropy**

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

- **Joint entropy**

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

- **Conditional entropy**

$$H(Y|X) = -\mathbb{E}\left\{\log\, p\left(Y|X\right)\right\} = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

- **Chain rule**

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

Summary

- **Mutual information**

$$I(X;Y)=\sum_x\sum_y P(x,y)\log\frac{P(x|y)}{P(x)}$$

- **Important inequalities and properties**
  - **Jensen's inequality**

$$\mathbb{E}\Big\{f(X)\Big\}\geq f\Big(\mathbb{E}\{X\}\Big)$$

  - **Uniform maximizes entropy**

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

  - **Nonnegativity of entropy and mutual information**
  - **Convexity of entropy and mutual information**