# Fundamentals of Information Theory

## Basic Concepts

**Yayu Gao**
School of Electronic Information and Communications
Huazhong University of Science and Technology
Email: yayugao@hust.edu.cn

# Outline

- Model of communication systems
- How to characterize the information source?
- How much information a message contains?
- What is entropy?
- Joint and conditional entropy
- Relative entropy and mutual information
- Entropies in communications
- Chain Rules
- Jensen’s Inequality and Log Sum Inequality
- Data processing inequality
- Entropy rate: from single-outcome to sequence-outcome
- What is a Markov source?
- Differential Entropy: from discrete to continuous

# 本节学习目标

1. 熟练掌握链式法则的运用
   - 写出熵的链式法则
   - 写出互信息的链式法则
   - 写出相对熵的链式法则
2. 能够写出以下的证明过程
   - Jensen’s inequality
   - Information inequality
   - Non-negativity of mutual information
   - Uniform PMF maximizes entropy
   - Conditioning reduces entropy
   - Independence bound on entropy
   - Log sum inequality
   - Data processing inequality
3. 记住entropy & mutual information的凹凸性

**重难点：**
- 链式法则的展开
- 三个不等式的写法及证明
- 三个不等式在信息论中的应用

# 09

# Chain Rules

# Chain rule: Motivation

How to compute the entropies of the composition of two or more random variables?

- In calculus, the chain rule is a formula for computing the derivative of the composition of two or more functions.
  - Let $y = f(u)$ and $u = g(x)$.

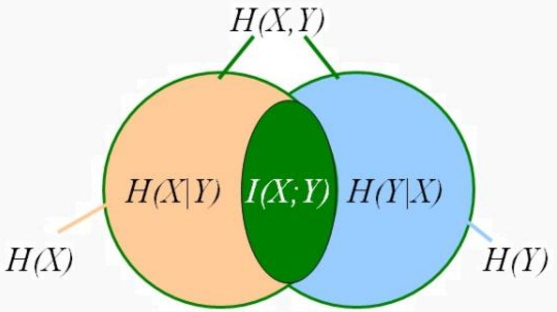
$$[f(g(x)]' = f'(g(x))g'(x)$$

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

In the information theory, the chain rule is a formula for computing the entropies of the composition of two or more random variables.

# Chain rule

$$\boxed{H(X,Y) = H(X) + H(Y|X)}$$




- Proof:

$$\begin{aligned}
H(X,Y) &= -\sum_x \sum_y p(x,y)\log[p(x,y)] \\
&= -\sum_x \sum_y p(x,y)\log[p(x)p(y|x)] \\
&= -\sum_x \sum_y p(x,y)\log[p(x)] - \sum_x \sum_y p(x,y)\log[p(y|x)] \\
&= -\sum_x p(x)\log[p(x)] - \sum_x \sum_y p(x,y)\log[p(y|x)] \\
&= H(X) + H(Y|X)
\end{aligned}$$

- Corollary $\boxed{H(X,Y|Z) = H(X|Z) + H(Y|X,Z)}$

# Chain rule: Entropy

- Chain rules can be derived by repeated applications of two-variable expansion rules.

$$\boxed{H(X,Y) = H(X) + H(Y|X)}$$

- **Entropy**

$$\boxed{H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, X_{i-2}, \ldots, X_1)}$$

- Example

$$\begin{aligned} H(X_1, X_2, X_3) &= \sum_{i=1}^{3} H(X_i | X_{i-1}, X_{i-2}, \ldots, X_1) \\ &= H(X_1) + H(X_2 | X_1) + H(X_3 | X_2, X_1) \end{aligned}$$

# Revisiting Example #1

- Joint $p.m.f.$ is:

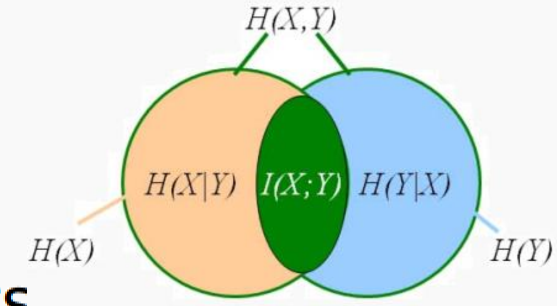
| $Y \backslash X$ | 1 | 2 | 3 | 4 | $p(y)$ |
|---|---|---|---|---|---|
| 1 | $1/8$ | $1/16$ | $1/32$ | $1/32$ | $1/4$ |
| 2 | $1/16$ | $1/8$ | $1/32$ | $1/32$ | $1/4$ |
| 3 | $1/16$ | $1/16$ | $1/16$ | $1/16$ | $1/4$ |
| 4 | $1/4$ | $0$ | $0$ | $0$ | $1/4$ |
| $p(x)$ | $1/2$ | $1/4$ | $1/8$ | $1/8$ | |

- What is $H(X)$, $H(Y)$, $H(X|Y)$, $H(Y|X)$, $H(X,Y)$, $I(X;Y)$?

- **How many at least to obtain all of them?**

# Solution of example #1

$$H(X) = H(1/2, 1/4, 1/8, 1/8) = 1.75 \text{ bits}$$
$$H(Y) = H(1/4, 1/4, 1/4, 1/4) = 2 \text{ bits}$$
$$H(X|Y) = \sum_i Pr(Y = i) H(X|Y = i) = 1.375 \text{ bits}$$




$$H(X, Y) = H(Y) + H(X|Y) = 2 + 1.375 = 3.375 \text{ bits } (\text{chain rule})$$
$$H(Y|X) = H(X, Y) - H(X) = 3.375 - 1.75 = 1.625 \text{ bits } (\text{chain rule})$$

$$H(X) - H(X|Y) = 1.75 - 1.375 = 0.375 \text{ bits}$$
$$H(Y) - H(Y|X) = 2 - 1.625 = 0.375 \text{ bits}$$

$$\boxed{I(X; Y) = H(X) - H(X|Y)}$$
$$\boxed{I(X; Y) = H(Y) - H(Y|X)}$$

# Chain rule: Mutual information

- **Mutual information**

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

- Example

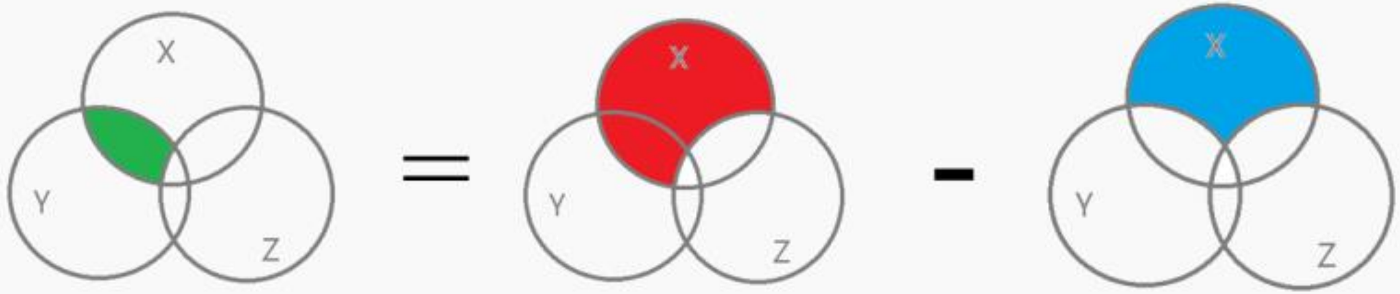
$$I(X_1, X_2, X_3; Y) = \sum_{i=1}^{3} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1)$$
$$= I(X_1; Y) + I(X_2; Y | X_1) + I(X_3; Y | X_2, X_1)$$

- What is $I(X_2; Y | X_1)$ ?
  - Conditional mutual information

# Conditional mutual information

- The conditional mutual information of random variables X and Y given Z is defined by

$$
\begin{aligned}
I(X;Y|Z) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y,z)\log\left[\frac{p(x,y|z)}{p(x|z)p(y|z)}\right] \\
&= E_{p(x,y,z)}\left\{\log\left[\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right]\right\} \\
&= {\color{red} H(X|Z) - H(X|Y,Z)}
\end{aligned}
$$




- **Can you prove it?**

# Chain rule: Relative entropy

- Relative entropy between two joint distributions can be expanded as the sum of a relative entropy and a conditional relative entropy.

$$\boxed{D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))}$$

- Conditional relative entropy

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

- **Can you prove it?**

# 10

# Jensen's Inequality
# and
# Log Sum Inequality

# Motivation

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

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

max H($X$)

Concavity of Entropy

Source coding theorem

Jensen's Inequality

Log Sum Inequality

Properties of entropies


Convexity of Mutual Information

Channel coding theorem

If $f$ is a convex function, then $E[f(X)] \geq f(E[X])$.

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

$$C = \max_{p(x)}\{I(X;Y)\}$$

Johan Jensen (1859-1925)
Danish mathematician

# Let's begin with **convexity**

- What is **convex set**?

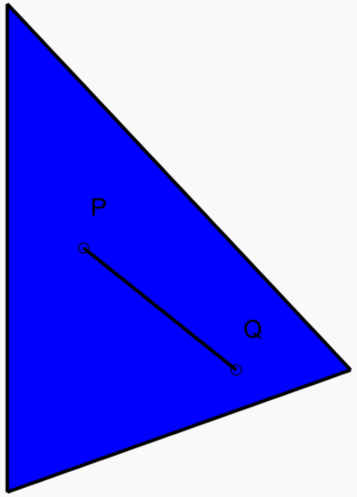



Figure 1: A Convex Set

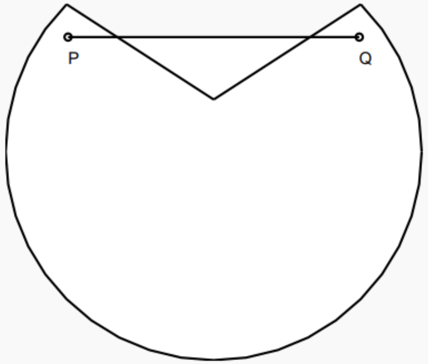
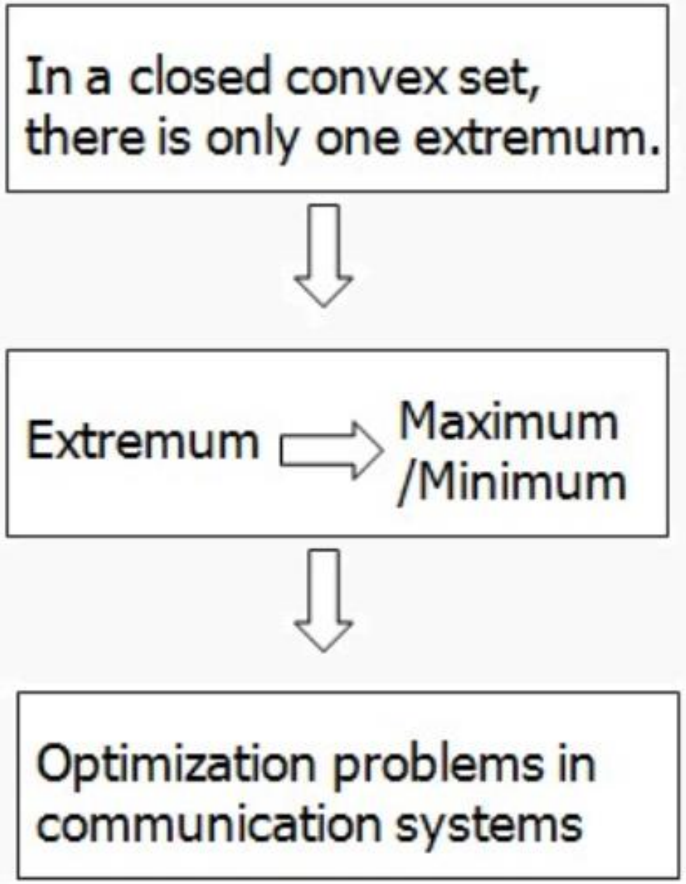
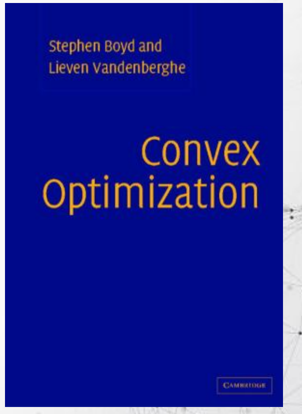


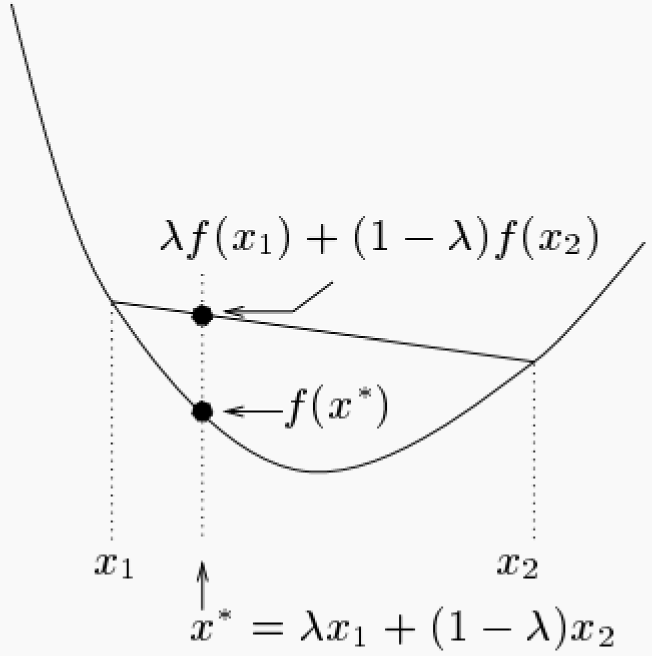

Figure 2: A Non-convex Set







# What is convex function?




- Convex functions lie below any chord.
- Notation
  - Convex
  - Concave upwards
  - Concave upwards
  - Concave up
  - Convex cup

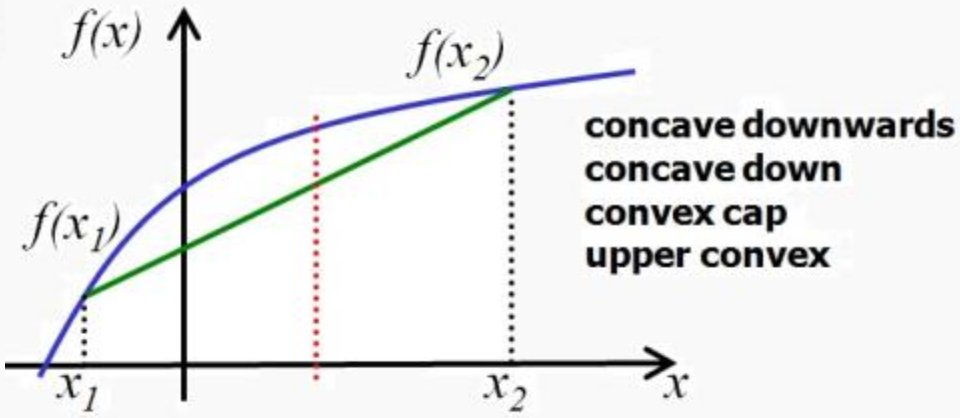
- Function $f(x)$ is convex over $(a, b)$ if

$$\forall x_1, x_2 \in (a, b), 0 \leq \lambda \leq 1 \ f(\lambda \cdot x_1 + (1-\lambda) \cdot x_2) \leq \lambda \cdot f(x_1) + (1-\lambda) \cdot f(x_2)$$

- Function $f(x)$ is strictly convex over $(a, b)$ if it is convex and

$$\forall x_1, x_2 \in (a, b), \ 0 \leq \lambda \leq 1 \ f(\lambda \cdot x_1 + (1-\lambda) \cdot x_2) = \lambda \cdot f(x_1) + (1-\lambda) \cdot f(x_2) \Leftrightarrow \lambda = 0 \text{ or } \lambda = 1$$

# What is concave function?




- Convex functions lie above any chord.

- Function $f(x)$ is concave over $(a, b)$ if

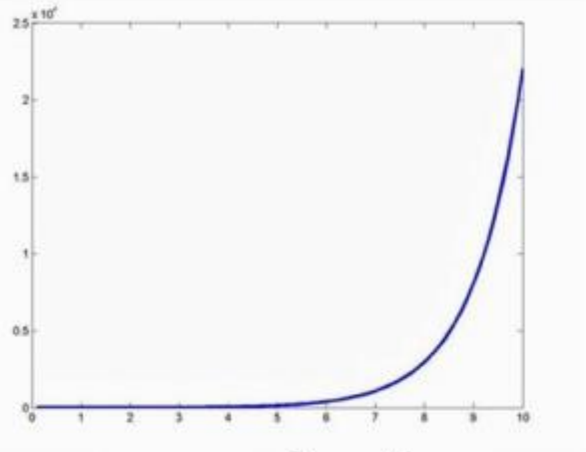
$$\forall x_1, x_2 \in (a, b), 0 \le \lambda \le 1$$
$$f(\lambda \cdot x_1 + (1-\lambda) \cdot x_2)) \ge \lambda \cdot f(x_1) + (1-\lambda) \cdot f(x_2)$$

- Function $f(x)$ is strictly concave over $(a, b)$ if it is concave and

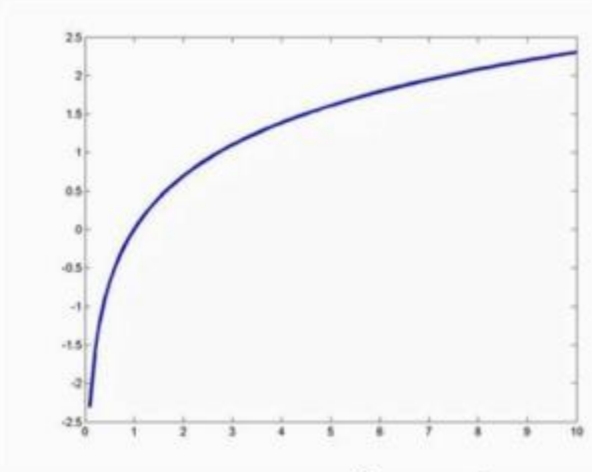
$$\forall x_1, x_2 \in (a, b), 0 \le \lambda \le 1$$
$$f(\lambda \cdot x_1 + (1-\lambda) \cdot x_2) = \lambda \cdot f(x_1) + (1-\lambda) \cdot f(x_2) \Leftrightarrow \lambda = 0 \text{ or } \lambda = 1$$

# How to know a function is convex or concave?

- Examples of convex and concave functions



convex function
$f(x)=e^x$



concave function
$f(x)=ln(x)$

Test of convexity and concavity If function $f(x)$ has a second derivative $f''(x)$, which is non-negative (positive) everywhere, then $f(x)$ is convex (strictly convex).

# Jensen's inequality: Preview

If $f$ is convex, then for $r.v. X$, $E[f(X)] \geq f(E[X])$.

If $f$ is strictly convex, "=" holds when $X = E[X]$ with probability 1.

- It is used very widely in information theory.

- To **prove some of the properties of entropy and relative entropy**.

- Most basic theorems are proved based on Jensen's inequality.

# Jensen's inequality: Proof

If $f$ is convex, then for $r.v. X$, $E[f(X)] \geq f(E[X])$.

If $f$ is strictly convex, "=" holds when $X = E[X]$ with probability 1.

- Sketch of the proof: We prove this for discrete distributions by the mathematical induction on the number of the mass points.
- $n = 2$, the inequality becomes $p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$. It holds by convexity.
- Suppose the theorem is true for distributions with $n - 1$ mass points.

$$\sum_{i=1}^{n-1} q_i f(x_i) \geq f\left(\sum_{i=1}^{n-1} q_i x_i\right), \sum_{i=1}^{n-1} q_i = 1.$$

- Then, prove the inequality holds for $n$.

# Jensen's inequality: Proof

If $f$ is convex, then for $r.v.X$, $E[f(X)] \geq f(E[X])$.

If $f$ is strictly convex, "$=$" holds when $X = E[X]$ with probability 1.

$$
\begin{aligned}
E[f(X)] &= \sum_{i=1}^{n} p_i f(x_i) = p_n f(x_n) + \sum_{i=1}^{n-1} p_i f(x_i) \\
&= p_n f(x_n) + (1 - p_n) \sum_{i=1}^{n-1} \frac{p_i}{1 - p_n} f(x_i) \\
&\geq p_n f(x_n) + (1 - p_n) f\left( \sum_{i=1}^{n-1} \frac{p_i}{1 - p_n} x_i \right) \\
&\geq f\left( p_n x_n + (1 - p_n) \sum_{i=1}^{n-1} \frac{p_i}{1 - p_n} x_i \right) = f\left( \sum_{i=1}^{n} p_i x_i \right) = f(E[X])
\end{aligned}
$$

# Relative-entropy properties proved by Jensen's inequality

- Theorem: Information inequality

Let $p(x)$, $q(x)$, $x \in \mathcal{X}$, be two $p.m.f.$'s. Then,

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

$$D(p(x)||q(x)) = 0 \Leftrightarrow p(x) = q(x).$$

Proof:

Let $\mathcal{A} = \{x : p(x) > 0\}$ be the support set of $p(x)$, then

If $f$ is a convex function, then $E[f(X)] \geq f(E[X])$.

$$\begin{aligned}
-D(p(x)||q(x)) &= -\sum_{x\in\mathcal{A}} p(x) \log \left[\frac{p(x)}{q(x)}\right] \\
&= \sum_{x\in\mathcal{A}} p(x) \log \left[\frac{q(x)}{p(x)}\right] \leq \log \left[\sum_{x\in\mathcal{A}} p(x) \frac{q(x)}{p(x)}\right] \text{ (by Jensen's inequality)} \\
&= \log \left[\sum_{x\in\mathcal{A}} q(x)\right] \leq \log \left[\sum_{x\in\mathcal{X}} q(x)\right] = \log 1 = 0
\end{aligned}$$

# Relative-entropy properties proved by Jensen's inequality

- Corollary: Non-negativity of mutual information

$$I(X;Y) \geq 0.$$
$$I(X;Y) = 0 \Leftrightarrow X \text{ and } Y \text{ are independent.}$$

Proof:

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$$

With equality if and only if $p(x,y) = p(x)p(y)$, i.e., $X$ and $Y$ are independent.

# Entropy properties proved by Jensen's inequality

- Theorem: Uniform PMF maximizes the entropy

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

$$H(X) = \log(|\mathcal{X}|) \Leftrightarrow p(x) = q(x) = 1/|\mathcal{X}|$$

- Theorem: Conditioning reduces entropy

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

- Theorem: Independence bound on entropy

$$H(X_1, X_2, \ldots, X_n) \leq \sum_i H(X_i)$$

$$H(X_1, X_2, \ldots, X_n) = \sum H(X_i) \Leftrightarrow X_i \text{ are independent with each other.}$$

# Theorem: uniform PMF maximizes the entropy

$$\boxed{\begin{aligned}&H(X) \le \log|\mathcal{X}|;\\&H(X) = \log|\mathcal{X}| \iff p(x) = q(x) = \tfrac{1}{|\mathcal{X}|}.\end{aligned}}$$

Proof: Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform $p.m.f.$ over $\mathcal{X}$. Let $p(x)$ be the $p.m.f.$ for $r.v.X$. Then,

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

# Theorem: conditioning reduces entropy

$$\boxed{H(X|Y) \le H(X)}$$

Proof:

$$0 \le I(X;Y) = H(X) - H(X|Y).$$

- Comments:
  - Knowing another *r.v.* $Y$ can only reduce the uncertainty in $X$.
  - This is true only on the average.

# Theorem: independence bound on entropy

$$H(X_1, X_2, \ldots, X_n) \le \sum_i H(X_i).$$

$$H(X_1, X_2, \ldots, X_n) = \sum_i H(X_i) \iff X_i \text{ are independent with each other.}$$

Proof: By the chain rule for entropy, we apply the theorem of conditioning reduces entropy.

$$H(X_i | X_{i-1}, X_{i-2}, \ldots, X_1) \le H(X_i)$$

$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, X_{i-2}, \ldots, X_1)$$

$$\le \sum_{i=n}^{n} H(X_i)$$

# Revisiting Motivation

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

en's Inequality

Log Sum Inequality

roperties of entropies

Concavity of Entropy

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

Source coding theorem

max H($X$)


Convexity of Mutual Information

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

Channel coding theorem

$$C = \max_{p(x)}\{I(X;Y)\}$$

If $f$ is a convex function, then $E[f(X)] \geq f(E[X])$.

Johan Jensen (1859-1925)
Danish mathematician

# Log sum inequality: Preview

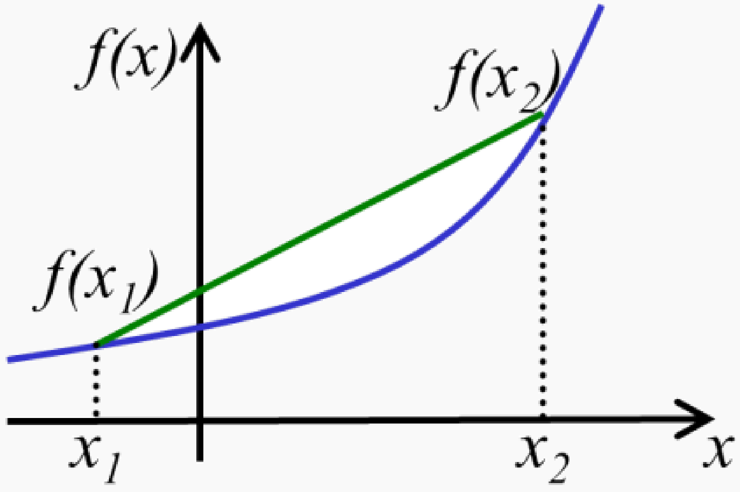
For non-negative numbers, $a_i$ and $b_i$, $(i = 1, 2, \ldots, n)$,

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

With equality, if and only if $\frac{a_i}{b_i} = \text{constant}$.

- Log sum inequality is proved based on Jensen's inequality.
- Beauty of Math: a "smart" selection of function.
- It is used to prove several theorems in information theory.

# Log sum inequality: Proof




Proof: (a brief sketch)

- Assume $a_i$ and $b_i$ are positive.
- Construct $f(t) = t \log t$.
- The function $f(t) = t \log t$ is strictly convex for all positive $t$.
- Construct $\alpha_i = \frac{b_i}{\sum_j b_j}$ and $t_i = \frac{a_i}{b_i}$.
- By Jensen's inequality, $\sum \alpha_i f(t_i) \geq f(\sum \alpha_i t_i)$.

# Log sum inequality: Applications

Extremum ⟹ Maximum /Minimum

- Theorem: convexity of relative entropy

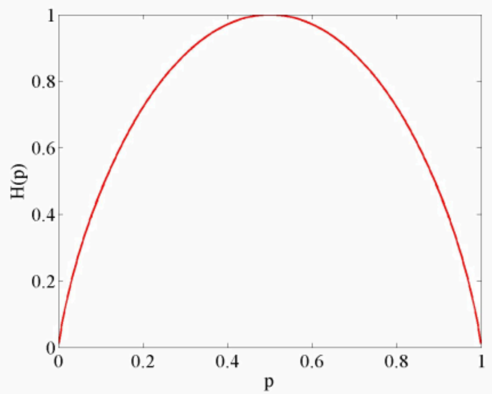
$D(p||q)$ is convex in the pair $(p, q)$;
$D[\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2] \le \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$.

- Theorem: convexity of mutual information

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

  - $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.

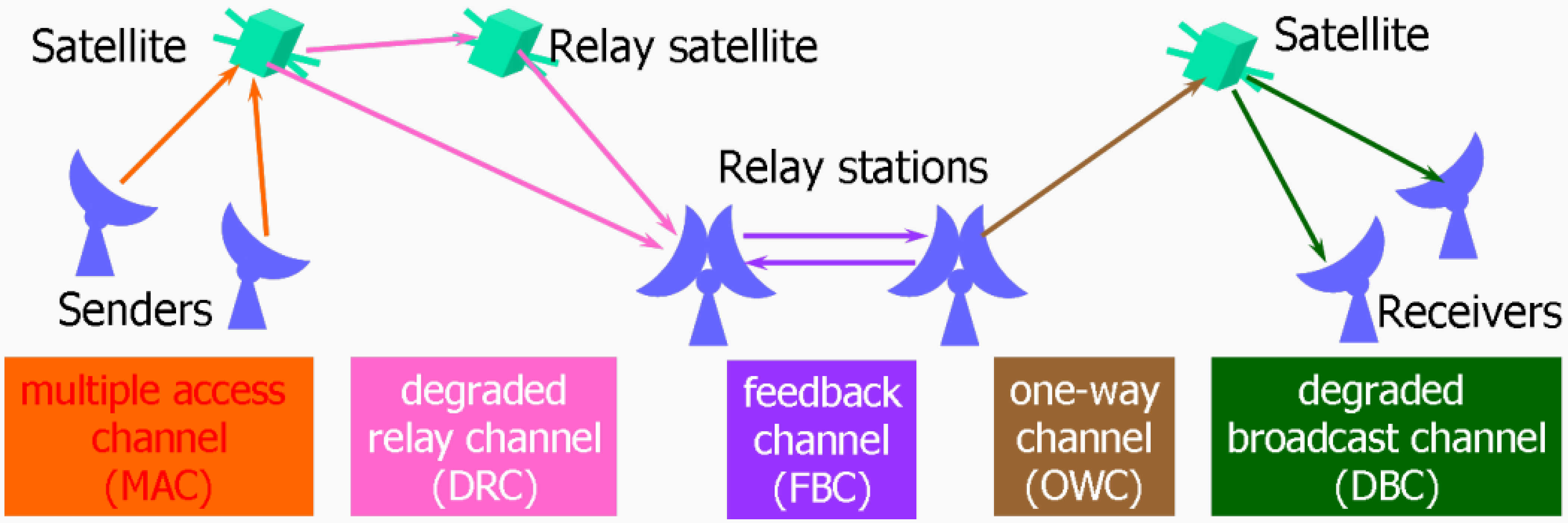
- Theorem: concavity of entropy
  - $H(p)$ is a concave function of $p$.



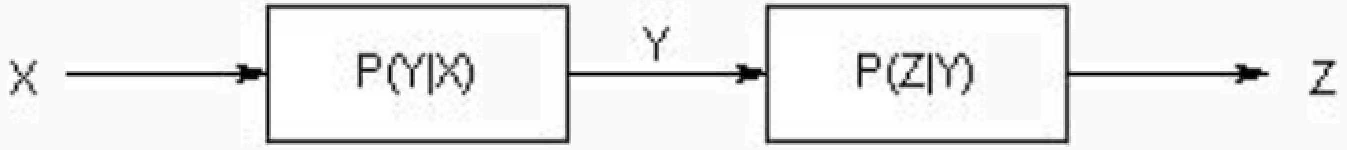
# 11

# Data Processing Inequality

# End-to-End Communication Systems



# Data processing inequality




- Note that by the chain rule,

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

- Markov Chain: Random variables $X$, $Y$, $Z$ form a Markov chain $(X \to Y \to Z)$, if

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

- Consequence: Markovity implies conditional independence because

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y).$$

# Data processing inequality: theorem

If $X \to Y \to Z$, then

$$\boxed{I(X;Y) \geq I(X;Z).}$$

Proof: applying the chain rule,

- $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$,
- $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$,
- $I(X;Z|Y) = 0$ and $I(X;Y|Z) \geq 0$.

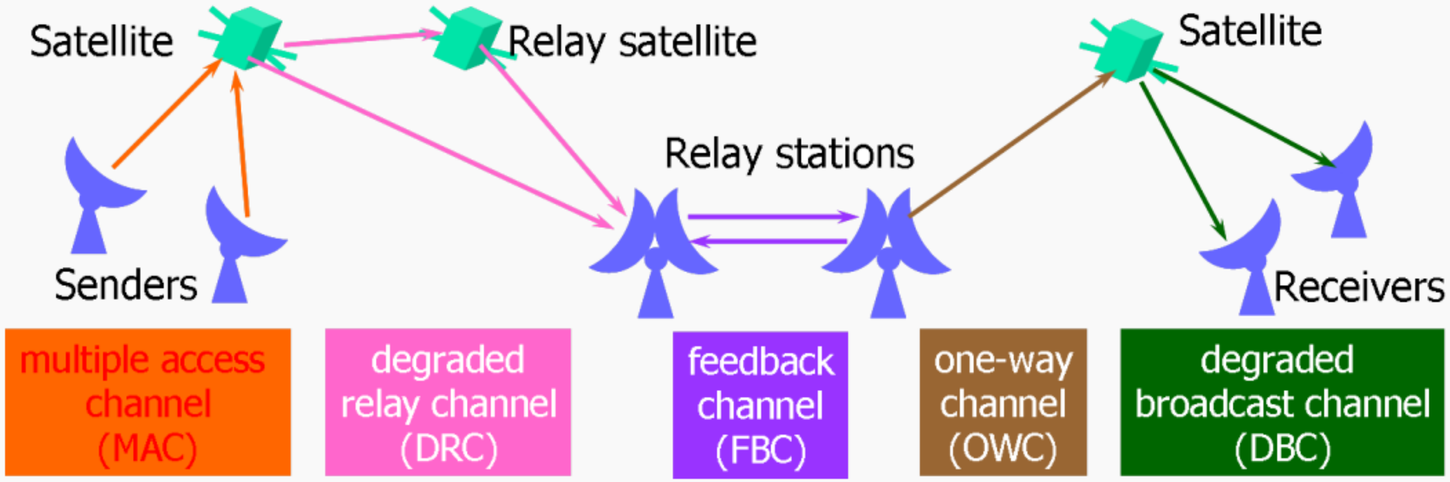
Thus, we have $I(X;Y) \geq I(X;Z)$.

# Data processing inequality: comments

If $X \to Y \to Z$, then

$$\boxed{I(X;Y) \geq I(X;Z).}$$

- Manipulation of data cannot increase its information.



# Summary

- Chain Rules

$$\boxed{H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, X_{i-2}, \ldots, X_1)}$$

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

- Jensen’s Inequality and Log Sum Inequality
  - Information inequality
  - Non-negativity of mutual information
  - Uniform PMF maximizes entropy
  - Conditioning reduces entropy
  - Independence bound on entropy
  - Concavity of entropy
  - Convexity of  mutual information
- Data processing inequality
  - Manipulation of data cannot increase its information.

# 本节学习目标

1. **熟练掌握链式法则的运用**
   - 写出熵的链式法则
   - 写出互信息的链式法则
   - 写出相对熵的链式法则
2. **能够写出以下的证明过程**
   - Jensen’s inequality
   - Information inequality
   - Non-negativity of mutual information
   - Uniform PMF maximizes entropy
   - Conditioning reduces entropy
   - Independence bound on entropy
   - Log sum inequality
   - Data processing inequality
3. **记住entropy & mutual information的凹凸性**

**重难点：**

- 链式法则的展开
- 三个不等式的写法及证明
- 三个不等式在信息论中的应用

# Thank you!

My Homepage

**Yayu Gao**
**School of Electronic Information and Communications**
**Huazhong University of Science and Technology**
**Email: yayugao@hust.edu.cn**