Convexity

Definition (Convexity)

A function $f(x)$ is said to be *convex* over an interval (a, b) if $\forall x_1, x_2 \in (a, b)$ and $0 \le \lambda \le 1$,

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

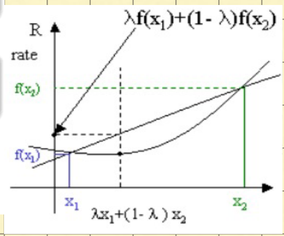
A function f is called *strictly convex* if equality holds only if $\lambda = 0$ or $\lambda = 1$.

仅$\lambda=0$或1时等号成立，则f为严格凸函数。
（即只有两端是等的）

Definition (Concavity)

A function f is *concave* if $-f$ is convex.

A function is convex if it always lies below any chord. A function is concave if it always lies above any chord.



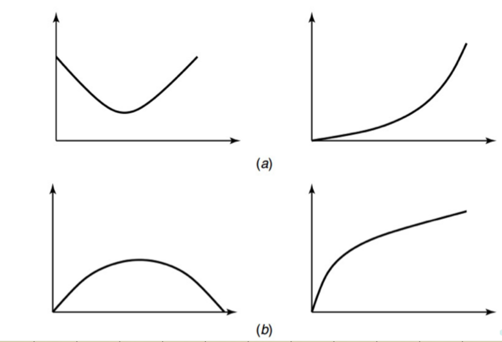
Example

Example

$$f(x) = x^2, \quad |x|, \quad e^x, \quad x\log x \quad (x > 0)$$

$$g(x) = \log x, \quad \sqrt{x}, \quad (x \ge 0)$$

$f(x)$是凸函数
$g(x)$是凹函数
二阶导≥ 0即为凸函数



Jensen's Inequality

Theorem 2.6.2 (Jensen's Inequality)

If f is a convex function and X is a random variable,

$$E[f(X)] \ge f(E[X]).$$

Moreover, if f is strictly convex, $E[f(X)] = f(E[X])$ implies that $X = E[X]$ with probability 1 (i.e., X is a constant).

Proof.

By mathematical induction.

- $k = 2$: $p(x_1)f(x_1) + p(x_2)f(x_2) \ge f(p(x_1)x_1 + p(x_2)x_2)$.
- Hypothesis: $\sum_{i=1}^{k-1} p(x_i)f(x_i) \ge f\left(\sum_{i=1}^{k-1} p(x_i)x_i\right)$.
- Induction: $\sum_{i=1}^{k} p(x_i)f(x_i)$.

X: $x_1, \dots, x_n$
$p(x_1), \dots, p(x_n)$

$E[f(X)] = \sum_{i=1}^{n} p(x_i) f(x_i)$

$f[E(X)] = f(\sum_{i=1}^{n} p(x_i) x_i)$

若f为凸，则$E[f(X)] \ge f(E(X))$

若f为严格凸且$E[f(x)] = f[E(x)]$，则$X = E[X]$
（X为常数）

①当$|X| = 2$时，$p(x_1)f(x_1) + p(x_2)f(x_2) \ge f(p(x_1)x_1 + p(x_2)x_2)$
由于f为凸函数，显然成立。

②假设当$|X| = k-1$时，$\sum_{i=1}^{k-1} p(x_i)f(x_i) \ge f(\sum_{i=1}^{k-1} p(x_i)x_i)$成立。
则当$|X| = k$时，

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

令$a = \sum_{i=1}^{k-1} p(x_i)$，则$= a\sum_{i=1}^{k-1} \frac{p(x_i)}{a} f(x_i) + (1-a)f(x_k)$

$$\ge a f(\sum_{i=1}^{k-1} \frac{p(x_i)}{a} x_i) + (1-a)f(x_k)$$

由于f为凸函数，则$\ge f(a \cdot \sum_{i=1}^{k-1} \frac{p(x_i)}{a} x_i + (1-a)f(x_k))$

$$= f(\sum_{i=1}^{k} p(x_i) x_i)$$

综上，Jensen's Inequality 成立。

Information Inequality

Theorem 2.6.3 (*Information Inequality*)

Let $p(x)$, $q(x)$, $x \in X$, be two probability mass functions. Then

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

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

Proof.

Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned} -D(p\|q) &= -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} \\ &= \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \qquad E_p(\log\frac{q(x)}{p(x)}) \\ &\leq \log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} \quad \text{(Jensen's Inequality)} \quad \leq \log E_p(\frac{q(x)}{p(x)}) \\ &= \log\sum_{x\in A} q(x) \\ &\leq \log\sum_{x\in\mathcal{X}} q(x) = 0 \end{aligned}$$

log是concave的

Corollaries

Corollary (*Nonnegativity of mutual information*)

For any two random variables, X, Y,

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

with equality iff X and Y are independent.

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

Corollary

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

with equality iff $p(y|x) = q(y|x)$ for all y and x such that $p(x) > 0$.

在支撑集上是同分布的

Corollary

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

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

The maximum entropy distribution

Theorem 2.6.4

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

均匀分布的熵最大，为 $\log|\mathcal{X}|$

Proof.

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

$$0 \leq D(p\|u) = \sum p(x)\log\frac{p(x)}{u(x)} = \log|\mathcal{X}| - H(X).$$

$$\begin{aligned} D(p\|u) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} \\ &= \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{u(x)} + \sum_{x\in\mathcal{X}} p(x)\log p(x) \qquad u(x) = \frac{1}{|\mathcal{X}|} \\ &= \log|\mathcal{X}|\sum_{x\in\mathcal{X}} p(x) - H(X) \\ &= \log|\mathcal{X}| - H(X) \geq 0 \quad \Rightarrow H(X) \leq \log|\mathcal{X}| \end{aligned}$$

对离散随机变量，均匀分布随机性最大

对连续随机变量，正态分布随机性最大

Conditioning reduces entropy

Theorem 2.6.5 (*Conditioning reduces entropy*)

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

with equality iff X and Y are independent.

$I(X;Y) = H(X) - H(X|Y) \ge 0 \Rightarrow H(X) \ge H(X|Y)$

Theorem 2.6.6 (*Independence bound on entropy*)

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$, then

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

with equality iff the X_i's are independent.

Data-processing inequality 信息传递

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 join probability mass function can be written as

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

- $X \to Y \to Z \Rightarrow p(x, z|y) = p(x|y)p(z|y)$
- $X \to Y \to Z \Rightarrow Z \to Y \to X$
- If $Z = f(Y)$, then $X \to Y \to Z$.

在给定Y的情况下，X与Z无关：$p(z|x,y) = p(z|y)$

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

给定系统当前状态Y，系统未来状态Z与过去状态X没有关系.

proof① $p(x,y,z) = p(x)p(y,z|x)$

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

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

proof②. $p(x,y,z) = p(z)p(x,y|z)$

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

由于 $p(x|y,z) = p(x|y)$

$\Leftrightarrow \dfrac{p(x,y,z)}{p(y,z)} = \dfrac{p(x,y)}{p(y)}$

$\Leftrightarrow \dfrac{p(x,y,z)}{p(x,y)} = \dfrac{p(y,z)}{p(y)}$

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

因为 $X \to Y \to Z$，故上式成立，所以 $p(x,y,z) = p(z)p(y|z)p(x|y)$

即 $Z \to Y \to X$.

$X_1, X_2, \ldots, X_n$ 形成马尔可夫链，当

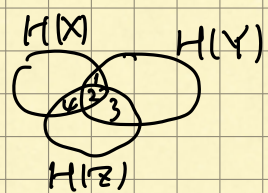
$p(x_i | x_1, \ldots x_{i-1}, x_{i+1}, \ldots, x_n) = p(x_i | x_{i-1}) \quad \forall i$

$p(x_1, \ldots x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2) \cdots p(x_n|x_{n-1})$

Theorem 2.8.1 (*Data-processing inequality*)

If $X \to Y \to Z$, then $I(X;Y) \ge I(X;Z)$.

一般情况下，



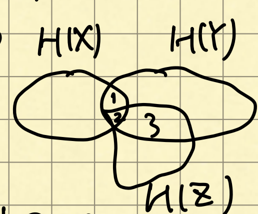
Proof.

By the chain rule, we expand $I(X; Y, Z)$ in two ways:

$$I(X; Y, Z) = I(X; Z) + \underbrace{I(X; Y|Z)}_{\ge 0}$$
$$= I(X; Y) + \underbrace{I(X; Z|Y)}_{=0}$$

Since $X \to Y \to Z$, we have $I(X; Z|Y) = 0$. Since $I(X; Y|Z) \ge 0$, we have $I(X; Y) \ge I(X; Z)$. □

但若 $X \to Y \to Z$，则为



不存在区域4，同时 ①≤①+②.

若 $X_1, X_2, \ldots, X_n$ 形成马尔可夫链，则

$I(X_1; X_i) \ge I(X_1; X_j), \ i \le j$

即传得越远，与源头的互信息越少.

Corollaries

Corollary

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

Corollary

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

Feno's inequality

发射，接收，猜测：$X \to Y \to \hat{X} = f(Y)$

Problem 2.5 (*Zero conditional entropy*)

Show that if $H(X|Y) = 0$, then X is a function of Y, i.e., for all y with $p(y) > 0$, there is only one possible value of x with $p(x,y) > 0$.

对任意的y，只有唯一一个x能与这个y同时出现.

Proof.

Assume that there exists an y, say y_0 and two different values of x, say x_1 and x_2 such that $p(y_0, x_1) > 0$ and $p(y_0, x_2) > 0$. Then $p(y_0) \geq p(y_0, x_1) + p(y_0, x_2) > 0$, and $p(x_1|y_0)$ and $p(x_2|y_0)$ are not equal to 0 or 1. Thus,

$$
\begin{aligned}
H(X|Y) &= -\sum_y p(y) \sum_x p(x|y) \log p(x|y) = \sum_y \sum_x p(x,y) \log \frac{1}{p(x|y)} \\
&\geq p(y_0)\left(-p(x_1|y_0)\log p(x_1|y_0) - p(x_2|y_0)\log p(x_2|y_0)\right) \\
&> 0
\end{aligned}
$$

since $-t\log t \geq 0$ for $0 \leq t \leq 1$, and is strictly positive for $t \neq 0, 1$, which is a contradiction to $H(X|Y) = 0$. □

- The conditional entropy of a random variable X given another random variable Y is zero ($H(X|Y) = 0$) iff X is a function of Y. Hence we can estimate X from Y with zero probability of error iff $H(X|Y) = 0$.
- We can estimate X with a low probability of error P_e only if the conditional entropy $H(X|Y)$ is small. *Fano's inequality* quantifies this idea.

当$H(X|Y)=0$时，可以从Y中估计X.

当$H(X|Y)$较小时，可以以错误率P_e从Y中估计X.

Why do we need to related P_e to entropy $H(X|Y)$? When we have a communication system, we send X, but receive a corrupted version Y. We want to infer X from Y. Our estimate is $\hat{X}$ and we will make a mistake as

$$P_e = \Pr[\hat{X} \neq X]$$

Markov chain $X \to Y \to \hat{X}$.

Problem

A random variable Y is related to another random variable X with a distribution $p(x)$. From Y, we calculate a function $g(Y) = \hat{X}$, where $\hat{X}$ is an estimate of X and takes on values in $\hat{\mathcal{X}}$. We observe that $X \to Y \to \hat{X}$ forms a Markov chain. How to bound the estimate error probability $P_e = \Pr[\hat{X} \neq X]$?

Theorem 2.11.1

For Markov chain $X \to Y \to \hat{X}$ *, with* $P_e = \Pr\{X \neq \hat{X}\}$*, we have*

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

This inequality can be weakened to

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

or

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

Remark: $\hat{X}$ can be treated as an estimation of X based on Y.

Proof.

Define an error random variable as

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

Using the chain rule for entropies to expand $H(E, X|\hat{X})$ in two different ways, we have

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} = \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log(|X|-1)}.$$

Since conditioning reduces entropy, $H(E|\hat{X}) \leq H(E) = H(P_e)$. Since E is a function of X and $\hat{X}$, the conditional entropy $H(E|X, \hat{X})$ is equal to 0. We now look at $H(X|E, \hat{X})$. By the equation $H(X|Y) = \sum_y p(y)H(X|Y = y)$, we have

$$H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{\Pr[\hat{X} = \hat{x}, E = 0] H(X|\hat{X} = \hat{x}, E = 0) + \Pr[\hat{X} = \hat{x}, E = 1] H(X|\hat{X} = \hat{x}, E = 1)\}.$$

Proof.

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} = \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log(|X|-1)}.$$

$$H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{\Pr[\hat{X} = \hat{x}, E = 0] H(X|\hat{X} = \hat{x}, E = 0) + \Pr[\hat{X} = \hat{x}, E = 1] H(X|\hat{X} = \hat{x}, E = 1)\}.$$

By definition of E, X is conditionally deterministic given $\hat{X} = \hat{x}$ and $E = 0$, then $H(X|\hat{X} = \hat{x}; E = 0) = 0$. If $\hat{X} = \hat{x}$ and $E = 1$, then X must take a value in the set $\{x \in \mathcal{X} : x \neq \hat{x}\}$ which contains $|\mathcal{X}| - 1$ elements. Then $H(X|\hat{X} = \hat{x}, E = 1) \leq \log(|\mathcal{X}| - 1)$.

$$\begin{aligned} H(X|E, \hat{X}) &\leq \sum_{\hat{x} \in \mathcal{X}} \Pr[\hat{X} = \hat{x}, E = 1] \log(|\mathcal{X}| - 1) \\ &= \Pr[E = 1] \log(|\mathcal{X}| - 1) \\ &= P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

□

Proof.

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} = \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log(|X|-1)}.$$

$$H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{\Pr[\hat{X} = \hat{x}, E = 0] H(X|\hat{X} = \hat{x}, E = 0) + \Pr[\hat{X} = \hat{x}, E = 1] H(X|\hat{X} = \hat{x}, E = 1)\}.$$

$$\begin{aligned} H(X|E, \hat{X}) &\leq \sum_{\hat{x} \in \mathcal{X}} \Pr[\hat{X} = \hat{x}, E = 1] \log(|\mathcal{X}| - 1) \\ &= \Pr[E = 1] \log(|\mathcal{X}| - 1) \\ &= P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

By the data-processing inequality, we have $I(X; \hat{X}) \leq I(X; Y)$ and therefore $H(X|\hat{X}) \geq H(X|Y)$. □

Corollary

Corollary

For any two random variables X and Y, let $p = Pr(X \neq Y)$.

$$H(p) + p \log(|\mathcal{X}| - 1) \geq H(X|Y).$$

Proof.

Let $\hat{X} = Y$ in Fano's inequality. □

Application of Fano's inequality

- Prove converse in many theorems (including channel capacity)
- Compressed sensing signal model

$$y = Ax + w$$

where $A \in \mathcal{R}^{M \times d}$: projection matrix for dimension reduction. Signal x is sparse. Want to estimate x from y.

Fano's inequality

Lemma 2.10.1

If X and X' are i.i.d. with entropy $H(X)$,

$$\Pr[X = X'] \geq 2^{-H(X)},$$

with equality iff X has a uniform distribution.

Corollary

Let X, X' be independent with $X \sim p(x)$, $X' \sim r(x)$, $x, x' \in \mathcal{X}$. Then

$$\Pr\left[X = X'\right] \geq 2^{-H(p)-D(p\|r)}$$
$$\Pr\left[X = X'\right] \geq 2^{-H(r)-D(r\|p)}$$