第 2 讲 熵的更多性质

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

熵，相对熵与互信息的链式法则

Jensen 不等式

Jensen 不等式的应用

对数和不等式及其应用

数据处理不等式

参考资料

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

- T. M. Cover and J. A. Thomas, Elements of Information Theory, 2.1-2.6.
- R. W. Yeung, Information and Network Coding, 2.5-2.7.
- F. Alajaji, P. Chen, An Introduction to Single-User Information Theory, 2.1-2.3.

熵，相对熵与互信息的链式法则

Jensen 不等式

Jensen 不等式的应用

对数和不等式及其应用

数据处理不等式

回顾

定理 1.1: 链式法则

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

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 1.2: 熵的链式法则

设随机变量 $X_1, X_2, \cdots, X_n$ 服从 $p(x_1, x_2, \cdots, x_n)$，则

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

证明.

反复应用定理1.1. $\square$

定义 1.3

随机变量 X 和 Y 在给定随机变量 Z 时的条件互信息定义为

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}$$

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

注意到 $I(X;Y|Z)$ 关于 X 和 Y 对称.

定理 1.4: 互信息的链式法则

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

证明.

$$\begin{aligned} & I(X_1, X_2, \cdots, X_n; Y) \\ = \quad & H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n | Y) \end{aligned}$$

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 1.5: 互信息的链式法则

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

证明.

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

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 1.6: 互信息的链式法则

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

证明.

$$\begin{aligned}
& I(X_1, X_2, \cdots, X_n; Y) \\
= \;& H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n | Y) \\
= \;& \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1) - \sum_{i-1}^{n} H(X_i | X_{i-1}, \cdots, X_1, Y) \\
= \;& \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \cdots, X_1). \quad \square
\end{aligned}$$

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定义 1.7

对于联合概率密度函数 $p(x,y)$ 和 $q(x,y)$，条件相对熵 $D(p(y|x)\|q(y|x))$ 定义为条件概率密度函数 $p(y|x)$ 和 $q(y|x)$ 之间的平均相对熵，其中取平均是关于概率密度函数 $p(x)$ 而言的. 更确切地，

$$
\begin{aligned}
D(p(y|x)\|q(y|x)) &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\
&= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}.
\end{aligned}
$$

定理 1.8: 相对熵的链式法则

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

证明.

$$D(p(x,y)\|q(x,y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)}$$

定理 1.9: 相对熵的链式法则

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

证明.

$$\begin{aligned} D(p(x,y)\|q(x,y)) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \end{aligned}$$

定理 1.10: 相对熵的链式法则

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

证明.

$$\begin{aligned}
D(p(x,y)\|q(x,y)) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)}
\end{aligned}$$

定理 1.11: 相对熵的链式法则

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

证明.

$$\begin{aligned}
D(p(x,y)\|q(x,y)) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} \\
&= D(p(x)\|q(x)) + D(p(y|x)\|q(y|x)). \quad \square
\end{aligned}$$

熵，相对熵与互信息的链式法则

Jensen 不等式

Jensen 不等式的应用

对数和不等式及其应用

数据处理不等式

定义 2.1

若对于任意的 $x_1, x_2 \in (a,b)$ 及 $0 \le \lambda \le 1$，满足

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

则称函数 $f(x)$ 在区间 (a,b) 上是凸的. 如果仅当 $\lambda = 0$ 或 $\lambda = 1$，上式等号成立，则称函数 f 是严格凸的.

定义 2.2

如果 $-f$ 为凸函数，则称函数 f 是凹的.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 2.3

如果函数 f 在某个区间上存在非负 (正) 的二阶导数，则 f 为该区间的凸函数（严格凸函数）.

定理 2.4

若给定凸函数 f 和一个随机变量 X，则

$$Ef(X) \geq f(EX). \tag{2.1}$$

进一步，若 f 是严格凸的，那么式 (3.1) 中的等式蕴含 $X = EX$ 的概率为 1（即 X 是一个常量）.

熵，相对熵与互信息的链式法则

Jensen 不等式

Jensen 不等式的应用

对数和不等式及其应用

数据处理不等式

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 3.1

若给定凸函数 f 和一个随机变量 X，则

$$Ef(X) \geq f(EX). \tag{3.1}$$

进一步，若 f 是严格凸的，那么式 (3.1) 中的等式蕴含 $X = EX$ 的概率为 1（即 X 是一个常量）.

定理 3.2: 信息不等式

设 $p(x), q(x)(x \in \mathcal{X})$ 为两个概率密度函数，则

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

当且仅当对任意的 x，$p(x) = q(x)$，等号成立.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

证明.

设 $A=\{x:p(x)>0\}$ 为 $p(x)$ 的支撑集，则

$$
\begin{aligned}
D(p\|q) &= \sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} \\
&= -\sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \\
&\geq -\log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} \quad (\text{Jensen 不等式，利用到 } f(t)=-\log t \text{ 是凸函数}) \\
&= -\log\sum_{x\in A} q(x) \\
&\geq -\log\sum_{x\in\mathcal{X}} q(x) \\
&= -\log 1 = 0.
\end{aligned}
$$

讨论上面式中等号成立的条件，我们知等号成立当且仅当对任意的 x，有 $p(x)=q(x)$，$D(p\|q)=0$. □

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

注记 3.3

我们可以利用下面的不等式给出上面结果的另一个证明：对任意 $a>0$，$\ln a \geq 1-\frac{1}{a}$ （这个不等式也被称为**基本不等式**），等号成立当且仅当 $a=1$. 这个不等式可以通过微积分的技巧证明. 我们记 B 为 $q(x)$ 的支撑集，于是

$$
\begin{aligned}
D(p\|q) &= (\log e)\sum_{x\in B} p(x)\ln\frac{p(x)}{q(x)} \\
&\geq (\log e)\sum_{x\in B} p(x)(1-\frac{q(x)}{p(x)}) \\
&= (\log e)[\sum_{x\in B} p(x) - \sum_{x\in B} q(x)] \\
&\geq 0,
\end{aligned}
$$

讨论上面式中等号成立的条件,我们知等号成立当且仅当对任意的 x,有 $p(x)=q(x)$，$D(p\|q)=0$.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

推论 3.4: 互信息的非负性

对任意两个随机变量 X 和 Y，

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

当且仅当 X 与 Y 相互独立，等号成立.

证明.
$I(X;Y) = D(p(x,y)\|p(x)p(y)) \geq 0$，当且仅当 $p(x,y) = p(x)p(y)$ （即 X 与 Y 为相互独立），等号成立. $\square$

推论 3.5

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

当且仅当对任意的 y 以及满足 $p(x) > 0$ 的 x，有 $p(y|x) = q(y|x)$，等号成立.

推论 3.6

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

当且仅当对给定随机变量 Z，X 和 Y 是条件独立的，等号成立.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 3.7

$H(X) \leq \log|\mathcal{X}|$，其中 $\mathcal{X}$ 表示 X 的字母表 $\mathcal{X}$ 中的元素个数，当且仅当 X 服从 $\mathcal{X}$ 上均匀分布，等号成立.

证明.
设 $u(x) = \frac{1}{|\mathcal{X}|}$ 为 $\mathcal{X}$ 上均匀分布的概率密度函数，$p(x)$ 是随机变量 X 的概率密度函数. 于是

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

因而由相对熵的非负性，

$$0 \leq D(p\|u) = \log|X| - H(X).$$

□

推论 3.8

对于任意非负实数，存在一个随机变量，使得其熵恰好为该数.

证明.

考虑定义在有限字母表 $\mathcal{X}$ 上的随机变量 X. 由上面的定理我们知当 X 在 $\mathcal{X}$ 上一致分布的时候，$H(X) = \log|\mathcal{X}|$ 可以达到. 当 X 为一个确定值的时候 $H(X) = 0$ 可以达到. 对于 $0 \le a \le |\mathcal{X}|^{-1}$，记

$$g(a) = H(\{1 - (|\mathcal{X}| - 1)a, a, \cdots, a\}) = -l(1 - (|\mathcal{X}| - 1)a) - (|\mathcal{X}| - 1)l(a),$$

其中 $l : [0, \infty) \to \mathbb{R}$ 由下定义：

$$l(a) = \begin{cases} a \log a & \text{如果 } a > 0 \\ 0 & \text{如果 } a = 0. \end{cases}$$

注意到 $g(a)$ 是 a 的一个连续函数，满足 $g(0) = 0$ 以及 $g(|\mathcal{X}|^{-1}) = \log|\mathcal{X}|$. 由介值定理，我们知对任意 b，$0 < b < \log|\mathcal{X}|$，我们可以找到一个 a，使得得到的随机变量 X 满足 $H(X) = b$.
由上讨论我们知只要让 $|\mathcal{X}|$ 充分大，我们可以使 $H(X)$ 达到任意的正值，从而得证. $\square$

由上面的讨论我们知如果一个随机变量的字母表是有限的，那么它的熵一定是有限的. 若字母表为无限的, 上一讲里我们已经看到了一个无穷大的例子. 这一讲, 我们通过下面两个新例子来说明字母表为无限的，其熵既可能是有限的，也可能是无限的..

例 3.9

设 X 为一个取值于自然数的随机变量，且 $P\{X=i\}=2^{-i}$. 则

$$H_2(X)=\sum_{i=1}^{\infty} i2^{-i}=2,$$

为一有限值.

例 3.10

设 Y 是取值于整数对的子集

$$\{(i,j) : 1 \le i < \infty,\ 1 \le j \le \frac{2^{2^i}}{2^i}\}$$

满足对任意 i，j，

$$P\{Y = (i,j)\} = 2^{-2^i}$$

的一个随机变量. 则

$$H_2(Y) = -\sum_{i=1}^{\infty} \sum_{j=1}^{2^{2^i}/2^i} 2^{-2^i} \log_2 2^{-2^i} = \sum_{i=1}^{\infty} 1 = \infty,$$

故其熵为无穷.

定理 3.11: 条件作用使熵减小

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

当且仅当 X 与 Y 相互独立时，等号成立.

证明.

$0 \le I(X;Y) = H(X) - H(X|Y).$ □

例 3.12

设随机变量 (X, Y) 有如下的联合分布：

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

于是 $H(X) = H(\frac{1}{8}, \frac{7}{8}) = 0.544$比特，$H(X|Y = 1) = 0$比特，$H(X|Y = 2) = 1$比特. 于是我们可以得到 $H(X|Y) = \frac{3}{4}H(X|Y = 1) + \frac{1}{4}H(X|Y = 2) = 0.25$比特.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

熵的独立界

定理 3.13: 熵的独立界

设 X_1，$X_2, \cdots, X_n$ 服从 $p(x_1, x_2, \cdots, x_n)$，则

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

当且仅当 $X_1, X_2, \cdots, X_n$ 相互独立，等号成立.

证明.

由熵的链式法则，我们有

$$\begin{aligned} H(X_1, X_2, \cdots, X_n) &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1) \\ &\leq \sum_{i=1}^{n} H(X_i). \end{aligned}$$

等号成立当且仅当对任意 i 有 X_i 和 $X_{i-1}, \cdots, X_1$ 独立. 于是对任意 $x_1, \cdots, x_n$, 有

$$\begin{aligned} p(x_1, \cdots, x_n) &= p(x_1, x_2, \cdots, x_{n-1}) p(x_n) \\ &= p(x_1, x_2, \cdots, x_{n-2}) p(x_{n-1}) p(x_n) \\ &\vdots \\ &= p(x_1) p(x_2) \cdots p(x_n), \end{aligned}$$

也就是说, $X_1, X_2, \cdots, X_n$ 相互独立. $\square$

熵，相对熵与互信息的链式法则

Jensen 不等式

Jensen 不等式的应用

对数和不等式及其应用

数据处理不等式

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 4.1: 对数和不等式

对于非负数 $a_1, a_2, \cdots, a_n$ 和 $b_1, b_2, \cdots, b_n$,

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

当且仅当 $\frac{a_i}{b_i} =$ 常数，等号成立.

这里我们约定 $0 \log 0 = 0$，$a \log \frac{a}{0} = \infty$(当$a > 0$)，$0 \log \frac{0}{0} = 0$.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

证明.

不失一般性，我们可以假定 $a_i > 0$，$b_i > 0$. 由于对任意的正数 t 有 $f''(t) = \frac{1}{t}\log e > 0$，可知函数 $f(t) = t\log t$ 严格凸. 从而由 Jensen 不等式，我们有

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right).$$

其中 $\alpha_i \geq 0$，$\sum_i \alpha_i = 1$. 令 $\alpha_i = \frac{b_i}{\sum_{j=1}^n b_j}$，$t_i = \frac{a_i}{b_i}$，可得

$$\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_j} \log \sum \frac{a_i}{\sum b_j}.$$

这就是对数和不等式. $\square$

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 4.2: 相对熵的凸性

$D(p\|q)$ 关于对 (p,q) 是凸的，即，如果 (p_1,q_1) 和 (p_2,q_2) 为两对概率密度函数，则对所有的 $0\le\lambda\le1$，有

$$D(\lambda p_1+(1-\lambda)p_2\|\lambda q_1+(1-\lambda)q_2)\le\lambda D(p_1\|q_1)+(1-\lambda)D(p_2\|q_2).$$

证明.
由对数不等式我们有：

$$
\begin{aligned}
(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \\
\leq \lambda p_1(x) \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}
\end{aligned}
$$

对 x 求和我们可以得到所需要的结论. $\square$

定理 4.3: 熵的凹性

$H(p)$ 是关于 p 的凹函数.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 4.4: 熵的凹性

$H(p)$ 是关于 p 的凹函数.

证明.
我们有

$$H(p) = \log|\mathcal{X}| - D(p\|u).$$

其中 u 为 $|\mathcal{X}|$ 个结果的均匀分布. 从而 H 的凹性可由 D 的凸性直接得到. $\square$

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 4.5

设 $(X,Y)\sim p(x,y)=p(x)p(y|x)$. 如果固定 $p(y|x)$，则互信息 $I(X;Y)$ 是关于 $p(x)$ 的凹函数；而如果固定 $p(x)$，则互信息 $I(X;Y)$ 是关于 $p(y|x)$ 的凸函数.

证明.

为了证明第一部分，将互信息展开

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X=x).$$

如果固定 $p(y|x)$，则 $p(y)$ 时关于 $p(x)$ 的线性函数. 因而，关于 $p(y)$ 的凹函数 $H(Y)$ 也是 $p(x)$ 的凹函数. 上式中的第二项是关于 $p(x)$ 的凹函数. 因此，它们的差仍是关于 $p(x)$ 的凹函数.

证明.

为证明第二部分，我们先固定 $p(x)$，同时考虑两个不同的条件分布 $p_1(y|x)$ 和 $p_2(y|x)$. 对应的联合分布为 $p_1(x,y)=p(x)p_1(y|x)$ 和 $p_2(x,y)=p(x)p_2(y|x)$，同时对应的边缘分布为 $p(x)$，$p_1(y)$ 以及 $p(x)$，$p_2(y)$. 我们考虑条件分布

$$p_\lambda(y|x)=\lambda p_1(y|x)+(1-\lambda)p_2(y|x),$$

为 $p_1(y|x)$ 和 $p_2(y|x)$ 的一个凸组合，其中 $0\le\lambda\le 1$. 对应的联合分布为

$$p_\lambda(x,y)=\lambda p_1(x,y)+(1-\lambda)p_2(x,y).$$

且 Y 的分布也是一个凸组合

$$p_\lambda(y)=\lambda p_1(y)+(1-\lambda)p_2(y).$$

因此，如果设 $q_\lambda(x,y)=p(x)p_\lambda(y)$ 为边际分布的乘积，则有

$$q_\lambda(x,y)=\lambda q_1(x,y)+(1-\lambda)q_2(x,y).$$

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

证明.
由于互信息是联合分布和边际分布乘积的相对熵，有

$$I_\lambda(X;Y) = D(p_\lambda(x,y)\|q_\lambda(x,y)).$$

相对熵 $D(p\|q)$ 为关于二元对 (p,q) 的凸函数，由此可知，

$$\begin{aligned}
I(X;Y) &= D(p_\lambda(x,y)\|q_\lambda(x,y)) \\
&= D(\lambda p_1(x,y) + (1-\lambda)p_2(x,y)\|\lambda q_1(x,y) + (1-\lambda)q_2(x,y)) \\
&\leq \lambda D(p_1(x,y)\|q_1(x,y)) + (1-\lambda)D(q_1(x,y)\|q_2(x,y)) \\
&= \lambda I_1(X;Y) + (1-\lambda)I_2(X;Y).
\end{aligned}$$

于是互信息是条件分布的凸函数. $\square$

熵，相对熵与互信息的链式法则

Jensen 不等式

Jensen 不等式的应用

对数和不等式及其应用

数据处理不等式

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

定理 5.1: 数据处理不等式

若 $X \to Y \to Z$，则有 $I(X;Y) \geq I(X;Z)$.

上面结果说明，数据处理过程中只会失掉一些信息，绝不会创造出新的信息.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

证明.
由链式法则，将互信息以两种不同方式展开：

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

由于在 Y 给定的条件下，X 和 Z 是条件独立的，因此有 $I(X;Z|Y) = 0$. 又由于 $I(X;Y|Z) \geq 0$, 则有

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

当且仅当 $I(X;Y|Z) = 0$ (即 $X \to Z \to Y$ 构成马尔可夫链) 等号成立. 类似地可以证明 $I(Y;Z) \geq I(X;Z)$. $\square$

推论 5.2

特别地，如果 $Z = g(Y)$，则 $I(X;Y) \geq I(X;g(Y))$.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

推论 5.3

如果 $X \to Y \to Z$，则 $I(X;Y|Z) \leq I(X;Y)$.

推论 5.4

如果 $X \to Y \to Z$，则 $I(X;Y|Z) \le I(X;Y)$.

证明.
在数据处理不等式的证明中我们看到

$$I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y).$$

由马尔可夫性 $I(X;Z|Y) = 0$，又我们有 $I(X;Z) \ge 0$，从而

$$I(X;Y|Z) \le I(X;Y).$$

□

注记 5.5

注意当 X, Y, Z 不构成马尔可夫链时，有可能 $I(X;Y|Z) > I(X;Y)$.

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

熵，相对熵与互信息的链式法则
Jensen 不等式
Jensen 不等式的应用
对数和不等式及其应用
数据处理不等式

注记 5.6

注意当 X, Y, Z 不构成马尔可夫链时，有可能 $I(X;Y|Z) > I(X;Y)$.

设 X, Y 是相互独立的二元随机变量，$Z = X + Y$. 则 $I(X;Y) = 0$，但 $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) = P(Z=1)H(X|Z=1) = \frac{1}{2}$比特. (注意到当 $Z = 0$ 或 2 时，X 的值就确定下来了).