



# 第2章 微分熵、相对熵与互信息 (第二部分)

樊平毅 教授

清华大学电子工程系 WIST LAB.

Email: [fpv@tsinghua.edu.cn](mailto:fpv@tsinghua.edu.cn)

---

2022年9月14日

- 1 定义 .....
- ~~2 连续随机变量的 AEP .....~~
- 3 微分熵与离散熵的关系 .....
- 4 联合微分熵与条件微分熵 .....
- 5 相对熵与互信息 .....
- 6 微分熵、相对熵以及互信息的性质 .....

在后面的章节讨论

## 定义 2.10. 微分熵

设  $X$  是一连续的随机变量，其概率密度函数为  $f(x)$ ，定义

$$h(X) = - \int_S f(x) \log f(x) dx \quad (2.139)$$

为随机变量  $X$  的微分熵，其中  $S$  是随机变量  $X$  的定义域。



**例 2.8** 设  $X$  是在  $[0, a]$  上均匀分布的随机变量，求其微分熵。

$$h(X) = - \int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a$$

**例 2.9** 设  $X$  是服从高斯分布的随机变量，其分布密度函数为  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}$ ，求其微分熵。

解

$$h(X) = - \int_{-\infty}^{\infty} f(x) \left( -\frac{x^2}{2\sigma^2} \log e - \log \sqrt{2\pi}\sigma \right) dx \quad (2.141)$$

$$= \frac{\log e}{2\sigma^2} E\{X^2\} + \frac{1}{2} \log(2\pi\sigma^2) \quad (2.142)$$

$$= \frac{1}{2} \log(2\pi e\sigma^2) \quad (2.143)$$

考虑连续随机变量  $X$ ，其概率密度为  $f(x)$ 。我们将随机变量的取值范围分成长度为  $\Delta$  的小区间，利用中值定理，我们知：在每个小区间内存在一  $x_i$  点，使得

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx \quad (2.144)$$

考虑量化随机变量  $X^\Delta$ ，其定义为  $X^\Delta = x_i$ ， $i\Delta \leq x < (i+1)\Delta$ ，于是有

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

则量化随机变量  $X^\Delta$  的熵为

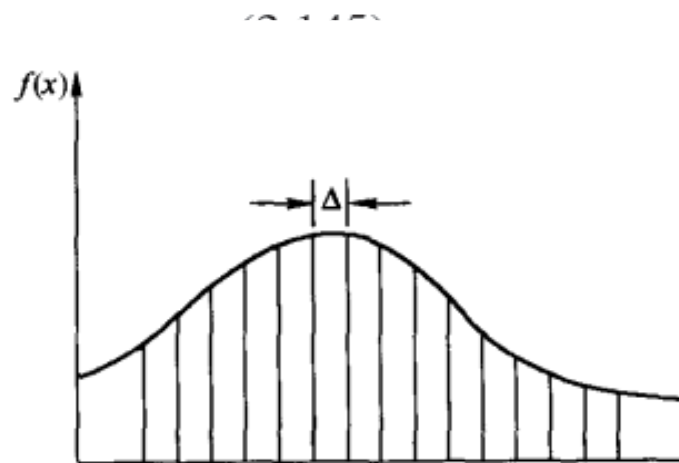
$$\begin{aligned} H(X^\Delta) &= -\sum_i p_i \log p_i = -\sum_i f(x_i)\Delta \log(f(x_i)\Delta) \\ &= -\sum_i \Delta f(x_i) \log f(x_i) - \log \Delta \end{aligned}$$

注意，在上式推导中利用了  $\sum_i f(x_i)\Delta = \int_{-\infty}^{\infty} f(x)dx = 1$ 。

基于上述推导，我们知

$$\lim_{\Delta \rightarrow 0} (H(X^\Delta) + \log \Delta) = h(X) \quad (2.148)$$

显然， $\Delta$  为量化间隔，对连续随机变量  $X$  进行  $n$  比特量化，那么得到的量化随机变量的信息熵近似为  $h(X) + n$ 。



定理

如果随机变量  $X$  的密度函数  $f(x)$  是黎曼可积的, 那么

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X), \text{ 当 } \Delta \rightarrow 0$$

如果随机变量  $X$  的密度函数  $f(x)$  是黎曼可积的, 那么

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X), \text{ 当 } \Delta \rightarrow 0$$

熵大约为  $h(X) + n$ .

1. 如果  $X$  服从  $[0,1]$  上的均匀分布, 取  $\Delta=2^{-n}$ , 则  $h=0$ ,  $H(X^\Delta)=n$ , 于是, 在精确到  $n$  位的意义下, 使用  $n$  比特足以描述  $X$ 。

2. 如果  $X$  服从  $\left[0, \frac{1}{8}\right]$  上的均匀分布, 那么在二进制表示中,  $X$  取值的小数点右边的前 3 位必定为 0。因而, 在精确到  $n$  位的意义下, 描述  $X$  仅需  $n-3$  比特, 这与  $h(X)=n-3$  相一致。

3. 如果  $X \sim \mathcal{N}(0, \sigma^2)$  且  $\sigma^2=100$ , 那么, 在精确到  $n$  位的意义下, 描述  $X$  需要的平均长度为  $n + \frac{1}{2} \log(2\pi e\sigma^2) = n + 5.37$  比特。

一个离散随机变量的微分熵可以看成  $-\infty$ 。注意到  $2^{-\infty}=0$


- 这个推导过程说明，对于一个信号，当两个不同的信息处理者采用完全相同的量化方式，得到的微分性信息熵才能一致，否则就会有较大的差距，差距来自量化间隔；
- 对于信号放大或缩小，差距体现在放大系数的对数值上，这就是隐含了量化采用了通用量化器，而没有考虑同一信号的尺度放缩现象；
- 从另一格角度讲，在数据处理，特别是，图像数据处理中，如果同一幅图像采用的尺寸不同，需要的相对信息量也不同，不同来源于尺度比的对数值；这与人们的实际印象有差距；所以，在类似问题的处理上，应仔细考虑这一现象。那么如何避免，也就是说，需要规定一个尺度一致的规则；
- 反映在矩阵变换上，也有类同的现象；所以通常采用U变换进行信息处理或调整；



### 定义 2.11. 联合微分熵

设连续随机变量  $X$  与  $Y$  的联合概率密度函数为  $f(x, y)$ , 则

$$h(X, Y) = - \iint_{S_X \times S_Y} f(x, y) \log f(x, y) dx dy \quad (2.149)$$

称为连续随机变量  $X$  与  $Y$  的联合微分熵。 

### 定义 2.12. 条件微分熵

设连续随机变量  $X$  与  $Y$  的联合概率密度函数为  $f(x, y)$ , 定义

$$h(X|Y) = \iint_{S_X \times S_Y} f(x, y) \log f(x|y) dx dy \quad (2.150)$$



定理 设  $X_1, X_2, \dots, X_n$  服从均值为  $\mu$ , 协方差矩阵为  $K$  的多元正态分布, (使用  $\mathcal{N}_n(\mu, K)$  或  $\mathcal{N}(\mu, K)$  来记该分布。 则

$$h(X_1, X_2, \dots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K|$$

其中  $|K|$  表示  $K$  的行列式。

注解:  $X_1, X_2, \dots, X_n$  的联合概率密度函数为

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}$$

$X_1, X_2, \dots, X_n$  的联合概率密度函数为

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T K^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$h(f) = - \int f(\mathbf{x}) \left[ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T K^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \ln(\sqrt{2\pi})^n |K|^{\frac{1}{2}} \right] d\mathbf{x}$$

$$= \frac{1}{2} E \left[ \sum_{i,j} (X_i - \mu_i)(K^{-1})_{ij}(X_j - \mu_j) \right] + \frac{1}{2} \ln(2\pi)^n |K|$$

$$= \frac{1}{2} E \left[ \sum_{i,j} (X_i - \mu_i)(X_j - \mu_j)(K^{-1})_{ij} \right] + \frac{1}{2} \ln(2\pi)^n |K|$$

$$= \frac{1}{2} \sum_{i,j} E[(X_j - \mu_j)(X_i - \mu_i)] (K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K|$$

$$= \frac{1}{2} \sum_j \sum_i \underline{K_{ji}} (K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K|$$

换位

$$= \frac{1}{2} \sum_j (KK^{-1})_{jj} + \frac{1}{2} \ln(2\pi)^n |K|$$

$$= \frac{1}{2} \sum_j I_{jj} + \frac{1}{2} \ln(2\pi)^n |K|$$

$$= \frac{n}{2} + \frac{1}{2} \ln(2\pi)^n |K|$$

$$= \frac{1}{2} \ln(2\pi e)^n |K| \quad \text{奈特}$$

$$= \frac{1}{2} \log(2\pi e)^n |K| \quad \text{比特} \quad \text{二维到高维}$$

**定义 2.13. 相对微分熵**

设  $f(x)$  与  $g(x)$  是概率密度函数，定义

$$D(f \parallel g) = \int_{S_X} f(x) \log \frac{f(x)}{g(x)} dx \quad (2.151)$$



**注**  $D(f \parallel g)$  是有限的必要条件是密度函数  $f(x)$  的支撑域（即函数的定义域部分）包含在密度函数  $g(x)$  的支撑域内。

**定义 2.14. 相对微分熵**

设连续随机变量  $X$  与  $Y$  的联合概率密度函数为  $f(x, y)$ ，定义

$$I(X; Y) = \iint_{S_X \times S_Y} f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy = D(f(x, y) \parallel f(x)f(y)) \quad (2.152)$$



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

$$I(X; Y) = D(f(x, y) \| f(x)f(y))$$

$$\begin{aligned} I(X^\Delta; Y^\Delta) &= H(X^\Delta) - H(X^\Delta|Y^\Delta) \\ &\approx h(X) - \log\Delta - (h(X|Y) - \log\Delta) \\ &= I(X; Y) \end{aligned}$$

离散与量化连续的关系

**注意**  $D(f \parallel g)$  与  $I(X;Y)$  的定义与讨论离散随机变量的情况类似。事实上,

$$\begin{aligned} I(X^\Delta; Y^\Delta) &= H(X^\Delta) - H(X^\Delta|Y^\Delta) \\ &\approx h(X) - \log \Delta - (h(X|Y) - \log \Delta) \\ &= h(X) - h(X|Y) \\ &= I(X;Y) \end{aligned}$$

即两个连续随机变量的互信息等于它们量化随机变量的互信息的极限。

**例 2.10** 设  $X, Y$  是正态分布的随机变量,  $E\{X\} = E\{Y\} = 0$ ,  $\text{Var}\{X\} = \sigma_1^2$ ,  $\text{Var}\{Y\} = \sigma_2^2$ ,  $\rho = \frac{E\{XY\}}{\sigma_1\sigma_2}$ , 求  $h(X)$ ,  $h(Y)$ ,  $I(X, Y)$ ?

**解** 根据已知条件, 可得  $X, Y$  的联合概率密度为

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} - \rho\frac{2xy}{\sigma_1\sigma_2}}{2(1-\rho^2)}\right\} \quad (2.165)$$

相应的边沿概率密度为

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{x^2}{2\sigma_1^2}\right\} \quad (2.166)$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{y^2}{2\sigma_2^2}\right\} \quad (2.167)$$

条件概率密度为

$$f_{X|Y}(x|y) = f_{XY}(x, y) / f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp \left\{ -\frac{\left(x - \rho \frac{\sigma_1 y}{\sigma_2}\right)^2}{2\sigma_1^2(1-\rho^2)} \right\} \quad (2.168)$$

利用微分熵的定义得：

$$h(X) = \log \sqrt{2\pi e \sigma_1^2} \quad (2.169)$$

$$h(X|Y) = \log \sqrt{2\pi e \sigma_1^2(1-\rho^2)} \quad (2.170)$$

$$I(X; Y) = h(X) - h(X|Y) = \log \sqrt{1/(1-\rho^2)} \quad (2.171)$$

当  $\rho=0$  时， $X$  与  $Y$  相互独立以及互信息为 0。

当  $\rho=\pm 1$  时， $X$  与  $Y$  完全相关且互信息为无穷大。



- 公式反映了互相关系数与**SNR**的关系，特别是高斯白噪声信道信息传输；
- 多个信号组合在一起进行解调或解码，往往优于单符号解调或解码，其核心在于多个信号内部可能有相关性，会提升联合后的信息功率，及信息谐振；从而使得系统的分段信噪比大于平均信噪比；因为对应的多符号噪声是白噪声，相互独立，噪声功率是满足加法规则，简单的线性加法；
- 信道纠错编码，可能利用了信息谐振的思想；那么，在构造信息谐振中是否有新的思路，可以帮助优化系统设计呢？

## 性质

- (1)  $h(X|Y) = h(X, Y) - h(Y)$
- (2)  $I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$
- (3)  $D(f \parallel g) \geq 0$ , 等号成立的条件为  $f(x) = g(x)$  几乎处处成立。
- (4)  $I(X; Y) \geq 0$ , 等号成立的条件  $X$  与  $Y$  统计独立。
- (5)  $h(X|Y) \leq h(X)$ , 等号成立的条件  $X$  与  $Y$  统计独立。
- (6)  $h(X + a) = h(X)$ , 其中  $a$  为一常数。
- (7)  $h(aX) = h(X) + \log|a|$ ,  $a \neq 0$ 。
- (8) 若  $A$  为  $n \times n$  的方阵,  $\vec{X}$  为  $n$  维随机向量, 则

$$h(A\vec{X}) = h(\vec{X}) + \log|A|$$

其中  $|A|$  为矩阵的行列式的绝对值。

## 定理 8.6.1

$$D(f \parallel g) \geq 0$$

当且仅当  $f = g$ , 几乎处处(a.e.)等号成立。

证明: 设  $f$  的支撑集为  $S$ 。则

$$\begin{aligned} -D(f \parallel g) &= \int_S f \log \frac{g}{f} \\ &\leq \log \int_S f \frac{g}{f} \quad (\text{由 Jensen 不等式}) \\ &= \log \int_S g \\ &\leq \log 1 = 0 \end{aligned}$$

当且仅当 Jensen 不等式中的等号成立, 即当且仅当  $f = g$  a.e. 等号成立。

**推论**  $I(X; Y) \geq 0$ , 当且仅当  $X$  与  $Y$  相互独立等号成立。

**推论**  $h(X|Y) \leq h(X)$ , 当且仅当  $X$  与  $Y$  相互独立等号成立。

**定理 8.6.2** (微分熵的链式规则)

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

**证明:** 利用概率密度的分解形式直接可得所要结论

**推论**

$$h(X_1, X_2, \dots, X_n) \leq \sum h(X_i)$$

当且仅当  $X_1, X_2, \dots, X_n$  相互独立等号成立。

应用(阿达玛(Hadamard)不等式) 设  $\mathbf{X} \sim \mathcal{N}(0, \mathbf{K})$  是一个多元正态分布, 那么将熵的定义公式代入上面的不等式中, 我们就可以得到

$$|\mathbf{K}| \leq \prod_{i=1}^n K_{ii} \quad (8-64)$$

定理:

平移变换不会改变微分熵。

$$h(\mathbf{X} + \mathbf{c}) = h(\mathbf{X})$$

高维情况

$$h(a\mathbf{X}) = h(\mathbf{X}) + \log|a| \quad \longrightarrow \quad h(\mathbf{A}\mathbf{X}) = h(\mathbf{X}) + \log|\det(\mathbf{A})|$$

证明: 令  $\mathbf{Y} = \mathbf{a}\mathbf{X}$ 。则  $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$ , 且经过积分变量替换, 有

$$\begin{aligned} h(a\mathbf{X}) &= - \int f_Y(y) \log f_Y(y) dy \\ &= - \int \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy \\ &= - \int f_X(x) \log f_X(x) dx + \log|a| \\ &= h(\mathbf{X}) + \log|a| \end{aligned}$$

设随机向量  $\mathbf{X} \in \mathbf{R}^n$  的均值为零, 协方差矩阵为  $K = E\mathbf{X}\mathbf{X}^t$  (即  $K_{ij} = EX_iX_j$ ),

则  $h(\mathbf{X}) \leq \frac{1}{2} \log(2\pi e)^n |K|$ , 当且仅当  $\mathbf{X} \sim \mathcal{N}(0, K)$  等号成立

证明: 设  $g(\mathbf{x})$  是对任意的  $i$  和  $j$  均满足  $\int g(\mathbf{x}) x_i x_j d\mathbf{x} = K_{ij}$  的密度函数

令  $\phi_K$  是  $\mathcal{N}(0, K)$  随机向量  $\log \phi_K(\mathbf{x})$  是一个二次型

并且  $\int x_i x_j \phi_K(\mathbf{x}) d\mathbf{x} = K_{ij} \quad 0 \leq D(g \parallel \phi_K)$

$$= \int g \log(g/\phi_K)$$

$$= -h(g) - \int g \log \phi_K$$

$$= -h(g) - \int \phi_K \log \phi_K$$

$$= -h(g) + h(\phi_K)$$

替换  $\int g \log \phi_K = \int \phi_K \log \phi_K$

二次型  $\log \phi_K(\mathbf{x})$  关于  $g$  和  $\phi_K$  具有相同的矩

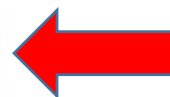
(估计误差与微分熵) 对任意随机变量  $X$  及其估计  $\hat{X}$ ,

$$E(X - \hat{X})^2 \geq \frac{1}{2\pi e} e^{2h(X)}$$

其中等号成立的充分必要条件是  $X$  为高斯分布而  $\hat{X}$  为其均值

证明: 令  $\hat{X}$  为  $X$  的一个估计,

$$\begin{aligned} E(X - \hat{X})^2 &\geq \min_{\hat{X}} E(X - \hat{X})^2 \\ &= E(X - E(X))^2 \\ &= \text{var}(X) \\ &\geq \frac{1}{2\pi e} e^{2h(X)} \end{aligned}$$



是因为  $X$  的均值是最佳估计

高斯分布在给定 方差的条件下具有最大熵

**推论** 当边信息  $Y$  以及估计  $\hat{X}(Y)$  已知时, 可以推出

$$E(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}$$

这个公式可以用于信号估计误差界分析,  
在统计学和信息论之间建立了一个桥梁;  
此公式与  $X$ ,  $Y$  的分布无关, 是一个通用的结果!!!



## (1)幅度受限

### 定理 2.7. 幅度受限

假设  $X$  的取值范围受限于有限区间  $[a, b]$ , 则  $X$  服从均匀分布时, 其微分熵达到最大值, 最大值为  $\log(b - a)$ 。



### 定理 2.8. 方差受限

设  $X$  的均值为  $\mu < \infty$ , 方差为  $\sigma^2 < \infty$ , 则  $X$  服从高斯分布时, 其微分熵达到最大值, 最大值为  $\log \sqrt{2\pi e \sigma^2}$ 。



**证明** 因为  $X$  的取值范围受限于有限区间  $[a, b]$ , 则有

$$\int_a^b f_X(x) dx = 1$$

定义拉格朗日函数


$$F(f_X(x)) = - \int_a^b f_X(x) \log f_X(x) dx + \lambda \left( \int_a^b f_X(x) dx - 1 \right)$$

$$\text{令 } \frac{\partial F}{\partial f_X} = 0, \quad \int_a^b \left( -\log f_X(x) - \frac{1}{\log e} + \lambda \right) dx = 0$$

  $f_X(x) = c$ , 其中  $c$  为待定一常数.

$$-\log f_X(x) - \frac{1}{\log e} + \lambda = 0$$

$$\int_a^b f_X(x) dx = 1$$

  $f_X(x) = \frac{1}{b-a},$

## (2) 幅度为半开区间

假设  $X$  的取值为  $[0, \infty)$ ，并且均值有限，即  $\int_0^\infty x f_X(x) dx = \mu > 0$ ，可以证明当  $X$  服从参数为  $\mu$  的负指数分布时，其微分熵达到最大值。构造一个拉格朗日函数

$$L(f_X) = - \int_0^\infty [f_X(x) \log f_X(x) + \lambda_1 f_X(x) + \lambda_2 x f_X(x)] dx \quad (2.178)$$

对  $L(f_X)$  关于  $f_X(x)$  求偏导，并令其等于 0 得：

$$\frac{\partial L(f_X)}{\partial f_X} = - \int_0^\infty [\log f_X(x) + \frac{1}{\log e} + \lambda_1 + \lambda_2 x] dx = 0 \quad (2.179)$$

于是有

$$\log f_X(x) + \frac{1}{\log e} + \lambda_1 + \lambda_2 x = 0 \quad (2.180)$$

利用条件

$$\int_0^\infty x f_X(x) dx = \mu > 0 \quad (2.181)$$

和

$$\int_0^\infty f_X(x) dx = 1 \quad (2.182)$$

得：

$$f_X(x) = \frac{1}{\mu} e^{-x/\mu} \quad (2.183)$$

相应的最大微分熵为

$$h_{\max}(X) = \int_0^\infty (\log \mu + \frac{x}{\mu}) \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) dx = 1 + \log(\mu) \quad (2.184)$$

## (2) 幅度为半开区间

假设  $X$  的取值为  $[0, \infty)$ , 并且均值有限, 即  $\int_0^\infty x f_X(x) dx = \mu > 0$ , 可以证明当  $X$  服从参数为  $\mu$  的负指数分布时, 其微分熵达到最大值。构造一个拉格朗日函数

$$L(f_X) = - \int_0^\infty [f_X(x) \log f_X(x) + \lambda_1 f_X(x) + \lambda_2 x f_X(x)] dx \quad \int_0^\infty x f_X(x) dx = \mu > 0$$

$$\frac{\partial L(f_X)}{\partial f_X} = - \int_0^\infty [\log f_X(x) + \frac{1}{\log e} + \lambda_1 + \lambda_2 x] dx = 0$$

$$\log f_X(x) + \frac{1}{\log e} + \lambda_1 + \lambda_2 x = 0 \quad \longrightarrow$$

$$f_X(x) = \frac{1}{\mu} e^{-x/\mu}$$

$$h_{\max}(X) = \int_0^\infty (\log \mu + \frac{x}{\mu}) \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) dx = 1 + \log(\mu)$$

## 定理 2.8. 方差受限

设  $X$  的均值为  $\mu < \infty$ , 方差为  $\sigma^2 < \infty$ , 则  $X$  服从高斯分布时, 其微分熵达到最大值, 最大值为  $\log \sqrt{2\pi e \sigma^2}$ .



定义拉格朗日函数

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

$$\int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx = \sigma^2$$

$$\diamond \frac{\partial F}{\partial f_X} = 0,$$

$$F(f_X(x)) = - \int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx + \lambda \left( \int_{-\infty}^{\infty} f_X(x) dx - 1 \right) + \eta \left( \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx - \sigma^2 \right)$$

$$\int_{-\infty}^{\infty} \left( -\log f_X(x) - \frac{1}{\log e} + \lambda + \eta(x - \mu)^2 \right) dx = 0$$

$$-\log f_X(x) - \frac{1}{\log e} + \lambda + \eta(x - \mu)^2 = 0$$

$$f_X(x) = c_1 e^{c_2(x-\mu)^2}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 如果将信息熵看作信息不确定度，那么对具有相同的方差的信号，高斯变量具有最高的信息不确定度，也是最差的一类分布；
- 同理将分布区间圈定在半开区间  $(0, +\infty)$  那么指数分布的变量也对应着最大不确定度的量，也是最差的一类分布；
- 在一个固定区间内，均匀分布对应的是否最大不确定度的变量？离散的情况是这样，连续的情况，需要仔细考虑一下，需要推导一下结果；
- 另外 **KL divergence** 是用于区分两个不同的分布，只要分布有差异，其结果就会显现；那么它如何体现这种差异的敏感性，历史上的讨论并不多，需要仔细推敲和分析。

对于一般的平均功率受限连续随机信号，其微分熵满足

$$h(X) \leq \log \sqrt{2\pi e P} \quad (2.195)$$

这意味着  $P \geq \frac{2^{2h(X)}}{2\pi e}$ 。

在历史上，Shannon 将连续随机变量  $X$  的熵指数函数  $N(X) = \frac{2^{2h(X)}}{2\pi e}$  定义为熵功率函数，并有著名的熵功率不等式。

定理 (Shannon): 对任意两个独立的连续随机变量  $X$  和  $Y$ ，则  $N(X + Y) \geq N(X) + N(Y)$ 。

(4) 假定每个信息比特的平均功率为  $E_b$ ，因为  $P_S$  表示每秒内信号的平均能量，则  $P_S = E_b R_b$ ， $R_b$  表示整个信号频带内所承载的信息速率，单位为 (比特/秒)。于是有

$$C = B \log \left( 1 + \frac{E_b}{N_0} \frac{R_b}{B} \right). \quad (2.224)$$

- 关于EPI 公式的基本认识;
- 对于高斯随机变量, 这个式子等号成立; 对于非高斯随机变量, 我们的推测如下:

- 我们观察左边, 当独立随机变量的个数不断增加时, 其平均功率也在增加;

考虑一个特殊的情况, 左端每个随机变量的平均功率 $P$ 相同, 那么 $n$ 个随机变量的功率为 $nP$ , 其熵指数函数的值也近似为 $nP$ , 这是由于大数定律, 左端的随机变量越来越象高斯变量; 而右边, 如果不是高斯变量, 其熵指数函数的值为 $P - \delta < P$ ,  $P > \delta > 0$ 右边的求和结果就是 $n(P - \delta)$

显然, 左边的结果大于右边;

这也就是为什么 Shannon可以猜对EPI公式;



- 这部分主要讨论随机变量是连续形式的微分熵，
- 相对熵和互信息等之间的关系；
- 也在离散变量和连续变量之间建立了内在的联系，量化是唯一途径；
- 量化保持了一些统计量的不变特征，这是非常重要的；
- 作为应用，也讨论一些不等式，
- 例如 Hardamada 不等式， 均值估计误差不等式等；

# 内容扩展部分

## 定义 2.17. Rényi熵

设  $X$  是一个离散的随机变量（随机向量），其定义空间为一个字符集（向量空间） $\Xi$ 。如果用  $p(x) = P(X = x)$ ,  $x \in \Xi$ ，表示相应的概率分布函数，则定义

$$H_{\alpha}(p) = \frac{1}{1-\alpha} \log \sum_{i=1}^n p_i^{\alpha}, \quad (2.266)$$

为离散随机变量（随机向量） $X$  的  $\alpha$  阶 Rényi 熵，其中  $0 < \alpha < \infty, \alpha \neq 1$ 。



随着  $\alpha \rightarrow 0$ ，Rényi 熵逐渐对全部可能的元素均等处理，而不考虑它们的概率值。特别的，此时可得

$$H_0(p) = \lim_{\alpha \rightarrow 0} \frac{1}{1-\alpha} \log \sum_{i=1}^n p_i^{\alpha} = \log n \quad (2.267)$$

这也是著名的 Hartley 熵，其可作为密度函数支撑集度量的对数值。

如果  $\alpha$  趋近于 1，Rényi 熵则会依据发生的概率值对待全部可能的元素，此时有

$$H_1(\mathbf{p}) = \lim_{\alpha \rightarrow 1} \frac{1}{1 - \alpha} \log \sum_{i=1}^n p_i^\alpha = - \sum_{i=1}^n p_i \log p_i \quad (2.268)$$

而这正是 Shannon 熵。

而当  $\alpha$  趋近于无穷大时，Rényi 熵将会聚焦于那些发生概率最大的事件，这意味着

$$H_\infty(\mathbf{p}) = \lim_{\alpha \rightarrow \infty} \frac{1}{1 - \alpha} \log \sum_{i=1}^n p_i^\alpha = - \log p^{\max} \quad (2.269)$$

其中  $p^{\max} = \max\{p_1, p_2, \dots, p_n\}$ 。

问题： Renyi 熵的特点是什么？  
为什么 Renyi 熵在古典信息论中没有多少应用？

- (a) Rényi熵非负;
- (b) 均匀分布时, 有  $H(u) = H_\alpha(p) = \log n$ ;
- (c) 对于满足  $\min_i p_i > 0$  的任意分布  $\mathbf{p} = (p_1, p_2, \dots, p_n)$ , 有  $H(\mathbf{p}) < H(u)$  和  $H_\alpha(\mathbf{p}) < H_\alpha(u)$ , 这意味着作为概率分布不确定性的指示器, Shannon熵和Rényi熵都在随机变量服从均匀分布时取得最大值;
- (d) 对于两个独立概率分布  $\mathbf{p} = (p_1, p_2, \dots, p_n)$  和  $\mathbf{q} = (q_1, q_2, \dots, q_n)$ , 有  $H(\mathbf{p}, \mathbf{q}) = H(\mathbf{p}) + H(\mathbf{q})$ , 以及  $H_\alpha(\mathbf{p}, \mathbf{q}) = H_\alpha(\mathbf{p}) + H_\alpha(\mathbf{q})$ 。

有关Rényi熵的研究, 感兴趣的读者可以进一步阅读Rényi最早的论文《On measures of entropy and information》[70], 以及其他有关论文或专著, 例如[3-4, 71]。

- Renyi Divergence 定义:

For finite alphabets, the *Rényi divergence* of positive order  $\alpha \neq 1$  of a probability distribution  $P = (p_1, \dots, p_n)$  from another distribution  $Q = (q_1, \dots, q_n)$  is

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \ln \sum_{i=1}^n p_i^\alpha q_i^{1-\alpha}, \quad (1)$$

观察结果:

the uniform distribution  $U = (1/n, \dots, 1/n)$ :

$$H_\alpha(P) = H_\alpha(U) - D_\alpha(P\|U) = \ln n - D_\alpha(P\|U).$$

The *Rényi entropy*

$$H_\alpha(P) = \frac{1}{1 - \alpha} \ln \sum_{i=1}^n p_i^\alpha$$

- 连续微分型Renyi 熵定义:

$$h_{\alpha}(P) = \frac{1}{1-\alpha} \ln \int (p(x))^{\alpha} dx$$

观察结果:  $h_{\alpha}(P) = \ln n - D_{\alpha}(P\|U_I)$

where  $U_I$  denotes the uniform distribution on  $I$

**Definition 2** (Simple Orders). For any simple order  $\alpha$ , the Rényi divergence of order  $\alpha$  of  $P$  from  $Q$  is defined as

$$D_{\alpha}(P\|Q) = \frac{1}{\alpha-1} \ln \int p^{\alpha} q^{1-\alpha} d\mu, \quad (9)$$

where, for  $\alpha > 1$ , we read  $p^{\alpha} q^{1-\alpha}$  as  $\frac{p^{\alpha}}{q^{\alpha-1}}$  and adopt the conventions that  $0/0 = 0$  and  $x/0 = \infty$  for  $x > 0$ .

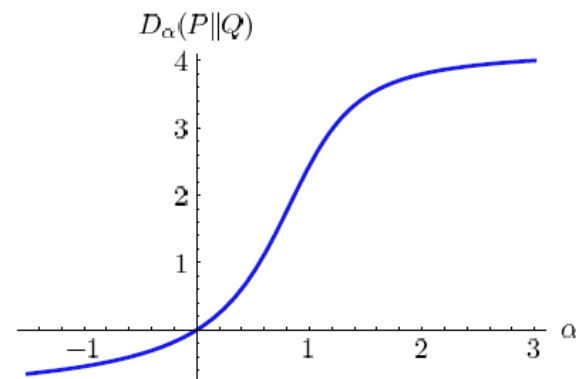


Fig. 1. Rényi divergence as a function of its order for fixed distributions

具体例子:

$$\begin{aligned} D_{\alpha}(\mathcal{N}(\mu_0, \sigma_0^2) \parallel \mathcal{N}(\mu_1, \sigma_1^2)) \\ = \frac{\alpha(\mu_1 - \mu_0)^2}{2\sigma_{\alpha}^2} + \frac{1}{1-\alpha} \ln \frac{\sigma_{\alpha}}{\sigma_0^{1-\alpha} \sigma_1^{\alpha}}, \end{aligned}$$

provided that  $\sigma_{\alpha}^2 = (1-\alpha)\sigma_0^2 + \alpha\sigma_1^2 > 0$  [20, p. 45].



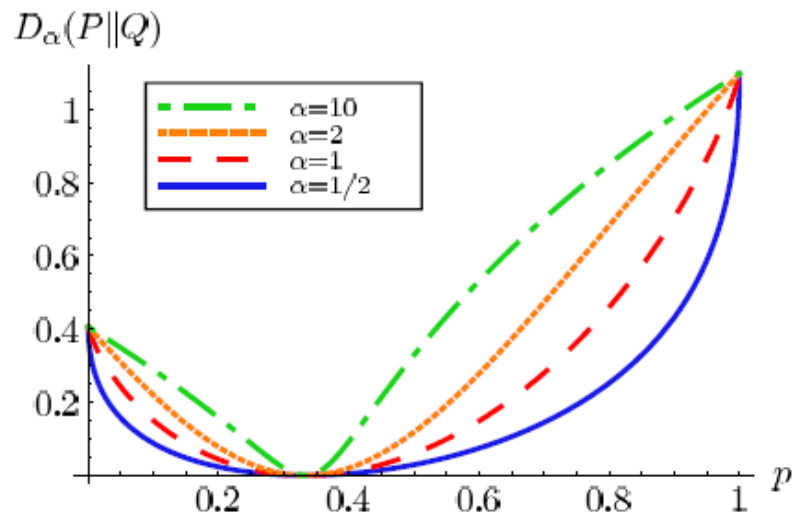


Fig. 2. Rényi divergence as a function of  $P = (p, 1-p)$  for  $Q = (1/3, 2/3)$

**Theorem 11.** For any order  $\alpha \in [0, 1]$  Rényi divergence is jointly convex in its arguments. That is, for any two pairs of probability distributions  $(P_0, Q_0)$  and  $(P_1, Q_1)$ , and any  $0 < \lambda < 1$

$$D_\alpha((1-\lambda)P_0 + \lambda P_1 \| (1-\lambda)Q_0 + \lambda Q_1) \leq (1-\lambda)D_\alpha(P_0 \| Q_0) + \lambda D_\alpha(P_1 \| Q_1). \quad (23)$$

Equality holds if and only if

$$\alpha = 0: D_0(P_0 \| Q_0) = D_0(P_1 \| Q_1),$$

$$p_0 = 0 \Rightarrow p_1 = 0 \text{ (} Q_0\text{-a.s.) and}$$

$$p_1 = 0 \Rightarrow p_0 = 0 \text{ (} Q_1\text{-a.s.);}$$

$$0 < \alpha < 1: D_\alpha(P_0 \| Q_0) = D_\alpha(P_1 \| Q_1) \text{ and}$$

$$p_0 q_1 = p_1 q_0 \text{ (}\mu\text{-a.s.)};$$

$$\alpha = 1: p_0 q_1 = p_1 q_0 \text{ (}\mu\text{-a.s.)}$$

阶数大于1，其凸性不再成立



- Square Hellinger Distance

$$\text{Hel}^2(P, Q) = \sum_{i=1}^n (p_i^{1/2} - q_i^{1/2})^2$$

$\chi^2$ -divergence

$$\chi^2(P, Q) = \sum_{i=1}^n \frac{(p_i - q_i)^2}{q_i}$$

**Total Variation Distance:**  $V(P, Q) = \sum_{i=1}^n |p_i - q_i|$

基本关系式:

$$\text{Hel}^2(P, Q) \leq D_{1/2}(P\|Q) \leq D_1(P\|Q) \leq D_2(P\|Q) \leq \chi^2(P, Q).$$

**Pinsker 不等式:**

$$\frac{\alpha}{2} V^2(P, Q) \leq D_\alpha(P\|Q) \quad \text{for } \alpha \in (0, 1].$$

恒等式:

$$D_2(P\|Q) = \ln(1 + \chi^2(P, Q)) \quad D_{1/2}(P\|Q) = -2 \ln \left( 1 - \frac{\text{Hel}^2(P, Q)}{2} \right)$$

与统计量的关系:

$$\lim_{\theta' \rightarrow \theta} \frac{1}{(\theta - \theta')^2} D_{\alpha}(P_{\theta} \| P_{\theta'}) = \frac{\alpha}{2} J(\theta) \quad \text{for } \alpha \in (0, \infty).$$

$J(\theta) = \mathbb{E} \left[ \left( \frac{d}{d\theta} \ln p_{\theta} \right)^2 \right]$  denotes the *Fisher information*

KL Divergence:

$$\lim_{\theta' \rightarrow \theta} \frac{1}{(\theta - \theta')^2} D(P_{\theta} \| P_{\theta'}) = \frac{1}{2} J(\theta).$$

推广: 从正阶数推广到负阶数

**Lemma 10** (Skew Symmetry). For any  $\alpha \in (-\infty, \infty)$ ,  $\alpha \notin \{0, 1\}$

$$D_{\alpha}(P \| Q) = \frac{\alpha}{1 - \alpha} D_{1-\alpha}(Q \| P). \quad (82)$$

- 如何应用信息熵进行数据推理？ 具体涉及到哪些关于熵的数学公式？
- 如何推广Renyi 熵理论？
- Renyi 熵理论可能的应用场景是什么？ 可解决什么问题？ 哪些问题难以解决？
- 信息表示理论和信道编码理论存在哪些共同点？ 哪些差异点
- 信息压缩理论和信息表示理论的关系是什么？
- 连续与离散在熵理论处理上的最大差异表现在哪些方面？