EE142
Fundamentals of Information Theory

Lecture notes

Zhou Shouchen

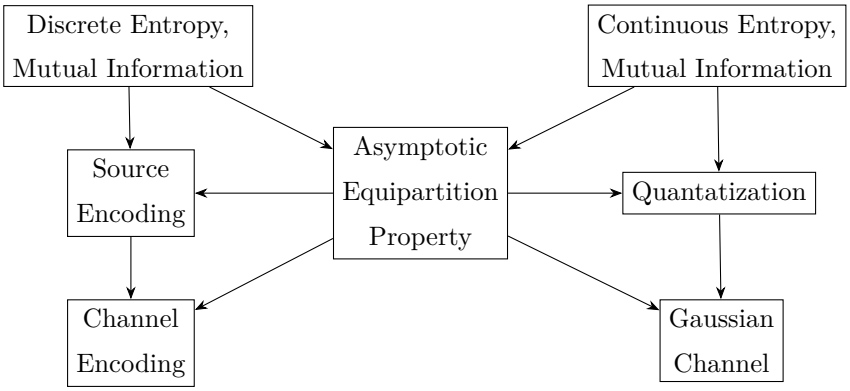
2021533042

LaTeX source code on GitHub

2025 年 1 月 11 日

目录

第一章　Summary



A.E.P. 将所有的章节连接起来.

主要围绕香农三大定理展开: (注意成立的条件! 所有事情要满足其条件才行!)

1. Lossless Source Coding Theorem (前提: $x^n \overset{i.i.d.}{\sim} p(x)$)

$$L^* = H(X)$$

2. Channel Coding Theorem (前提: Memoryless Channel)

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

3. Rate Distortion Theorem (前提: $x^n \overset{i.i.d.}{\sim} p(x)$)

$$R(D) = \min_{p(x),\mathbb{E}(X)\leq p^2} I(X;\hat{X})$$

Some interesting topics but not covered in class: §11.3 Universal Source Coding, §11.10 Fisher Information, §15.10 General Multiterminal Networks.

. . . . . A lot of interesting topics . . . . . .

第二章　Entropy, Relative Entropy, Mutual Information

Book Chapter2. (P39)

$\log x$ 若无特殊说明, 默认为 $\log_2 x$, $0 \log 0 = 0$.

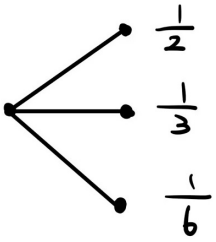
离散型随机变量 $\mathcal{X}$ 看作是有限的, i.e. $|\mathcal{X}| < +\infty$.

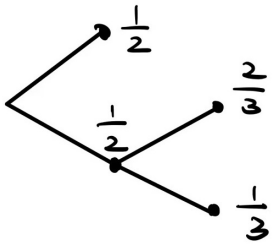
2.1　Entropy

Symmetric functions $H_m(p_1, p_2, \ldots, p_m)$ 满足以下条件:

1. Normalization: $H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$
2. Continuity: $H_2(p, 1-p)$ is a continuous function of p
3. Grouping: $H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$

具有如下图所示的拆分/合并性质:





$$H(X) = H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right)$$

具体简要推导见 Appendix 11.1, 最后能够推导出 $H(X)$ 的唯一形式为

$$H(X) = k \cdot \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$

定义 2.1.1. 事件 x 发生的概率为 $p(x)$, 则 x 的信息量为 $\log \frac{1}{p(x)}$.

定义 2.1.2. 离散型随机变量 X 的熵 *(entropy)* $H(X)$:

$$\begin{aligned} H(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \mathbb{E}\left[\log \frac{1}{p(x)}\right] \end{aligned}$$

$H(X)$ 物理意义: $X = x$ 事件发生的概率 $p(x)$, 信息量为 $\log \frac{1}{p(x)}$.
$H(X)$: 所有事件发生的期望信息量.

命题 2.1.3. $H(X) \geq 0$, 当且仅当 $p(x) = 1$ 时, $H(X) = 0$.
$p(x) = 1$ 时, 事件是确定的 *(deterministic)*, 信息量为 *0*.

定义 2.1.4. $X \in \mathcal{X}, Y \in \mathcal{Y}; |\mathcal{X}|, |\mathcal{Y}| < \infty$*(*离散型随机变量*)*.
X 和 Y 的联合熵 *(joint entropy)* $H(X, Y)$:

$$\begin{aligned} H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)} \\ &= \mathbb{E}\left[\log \frac{1}{p(x, y)}\right] \end{aligned}$$

命题 2.1.5. $0 \leq H(X) \leq \log |\mathcal{X}|$.
X 为冲激函数时取 0 *(deterministic)*, X 为均匀分布时取 $\log |\mathcal{X}|$.
prove 均匀分布时取得最大值 $\log |\mathcal{X}|$:

Let $u(x) = \dfrac{1}{|\mathcal{X}|}$, *then*

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

当且仅当 $p(x) = u(x)$ 时等号成立, *i.e.* $H(X)$ 取得最大值 $\log|\mathcal{X}|$.

定义 2.1.6. 条件熵 *(conditional entropy)* $H(Y|X)$:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(y|x)} \\ &= \mathbb{E}\left[\log\frac{1}{p(y|x)}\right] \end{aligned}$$

定理 2.1.7. *chain rule* 剥洋葱: $H(X,Y) = H(X) + H(Y|X)$

例 2.1.8. *Find* $H(X), H(Y), H(X|Y), H(X,Y)$.

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

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

$$H(Y) = H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = \log 4 = 2 bits \ (uniform\ distribution)$$

$$\begin{aligned} H(X|Y) &= \sum_{y\in\mathcal{Y}} p(y) H(X|Y=y) \\ &= \frac{1}{4}\left(H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + H(1)\right) \\ &= \frac{11}{8} bits \end{aligned}$$

$$H(X,Y) = H(Y) + H(X|Y) = 2 + \frac{11}{8} = \frac{27}{8}\ bits.$$

命题 2.1.9.

$$\begin{aligned} H(X,Y) &= H(X) + H(Y) \quad (X \perp Y) \\ H(X,Y) &= H(X) \qquad\qquad\quad (Y = X) \end{aligned}$$

2.2 Relative Entropy and Mutual Information

概率论衡量两个变量相关程度 (概率论方法):

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \in [-1,1]$$

只能刻画**线性**相关性, 且正负相关程度相同 (正负相关).

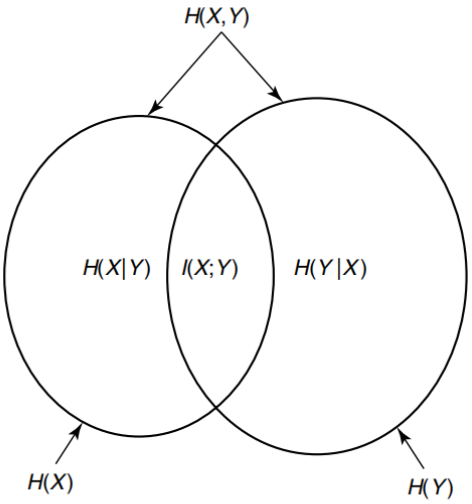
命题 2.2.1. X,Y 独立, 则 $\rho_{X,Y} = 0$. 但是 $\rho_{X,Y} = 0$ **不一定**独立.

Gaussian 分布独立 $\Leftrightarrow$ 不相关.

信息论衡量方法 (用 bit 衡量):

定义 2.2.2. $I(X;Y)$: X,Y 之间的互信息 *(mutual information)*.

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



Relationship between entropy and mutual information.

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

proof:

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)p(y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= \sum_{x,y} p(x,y) \log \frac{1}{p(x)} - \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)} \\
&= \sum_{x} p(x) \log \frac{1}{p(x)} - H(X|Y) \quad (p(x,y) \text{ 对 } y \text{ 求和, 求出 margin distribution } p(x)) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

定义 2.2.3. 两个分布 $p(x), q(x)$ 的相对熵 *Relative Entropy(KL-Divergence):*

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

$x \in \mathcal{X}$: 不考虑两个 *support* 不同的分布.

物理意义: 两个分布之间的距离.

命题 2.2.4. $D\left(p(x)\|q(x)\right) \geq 0.$

proof:

$$\begin{aligned}
-D\left(p(x)\|q(x)\right) &= \sum_x p(x)\log\frac{q(x)}{p(x)} \\
&= \mathbb{E}_{x\sim p(x)}\left[\log\frac{q(x)}{p(x)}\right] \\
&\leq \log\mathbb{E}_{x\sim p(x)}\left[\frac{q(x)}{p(x)}\right] \qquad \textit{(Jensen's Inequality)} \\
&= \log\sum_x p(x)\frac{q(x)}{p(x)} \\
&= 0
\end{aligned}$$

i.e. $D\left(p(x)\|q(x)\right) \geq 0.$

当且仅当 $p(x) = q(x)$ 时等号成立 (*Jensen's Inequality* 成立条件: 函数是线性的).

命题 2.2.5. $I(X;Y) = I(Y;X)$

$D\left(p(x)\|q(x)\right) \neq D\left(q(x)\|p(x)\right)$

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

命题 2.2.6.

$$0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}$$

1. $I(X;Y) \geq 0$: 当且仅当X,Y 独立时 $(X \perp Y)$等号成立.

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

当且仅当 $p(x,y)=p(x)p(y)$ 时等号成立, 即 X,Y 独立.

2. $I(X;Y)\le \min\{H(X),H(Y)\}$:

Since $H(X)\ge 0$, similarly, $H(X|Y)\ge 0$.

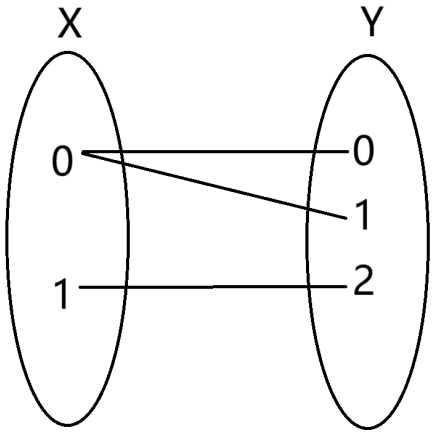
$$I(X;Y)=H(X)-H(X|Y)\le H(X)$$

当且仅当 $H(X|Y)$=0 时等号成立.

同理 $I(X;Y)\le H(Y)$, 当且仅当 $H(Y|X)$=0 时等号成立.

命题 2.2.7. 即使 $H(X|Y)=0$, 也无法得出 X,Y 有关系.

如图 $H(X|Y)=0, H(Y|X)\neq 0$.



命题 2.2.8. *Conditioning Reduces Entropy(Information can't hurt):*

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

proof:

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

当且仅当 $I(X;Y)=0$, 即 X,Y 独立时取等.

将 X, Y 两个分布的性质拓展到 n 个分布:

多元 KL 散度:

$$D\left(p(x,y)\|q(x,y)\right)=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{q(x,y)}$$

$$D\left(p(y|x)\|q(y|x)\right)=\sum_{x,y}p(x,y)\log\frac{p(y|x)}{q(y|x)}$$

无论 KL 散度的形式如何, log 前都是 $p(x,y)$!!!

命题 2.2.9. *1. Chain Rule:*

<1> Entropy's Chain Rule:

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

<2> Mutual Information's Chain Rule:

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

proof:

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

<3> KL-Divergence's Chain Rule:

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

proof:

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

$$\Rightarrow D\left(p(x_1,x_2,\cdots,x_n)\|q(x_1,x_2,\cdots,x_n)\right) = \sum_{i=1}^{n} D\left(p(x_i|x_1,\ldots,x_{i-1})\|q(x_i|x_1,\ldots,x_{i-1})\right)$$

2. Mutual Information ⇒ Conditional Mutual Information:
$I(X;Y|Z)$*: given* Z*,* X*,*Y 的互信息.

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

已知 $H(X) \geq H(X|Y)$, 但是 $I(X;Y)$ 和 $I(X;Y|Z)$ 大小关系不确定.

例 2.2.10. $I(X;Y|Z) > I(X;Y)$
$X, Y \overset{i.i.d.}{\sim} \text{Bern}\left(\frac{1}{2}\right)$, $Z = X + Y$
$X \perp Y \Rightarrow I(X;Y) = 0$

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= H(X|Z) \quad \textit{(X=Z-Y, deterministic)} \\ &= P(Z=0)H(X|Z=0) + P(Z=1)H(X|Z=1) + P(Z=2)H(X|Z=2) \\ &= 0 + \frac{1}{2}H(X|Z=1) + 0 \\ &> 0 \end{aligned}$$

$I(X;Y|Z) > I(X;Y)$

例 2.2.11. *$I(X;Y|Z) \le I(X;Y)$ Construct Markov Chain: $X \to Y \to Z$*

$$\begin{aligned} p(x,y,z) &= p(x)p(y|x)p(z|y) \\ I(X;Y,Z) &= I(Y,Z;X) \\ &= I(Y;X) + I(Z;X|Y) \\ &= I(Z;X) + I(Y;X|Z) \end{aligned}$$

Since $Z \perp X|Y \Rightarrow I(Z;X|Y) = 0$
So $I(Y;X) = I(Z;X) + I(Y;X|Z) \ge I(Y;X|Z)$
所以 $I(X;Y) \ge I(X;Y|Z)$

proof: $Z \perp X|Y \Rightarrow I(Z;X|Y) = 0$:

$$\begin{aligned} I(Z;X|Y) &= H(Z|Y) - H(Z|X,Y) \\ H(Z|X,Y) &= \sum_{x,y,z} p(x,y,z) \log \frac{1}{p(z|x,y)} \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{1}{p(z|y)} = \sum_{y,z} p(y,z) \log \frac{1}{p(z|y)} \\ &= H(Z|Y) \\ \Rightarrow I(Z;X|Y) &= 0 \end{aligned}$$

命题 2.2.12. $I(X;Y_1,\cdots,Y_n) = \sum_{i=1}^{n} I(X;Y_i|Y_1,\cdots,Y_{i-1})$
$I(X_1,\cdots,X_n;Y_1,\cdots,Y_n)$*:* 将上面的 X 替换成 $X_1,\cdots,X_n$*,* 然后再展开*.*
Specifically: When $(X_1,Y_1) \perp (X_2,Y_2) \perp \cdots \perp (X_n,Y_n)$, then

$$I(X_1,\cdots,X_n;Y_1,\cdots,Y_n) = \sum_{i=1}^{n} I(X_i;Y_i)$$

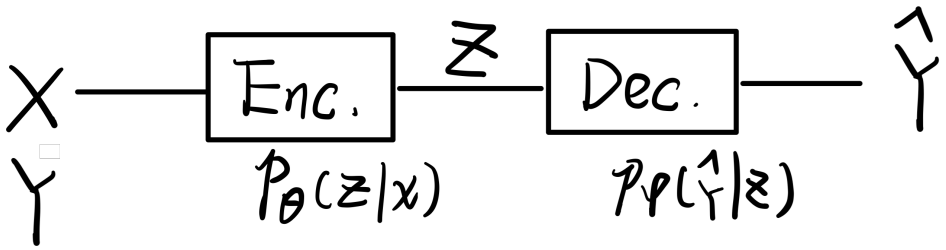
例 2.2.13. 我们知道 $H(X) \ge H(X|Y)$*,* 但是 $H(X)$ 和 $H(X|Y=y)$ 的大小不确定*.*

Y \ X	0	1
0	0	$\frac{3}{4}$
1	$\frac{1}{8}$	$\frac{1}{8}$

$$H(X) = H\left(\frac{1}{8}, \frac{7}{8}\right) \in (0, \log|\mathcal{X}|) = (0, 1)$$
$$H(X|Y=0) = H(0,1) = 0$$
$$H(X|Y=1) = H\left(\frac{1}{2}, \frac{1}{2}\right) = \log|\mathcal{X}| = 1$$

命题 2.2.14. 从条件熵的角度推导出 *cross entropy.* 如下图所示, 网络满足马尔可夫链*:*

$$Y \to X \to Z \to \hat{Y}$$



$$\min_{p(\hat{y},z)} \quad H(\hat{Y}|Z)$$

使网络确定到正确的答案*.*

$$
\begin{aligned}
H(\hat{Y}|Z) &= \sum_{\hat{y},z} p(\hat{y},z) \log \frac{1}{p(\hat{y}|z)} \\
&= \sum_{x,y,z,\hat{y}} p(x,y,z,\hat{y}) \log \frac{1}{p(\hat{y}|z)} \\
&= \sum_{x,y} p(x,y) p_\theta(z|x) p_\phi(\hat{y}|z) \log \frac{1}{p(\hat{y}|z)} \quad \textit{(Monte Carlo sample} \rightarrow (x,y)\textit{)}
\end{aligned}
$$

红色部分为网络 *trainable* 的部分，为了使网络确定到正确的答案，需要最小化剩余的部分，*i.e.*

$$\sum_{x,y} p(x,y) \log \frac{1}{p(\hat{y}|z)}$$

Which is the cross entropy between $p(x,y)$ *and* $p(\hat{y}|z)$.

2.3 Inequalities

1. $\ln x \leq x - 1$
2. Convex function <1>. $f(x)$ is convex $\Leftrightarrow \forall x_1, x_2, \lambda \in [0,1]$,

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

<2>. $f''(x) \geq 0 \Rightarrow f(x)$ is convex.

proof: From Taylor's Theorem, $f''(x) \geq 0$:

$$f(y) \geq f(x) + f'(x)(y-x)$$

Let $z = \lambda x + (1-\lambda)y$:

$$1.\ f(x) \geq f(z) + f'(z)(x-z)$$
$$2.\ f(y) \geq f(z) + f'(z)(y-z)$$

$$
\begin{aligned}
\lambda \cdot 1. + (1-\lambda) \cdot 2. \Rightarrow \lambda f(x) + (1-\lambda) f(y) &\geq f(z) + f'(z)(\lambda x + (1-\lambda)y - z) \\
&= f(z) = f(\lambda x + (1-\lambda)y)
\end{aligned}
$$

So $f(x)$ is convex.

3. Jensen's Inequality

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

4. Sum-log Inequality(求和不等式) $\forall a_1, \cdots, a_n, b_1, \cdots, b_n > 0$:

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

With equality if and only if $a_i = \lambda b_i$ for all i, with same λ.

Specifically, when $n = 2$:

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2}$$

proof: Let $f(t) = t \log t$, then $f''(t) = \frac{1}{t} \geq 0$. Let $p_i = \frac{b_i}{\sum_{i=1}^{n} b_i}$,$t_i = \frac{a_i}{b_i}$.

$$\sum_{i=1}^{n} p_i f(t_i) \geq f\left(\sum_{i=1}^{n} p_i t_i\right)$$

$$\sum_{i=1}^{n} \left[\frac{b_i}{\sum_{j=1}^{n} b_j} \left(\frac{a_i}{b_i} \log \frac{a_i}{b_i}\right)\right] \geq \left[\sum_{i=1}^{n} \left(\frac{b_i}{\sum_{j=1}^{n} b_j} \cdot \frac{a_i}{b_i}\right)\right] \log \left[\sum_{i=1}^{n} \left(\frac{b_i}{\sum_{j=1}^{n} b_j} \cdot \frac{a_i}{b_i}\right)\right]$$

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

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

命题 2.3.1. *$H(\vec{p})$ is concave in $\vec{p}$.*

$\forall \vec{p}_1, \vec{p}_2, \lambda \in [0,1]$:

$$\begin{aligned} H(\lambda\vec{p}_1 + (1-\lambda)\vec{p}_2) &= -\sum_{x\in\mathcal{X}} [\lambda p_1(x) + (1-\lambda)p_2(x)] \log [\lambda p_1(x) + (1-\lambda)p_2(x)] \\ &= -\sum_{x\in\mathcal{X}} [\lambda p_1(x) + (1-\lambda)p_2(x)] \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda + (1-\lambda)} \\ &= -\sum_{x\in\mathcal{X}} [a_1(x) + a_2(x)] \log \frac{a_1(x) + a_2(x)}{b_1 + b_2} \end{aligned}$$

Where $a_1(x) = \lambda p_1(x)$, $a_2(x) = (1-\lambda)p_2(x)$, $b_1 = \lambda$, $b_2 = 1-\lambda$, since $a_1(x), b_1, a_2(x), b_2 > 0$, and the log-sum Inequality:

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2}$$

So

$$[a_1(x) + a_2(x)] \log \frac{a_1(x) + a_2(x)}{b_1 + b_2} \leq a_1(x) \log \frac{a_1(x)}{b_1} + a_2(x) \log \frac{a_2(x)}{b_2}$$

i.e.

$$\begin{aligned} H(\lambda\vec{p}_1 + (1-\lambda)\vec{p}_2) &\geq -\sum_{x\in\mathcal{X}} \left[\lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{1-\lambda} \right] \\ &= \lambda \left(-\sum_{x\in\mathcal{X}} p_1(x) \log p_1(x) \right) + (1-\lambda) \left(-\sum_{x\in\mathcal{X}} p_2(x) \log p_2(x) \right) \\ &= \lambda H(\vec{p}_1) + (1-\lambda) H(\vec{p}_2) \end{aligned}$$

So $H(\vec{p})$ is concave in $\vec{p}$.

命题 2.3.2.

1. *$I(X;Y)$ is the function of $p(x,y)$, but it is non-convex and non-concave in $p(x,y)$.*
2. *If $p(y|x)$ is fixed(given), $I(X;Y)$ is concave in $p(x)$.*
3. *$X \sim p(x)$, $H(X)$ is concave in $p(x)$.*

4. If $p(x)$ is fixed(given), $I(X;Y)$ is convex in $p(y|x)$.

Lemma: If $f(x)$ is convex, $g(x)$ is an increasing linear function, then $f(g(x))$ is convex.

2. If $p(y|x)$ is fixed(given), $I(X;Y)$ is concave in $p(x)$.

proof:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= \sum_y \underbrace{p(y)\log\frac{1}{p(y)}}_{x\log x \text{ is convex}} - \underbrace{\sum_x p(x)H(Y|X=x)}_{\text{linear to } p(x)} \end{aligned}$$

由于 $p(y) = \sum_x p(x)p(y|x)$, 所以 $p(y)$ 是关于 $p(x)$ 的线性函数, 所以 $H(Y)$ 关于 $p(y)$ 是 concave, 关于 $p(x)$ 也是 concave 的.

So $I(X;Y)$ is concave in $p(x)$.

3. $X \sim p(x)$, $H(X)$ is concave in $p(x)$.

Define Θ: Time sharing random variable.

$$\Theta = \begin{cases} 1, & \text{w.p. } \lambda \\ 2, & \text{w.p. } 1-\lambda \end{cases}$$

Let $Z = X_\Theta$, where $X_1 \sim p_1(x)$, $X_2 \sim p_2(x)$, so $Z \sim \lambda p_1(x) + (1-\lambda)p_2(x)$.

And since that $H(Z) \geq H(Z|\Theta)$, so we have

$$\begin{aligned} H(Z) &\geq H(Z|\Theta) \\ &= p(\Theta=1)H(Z|\Theta=1) + p(\Theta=2)H(Z|\Theta=2) \\ &= \lambda H(X_1) + (1-\lambda)H(X_2) \end{aligned}$$

i.e.

$$H(p_\lambda) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$$

So $H(X)$ is concave in $p(x)$.

4. If $p(x)$ is fixed(given), $I(X;Y)$ is convex in $p(y|x)$.

define: 3 个概率分布: $p_1(y|x), p_2(y|x), p_\lambda(y|x)$, where

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

$$I_i(X;Y) = \sum_{x,y} p_i(x,y) \log \frac{p_i(x,y)}{p_i(x)p_i(y)}, \quad i = 1,2$$

$$I_\lambda(X;Y) = \sum_{x,y} p_\lambda(x,y) \log \frac{p_\lambda(x,y)}{p_\lambda(x)p_\lambda(y)}$$

From log-sum Inequality with $n = 2$:

$$(a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2} \leq a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}$$

$$\begin{aligned}
I_\lambda(X;Y) &= \sum_{x,y} p_\lambda(x,y) \log \frac{p_\lambda(x,y)}{p_\lambda(x)p_\lambda(y)} \\
&= \sum_{x,y} p_\lambda(y|x)p_\lambda(x) \log \frac{p_\lambda(y|x)}{p_\lambda(y)} \\
&= \sum_x p(x) \sum_y \left(\underbrace{\lambda p_1(y|x)}_{a_1} + \underbrace{(1-\lambda)p_2(y|x)}_{a_2} \right) \log \frac{\lambda p_1(y|x) + (1-\lambda)p_2(y|x)}{\underbrace{\lambda p_1(y)}_{b_1} + \underbrace{(1-\lambda)p_2(y)}_{b_2}} \\
&\leq \sum_x p(x) \sum_y \left[\lambda p_1(y|x) \log \frac{\lambda p_1(y|x)}{\lambda p_1(y)} + (1-\lambda)p_2(y|x) \log \frac{(1-\lambda)p_2(y|x)}{(1-\lambda)p_2(y)} \right] \\
&= \lambda \sum_x p(x) \sum_y p_1(y|x) \log \frac{p_1(y|x)}{p_1(y)} + (1-\lambda) \sum_x p(x) \sum_y p_2(y|x) \log \frac{p_2(y|x)}{p_2(y)} \\
&= \lambda I_1(X;Y) + (1-\lambda) I_2(X;Y)
\end{aligned}$$

So we have proved that $I(X;Y)$ is convex in $p(y|x)$.

命题 2.3.3. *Markov Chain:* $X \to Y \to Z$, 看作是双向的, *i.e.* 这等价于 $X \leftrightarrow Y \leftrightarrow Z$.

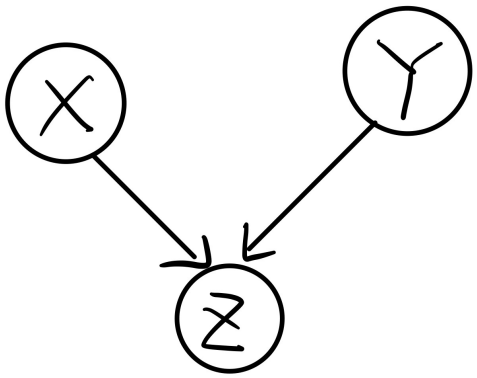
$$X \perp Z|Y$$

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

$$H(Z|XY) = H(Z|Y)$$

命题 2.3.4. *Independent $\not\Rightarrow$ Conditional Independent.*

$$X \perp Y; X \not\perp Y|Z$$



命题 2.3.5. *If $Z = g(Y)$, then $X \to Y \to Z$.*

命题 2.3.6. *Data Processing Inequality(数据处理不等式): For a Markov Chain*

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

$$X \to Y \to Z \to M, I(X;M) \leq I(Y;Z)$$

物理意义: 数据在不断处理的过程中, 相关性不断下降.

处理后的相关性 $\leq$ 处理前的相关性.

传递信息的过程中, 最好的情况是信息量不减少, 后面的数据不会有办法与更往前的数据有更强的相关性.

proof:

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

Since $I(X;Z|Y) = 0$, so $I(X;Y) = I(X;Z) + I(X;Y|Z) \geq I(X;Z)$. i.e.

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

命题 2.3.7. *With Markov Chain* $X \to Y \to Z$

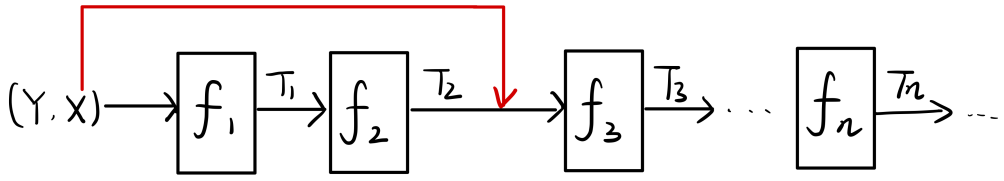
$$I(X;Y) \geq I(X;Y|Z)$$

In general, 这两个互信息无法进行比较, 但是满足以上 *Markov Chain* 时成立*!*

由上个定理的证明*:* $I(X;Y) = I(X;Z) + I(X;Y|Z)$, 由于 $I(X;Z) \geq 0$*:*

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

物理意义*:* X, Z 的互信息不能完全表达出 X, Y 的互信息, 剩余的部分由 $X, Y|Z$ 的互信息来表达.



不加红色的跳层连接: $X \to T_1 \to T_2 \to \cdots \to T_n$

加红色的跳层连接: $(X, T_2) \to T_3 \to \cdots$

在没有保障的情况下, T_n 与 Y 的互信息随 n 的增加不断减少. 为了保障 T_n 与 Y 的互信息: sufficient statistics(充分统计量), 保留足够的信息.

最极端的情况: 保存的信息 s.t. $\hat{Y} = Y$, 保留了全部信息.

2.4 Sufficient Statistics

Background: $X_1, \cdots, X_n \sim \{f_\theta(X)\} = \{\mathcal{N}(\theta, 1)\}$(a family of distribution, 一族元素), try to estimate the unknown parameter θs with the samples $X_1, \cdots, X_n$.

用 MLE 估计 θ 时, 我们知道:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n}$$

所以拥有全部样本 $X_1, \cdots, X_n$, 以及 $T(X_1, \cdots, X_n) = \dfrac{\sum_{i=1}^{n} X_i}{n}$, 对于预测 θ 的效果是相同的 $\Rightarrow$ 大大减少了数据储存量.

两者对参数的估计效果相同: $T(X_1, \cdots, X_n)$ 对变量的操作没有信息损失!

定义 2.4.1. *$T(X_1, \cdots, X_n)$ is the sufficient statistic(s.s.), if the Markov chain holds:*

$$\theta \leftrightarrow T\left(X_1, \cdots, X_n\right) \leftrightarrow \left(X_1, \cdots, X_n\right)$$

由于 $X \leftrightarrow Y \leftrightarrow g(Y)$ 天然成立, 所以

$$\theta \leftrightarrow \left(X_1, \cdots, X_n\right) \leftrightarrow T\left(X_1, \cdots, X_n\right)$$

From the data processing Inequality, we have:

$$\begin{aligned} I\left(\theta; T\left(X_1, \cdots, X_n\right)\right) &\geq I\left(\theta; \left(X_1, \cdots, X_n\right)\right) \\ I\left(\theta; \left(X_1, \cdots, X_n\right)\right) &\geq I\left(\theta; T\left(X_1, \cdots, X_n\right)\right) \\ I\left(\theta; \left(X_1, \cdots, X_n\right)\right) &= I\left(\theta; T\left(X_1, \cdots, X_n\right)\right) \end{aligned}$$

例 2.4.2. *$X_1, \cdots, X_n \overset{i.i.d.}{\sim} Bern(\theta)$, n is known. The sufficient statistic is*

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

prove:

$$P\left[(X_1, \cdots, X_n) = (x_1, \cdots, x_n) \middle| \sum_{i=1}^{n} X_i = k\right] = \frac{1}{\binom{n}{k}}$$

与 θ 无关, 所以 $T(X_1, \cdots, X_n)$ 是充分统计量.

例 2.4.3. $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$.

$$\mu \leftrightarrow \text{样本均值} \leftrightarrow (X_1, \cdots, X_n)$$

$$\sigma \leftrightarrow \text{样本方差} \leftrightarrow (X_1, \cdots, X_n)$$

例 2.4.4. $f_\theta = Unif(\theta, \theta+1)$

$T(X_1, \cdots, X_n) = \{\min\{X_1, \cdots, X_n\}, \max\{X_1, \cdots, X_n\}\}$

simply prove: Since $X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{Unif}(\theta, \theta+1)$*, so the PDF:*

$$f(x_i \mid \theta) = \mathbb{I}_{\theta \le x_i \le \theta+1}$$

The joint distribution is:

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \mathbb{I}_{\theta \le \min\{x_1,\ldots,x_n\} \,\&\, \max\{x_1,\ldots,x_n\} \le \theta+1}$$

when $\theta \le \min\{x_1, \ldots, x_n\}, \{x_1, \ldots, x_n\} \le \theta + 1$, $g(T(x_1, \ldots, x_n), \theta) = 1$.
So $T(X) = \{\min\{X_1, \ldots, X_n\}, \max\{X_1, \ldots, X_n\}\}$ *is a sufficient statistic.*

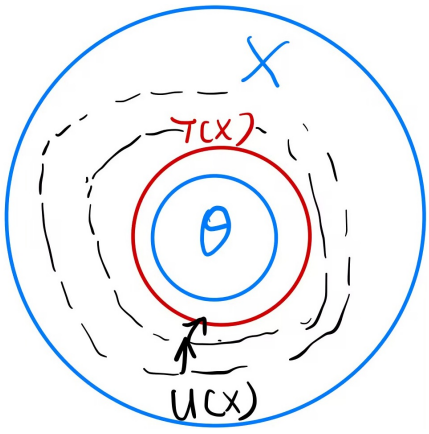
命题 2.4.5. 充分统计量可能不唯一. *e.g.* $\forall k$

$$\theta \leftrightarrow X \leftrightarrow X + k$$

$$\theta \leftrightarrow X + k \leftrightarrow X$$

定义 2.4.6. 所以引入最小充分统计量 *(minimal sufficient statistic): $T(X)$ is the minimal sufficient statistic, if for any other sufficient statistic $U(X)$ has:*

$$\theta \leftrightarrow T(X) \leftrightarrow U(X) \leftrightarrow X$$



命题 2.4.7. *$T(X), U(X)$ are the sufficient statistic, so*

$$I(\theta; T(X)) = I(\theta; X)$$
$$I(\theta; U(X)) = I(\theta; X)$$

$T(X)$ is the minimal sufficient statistic, so from data processing inequality:

$$I(X; T(X)) \leq I(X; U(X))$$

understanding: 从上面的关系图来看, $T(X)$ 是对 $U(X)$ 进行不断提纯, 仅可能的去掉和 X 相关的信息, 只保留和 θ 相关的信息.

e.g. 图像分类, 最理想的情况下, (X, Y) 的最小充分统计量是 $T(X) = Y$.

第三章　Entropy Rates of Stochastic Processes*

Book Chapter4. (P97)

3.1 Stochastic Processes

Random variable $X \to$ Random process $X(t)$.
每一个时刻 t 都是一个随机变量 $X(t)$.

$$\begin{aligned} \text{Mean:} \qquad & \mu_X(t) = \mathbb{E}[X(t)] \\ \text{Autocorrelation (自相关):} \quad & R_X(t_1, t_2) = \mathbb{E}[X(t_1)X^*(t_2)] \\ \text{Cross-correlation(互相关):} \quad & R_{XY}(t_1, t_2) = \mathbb{E}[X(t_1)Y^*(t_2)] \end{aligned}$$

$$\begin{aligned} R_X(t_2, t_1) &= R_X^*(t_1, t_2) \\ R_{XY}(t_1, t_2) &= R_{YX}^*(t_2, t_1) \end{aligned}$$

定义 3.1.1. *Stationary Process(平稳过程): for all time shift τ, time duration k, sample time $t_1, \ldots, t_k$:*

$$F_X(x_{t_1+\tau}, \ldots, x_{t_k+\tau}) = F_X(x_{t_1}, \ldots, x_{t_k})$$

PDF 或 CDF 不随时间变化而变化. (Imply constant mean and variance if exist).

定义 3.1.2. *Wide-Sense Stationary Process(WSS)* 宽平稳过程*:*

1. $\mu_X(t)$ *is constant.*

2. $R_X(t_1, t_2) = R_X(t_1 + \tau, t_2 + \tau)$*, where* $\tau = t_2 - t_1$*.*

Jointly WSS: $R_{XY}(t_1, t_2) = R_{XY}(\tau)$*,* **只和时间差 *(duration)*$\tau$ 有关.**

3.2 Entropy Rates of Stochastic Processes

定义 3.2.1. *The* ***entropy rate*** *of a stochastic process* $\{X_i\}$ *is defined as*

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

随机过程 $X(t)$ 的不确定度为 $H(X(t))$, 整个随机过程的不确定度为 $H(\mathcal{X})$. $H(\mathcal{X})$ 平衡整个过程中每个变量的不确定度 (信息量), 极限可能不存在 (e.g. 趋于极限时 $H(\mathcal{X})$ 可能在振荡, 而不是收敛到具体值). 但 $H(\mathcal{X})$ 一定有界:

$$0 \le H(X_i) \le \log|\mathcal{X}| \Rightarrow 0 \le H(\mathcal{X}) \le \log|\mathcal{X}|$$

例 3.2.2. *1.* $X_1, X_2, \cdots, X_n$ *are i.i.d., then:*

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) = H(X_1) = \ldots = H(X_n)$$

2. $X_1 \perp X_2 \perp \ldots \perp X_n$*, then:*

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i)$$

定义 3.2.3. *Define a related quantity for entropy rate:*

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \cdots, X_1)$$

定理 3.2.4. *For a stationary stochastic process* $\{X(t)\}$*,* $H(\mathcal{X})$ *limit exists:*

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \cdots, X_1)$$

例 3.2.5. *For a stationary Markov chain:* $X_1 \to X_2 \to \cdots \to X_n$ ***(记忆为 1!)***

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, \ldots, X_1) = H(X_n | X_{n-1}) = \ldots = H(X_2 | X_1)$$

第四章　Data Compression

Book Chapter5. (P129)

4.1　Codes

定义 4.1.1. *Discrete Memoryless Source (DMS)* 离散无记忆信源

1. 压缩后有唯一解码

2. 无损压缩

3. 压缩后用尽可能少的比特表示

Let $c(x)$ denote the codeword for x, $l(x)$ denote the length of $c(x)$, $p_X(x)$ denote the probability of x.

假设所有的符号 $X_i \overset{i.i.d.}{\sim} p_X(x)$, 则平均码长为

$$\mathbb{E}[L] = \overline{L} = \sum_x p_X(x) l(x)$$

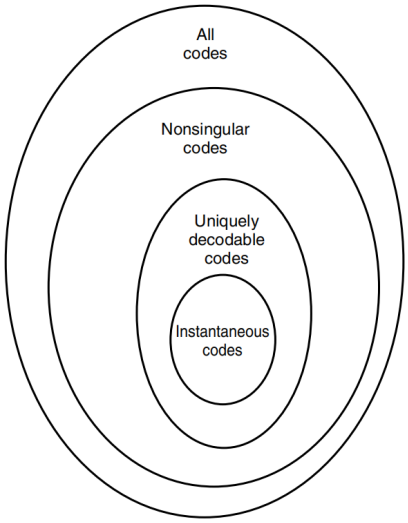
Classes of codes:

1. Nonsingular codes: 无歧义解码

$$x \neq x' \Rightarrow c(x) \neq c(x')$$

2. uniquely decodable codes: 有唯一解码

3. Instantaneous codes / Prefix codes(前缀码): 无 comma 可以及时解码
   e.g. codebook=$\{11, 110\}$, 解码时遇到"11" 无法及时解码

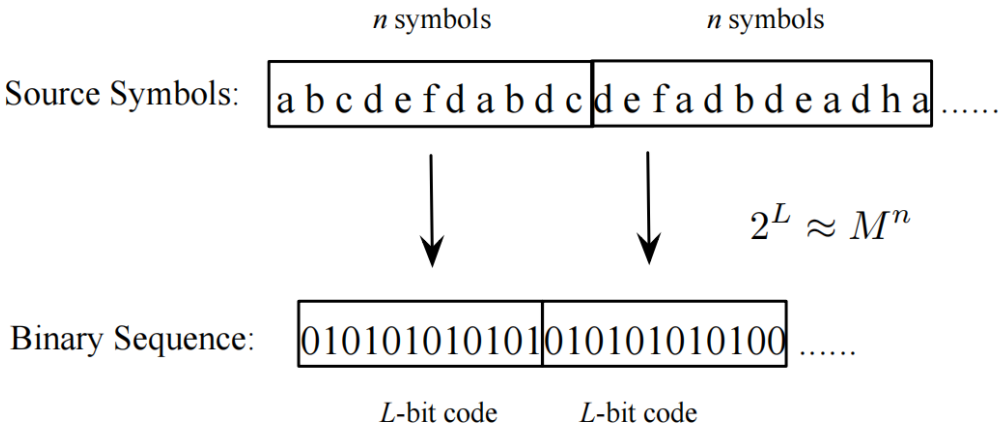


4. Fixed length codes: 定长码

   Alphabet size: M, code length: L, 则 $2^L \geq M$ 以表示所有的符号, i.e.

$$L = \lceil \log M \rceil \Rightarrow \log M \leq L < \log M + 1$$

   Code-blocking: 将 n 个 symbol 组合成一个 block



$$L = \lceil \log M^n \rceil, \overline{L} = \frac{L}{n} \Rightarrow \log M \leq L < \log M + \frac{1}{n}$$

   当 n 足够大是, 保证平均码长最优 $\overline{L} \to \log M$, 但是空间 (码本大小) 指数级增加: M^n.

   定长码都是非前缀码!

5. Variable length codes: 变长码
   无 comma 时可能会有歧义 (非前缀码可以避免).

4.2 Prefix-free Code

定理 4.2.1. *Kraft Inequality: Every* ***prefix-free code*** *for an alphabet* $\mathcal{X}$ *with lengths* $l(x)$ *satisfies*

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} \le 1$$

当且仅当长度取到最优时等号成立. *(i.e.* $l(x) = \log \dfrac{1}{p_X(x)}$, $l(x)$ 为整数*)*

Proof: 将所有 codebook 中的 01 binary sequence 转成数字:

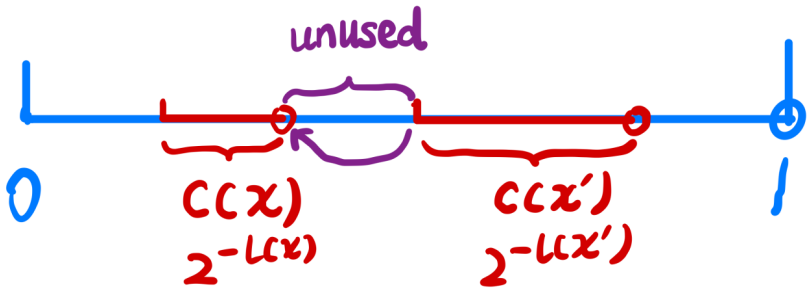
$$(0.c(x)) = (0.y_1 \dots y_l)_2 \to \left(\sum_{m=1}^{l} y_m 2^{-m}\right)_{10} \in [0,1)$$

Prefix-free 可以保证不存在 $(y_1 \dots y_l 000\cdots) \sim (y_1 \dots y_l 111\cdots)$ 之间的所有情况. 这个区间的 code 转换成的数字对应范围为

$$\left[\sum_{m=1}^{l} y_m 2^{-m}, \sum_{m=1}^{l} y_m 2^{-m} + \sum_{i=l+1}^{+\infty} 2^{-i}\right) = \left[\sum_{m=1}^{l} y_m 2^{-m}, \sum_{m=1}^{l} y_m 2^{-m} + 2^{-l}\right)$$

即占据一段长度为 $2^{-l(x)}$ 的左闭右开的区间. prefix-free 的 code 对应的区间必然没有 overlap, 即

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} \le 1$$



Optimal prefix-free code: 考虑优化问题, 先忽略码字长度必须为整数的限制:

$$\begin{aligned} \overline{L}_{\min} &= \min_{l_1,\dots,l_M} \sum_{i=1}^{M} p_i l_i \\ s.t. \qquad & \sum_{i=1}^{M} 2^{-l_i} \le 1 \end{aligned}$$

Lagrange function

$$\begin{aligned} \mathcal{L}(l_1,\dots,l_M,\lambda) &= \sum_{i=1}^{M} p_i l_i + \lambda \left(\sum_{i=1}^{M} 2^{-l_i} - 1 \right) \\ \frac{\partial \mathcal{L}}{\partial l_i} = 0 &\Rightarrow p_i = (\lambda \ln 2)\, 2^{-l_i} \end{aligned}$$

因为我们想要让所有的 l_i 都尽可能小, 所以我们可以取 Kraft 不等式的等号, 并且将所有的 $p_i = (\lambda \ln 2)\, 2^{-l_i}$ 相加可得

$$\begin{aligned} \sum_{i=1}^{M} p_i = 1 = \lambda \ln 2 \sum_{i=1}^{M} 2^{-l_i} = \lambda \ln 2 \\ \Rightarrow \lambda = \frac{1}{\ln 2} \\ \Rightarrow p_i = 2^{-l_i} \end{aligned}$$

所以在忽略码字长度必须为整数的限制时, 最优的码字长度为 $-\log p_i$, 且

$$\overline{L}_{\min} = \sum_x p(x) l(x) = \sum_x p(x) \log \frac{1}{p(x)} = H(X)$$

所以 $H(X)$ 有了新的物理意义: 对 source code 进行压缩之后的最优期望码长.

现在考虑加上整数限制的情况:

定理 4.2.2. *给定 DMS X, 则 minimum expected codeword length for all prefix-free codes is*

$$H(X) \le \overline{L}_{\min} < H(X) + 1$$

证明该定理: 只需要存在一种编码方式使得 $\overline{L}_{\min} < H(X) + 1$ 即可, 但是需要所有的编码方式都有 $H(X) \leq \overline{L}_{\min}$.

1. Shannon code: $l_i = \lceil \log \frac{1}{p_i} \rceil$, we have $\log \frac{1}{p_i} \leq l_i < \log \frac{1}{p_i} + 1$, so

$$\begin{aligned}
\overline{L}_{\min} &= \sum_{i=1}^{M} p_i l_i \\
&= \sum_{i=1}^{M} p_i \lceil \log \frac{1}{p_i} \rceil \\
&< \sum_{i=1}^{M} p_i \left(\log \frac{1}{p_i} + 1 \right) \\
&= \sum_{i=1}^{M} p_i \log \frac{1}{p_i} + \sum_{i=1}^{M} p_i \\
&= H(X) + 1
\end{aligned}$$

2. 考虑所有情况:

$$\begin{aligned}
H(X) - \overline{L}_{\min} &= \sum_{i=1}^{M} p_i \log \frac{1}{p_i} - \sum_{i=1}^{M} p_i l_i = \sum_{i=1}^{M} p_i \log \left(\frac{2^{-l_i}}{p_i} \right) \\
&\leq \log \left(\sum_{i=1}^{M} p_i \frac{2^{-l_i}}{p_i} \right) \qquad \text{(Jensen's inequality, } \log x \text{ is concave)} \\
&= \log \left(\sum_{i=1}^{M} 2^{-l_i} \right) \\
&\leq \log 1 \qquad \text{(Kraft inequality)} \\
&= 0
\end{aligned}$$

取等条件:

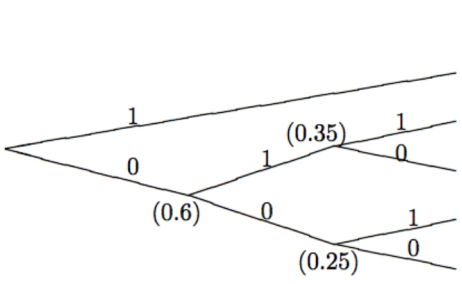
1. Jensen's Inequality: $2^{-l_i} = p_i$ (无法人为控制).
2. Kraft inequality: $\sum_{i=1}^{M} 2^{-l_i} = 1$ (可以通过调整编码方式来尽可能的控制).

block-coding: $l_i = \log \frac{1}{p_i} \Rightarrow l_i = \log \left(\frac{1}{p_i} \right)^n = \frac{1}{n} \log \frac{1}{p_i}$

可以渐进逼近最优 (Jensen's inequality 取等), 但 codebook 复杂度指数级增加.

4.3 Huffman Coding

To make $\bar{L}_{\min}$ small, if $p_i > p_j$, then $l_i \le l_j$. Huffman used this simple idea to construct the code:



p_j	symbol	codeword
0.4	1	1
0.2	2	011
0.15	3	010
0.15	4	001
0.1	5	000

Huffman 编码可能不唯一. e.g. 两个元素有相同的概率 / 合并时给哪个分支 0, 1 均可.

Property: .

Efficiency of the code: $\eta = \dfrac{H(X)}{L} \le 1$.

命题 4.3.1. *Huffman 编码是最优的 (最小的平均码长), 并且满足:*

1. *If $p_i > p_j$, then $l_i \le l_j$.*
2. *Optimal prefix-free code has a full tree.*
3. *The two least probable symbols have the same length:* 拥有最长编码长度的元素一定至少有 2 个 *(否则可以一块剪短一层).*

Huffman Code for encoding a block:

Let $Y = (x_1, \dots, x_n)$: n symbol block.

假设对 Y 使用 Huffman 编码, 则平均码长 $\overline{L}_n$ 满足:

$$H(Y) \le \overline{L}_n < H(Y) + 1$$

因为 $X_1, \dots, X_n \overset{i.i.d.}{\sim} P_X$, 所以

$$H(Y) = H(X^n) = H(X_1, \dots, X_n) = \sum_{i=1}^{n} H(X_i) = nH(X)$$

i.e. 每个 symbol 的平均码长 $\overline{L} = \dfrac{\overline{L}_n}{n}$ 满足:

$$nH(X) \le \overline{L}_n < nH(X) + 1$$
$$\Rightarrow H(X) \le \overline{L} < H(X) + \frac{1}{n}$$

证明 Huffman coding is optimal: 用数学归纳法, 假设有 M 个元素的 Huffman 编码是最优的 $\rightarrow$ 有 $M+1$ 个元素的 Huffman 编码是最优的.

Huffman is optimal for symbol-to-symbol coding with a known input probability distribution.

Proof:

- Huffman algorithm chooses an optimal code tree by starting with two least likely symbols, specifically M and $M-1$.
- Let X' be the reduced RV from X (Combining the two smallest probability symbols)
- Let $\bar{L}'$ be the expected length of X'. Then the optimal L satisfies

$$\bar{L} = \bar{L}' + p_{M-1} + p_M$$

(Extending the codeword $\mathcal{C}'(M-1)$ in to two sibling for $M-1$ and M)

- $\bar{L}_{\min} = \bar{L}'_{\min} + p_{M-1} + p_M$
- Using Huffman algorithm, an optimal code for X' yields an optimal code for X. Prove X'' to X' and so forth, down to a binary symbol.

4.4 Applications of Huffman Code

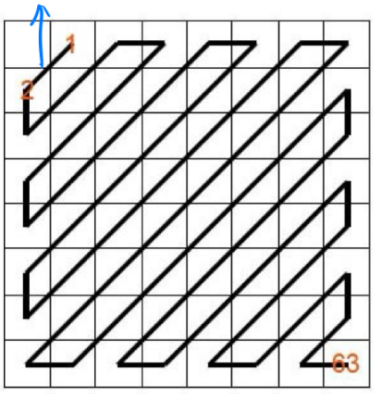
1. JPEG 图像压缩

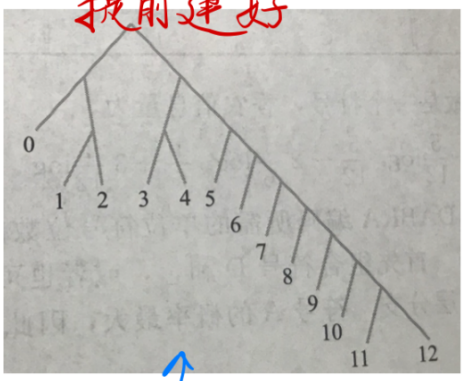
<1>. 离散余弦变换 (Discrete cosine transform DCT).

<2>. 量化 (Quantization). $z = \text{round}\left(\dfrac{y}{q}\right)$, q 为量化步长.

<3>. 对直流分量 (DC) 应用 DPCM Huffman tree(根据先验的现成的树), 交流分量 (AC) 用行程码编码.

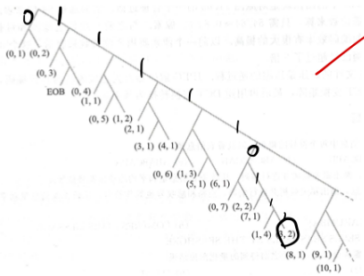
Huffman Code for JPEG

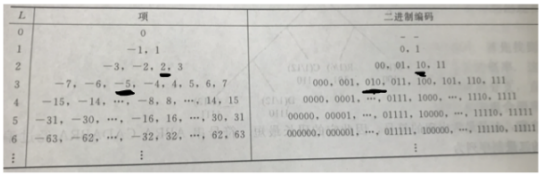




- In each block, the DC part y_{00} uses a DPCM Huffman tree
- Encode the difference between blocks
- the rest 63 AC terms uses run length encoding (RLC) with another Huffman tree and integer tabel

Huffman Code for JPEG





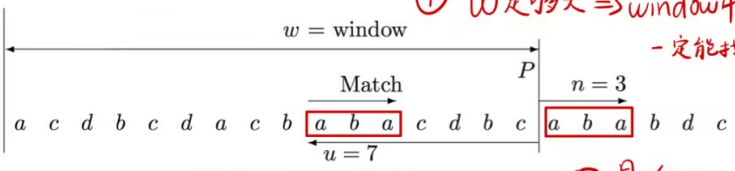
AC sequence -5,0,0,0,2

- present by $(0,3)-5(3,2)EOB$, where (n,L) means after n number of 0, there is a nonzero value with size $L=\lfloor \log_2 |y|+1 \rfloor$
- (0,3) after 0 number of 0, there is a value of size 3, its value is -5
- (0,3) in Huffman tree is 100, -5 in the integer table is 010,
- EOB (end of block) in tree is 1010 item -5,0,0,0,2 is encoded as (100)(010)(11111011)(1010)

Lempel-Ziv Data Compression

找不到 n?

① w足够大 ⇒ window中的seq. ⇒ $A_2^{(n)}$ 一定能找到 n.



② 量化

Set window size w 找不到n⇒扩大w

1. Encode the first w symbols in a fixed length code, without compression
2. Set pointer $P = w$
3. Find the largest $n \geq 2$ such that $x_{P+1}^{P+n} = x_{P+1-u}^{P+n-u}$ for some $u \in [1, w]$. x_{P+1}^{P+n} is encoded by encoding n and u (p. 53)
   - Encode n into a codeword from the unary-binary code
   - Encode $u \leq w$ using fixed-length code of length $\lceil \log w \rceil$
4. Set the pointer P to $P + n$ and go to step (3). Iterate forever

abab × 先前未出现

⇒ P ← P+n

w通常取2^k

Lempel-Ziv Data Compression

Unary-binary code (prefix-free)

n	prefix	base 2 expansion	codeword
1		1	1
2	0	10	010
3	0	11	011
4	00	100	00100
5	00	101	00101
6	00	110	00110
7	00	111	00111
8	000	1000	0001000

Prefix 0的个数 $\lceil \log_2 n \rceil$

2. Lempel-Ziv Data Compression

Lempel-Ziv 算法的思想是利用历史数据的重复性来估计真实的概率分布 $p(x)$, 而不去预测 $q(x)$, 达到自适应最小化.

若数据的真实分布为 $p(x)$, 根据先验得出的概率分布为 $q(x)$, 则使用 $q(x)$ 进行 Huffman coding 与最优解的差距:

$$\begin{aligned}\Delta &= \sum_x p(x)\log\frac{1}{q(x)} - \sum_x p(x)\log\frac{1}{p(x)} \\ &= \sum_x p(x)\log\frac{p(x)}{q(x)} \\ &= D(p\|q)\end{aligned}$$

当窗口 w 足够大时, window 中的 sequence 可以看作是 typical set: $A_\epsilon^{(n)}$, 可以视作 w.p. 1 的可以找得到 n.

所以引出下一章的内容: Asymptotic Equipartition Property(AEP).

第五章　Asymptotic Equipartition Property

Book Chapter3. (P83)

Asymptotic Equipartition Property(AEP): 等分渐进性. 可以看作大数定理在信息论中的体现.

5.1　Law of Large Numbers

几种收敛方式: Given a sequence $X_1, X_2, \ldots, X_n$, it converges to X:

1. In probability(依概率收敛):

$$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0 \Rightarrow \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) < \delta$$

2. In mean square(均方收敛): $\lim_{n\to\infty} \mathbb{E}\left[(X_n - X)^2\right] = 0$.

3. Almost surely / with probability 1: $P\left(\lim_{n\to\infty} X_n = X\right) = 1$.

Law of Large Numbers(LLN): 大数定律. $x_1, \ldots, x_n \overset{i.i.d.}{\sim} p(x)$, then $\frac{1}{n}\sum_{i=1}^{n} x_i$ converges to $\mathbb{E}[x]$.

(1) in probability: Weak LLN(Weak Law of Large Numbers).

(2) w.p. 1: Strong LLN(Strong Law of Large Numbers).

5.2 Typical Sequence, Typical Set

typical sequence 的物理意义:

例 5.2.1. *Suppose* $X \sim Bern\left(\frac{1}{3}\right), x_1, \ldots, x_n \overset{i.i.d.}{\sim} X$. 则下列序列*:*

1. $00\cdots0$ *(*18 个 0*)*

2. $1010\cdots10$ *(*9 个 1, 9 个 0*)*

3. $001001\cdots001$ *(*6 个 1, 12 个 0*)*

4. $1111\cdots11$ *(*18 个 1*)*

直觉上, 序列 3 每个数字的频率最符合预期 *(*与概率相同*)*, 更加典型.
序列 1 虽然生成出来的概率大, 但是有着大概率的序列的种类也更少, 因此未必典型.

定义 5.2.2. $x_1, \cdots, x_n \overset{i.i.d.}{\sim} p(x)$, *then* $(x_1, \ldots, x_n)$ *is* ϵ*-typical sequence if:*

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

当 $n \to +\infty$ 时, 可以让 $\epsilon \to 0$, 即

$$p(x_1, \ldots, x_n) \to 2^{-nH(X)}$$

定理 5.2.3. *Asymototic Equipartition Property(AEP): If* $x_1, \ldots, x_n \overset{i.i.d.}{\sim} p(x)$*:*

$$-\frac{1}{n}\log p(x_1, \ldots, x_n) \to H(X) \qquad \textit{in probability}$$

proof:

$$\begin{aligned}
-\frac{1}{n}\log p(x_1, \ldots, x_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i) \\
&= -\sum_{i=1}^{n}\frac{1}{n}y_i \qquad (\textit{Let } Y = \log p(X)) \\
&\to -\mathbb{E}_{X\sim p(x)}[Y] \qquad \textit{in probability (LLN)} \\
&= -\sum_{x} p(x)\log p(x) \\
&= H(X)
\end{aligned}$$

定义 5.2.4. *$A_\epsilon^{(n)}$ is a typical set: $\forall (x_1,\dots,x_n) \in A_\epsilon^{(n)}$ are ϵ-typical sequences.*

命题 5.2.5. *1. If $(x_1,\dots,x_n) \in A_\epsilon^{(n)}$, then*

$$H(X) - \epsilon \le -\frac{1}{n}\log p(x_1,\dots,x_n) \le H(X) + \epsilon$$

2. $P\left[(x_1,\dots,x_n) \in A_\epsilon^{(n)}\right] > 1-\epsilon$, for $n \to \infty$

3. $(1-\epsilon)2^{n(H(X)-\epsilon)} \le \left|A_\epsilon^{(n)}\right| \le 2^{n(H(X)+\epsilon)}$

当 $n \to +\infty$ 时, 可以让 $\epsilon \to 0$, 此时可以看作 $\left|A_\epsilon^{(n)}\right| = 2^{nH(X)}$, 即有

$$P\left[(x_1,\dots,x_n) \in A_\epsilon^{(n)}\right] = 1 \Rightarrow P(x_1,\dots,x_n) = \frac{1}{\left|A_\epsilon^{(n)}\right|} = 2^{-nH(X)}$$

即每个 typical sequence 均以此概率等概率出现

proof:

1. Typical set 的定义.

2. 由定理 5.2.3(AEP) 可得

$$P\left[\left|-\frac{1}{n}\log p(x_1,\dots,x_n) - H(X)\right| < \epsilon\right] > 1-\delta$$

取 $\delta = \epsilon$ 即可.

3. 合并上下界即可证明 3.

下界: 由 1. 和 2. 结合可得

$$\begin{aligned} 1-\epsilon &\le P\left[(x_1,\dots,x_n) \in A_\epsilon^{(n)}\right] \qquad (\text{由 2. 得}) \\ &= \sum_{(x_1,\dots,x_n)\in A_\epsilon^{(n)}} p(x_1,\dots,x_n) \\ &\le \sum_{(x_1,\dots,x_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \qquad (\text{由 1. 得}) \\ &= \left|A_\epsilon^{(n)}\right| 2^{-n(H(X)-\epsilon)} \end{aligned}$$

上界: 由 1. 可得

$$1 = \sum_{x^n\in\mathcal{X}^n} p(x^n) \ge \sum_{x^n\in A_\epsilon^{(n)}} p(x^n) \ge \left|A_\epsilon^{(n)}\right| 2^{-n(H(X)+\epsilon)}$$

当 $n \to +\infty$ 时, $\epsilon \to 0$, 可以理解为:

1. $-\frac{1}{n}\log p(x^n) = H(X)$
2. $p\left(x^n \in A_\epsilon^{(n)}\right) = 1$
3. $\left|A_\epsilon^{(n)}\right| = 2^{nH(X)}$

5.3 Consequences of the AEP: Data Compression

回到上一章的最后, 我们用 AEP 来解释 data compression.

长度为 n 的 sequence 的总数为 $|\mathcal{X}|^n$, typical set $A_\epsilon^{(n)}$ 的个数为

$$\left|A_\epsilon^{(n)}\right| \le 2^{n(H(X)+\epsilon)} \to 2^{nH(X)}$$

typical sequence 占总 sequence 的比例为

$$\frac{\left|A_\epsilon^{(n)}\right|}{|\mathcal{X}|^n} = \frac{2^{nH(X)}}{2^{\log|\mathcal{X}|^n}} = 2^{n(H(X)-\log|\mathcal{X}|)} \to 0$$

因为 $H(X) < \log|\mathcal{X}|$, 所以当 $n \to +\infty$ 时, typical sequence 占总 sequence 的比例趋近于 0.

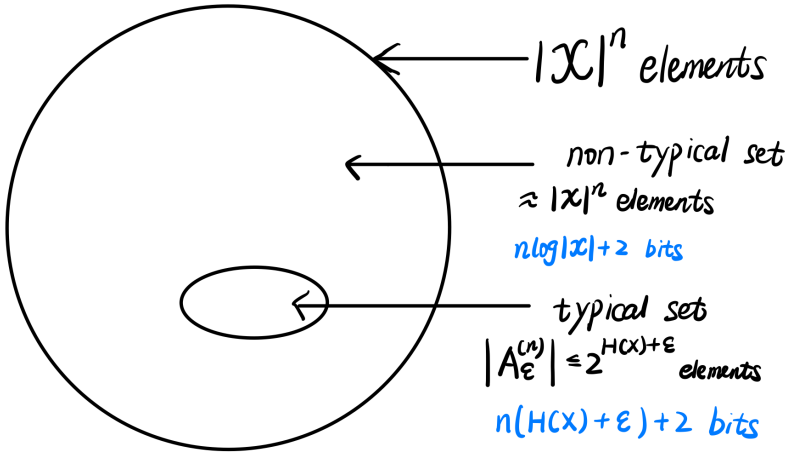
(当 X 为均匀分布时, 此时所有的 sequence 均为 typical sequence, 此时 typical sequence 占总 sequence 的比例为 1, 较为 trival 的情况, 不单独考虑).

这说明 A_ϵ^n 中的序列总数量较少 (占所有序列的比例趋近于 0), 但是整体出现的概率都很高 (任取一个序列时 typical sequence 的概率趋近于 1)

方案 1: 只对 typical set 中的序列进行定长码编码, 不对 non-typical set 中的元素进行编码. 这样的编码方式时有损的, 但是当 $n \to +\infty$ 时, 信息损失趋近于 0(遇到 non-typical sequence 的概率为 0).

定长码的码长为 $L = \left\lceil \log\left|A_\epsilon^{(n)}\right| \right\rceil \to nH(X)+1$. 此时平均码长为

$$\overline{L} = \frac{nH(X)+1}{n} = H(X) + \frac{1}{n} \overset{n\to+\infty}{\to} H(X)$$



方案 2: Asympototically Optimal coding 渐进最优的编码方式:

将 typical set 和 non-typical set 分别用不同长度的定长码的方式进行编码

typical set 中每个序列的编码长度:

$$L_n = \left\lceil \log \left| A_\epsilon^{(n)} \right| \right\rceil \leq \lceil n(H(X)+\epsilon) \rceil < n(H(X)+\epsilon) + 1 + 1$$

Non-typical set 中每个序列的编码长度:

$$L_n^{(c)} = \lceil \log |\mathcal{X}|^n \rceil < n \log |\mathcal{X}| + 1 + 1$$

其中 $+1$ 是用于区分此序列是 typical set 还是 non-typical set.

$$\overline{L} = \frac{(1-\epsilon)\left((H(X)+\epsilon)+2\right) + \epsilon\left(n\log|\mathcal{X}|+2\right)}{n} \to H(X)$$

As $n \to +\infty, \epsilon \to 0$. 其中上面出现 $(1-\epsilon)$ 的原因是 typical set 出现的概率是 $p(A_\epsilon^{(n)}) > 1-\epsilon$.

这些编码方式都是渐进最优 Asympototically Optimal $\neq$ optimal 的 (optimal 的还是 Huffman coding). 当 $n \to +\infty$ 时, 只是在理论分析上渐进最优, 但空间爆炸, encode, decode 费时, 所以实际不会使用.

5.4 Smallest Set*

在最开始的例子中, 我们提到 typical sequence 只取最典型的序列, 而概率最大的序列可能不在其中. 满足符合出现概率大的序列非常少. Smallest set 会加入这些出现概率大的序列. 但其中序列的数量在 $\delta \to 0, \epsilon \to 0$ 时与 $\left|A_\epsilon^{(n)}\right|$ 同阶.

定义 5.4.1. $\forall n = 1, 2, \ldots$, *let* $B_\delta^{(n)} \subset \mathcal{X}^n$ *be the smallest set with*

$$p\left(B_\delta^{(n)}\right) \geq 1 - \delta$$

命题 5.4.2. *Suppose* $x_1, \ldots, x_n \overset{i.i.d.}{\sim} p(x)$, *for* $\delta < \frac{1}{2}$, *and any* $\delta' > 0$, *if* $B_\delta^{(n)}$ *is a smallest set(i.e.* $p\left(B_\delta^{(n)}\right) \geq 1 - \delta$*), then*

$$\frac{1}{n} \log \left|B_\delta^{(n)}\right| > H(X) - \delta' \quad \textit{for } n \textit{ is sufficiently large}$$

Which means that $\left|B_\delta^{(n)}\right| \geq 2^{n(H(X)-\delta')}$ *for* n *is sufficiently large.*

And since $\left|A_\epsilon^{(n)}\right|$ has $2^{n(H(X)\pm\epsilon)}$ elements, so we can say that when $\delta \to 0, \epsilon \to 0$, $\left|B_\delta^{(n)}\right| = \left|A_\epsilon^{(n)}\right| = 2^{nH(X)}$.

5.5 Jointly Typical Sequence, Jointly Typical Set

Book §7.6 (P221)

$$X^n \triangleq (X_1, \ldots, X_n), Y^n \triangleq (Y_1, \ldots, Y_n)$$

定义 5.5.1. 基于 *joint distribution* $P_{X,Y}(x, y)$: $(X_i, Y_i) \overset{i.i.d.}{\sim} P_{X,Y}(x, y)$, 产生的 *sequence* (X^n, Y^n) *belongs the* ϵ*-jointly typical set* $A_\epsilon^{(n)}(P_{X,Y})$ *if*

1. $\left|-\frac{1}{n} \log p(X^n) - H(X)\right| < \epsilon$

2. $\left|-\frac{1}{n}\log p(Y^n) - H(Y)\right| < \epsilon$

3. $\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X,Y)\right| < \epsilon$

1. 2. 说明 X^n, Y^n 分别是 typical sequence, 但是无法说明 (X^n, Y^n) 是 jointly typical sequence.(推不出 3.)

类似于单变量的 AEP, 我们可以推演出 Joint AEP:

命题 5.5.2. *1.* $2^{-n(H(X,Y)+\epsilon)} \le p(x^n, y^n) \le 2^{-n(H(X,Y)-\epsilon)}$

2. $P\left((x^n, y^n) \in A_\epsilon^{(n)}(P_{X,Y})\right) \to 1$ *as* $n \to \infty$.

3. $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \le |A_\epsilon^{(n)}(P_{X,Y})| \le 2^{n(H(X,Y)+\epsilon)}$

4. $P\left[(\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)}(P_{X,Y})\right] \to 2^{-nI(X;Y)}$ *($\tilde{x}, \tilde{y}$ 独立产生)*

proof: 1. 由定义可得.

2. 由 jointly typical set 的定义可得: in probability:

$$-\frac{1}{n}\log p(x^n) \to H(X), \quad -\frac{1}{n}\log p(y^n) \to H(Y), \quad -\frac{1}{n}\log p(x^n, y^n) \to H(X,Y)$$

i.e. $\exists N_1, N_2, N_3 \in \mathbb{N}$, s.t.

$$\Pr\left[\left|-\frac{1}{n}\log p(x^n) - H(X)\right| \ge \epsilon\right] < \delta = \frac{\epsilon}{3} \qquad \text{for } n \ge N_1$$

$$\Pr\left[\left|-\frac{1}{n}\log p(y^n) - H(Y)\right| \ge \epsilon\right] < \delta = \frac{\epsilon}{3} \qquad \text{for } n \ge N_2$$

$$\Pr\left[\left|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| \ge \epsilon\right] < \delta = \frac{\epsilon}{3} \qquad \text{for } n \ge N_3$$

Then let $N = \max\{N_1, N_2, N_3\}$, we have $\forall n \ge N$:

$$p\left(x^n \notin A_\epsilon^{(n)}(p_X)\right) < \frac{\epsilon}{3}, \quad p\left(y^n \notin A_\epsilon^{(n)}(p_Y)\right) < \frac{\epsilon}{3}, \quad p\left(x^n, y^n \notin A_\epsilon^{(n)}(p_{X,Y})\right) < \frac{\epsilon}{3} \Rightarrow$$

$$\begin{aligned} p\left[(x^n, y^n) \in A_\epsilon^{(n)}(P_{X,Y})\right] &= 1 - p\left[(x^n, y^n) \notin A_\epsilon^{(n)}(P_{X,Y})\right] \\ &\ge 1 - p\left(x^n \notin A_\epsilon^{(n)}(P_X)\right) - p\left(y^n \notin A_\epsilon^{(n)}(P_Y)\right) \\ &\quad - p\left(x^n, y^n \notin A_\epsilon^{(n)}(P_{X,Y})\right) \qquad \text{(违反 jointly typical set 任意一条)} \\ &\ge 1 - \frac{\epsilon}{3} \times 3 \\ &= 1 - \epsilon \to 1 \qquad \text{as } n \to \infty \end{aligned}$$

3. 由 2. 可得下界:

$$\begin{aligned} 1-\epsilon &\le p\left[(x^n,y^n)\in A_\epsilon^{(n)}(P_{X,Y})\right] \qquad (\text{由 2. 可得}) \\ &= \sum_{(x^n,y^n)\in A_\epsilon^{(n)}(P_{X,Y})} p(x^n,y^n) \le \left|A_\epsilon^{(n)}(P_{X,Y})\right| 2^{-n(H(X,Y)-\epsilon)} \end{aligned}$$

i.e. $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \le \left|A_\epsilon^{(n)}(P_{X,Y})\right|$

上界:

$$\begin{aligned} 1 &= \sum_{(x^n,y^n)} p(x^n,y^n) \ge \sum_{(x^n,y^n)\in A_\epsilon^{(n)}(P_{X,Y})} p(x^n,y^n) \\ &\ge \left|A_\epsilon^{(n)}(P_{X,Y})\right| 2^{-n(H(X,Y)+\epsilon)} \end{aligned}$$

i.e. $\left|A_\epsilon^{(n)}(P_{X,Y})\right| \le 2^{n(H(X,Y)+\epsilon)}$

结合上下界可得

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \le \left|A_\epsilon^{(n)}(P_{X,Y})\right| \le 2^{n(H(X,Y)+\epsilon)}$$

4. Since $(\tilde{x}^n,\tilde{y}^n) \overset{i.i.d.}{\sim} p(x)p(y)$, we have

$$p\left[(\tilde{x}^n,\tilde{y}^n)\in A_\epsilon^{(n)}(P_{X,Y})\right] = \sum_{(\tilde{x}^n,\tilde{y}^n)\in A_\epsilon^{(n)}(P_{X,Y})} p(\tilde{x})p(\tilde{y}) = \left|A_\epsilon^{(n)}(P_{X,Y})\right| p(\tilde{x})p(\tilde{y})$$

结合 3. $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \le \left|A_\epsilon^{(n)}(P_{X,Y})\right| \le 2^{n(H(X,Y)+\epsilon)}$, 及 $\tilde{x}^n,\tilde{y}^n$ 是 typical sequence $(2^{-n(H(X)+\epsilon)} \le p(\tilde{x}^n) \le 2^{-n(H(X)-\epsilon)})$ 可得

$$p\left[(\tilde{x}^n,\tilde{y}^n)\in A_\epsilon^{(n)}(P_{X,Y})\right] \le 2^{-n((H(X)-\epsilon)+(H(Y)-\epsilon)-(H(X,Y)+\epsilon))} = 2^{-n(I(X;Y)-3\epsilon)}$$

$$p\left[(\tilde{x}^n,\tilde{y}^n)\in A_\epsilon^{(n)}(P_{X,Y})\right] \ge (1-\epsilon)2^{-n((H(X)+\epsilon)+(H(Y)+\epsilon)-(H(X,Y)-\epsilon))} = (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

合并可得

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \le p\left[(\tilde{x}^n,\tilde{y}^n)\in A_\epsilon^{(n)}(P_{X,Y})\right] \le 2^{-n(I(X;Y)-3\epsilon)}$$

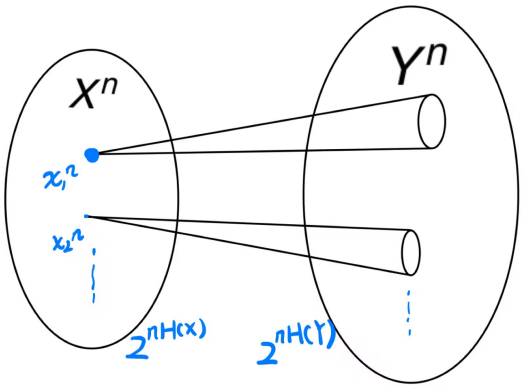
2 个序列基于 joint distribution $p_{X,Y}(x,y)$ 产生, 则这 2 个序列是 jointly typical 的概率趋近于 1.

2 个序列独立产生 (基于 $p_X(x)p_Y(y)$) 产生, 则这 2 个序列是 jointly typical 的概率趋近于 $0(2^{-n(I(X;Y))})$.

例 5.5.3. 若 $(\tilde{x}^n, \tilde{y}^n, \tilde{z}^n)$ 的产生方式为 $(\tilde{x}^n, \tilde{y}^n) \overset{i.i.d.}{\sim} p(x)p(y)$, $\tilde{z}^n \sim p(z^n|y^n)$, 则

$$\begin{aligned} p\left[(\tilde{x}^n, \tilde{y}^n, \tilde{z}^n) \in A_\epsilon^{(n)}(P_{X,Y,Z})\right] &= \sum_{(\tilde{x}^n, \tilde{y}^n, \tilde{z}^n) \in A_\epsilon^{(n)}} p(\tilde{x}^n, \tilde{y}^n) p(\tilde{z}^n|\tilde{y}^n, \tilde{x}^n) \\ &= \left|A_\epsilon^{(n)}(P_{X,Y,Z})\right| p(\tilde{x}^n, \tilde{y}^n) p(\tilde{z}^n|\tilde{y}^n) \\ &\approx 2^{n(H(X,Y,Z))} 2^{-n(H(X,Y))} 2^{-n(H(Z|Y))} \\ &= 2^{-n(I(X;Z|Y))} \end{aligned}$$

命题 5.5.4. $X^n \in A_\epsilon^{(n)}(P_X)$, $Y^n \in A_\epsilon^{(n)}(P_Y)$, 则 X^n 的数量为 $2^{nH(X)}$, Y^n 的数量为 $2^{nH(Y)}$, 两者的交集 *(jointly typical)* 数量为 $2^{nH(X,Y)}$.



抽中一对 (x^n, y^n), 他们是联合典型的概率为 $\dfrac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI(X;Y)}$.

换一个角度: 对于每个 X^n, 大约会有 $\dfrac{2^{nH(X,Y)}}{2^{nH(X)}} = 2^{nH(X|Y)}$ 个 Y^n 是与他联合典型的. 这样子每一个 Y^n 都会有一个 X^n 与他联合典型, 但是对于每一个 X^n, 他会有 $2^{nH(Y|X)}$ 个 Y^n 与他联合典型.

第六章　Channel Capacity

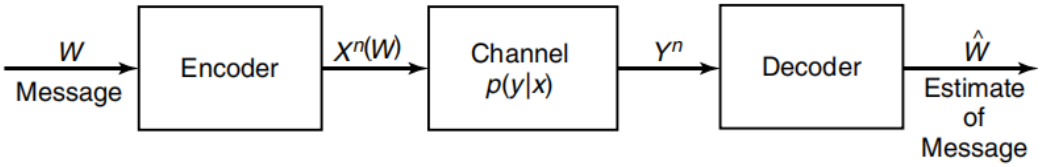
信道容量.　　Book Chapter7. (P209)

Source Encoding(信源编码): 为了减少信源冗余度而进行的信源符号变换. 针对信源输出符号序列的统计特性, 把信源输出符号序列变换为最短的码字序列. 目的是数据压缩, 模数转换.

Channel Encoding(信道编码): 为了对抗信道中的噪音和衰减, 通过增加冗余来提高抗干扰能力和纠错能力, 提高信道传输可靠性.

6.1　Channel Encoding

定义 6.1.1. *DMC(Discrete memoryless channel): Input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$.*



A (M, n) code consists of (here M takes 2^{nR}, M 是 index range, n 是 block length):

1. $W \in [1 : 2^{nR}]$, Uniformly distributed in $[1 : 2^{nR}]$
2. Encoder: $X^n = f(W)$, $[1 : 2^{nR}] \to \mathcal{X}^n$
3. Channel: $Y^n = p(Y^n|X^n)$
4. Decoder: $\hat{W} = g(Y^n), \mathcal{Y}^n \to [1 : 2^{nR}]$

Some definitions:

1. error probability: $\lambda(i) = \Pr\left[\hat{W} \neq i | W = i\right]$
2. $\lambda^{(n)} = \max_{i \in [1, w^{nR}]} \lambda(i)$
3. average error probability: $p_e^{(n)} = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda(i)$

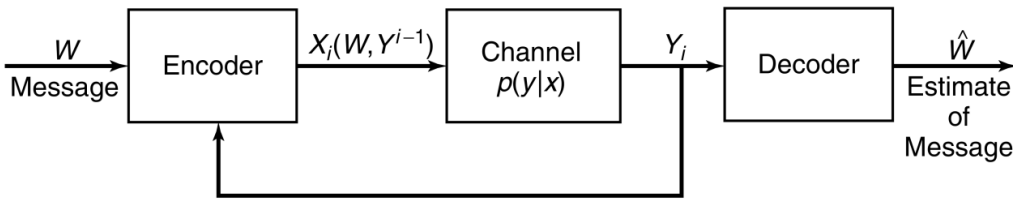
   (和 $p_e^{(n)} \to 0$ 相比, $\lambda(i) \to 0$ constrain 更强)
4. The rate R of an (M, n) code is $R = \frac{\log M}{n}$ bits per transmission

   当 $M = 2^{nR}$ 时理解: 信道编码把信源 index(W) 映射到 X^n, 用 n 次传输传送 W. 因为 W 从 $[1, \ldots, M]$ 中均匀采样, 在无损传输的条件下 ($p_e^{(n)} \to 0$), 平均每次用信道发送的 bit 数为 $\frac{H(W)}{n} = \frac{\log M}{n} = \frac{\log 2^{nR}}{n} = R$.
5. R is achievable if $\exists (n, R)$, s.t. $p_e^{(n)} \to 0$.
6. The capacity of a channel C is the maximum(supremum) of all achievable rate R.
7. Codebook: $\mathcal{C} = \left\{X^n(1), X^n(2) \ldots, X^n(2^{nR})\right\}$ ($X^n(i)$: 长度为 n 的第 i 个序列)

Memoryless: distribution of outputs depends only on the input of that time. 当前输出只与当前输入有关

$$p(y_k | x_k, y_{k-1}) = p(y_k | x_k) \Rightarrow p(y^n | x^n) = \prod_{k=1}^{n} p(y_k | x_k)$$

常数与任何变量都独立, X is deterministic, $I(X; Y) = 0$.



DMC with feedback 和没有 feedback 相比, 不会增加信道容量

$$p(y^n | x^n) = \prod_{k=1}^{n} p(y_k | x^n, y_1, \ldots, y_{k-1})$$

$$X_i = f(W, Y_1, \ldots, Y_{i-1})$$

6.2 Channel Capacity

The 'information' channel capacity of a DMC is defined as:

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

其中 $p(y|x)$ 是已知的, $p(x)$ 是变量.

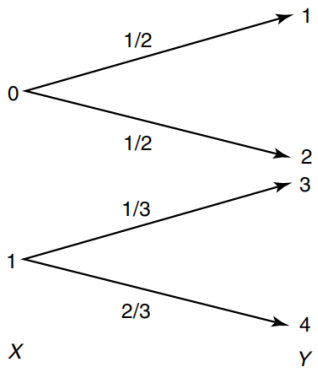
在第 1 章 Mutual Information 中, 我们已经证明了 $H(X)$ 是关于 $p(x)$ 的 concave function, 由于

$$I(X;Y) = H(Y) - H(Y|X) = -\sum_y p(y)\log p(y) - \sum_x p(x)H(Y|X=x)$$

其中 $p(y) = \sum_x p(x)p(y|x)$, 所以 $p(y)$ 是 $p(x)$ 的线性函数, 所以 $I(X;Y)$ 是 $p(x)$ 的 concave function.

$I(X;Y) = f(p(x))$ is a concave function of $p(x)$!

例 6.2.1. *Noisy channel with non-overlapping outputs.* $C = 1$ *bit per transmission.*

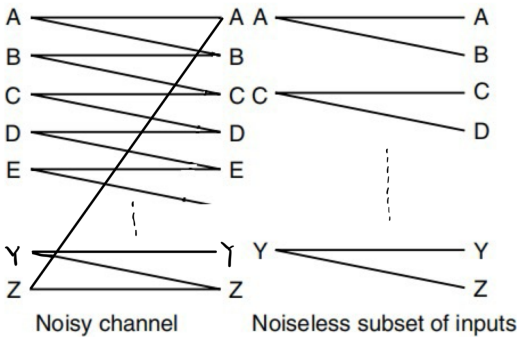


$$I(X;Y) = H(X) - H(X|Y) = H(X) \le \log 2 = 1$$

$$C = \max_{p(x)} I(X;Y) = 1$$

When $p(x) = \left(\frac{1}{2}, \frac{1}{2}\right)$, $C = 1$ *bit per transmission.*

由 Y 可以直接得到 X, 虽然由 *noisy*, 但是 *noisy* 没有 *overlap*, 不影响恢复出来 X.



例 6.2.2. *Noisy typewriter.*

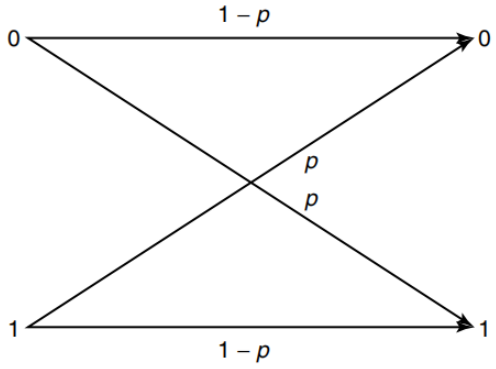
$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H\left(\frac{1}{2}, \frac{1}{2}\right) \leq \log 26 - \log 2 = \log 13$$

当 $X \sim DUnif(26)$ 时, $Y \sim DUnif(26)$. 此时 C 取到 $I(X;Y)$ 的最大值 $\log 13$ *bits per transmission.* 但是注意到, 若按上图右侧的选择方式 *(i.e.* $P(X = `A') = P(X = `C') = \ldots = P(X = `Y') = \frac{1}{13}, P(X = `B') = P(X = `D') = \ldots = P(X = `Z') = 0$*)* 仍能取到最大值.

可见达到信道容量的 $p(X)$ 不唯一.

Noisy typewriter 平时出现的情况较少, 但 BSC 较为常见. e.g. 在硬盘读写时出现错误.

例 6.2.3. *Binary Symmetric Channel(BSC).*

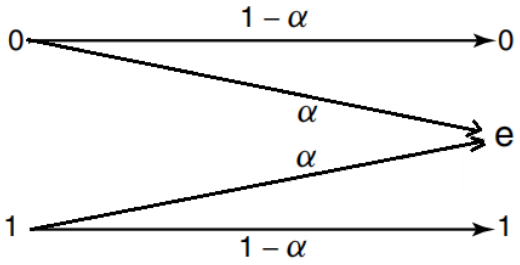


$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x H(Y|X=x)p(X=x) \\ &= H(Y) - H\left(p, 1-p\right) \\ &\leq \log 2 - H\left(p, 1-p\right) \\ &= 1 - H\left(p, 1-p\right) \end{aligned}$$

When $p(X=0) = p(X=1) = \frac{1}{2}$, $P(Y=0) = P(Y=1) = \frac{1}{2}$, *which makes the inequality take the equality.*

So $C = \max\limits_{p(x)} I(X;Y) = 1 - H\left(p, 1-p\right)$.

例 6.2.4. *Binary Erasure Channel(BEC).* 可以看作有 p 的概率丢包.



设 $p(X=0) = \pi, p(X=1) = 1-\pi$. 则

$$p(Y=e) = \sum_x p(Y=e|X=x)p(X=x) = \alpha$$

$$p(X=0|Y=e) = \frac{p(Y=e|X=0)p(X=0)}{p(Y=e)} = \pi$$

< 法一 >.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - \sum_y H(X|Y=y)p(Y=y) \\ &= H\left(\pi, 1-\pi\right) - \alpha H\left(\pi, 1-\pi\right) \\ &= H\left(\pi, 1-\pi\right)(1-\alpha) \\ &\leq 1-\alpha \end{aligned}$$

当 $p(x)=\left(\frac{1}{2},\frac{1}{2}\right)$ 时, $C=\max\limits_{p(x)} I(X;Y)=1-\alpha$.

< 法二 >. 设置辅助变量. E: *indictor* 是否丢包.

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

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

同时又有 $p(E=1)=p(Y=e)=\alpha$, 所以

$$\begin{aligned}
& H(Y,E)=H(E)+H(Y|E)=H(\alpha)+(1-\alpha)H(\pi)\\
\Rightarrow\quad & H(Y)=H(\alpha)+(1-\alpha)H(\pi)\\
\Rightarrow\quad & C=H(Y)-H(Y|X)\\
& \quad =H(\alpha)+(1-\alpha)H(\pi)-\pi H(\alpha)-(1-\pi)H(\alpha)\\
& \quad =(1-\alpha)H(\pi)\\
& \quad \leq 1-\alpha
\end{aligned}$$

当 $p(x)=\left(\frac{1}{2},\frac{1}{2}\right)$ 时, $C=\max\limits_{p(x)} I(X;Y)=1-\alpha$.

< 法三 >. $I(X;Y)=f(\pi)$ 是关于 π 的 *concave function.* 一阶导数为 *0* 时取得最大值.

例 6.2.5. *Symmetric Channel.* 若 $p(y|x)$ 的转移矩阵满足:

1. 每一行都是其他行的置换 *(permutation).*

2. 每一列都是其他列的置换 *(permutation).*

则 $C=\log|\mathcal{Y}|-H(\mathbf{r})$. *e.g.* 图中例子 $C=\log 3-H(0.5,0.3,0.2)$.

$$p(y|x)=\begin{array}{c}\diagdown\!\!\diagup\ \begin{matrix}\quad Y\end{matrix}\\ \begin{matrix}X\end{matrix}\end{array}\begin{pmatrix}0.3 & 0.2 & 0.5\\ 0.5 & 0.3 & 0.2\\ 0.2 & 0.5 & 0.3\end{pmatrix}$$

Weak Symmetric Channel. 若 $p(y|x)$ 的转移矩阵满足:

1. 每一行都是其他行的置换 *(permutation)*.
2. 每一列的和相等.

则 $C = \log|\mathcal{Y}| - H(\mathbf{r})$, $\mathbf{r}$ 是转移矩阵的行向量 *(*都是相同行向量的置换*)*.

Proof:

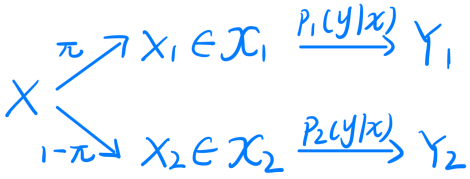
$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x)H(Y|X=x) \\ &= H(Y) - \sum_x p(x)H(\mathbf{r}) \\ &= H(Y) - H(\mathbf{r}) \\ &= \log|\mathcal{Y}| - H(\mathbf{r}) \end{aligned}$$

设每一列的和为同一个定值 α, 则当 $p(x) = \dfrac{1}{|\mathcal{X}|}$ 时

$$p(Y=y) = \sum_x p(Y=y|X=x)p(X=x) = \frac{\alpha}{|\mathcal{X}|}, \forall y \in \mathcal{Y}$$

所以所有的 $y \in \mathcal{Y}$, 概率相同. 即 $p(Y=y) = \dfrac{1}{|\mathcal{Y}|}$. 此时取得 $I(X;Y)$ 的最大值.

例 6.2.6. *Parallel Channel.*



两个平行的信道容量分别为 C_1, C_2, 则总容量满足

$$2^C = 2^{C_1} + 2^{C_2}$$

其中当 $p(X \in \mathcal{X}_1) = \dfrac{2^{C_1}}{2^{C_1} + 2^{C_2}}$, 且两个平行信道分别取得最大值时达到信道容量.

详见 *Homework4 Problem 5.*

In more general cases, 在开始时我们证明了 $I(X;Y)$ 是 $p(x)$ 的 concave function, 若没有闭式解, 我们可以用优化算法迭代求解.

通常采用 Frank-Wolfe grident search algorithm 或 Blahut-Arimoto 算法求 $I(X;Y)$ 最大值. 详见 §10.8 (In book P358).

命题 6.2.7. *Properties of channel capacity:*

1. $C \geq 0 \qquad (I(X;Y) \geq 0)$
2. $C \leq \log|\mathcal{X}|$, $C \leq \log|\mathcal{Y}|$.

Insight: 信道的 *input* 和 *output* 直接决定了 *capacity* 的上限, 可以看作输入和输出都会限制流量的宽度.

Proof: $I(X;Y) \leq \min\{H(X), H(Y)\} \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$

3. $I(X;Y)$ 是关于 $p(x)$ 的 *concave function.*

6.3 Fano's Inequality, Channel Coding Theorem(Shannon's Second Theorem)

简单理解: 从 jointly typical set 的角度. 在第四章的最后 (jointly AEP), 每个 X^n 平均会有 $\frac{2^{nH(X,Y)}}{2^{nH(X)}} = 2^{nH(Y|X)}$ 个 Y^n 与之 jointly typical, 即将所有的 Y^n 进行划分, 每份的大小为 $2^{nH(Y|X)}$, 可以分成 $\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X;Y)}$ 份. 这样平均每次用信道发送的 bit 数为 $\frac{2^{nI(X;Y)}}{n} = I(X;Y)$

在保证无损传输的情况下, 信道容量:

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

定义 6.3.1. *Channel Coding Theorem(Shannon's Second Theorem)*
对于一个 *rate* 为 R 的 *DMC*, 只要 $R < C$, 那么这个 *DMC* 是 *achievable* 的.
i.e. $\forall R < C, \exists$ *a sequence of* $\left(2^{nR}, n\right)$ *codes,* 使得最大错误概率 $\lambda^{(n)} \to 0$.
Conversely, 任意 *sequence of* $\left(2^{nR}, n\right)$ *codes,* $\lambda^{(n)} \to 0$, 那么 $R \leq C$.

所以证明需要从两个角度:

1. 可达性证明 (achievable proof): $\forall R < C$, $\exists (R, n)$, s.t. $\lambda^{(n)} \to 0$, 找到一种

达到 C 的方法.

2. converse proof: 所有方法都不可能超过 C. i.e. $R > C$ 时, $p_e^{(n)} \to 1 \Rightarrow R \leq C$ 时，$p_e^{(n)} \not\to 0$.

1. 可达性证明 achievable proof:

<1>. 码本构建 (randomly coding): fixed $p(x)$, 产生 2^{nR} 个长度为 n 的 codewords: $X^n(1), \ldots, X^n(2^{nR}) \overset{i.i.d.}{\sim} p(X^n) = \prod_{i=1}^{n} p(x_i)$.

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}$$

$X_i(w)$ 表示第 w 个 codeword 的第 i 个 bit(下标为 i).

码本构建好之后发给发送端和接收端 (is known by the encoder and the decoder). 同时, 认为 sender 和 receiver 知道 channel transition matrix $p(y|x)$.

<2>. Encoder: 随机生成一个数字 w, $p(W = w) = \frac{1}{2^{nR}}$, 从 codebook $\mathcal{C}$ 中找出第 w 行 $X^n(w)$, 发到信道当中.

<3> Decoder: reciver 接收由 $p(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w))$ 生成的 y^n, 从 codebook 中找出一条 message $\hat{W}$, s.t. $(X^n(\hat{W}), y^n) \in A_\epsilon^{(n)}(P_{X,Y})$, 且 no other index $W' \neq \hat{W}$ s.t. $(X^n(W'), y^n) \in A_\epsilon^{(n)}(P_{X,Y})$.

若找不到 $\hat{W}$, an error is decleared.

<4> Error probability analysis:

先忽略码本的随机性 ($p(\mathcal{C}) = \prod_{w}^{2^{nR}} \prod_{i=1}^{n} p(x_i)(w)$), 考虑单一码本的情况.

发生错误: $\mathcal{E} = \left\{\hat{W}(Y^n) \neq W\right\}$.

考虑定义: $\lambda(i) = \Pr\left[\hat{W} \neq i | W = i\right]$, 则 $\Pr[\mathcal{E}|W = i] = \lambda(i)$.

$\Rightarrow \Pr\left[\hat{W} \neq W\right] = \sum_{w=1}^{2^{nR}} p(W = w)\lambda(w) = \sum_{w=1}^{2^{nR}} \frac{1}{2^{nR}}\lambda(w)$.

由于 $W = 1, \ldots, 2^{nR}$ 是等概率产生的, 且地位相同, 所以 W.L.O.G. 取 $W = 1$ 时的情况进行分析:

定义事件 $E_i = \left\{(X^n(i), Y^n) \in A_\epsilon^{(n)}(P_{X,Y})\right\}$, i.e. the i-th codeword is joint typical with the received sequence Y^n.

由 jointly AEP, we can get that for sufficiently large n:

$$\Pr\left[E_1^c : (X^n(i), Y^n) \notin A_\epsilon^{(n)}(P_{X,Y})\right] \to 0 \Rightarrow \Pr\left[E_1^c : (X^n(i), Y^n) \notin A_\epsilon^{(n)}(P_{X,Y})\right] \le \epsilon$$

由于生成 codebook $\mathcal{C}$ 时, 每一行的产生是独立的, i.e. $X^n(1)$ 与 $X^n(2), \ldots, X^n(2^{nR})$ 是独立的, 所以 Y^n 与 $X^n(i)$ 也是独立的, 所以有:

$$\Pr\left[E_i : (X^n(i), Y^n) \in A_\epsilon^{(n)}(P_{X,Y})\right] \le 2^{-n(I(X;Y)-3\epsilon)}$$

由 jointly AEP, 当 $R < I(X;Y) - 3\epsilon$ 时

$$\begin{aligned}
\Pr\left[\hat{W} \neq 1 | W = 1\right] &= \Pr\left[E_1^c \cup E_2 \cup \ldots \cup E_{2^{nR}}\right] \\
&\le \Pr\left[E_1^c\right] + \sum_{w=2}^{2^{nR}} \Pr\left[E_w\right] \\
&\le \epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)} \\
&\le \epsilon + 2^{n(R-I(X;Y)+3\epsilon)} \\
&\le 2\epsilon \quad (\text{sufficiently large } n,\ R < I(X;Y) - 3\epsilon \Rightarrow 2^{n(R-I(X;Y)+3\epsilon)} \to 0 < \epsilon)
\end{aligned}$$

总结: 选择证明中取等时的分布作为真实分布 $p(x) \leftarrow p^*(x)$, 此时 $\forall R < I(X;Y) - 3\epsilon$, i.e. $R < C$, 错误概率 $p_e^{(n)} = \Pr\left[\hat{W} \neq W\right] \to 0 \Rightarrow \lambda^{(n)} \to 0$. (Actually, $\lambda^{(n)}$ 条件比 $p_e^{(n)}$ 更强, 但是 $p_e^{(n)}$ 更容易计算)

若将 codebook $\mathcal{C}$ 的随机性也考虑进去, 则会得到更加平均的结果:

$$\Pr\left[\hat{W} \neq W\right] = \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr[\mathcal{C}] \frac{1}{2^{nR}} \lambda_w(\mathcal{C}) = \lambda_1(\mathcal{C}) \Rightarrow \Pr(\mathcal{E}|\mathcal{C}^*) \le 2\epsilon$$

The best codebook $\mathcal{C}^*$ 中必然由超过一半的 indices i and their assciated codewords $X^n(i)$, 满足 $\lambda_i \le 4\epsilon$, 否则 $\Pr(\mathcal{E}|\mathcal{C}^*) > 2\epsilon$. 砍去另一半 codewords, 可以得到 rate 为 $R' = \frac{\frac{2^{nR}}{2}}{n} = R - \frac{1}{n}$, s.t. $\lambda^{(n)} \le 4\epsilon \to 0$. 所以可以使得 ϵ 不断的取小, 说明了 R 可以取到 C 的任意接近的地方.(取 $p^*(x)$ 时, $C = I(X;Y), R = I(X;Y) - 3\epsilon \to C$)

由此证明了 Achievability.

命题 6.3.2. *Zero-error codes.*

$$p_e^{(n)} \to 0 \Rightarrow R \leq C$$

proof:

Channel encoding 满足 Markov chain $W \to X^n(W) \to Y^n \to \hat{W}$, so from data-processing inequality:

$$I(W;Y^n) \leq I(X^n(W);Y^n)$$

由 DMC 的 memoryless, i.e. $p(Y_i|Y_1,\ldots,Y_{i-1},X^n) = p(Y_i|X_i)$, so

$$\begin{aligned}
I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) \\
&= \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1}) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},X^n) \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \qquad \text{(memoryless \& conditioning reduced entropy)} \\
&= \sum_{i=1}^{n} I(X_i;Y_i) \\
&\leq nC \qquad\qquad (Y_i \text{ are independent, } p_i(x) = p^*(x))
\end{aligned}$$

Since the channel is zero-error, so $\hat{W} = g(Y^n) = W$, i.e. $H(W|Y^n) = 0$.

$$\begin{aligned}
nR &= H(W) \\
&= H(W|Y^n) + I(W;Y^n) \\
&= I(W;Y^n) \qquad (H(W|Y^n) = 0) \\
&\leq I(X^n;Y^n) \qquad \text{(data-processing inequality)} \\
&\leq nC
\end{aligned}$$

So for any zero-error $\left(2^{nR}, n\right)$ code, for all n:

$$R \leq C$$

2. converse proof: 证明 $R > C$ 时, $p_e^{(n)} \to 1$ or $p_e^{(n)} \not\to 0$ as $n \to \infty$.

定义 6.3.3. *Fano's Inequality:* 是关于 p_e 的函数, 建立条件熵和错误概率 p_e 的关系.

For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, $p_e = P(\hat{X} \neq X)$, then

$$H(p_e) + p_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

It could be weakened to:

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

or

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

Proof:

Let E be the auxiliary variable indicate where the prediction is correct. i.e.

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

Then we have $p(E = 1) = p_e$, so

$$\begin{aligned} H\left(E, X|\hat{X}\right) &= \underbrace{H\left(E|\hat{X}\right)}_{\leq H(p_e)} + \underbrace{H\left(X|E, \hat{X}\right)}_{\leq p_e \log|\mathcal{X}|} \\ &= H\left(X|\hat{X}\right) + \underbrace{H\left(E|X, \hat{X}\right)}_{=0} \end{aligned}$$

其中由于 $E = \mathbb{I}_{X=\hat{X}}$, 所以 $H\left(E|X, \hat{X}\right) = 0$.

由于 conditioning reduces entropy, 所以 $H\left(E|\hat{X}\right) \leq H(E) = H(p_e)$.

同时有当 $E = 1$ 时, 给定 $\hat{X}$, X 的取值共有 $|\mathcal{X}| - 1$ 种, 所以有:

$$\begin{aligned} H\left(X|E, \hat{X}\right) &= \sum_{e=0,1} H\left(X|E = e, \hat{X}\right) P(E = e) \\ &\leq p_e \log\left(|\mathcal{X}| - 1\right) + (1 - p_e) \cdot 0 \\ &\leq p_e \log |\mathcal{X}| \end{aligned}$$

综上, 可以得到 Fano's Inequality:

$$H(p_e) + p_e \log |\mathcal{X}| \geq H(X|\hat{X})$$

weakened to 的版本可由 $H(p_e) \leq 1$, 以及数据处理不等式得到:

$$I(X;Y) \geq I(X;\hat{X}) \Rightarrow H(X) - H(X|Y) \geq H(X) - H(X|\hat{X}) \Rightarrow H(X|\hat{X}) \geq H(X|Y)$$

由于 Channel encoding 满足 Markov chain $W \to X^n(W) \to Y^n \to \hat{W}$, 所以应用 Fano's Inequality:

$$H(W|\hat{W}) \leq H(p_e^{(n)}) + p_e^{(n)} \log |\mathcal{W}|$$

其中 $\mathcal{W} = \{1, \ldots, 2^{nR}\}$, 所以 $|\mathcal{W}| = 2^{nR} \Rightarrow \log |\mathcal{W}| = nR$, 由于 W 是等概率产生的, 所以 $H(W) = \log |\mathcal{W}| = nR$. 所以有

$$\begin{aligned}
nR &= H(W) \\
&= H(W|\hat{W}) + I(W;\hat{W}) && \text{(definition of mutual information)} \\
&\leq H(p_e^{(n)}) + p_e^{(n)} nR + I(W;\hat{W}) && \text{(Fano's Inequality)} \\
&\leq 1 + p_e^{(n)} nR + nC && (I(W;\hat{W}) \leq I(X^n;Y^n) \leq nC \text{ proved in zero-error code part}) \\
\Rightarrow R &\leq \frac{1}{n} + p_e^{(n)} R + C \\
\Rightarrow p_e^{(n)} &\geq 1 - \frac{C}{R} - \frac{1}{nR}
\end{aligned}$$

取等条件: $W, \hat{W}$ 无信息丢失 (zero-error code), i.e. $X^n = f(W), Y^n = g(X^n)$ 是 W 的充分统计量, W 到 $\hat{W}$ 是一对一的映射. 由充分统计量 (s.s.), 此时可以看作构成 Markov Chain:

$$W \to X^n \to Y^n \to \hat{W}$$
$$W \to \hat{W} \to X^n \to Y^n$$

由于我们的前提假设是 $R > C$, 且 R, C 为常数, 所以 $\frac{C}{R} < 1$, 是常数. 当 $n \to \infty$ 时, $\frac{1}{nR} \to 0 < \epsilon$, 所以 $p_e^{(n)} \geq 1 - (\text{constant less than } 1) - \epsilon \not\to 0$.

所以我们证明了 $\forall R > C, p_e^{(n)} \not\to 0$, 即证明了 converse proof.

6.4 Hamming Code

Channel coding theorem 证明了存在编码方式使得 transmit imformation at rate $R < C$, 在 $n \to \infty$ 时可以达到任意小的 error probability $p_e^{(n)}$.

LDPC(Low Density Parity Check) code, polar code 等编码方式可以信道容量的任意接近, Hamming code 可以检验错误的位置 (发生错误的数量较少时可以纠正, 稍微多一点只能发现错误无法纠正, 再多无法发现).

编码时涉及到的加法运算均为模 2 加法.

定义 6.4.1. *Hamming Code.* H 为奇偶校验矩阵 *(parity check matrix).* H 的每一列是 $1, \dots, 7$ 对应的二进制.

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

例子中 $H_{3\times 7}$, $\text{rank}(H) = 3$, $\text{nullity}(H) = 7 - 3 = 4$. 所以 H 的零空间中共有 $2^{\text{nullity}(H)} = 2^4 = 16$ 个向量. 如图所示:

0000000	0100101	1000011	1100110
0001111	0101010	1001100	1101001
0010110	0110011	1010101	1110000
0011001	0111100	1011010	1111111

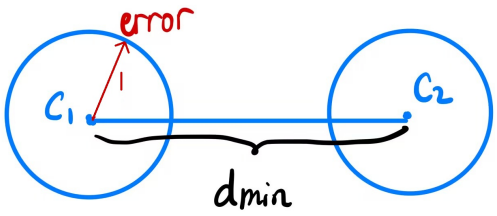
零空间的所有向量构成 codebook $\mathcal{C}$, 且零空间是一个线性空间. Sense the sum of any two codewords is also a codeword. 将 $\mathcal{C}$ 中的 codeword $\mathbf{c}_1, \mathbf{c}_2$ 看作列向量, 则 $Hc_1 = 0, Hc_2 = 0 \Rightarrow H(\mathbf{c}_1 + \mathbf{c}_2) = 0$.

其中每一个向量的前 k 位 (例子中 $k = 4$) 位信息位 information bits, 后 $n - k$ 位 (例子中 $n = 7, n - k = 3$) 为校验位 parity check bits. 一个 (n, k)-code 的传输速率为 $R = \dfrac{k}{n}$.

recall: 信道编码的目的是增加传输位数上的冗余, 以增加信息传输的可靠性. 所以我们尽可能的增强 code 的纠错能力.

定义 6.4.2. *Hamming Distance between codeword* $\mathbf{a}$ *and* $\mathbf{b}$*: 两个码字中不同的 bit 的个数. (Actually* $\|\mathbf{a}-\mathbf{b}\|_0$*).*

The minimum weight of a code: $\mathcal{C}$ 中除了全 0 码字外, 码字中 1 的个数的最小值. Actually, minimum weight $= d_{\min}$.
在 $(7,3)$-code 中, 任意两个码字之间的 Hamming Distance 最小值 $d_{\min}=3$.



Hamming code 可以纠正 $\left\lfloor \frac{d_{\min}-1}{2} \right\rfloor$ 个错误, 检测到 $d_{\min}-1$ 个错误. 例如 $(7,3)$-code 可以纠正 1 个错误. 但是如果有 2 个错误, 则无法纠正, 只能检测. 若有 ≥ 3 个错误, 则无法检测出来.

例 6.4.3. *若只发生一个错误: Let* $\mathbf{x}$ *be the input, 第* i *位发生错误, 则接收到的信号为* $\mathbf{r}=\mathbf{x}+\mathbf{e}_i$*. 则*

$$H\mathbf{r}=H(\mathbf{x}+\mathbf{e}_i)=H\mathbf{x}+H\mathbf{e}_i=H\mathbf{e}_i=\mathbf{h}_i$$

由于 H *矩阵第* i *列为* $\mathbf{h}_i$*, 转成 2 进制后即为列数本身, 则可直接找到第* i *位发生错误.*

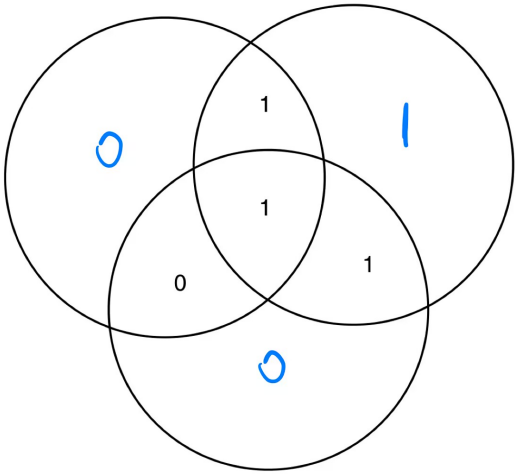
命题 6.4.4. *Hamming code 的构建: 通常若矩阵的行数为* l*, 则设列数 (Hamming code 的长度) 为:* $n=2^l-1$*, 信息位数* $k=nullity(H)=n-l=2^l-1-l$*. 通常设* $d_{\min}=3$*. 即* $(2^l-1,2^l-1-l,3)$*-code 或写作* $(2^l-1,2^l-1-l)$*-code.*

$$R=\frac{k}{n}=\frac{2^l-1-l}{2^l-1}$$

随着 l 的增加, R 逐渐从小接近到 1.
当 $R<C$ 时, 错误概率 $p_e^{(n)}\to 0$. $R>C$ 时, 错误概率 $p_e^{(n)}\not\to 0$.

例 6.4.5. *For a BSC using Hamming code,* $C = 1 - H(p), R = \frac{k}{n} = \frac{2^l - 1 - l}{2^l - 1}$, 当 $p = \frac{1}{2}$ 时, $C = 1 - H(\frac{1}{2}) = 0, R > 0$, 所以此时传输一定会出错.

可以用 Venn 图来理解 Hamming code: 假设 code 为 1101001, 中间有



交集的部分放置信息位 (1101), 边放置校验位001, 其中在放置校验位时, 保证每个大圆中 1 的个数为偶数. 若无法保证, 则检测到错误.
纠错方式: 在 Venn 图中更改某一位, 使得满足要求. 或者在码本中找与当前错误的码字 Hamming Distance 最小的码字进行替换.

6.5 Other codes*

- Reed-Muller Code: 利用 $\mathcal{C}$ 的线性性质.
- Hudmard Code: 参数选择上有限制 $(n = 2^m)$
- LDPC Code(Low-Density Parity-Check Code) :5G 中常用的编码方式, 逼近 Shannon channel capacity

Details in PPT Lecture 4 Channle Encoding.

第七章　Differential Entropy

连续熵 / 微分熵.　　Book Chapter8. (P269)

nat(aka. nit / nepit) 是信息论中的单位, 将以 nat 为单位的量中所有的 ln 都换成 log, 即可换算成单位 bit. $1\text{ nat} = \ln e\text{ nat} = \log e\text{ bit} = \dfrac{1}{\ln 2}\text{ bit}$.

7.1　Entropy, Relative Entropy, Mutual Information

$X \sim f(x)$, where $f(x)$ is the PDF.

$$h(X) = -\int_x f(x)\log f(x)dx = \mathbb{E}_{X\sim f(x)}\left[-\log f(x)\right]$$

$$\begin{aligned} h(X) &= -\int_x f(x)\log f(x)\,\mathrm{dx}\ \text{ bits} \\ &= -\int_x f(x)\ln f(x)\,\mathrm{dx}\ \text{ nats} \end{aligned}$$

不同于离散型随机变量 $0 \le H(X) \le \log|\mathcal{X}|$ 有界, 连续型随机变量 $h(X)$ 无界. $h(X)$ 是反常积分, 可以 < 0, 也可能不存在 $(\to \infty)$. 但 $h(X)$ 仍可以反映 X 的不确定性, $h(X)$ 越大, X 的不确定性越大.

例 7.1.1. $X \sim Unif(a,b)$, $f(x) = \dfrac{1}{b-a}$, $a \le x \le b$.

$$h(X) = -\int_a^b \frac{1}{b-a}\log\frac{1}{b-a}\,\mathrm{dx} = \log(b-a)$$

For example, when $a = 0, b = 0.5$, $h(X) = \log 0.5 = -1 < 0$.

例 7.1.2. $X \sim \mathcal{N}(\mu, \sigma^2)$, $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$.

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ln \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{dx} \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \left(\ln\left(\sqrt{2\pi}\sigma\right) + \frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{dx} \\ &= \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} (x-\mu)^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{dx} \\ &\quad + \ln\left(\sqrt{2\pi}\sigma\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{dx} \end{aligned}$$

For the first part(variance):

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{dx} = \sigma^2$$

For the second part(integral of PDF):

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{dx} = 1$$

So put them into the equation:

$$h(X) = \frac{1}{2\sigma^2}\sigma^2 + \ln\left(\sqrt{2\pi}\sigma\right) = \frac{1}{2}\ln(2\pi e\sigma^2) \ \textit{nats} = \frac{1}{2}\log(2\pi e\sigma^2) \ \textit{bits}$$

定义 7.1.3. *Joint Differential Entropy:* $X_1, \ldots, X_n \sim f(x_1, \ldots, x_n)$, *then*

$$h(X_1, \ldots, X_n) = -\int_{x^n} f(x^n) \log f(x^n) \, \mathrm{dx}^n$$

定义 7.1.4. *Conditonal Differential Entropy:*

$$\begin{aligned} h(X|Y) &= -\int_{x,y} f(x,y) \log f(x|y) \, \mathrm{dx} \, \mathrm{dy} \\ &= h(X,Y) - h(Y) \end{aligned}$$

定义 7.1.5. *Relative Entropy(KL Divergence):* 为了保证连续性质，定义

$0 \log \frac{0}{0} = 0.$

$$\begin{aligned} -D(f\|g) &= -\int_x f(x) \log \frac{f(x)}{g(x)} \, \mathrm{dx} \\ &= \mathbb{E}_{X\sim f(x)} \left[\log \frac{g(X)}{f(X)} \right] \\ &\leq \log \left(\mathbb{E}_{X\sim f(x)} \left[\frac{g(X)}{f(X)} \right] \right) \\ &= \log \left(\int_x f(x) \frac{g(x)}{f(x)} \, \mathrm{dx} \right) \\ &= \log \left(\int_x g(x) \, \mathrm{dx} \right) = \log 1 \\ &= 0 \end{aligned}$$

i.e. $D(f\|g) \geq 0$, *and* $D(f\|g) = 0$ *if and only if* $f = g$.

定义 7.1.6. *Mutual Information:*

$$I(X;Y) = \int_{x,y} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} \, \mathrm{dx} \, \mathrm{dy}$$

Similarly with discrete case:

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y)$$

由于 $I(X;Y) = KL(f(x,y)\|f(x)f(y)) \geq 0$, 所以 $h(X) \geq h(X|Y)$, *i.e. conditioning reduces entropy* 仍成立.

命题 7.1.7. 在指定范围 $[a,b]$ 内的所有分布中, 若对分布没有限制, 则当且仅当均匀分布 $X \sim Unif(a,b)$ 时熵最大 $\log(b-a)$.

proof: Let $u(x)$ *be the PDF of Unif*(a,b), $f(x)$ *be the PDF of any other distribution in* $[a,b]$. *Then*

$$\begin{aligned} 0 \leq D\left(f\|u\right) &= \int_a^b f(x) \ln \frac{f(x)}{u(x)} \, \mathrm{dx} \\ &= \int_a^b f(x) \ln u(x) \, \mathrm{dx} - \int_a^b f(x) \ln f(x) \, \mathrm{dx} \\ &= \ln(b-a) - h(f) \\ &= h(u) - h(f) \end{aligned}$$

命题 7.1.8. $\forall C \in \mathbb{R}$, $h(X+C) = h(X)$. 熵衡量不确定性, 只改变均值, 概率 (不确定性) 不变.

命题 7.1.9. 书上似乎有问题? 积分换元漏掉了绝对值导致的正负号变化.

$$h(aX) = h(X) + \log|a|$$

由 *change of variable:* $f(\mathbf{Y}) = f(\mathbf{X})\left|\dfrac{\partial \mathbf{X}}{\partial \mathbf{Y}}\right|$, 可得:

Let $Y = aX \Rightarrow f_Y(y) = f_X(x)\dfrac{1}{|a|} = \dfrac{1}{|a|} f_X\left(\dfrac{y}{a}\right)$.

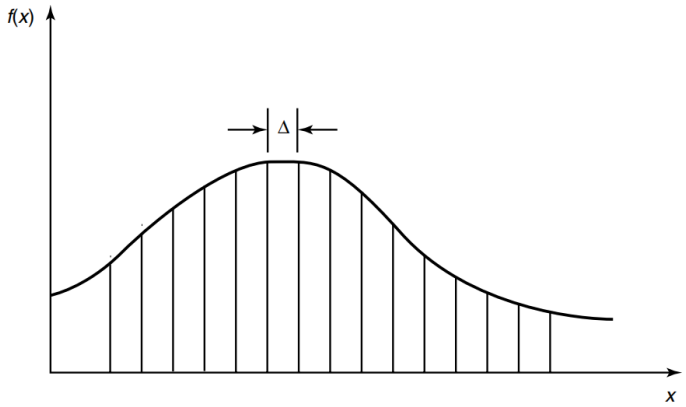
$$\begin{aligned} h(Y) &= -\int_y f_Y(y)\log f_Y(y)\,\mathrm{d}y \\ &= -\int_y \frac{1}{|a|} f_X\left(\frac{y}{a}\right)\log\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\mathrm{d}y \\ &= \frac{a}{|a|}\cdot\left(-\int_x f_X(x)\log\frac{1}{|a|} f_X(x)\,\mathrm{d}x\right) \\ &= (-1)^{\mathbb{I}_{a<0}}\left(h(X)+\log|a|\right) \end{aligned}$$

$a > 1$, $h(aX) > h(X)$, $0 < a < 1$, $h(aX) < h(X)$.

命题 7.1.10. 由上一个命题, 拓展到向量: $\mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times n}$, *then*

$$h(\mathbf{Ax}) = h(\mathbf{x}) + \log|\det(\mathbf{A})|$$

命题 7.1.11. 离散与连续的关系:



由积分中值定理: $f(x)$ 在 $[a,b]$ 上连续, 则 $\exists \epsilon \in [a,b]$, s.t.

$$(b-a)f(\epsilon) = \int_a^b f(x)\,\mathrm{dx}$$

将连续型随机变量进行量化, 每 Δ 为一个 bin / block, 记第 i 个 bin 为区间 $[i\Delta,(i+1)\Delta]$, 则 $\exists x_i \in [i\Delta,(i+1)\Delta]$, s.t.

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

Then we construct the quantization random variable X^Δ, s.t. $X^\Delta = x_i$ with probability p_i. Then we have:

$$\begin{aligned}
H(X^\Delta) &= -\sum_{i=-\infty}^{+\infty} p_i \log p_i \\
&= -\sum_{i=-\infty}^{+\infty} f(x_i)\Delta \log\left(f(x_i)\Delta\right) \\
&= -\sum_{i=-\infty}^{+\infty} f(x_i)\Delta \log f(x_i) - \sum_{i=-\infty}^{+\infty} \underbrace{f(x_i)\Delta}_{p_i} \log\Delta \\
&\to -\int_{-\infty}^{+\infty} f(x)\log f(x)\,\mathrm{dx} - \log\Delta \\
&= h(X) - \log\Delta
\end{aligned}$$

如果函数 $f(x)\log f(x)$ 黎曼可积, 则当 $\Delta \to 0$ 时

$$-\sum_{i=-\infty}^{+\infty} f(x_i)\Delta\log f(x_i) \to -\int_{-\infty}^{+\infty} f(x)\log f(x)\,\mathrm{dx}$$

所以有结论

$$H(X^\Delta) + \log\Delta \to h(X),\ \text{as } \Delta \to 0$$

如果在 $[a,b]$ 上, 用 n 个 bit 进行量化, 即 $\Delta = 2^{-n}$, 则

$$h(X) = H(X^\Delta) - n$$

7.2 AEP for continuous Random Variables

$h(X)$ 可正可负, 且越大 X 的不确定度越大. 如何用负值衡量不确定度? 需要用 AEP 解释.

定理 7.2.1. *$x_1, \ldots, x_n \overset{i.i.d.}{\sim} f(x)$, then*

$$-\frac{1}{n}\log f(x_1, \ldots, x_n) \to \mathbb{E}\left(-\log f(x)\right) = h(f) \text{ in probability}$$

定义 7.2.2. *$x_1, \ldots, x_n$ is a typical sequence if*

$$2^{-n(h(f)+\epsilon)} \le f(x_1, \ldots, x_n) \le 2^{-n(h(f)-\epsilon)}$$

定义 7.2.3. *$A_\epsilon^n(f(x))$ is a typical set, then*

$$A_\epsilon^n(f_x) = \left\{(x_1, \ldots, x_n) \mid \left|-\frac{1}{n}\log f(x_1, \ldots, x_n) - h(f)\right| \le \epsilon\right\}$$

Where $f(x^n) = \prod_{i=1}^{n} f(x_i)$

离散型随机变量用 $|A_\epsilon^n(P_X)|$ 来描述 typical set 的大小。连续型随机变量将 typical set 的区域看作是一个正方体, 其大小用 $\mathrm{Vol}\left(A_\epsilon^n(f_X)\right)$ 来描述, 看作是体积:

$$\mathrm{Vol}\left(A_\epsilon^n(f_X)\right) = \int_{A_\epsilon^n(f_X)} \mathrm{dx}^n$$

命题 7.2.4. *1.* $\Pr\left(A_\epsilon^n(f_X)\right) > 1-\epsilon$ *as n is sufficiently large.*
2. $(1-\epsilon)2^{n(h(f)-\epsilon)} \le \mathrm{Vol}\left(A_\epsilon^n(f_X)\right) \le 2^{n(h(f)+\epsilon)}$

则正方体的边长 d 为:

$$d^n = \mathrm{Vol}\left(A_\epsilon^n(f_X)\right) = 2^{nh(f)} \Rightarrow d = 2^{h(f)}$$

用 d 来衡量 X 的不确定度. $d \ge 0$, d 越大, X 的不确定度越大. $d = 0$ 时, X 的不确定度为 0.

$h(X_1) = -1, h(X_2) = 0$, X_2 的不确定度大于 X_1. 但用 d 衡量更直观一些, i.e. $d(X_2) > d(X_1) > 0$.

7.3 Variance Limited Variable's Entropy

1. Entropy of a multivariate normal distribution:

$X_1, \ldots, X_n \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})$, then $h\left(\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})\right) = \frac{1}{2}\log\left((2\pi e)^n|\mathbf{K}|\right)$ bits.

proof:

Since $f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |\mathbf{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$, we have

$$
\begin{aligned}
h(f) &= -\int f(\mathbf{x}) \ln f(\mathbf{x}) \, \mathrm{d}\,\mathbf{x} \\
&= -\int f(\mathbf{x}) \left[\ln\left(\frac{1}{\left(\sqrt{2\pi}\right)^n |\mathbf{K}|^{\frac{1}{2}}}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \mathrm{d}\,\mathbf{x} \\
&= \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim f(\mathbf{x})}\left[\sum_{i,j}(X_i-\mu_i)\left(\mathbf{K}^{-1}\right)_{ij}(X_j-\mu_j)\right] + \frac{1}{2}\left(\ln(2\pi)^n|\mathbf{K}|\right) \\
&= \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim f(\mathbf{x})}\left[\sum_{i,j}(X_i-\mu_i)(X_j-\mu_j)\left(\mathbf{K}^{-1}\right)_{ij}\right] + \frac{1}{2}\left(\ln(2\pi)^n|\mathbf{K}|\right) \\
&= \frac{1}{2}\sum_{i,j}\underbrace{\mathbb{E}_{\mathbf{x}\sim f(\mathbf{x})}\left[(X_j-\mu_j)(X_i-\mu_i)\right]}_{\mathrm{Cov}(X_i,X_j)=\mathbf{K}_{ij}}\left(\mathbf{K}^{-1}\right)_{ij} + \frac{1}{2}\left(\ln(2\pi)^n|\mathbf{K}|\right) \\
&= \frac{1}{2}\sum_{j}\sum_{i}\mathbf{K}_{ji}\left(K^{-1}\right)_{ij} + \frac{1}{2}\left(\ln(2\pi)^n|\mathbf{K}|\right) \\
&= \frac{1}{2}\sum_{j}\left(\mathbf{K}\mathbf{K}^{-1}\right)_{jj} + \frac{1}{2}\left(\ln(2\pi)^n|\mathbf{K}|\right) \\
&= \frac{1}{2}\sum_{j} I_{jj} + \frac{1}{2}\left(\ln(2\pi)^n|\mathbf{K}|\right) \\
&= \frac{n}{2} + \frac{1}{2}\left(\ln(2\pi)^n|\mathbf{K}|\right) \\
&= \frac{1}{2}\ln(2\pi e)^n|\mathbf{K}| \quad \text{nats}
\end{aligned}
$$

2. Mutual information of two Gaussian random variables:

Let $(X,Y) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$, $\rho \in [-1, 1]$.

Then $h(X) = h(Y) = \frac{1}{2}\log(2\pi e\sigma^2)$, $h(X,Y) = \frac{1}{2}\log((2\pi e)^2|\mathbf{K}|) = \frac{1}{2}\log((2\pi e)^2\sigma^4(1-\rho^2))$

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = -\frac{1}{2}\log\left(1-\rho^2\right)$$

If $\rho = 0$, $I(X;Y) = 0 \Rightarrow X \perp Y$.

If $\rho = \pm 1$, X and Y are perfectly correlated and $I(X;Y) = +\infty$.

ρ 刻画**线性**相关性. Gaussian 不相关 = 独立!!

3. Given $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$, $\mathrm{Cov}(\mathbf{X}) = \mathbf{K}$. Then $h(\mathbf{X}) \le \frac{1}{2}\log((2\pi e)^n|\mathbf{K}|)$ bits, iff $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})$.

令 $g(\mathbf{x})$ 是任意一个满足上述条件的 PDF, $\phi(\mathbf{x})$ 是 $\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})$ 的 PDF, 则

Lemma: $\mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}(\ln(\mathbf{X})) = \mathbb{E}_{\mathbf{X}\sim\phi(\mathbf{x})}(\ln(\mathbf{X}))$, i.e. $\int g(\mathbf{x})\ln\phi(\mathbf{x})\,\mathrm{d}\,\mathbf{x} = \int \phi(\mathbf{x})\ln\phi(\mathbf{x})\,\mathrm{d}\,\mathbf{x}$.

Since $\ln\phi(\mathbf{x}) = -\frac{1}{2}\ln((2\pi)^n|\mathbf{K}|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, so

$$\ln\phi(\mathbf{x}) = -\frac{1}{2}\ln((2\pi)^n|\mathbf{K}|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

$$\mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}[\ln\phi(\mathbf{x})] = -\frac{1}{2}\ln((2\pi)^n|\mathbf{K}|) - \frac{1}{2}\mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}\left[(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

关于二次型的期望, 有:

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}\left[(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] &= \mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}\left[\mathrm{Tr}\left((\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right] \\
&= \mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}\left[\mathrm{Tr}\left(\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right)\right] \\
&= \mathrm{Tr}\left(\mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}\left[\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]\right) \\
&= \mathrm{Tr}\left(\mathbf{K}^{-1}\mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]\right) \\
&= \mathrm{Tr}\left(\mathbf{K}^{-1}\mathbf{K}\right) \\
&= \mathrm{Tr}\left(I_d\right) \\
&= n
\end{aligned}$$

为一个常数, 所以有

$$\mathbb{E}_{\mathbf{X}\sim g(\mathbf{x})}[\ln\phi(\mathbf{x})] = \mathbb{E}_{\mathbf{X}\sim\phi(\mathbf{x})}[\ln\phi(\mathbf{x})]$$

i.e.

$$\int g(\mathbf{x})\ln\phi(\mathbf{x})\,\mathrm{d}\,\mathbf{x} = \int \phi(\mathbf{x})\ln\phi(\mathbf{x})\,\mathrm{d}\,\mathbf{x}$$

有了 Lemma, 结合 KL 散度, 我们就可以证明结论了:

$$\begin{aligned}
0 &\geq -D\left(g\|\phi\right) \\
&= -\int g(\mathbf{x}) \ln\left(\frac{g(\mathbf{x})}{\phi(\mathbf{x})}\right) \mathrm{d}\,\mathbf{x} \\
&= \int g(\mathbf{x}) \ln \phi(\mathbf{x}) \,\mathrm{d}\,\mathbf{x} - \int g(\mathbf{x}) \ln g(\mathbf{x}) \,\mathrm{d}\,\mathbf{x} \\
&= \int \phi(\mathbf{x}) \ln \phi(\mathbf{x}) \,\mathrm{d}\,\mathbf{x} - \int g(\mathbf{x}) \ln g(\mathbf{x}) \,\mathrm{d}\,\mathbf{x} \qquad \text{(Lemma)} \\
&= -h(\phi) + h(g) \quad \square
\end{aligned}$$

4. variance limitied 由 3. 的证明过程可知, 并没有用到他们均值相同的性质, 所以更一般的结论是:

<1>. Given $\mathrm{Var}(\mathbf{X}) = \mathbf{K}$, then $h(\mathbf{X}) \leq \frac{1}{2}\log\left((2\pi e)^n|\mathbf{K}|\right)$ bits, iff $\mathbf{X} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{K})$.

<2>. 若给定均值 μ, 且 support 为 $[0, +\infty)$, 则当指数分布 $f(x) = \frac{1}{\mu}e^{-\frac{x}{\mu}}$ 时熵最大 (似乎可以用变分的 Lagrangian 证明).

<3>. 若均值和方差均没有限制, support 有区间限制, 则均匀分布熵最大.

5. Estimation error and dofferential entropy.

对于分布 X, 以及他的 estimator $\hat{X}$, 则有

$$\mathbb{E}\left[\left(\hat{X} - X\right)^2\right] \geq \frac{1}{2\pi e}e^{2h(X)}$$

proof:

$$\mathbb{E}\left[\left(\hat{X} - X\right)^2\right] \geq \min_{\hat{X}} \mathbb{E}\left[\left(\hat{X} - X\right)^2\right] = \mathbb{E}\left[(\mathbb{E}(X) - X)^2\right] = \mathrm{Var}(X) \geq \frac{1}{2\pi e}e^{2h(X)}$$

第一个不等式: MMSE 的最优解.

第二个不等式: 当方差为 σ^2 时, $h(X) \leq \frac{1}{2}\log(2\pi e\sigma^2) \Rightarrow \sigma^2 \geq \frac{1}{2\pi e}e^{2h(X)}$.

Insight: X 越不确定, $h(X)$ 越大, 凭空猜 $\hat{X}$ 的误差越大.

推论: X 有一个观测 $\hat{X}(Y)$, 则

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

要想让 MSE 尽可能小, 则需要让 $h(X|Y) \to -\infty$, i.e. 观测和真实值之间没有不确定度.

Recall: 观测值为 Y, 真实值为 X, 则

MMSE: $\mathbb{E}(Y|X)$

LLSE: $L[Y|X] = \mathbb{E}(Y) + \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - \mathbb{E}(X))$

X, Y 为联合高斯时, LLSE=MMSE.

第八章　Gaussian Channel

Book Chapter9. (P287)

连续信号, 高斯信道的信道容量为 $C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$. 这是无线通讯领域内的重要结论. 这里的 $\frac{P}{N}$ 是信噪比, i.e. Signal to Noise Ratio (SNR).

8.1　Gaussian Channel

一个信号 $x(t)$ 的能量和功率为

$$E = \int_{-\infty}^{\infty} x^2(t)\,\mathrm{dt}$$

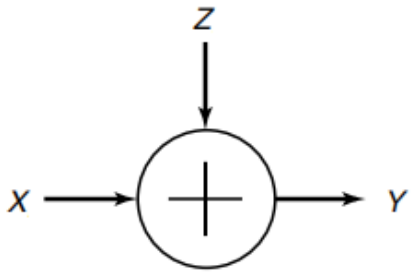
$$P = \lim_{T\to\infty} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x^2(t)\,\mathrm{dt}$$

当传输为离散时, 采样点为 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$, 由大数定理可得:

$$P = \lim_{T\to\infty} \frac{1}{T}\sum_{n=-\frac{T}{2}}^{\frac{T}{2}} X_i^2 \to \mathbb{E}(X^2)$$

信道在传输信号时有传输功率 P 的限制. 所以我们可以用 $\mathbb{E}(X^2) \le P$ 作为信号的传输功率限制. 因为平移不会改变信息熵, 所以不妨设 X 的均值为 0 以减少能量的浪费, 此时约束变为 $\mathrm{Var}(X) \le P$.

高斯信道如下图所示. 其中 X 为信道的输入, 传输过程中有噪声 $Z \sim \mathcal{N}(0, N)$, 且 $X \perp Z$. 信道的输出为 $Y = X + Z$. 高斯信道又叫 Additive White Gaussian Noise (AWGN) 信道.



由于信道容量 C 为每次可靠传输的 bit 数, 所以 $N=0$ 时 $C=+\infty$. 若 X 传输时没有功率限制, 则 $C=+\infty$. 但是在考虑真实的高斯信道时加上 $\mathrm{Var}(X)\le P$ 的限制.

由上一章可知, Gaussian 的方差为 σ^2 时的熵为 $h(X)=\frac{1}{2}\ln(2\pi e\sigma^2)$ nats. 所以当 $X\sim\mathcal{N}(0,P)$ 时, $Y\sim\mathcal{N}(0,P+N)$, $Y|X\sim\mathcal{N}(X,N)$, 此时有

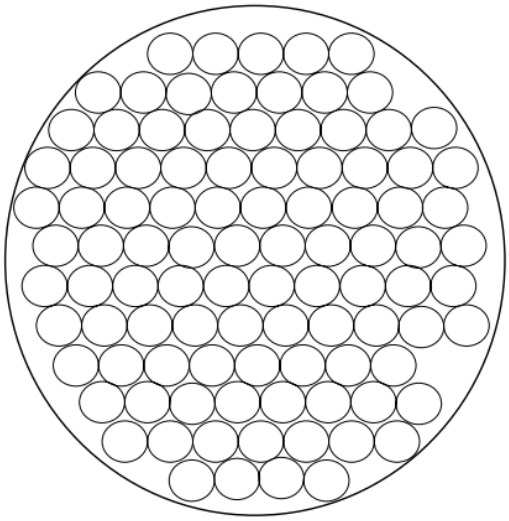
$$I(X;Y)=h(Y)-h(Y|X)=\frac{1}{2}\ln\left(1+\frac{P}{N}\right)\ \text{nats}$$

后面我们证明这其实就是高斯信道的信道容量.

例 8.1.1. *A non-optimal example: Let* $X=\begin{cases}\sqrt{P}, & w.p.\ \frac{1}{2}\\ -\sqrt{P}, & w.p.\ \frac{1}{2}\end{cases}$. X 有 $\mathbb{E}(X)=0,\mathrm{Var}(X)=P$. 若取 *Decoder* 为 $\hat{X}=\arg\max_{\hat{X}}\|Y-\hat{X}\|^2$, 则有 $\hat{X}=\begin{cases}\sqrt{P}, & if\ Y\ge 0\\ -\sqrt{P}, & if\ Y<0\end{cases}$. 此时有错误概率

$$\begin{aligned}P_e&=\Pr(\hat{X}\neq X)=\Pr(Y<0|X=\sqrt{P})\cdot\Pr(X=\sqrt{P})+\Pr(Y\ge 0|X=-\sqrt{P})\cdot\Pr(X=-\sqrt{P})\\&=\frac{1}{2}\Pr(Y<0|X=\sqrt{P})+\frac{1}{2}\Pr(Y\ge 0|X=-\sqrt{P})=\frac{1}{2}\Pr(Z>\sqrt{P})+\frac{1}{2}\Pr(Z<-\sqrt{P})\\&=\frac{1}{2}\left(1-\Phi\left(\frac{\sqrt{P}}{\sqrt{N}}\right)\right)+\frac{1}{2}\Phi\left(\frac{\sqrt{P}}{\sqrt{N}}\right)\\&=\frac{1}{2}\left(1-\Phi\left(\frac{\sqrt{P}}{\sqrt{N}}\right)\right)+\frac{1}{2}\left(1-\Phi\left(\frac{\sqrt{P}}{\sqrt{N}}\right)\right)\\&=1-\Phi\left(\sqrt{\frac{P}{N}}\right)\end{aligned}$$

和先前证明信道容量时的方法类似, 我们通过 Achievability 和 Converse 来证明信道容量. 我们先通过填充球理论 sphere packing(本质还是 AEP) 来直观感受:



如上图所示, 假设每一个小球的球心为一个 codeword, 球的范围为可以 decode 到该 codeword 的所有码字 (jointly AEP). 假设这些小球的半径为 r, 则一个 n 为球的体积可以写为 $\text{Vol} = C_n r^n$. 由于 noise 的方差为 N, 所以小球的半径为 $\sqrt{n(n+\epsilon)}$, 而 Y 的方差为 $\mathbb{E}(Y^2) = \mathbb{E}((X+Z)^2) = \mathbb{E}(X^2) + \mathbb{E}(Z^2) = P + N$, 所以整个大球的半径为 $\sqrt{n(P+N)}$. 因此最多有 $\frac{C_n(\sqrt{n(P+N)})^n}{C_n(\sqrt{n(n+\epsilon)})^n} = \left(1+\frac{P}{N}\right)^{\frac{1}{2}}$ 个小球. 所以 rate of the code(小球数量需要的 bit 数) is $\log\left(1+\frac{P}{N}\right)^{\frac{1}{2}} = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$.

关于大球的半径大小: typical set 的大小为 $\text{Vol}\left(A_\epsilon^{(n)}(f_Y)\right) = 2^{n(H(Y))} \leq 2^{n(\frac{1}{2}\log(2\pi e(P+N)))}$, 所以半径取 $\sqrt{n(P+N)}$. 小球的半径同理.

严谨证明: 1. 可达性证明: 大致和先前证明信道容量的方法类似, 要证明 $\forall R < C = I(X;Y)$ 时, 发生错误的概率 $P_e^{(n)} \to 0$.

通过构建码本, $X_i(w)$ 表示码字 w 的第 i 个 bit. 其中 $X_i(w) \overset{i.i.d.}{\sim} \mathcal{N}(0, P-\epsilon)$.

和先前一样, W.L.O.G. 令 w=1, 则 $P_e^{(n)} = \Pr\left(\hat{W} \neq 1 | W = 1\right)$, 然后分

类讨论不符合要求的情况即可. 不同之处: 增加了 $E_0 : \frac{1}{n}\sum_{j=1}^{n} X_j(1) > P$(不满足功率约束限制) 的情况.

摆烂了, 放图片了.

1. *Generation of the codebook*. We wish to generate a codebook in which all the codewords satisfy the power constraint. To ensure this, we generate the codewords with each element i.i.d. according to a normal distribution with variance $P-\epsilon$. Since for large n, $\frac{1}{n}\sum X_i^2 \to P-\epsilon$, the probability that a codeword does not satisfy the power constraint will be small. Let $X_i(w)$, $i=1,2,\ldots,n$, $w=1,2,\ldots,2^{nR}$ be i.i.d. $\sim \mathcal{N}(0, P-\epsilon)$, forming codewords $X^n(1), X^n(2), \ldots, X^n(2^{nR}) \in \mathcal{R}^n$.
2. *Encoding*. After the generation of the codebook, the codebook is revealed to both the sender and the receiver. To send the message index w, the transmitter sends the wth codeword $X^n(w)$ in the codebook.
3. *Decoding*. The receiver looks down the list of codewords $\{X^n(w)\}$ and searches for one that is jointly typical with the received vector. If there is one and only one such codeword $X^n(w)$, the receiver declares $\hat{W}=w$ to be the transmitted codeword. Otherwise, the receiver declares an error. The receiver also declares an error if the chosen codeword does not satisfy the power constraint.
4. *Probability of error*. Without loss of generality, assume that codeword 1 was sent. Thus, $Y^n = X^n(1) + Z^n$. Define the following events:

$$E_0 = \left\{ \frac{1}{n}\sum_{j=1}^{n} X_j^2(1) > P \right\} \tag{9.23}$$

and

$$E_i = \left\{ (X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)} \right\}. \tag{9.24}$$

Then an error occurs if E_0 occurs (the power constraint is violated) or E_1^c occurs (the transmitted codeword and the received sequence are not jointly typical) or $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (some wrong codeword is jointly typical with the received sequence). Let $\mathcal{E}$ denote the event $\hat{W} \neq W$ and let P denote the conditional probability given that $W=1$. Hence,

$$\Pr(\mathcal{E}|W=1) = P(\mathcal{E}) = P\left(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}\right)$$

$$\leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i), \tag{9.26}$$

by the union of events bound for probabilities. By the law of large numbers, $P(E_0) \to 0$ as $n \to \infty$. Now, by the joint AEP (which can be proved using the same argument as that used in the discrete case), $P(E_1^c) \to 0$, and hence

$$P(E_1^c) \leq \epsilon \quad \text{for } n \text{ sufficiently large.} \tag{9.27}$$

Since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent, so are Y^n and $X^n(i)$. Hence, the probability that $X^n(i)$ and Y^n will be jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP.
Now let W be uniformly distributed over $\{1,2,\ldots,2^{nR}\}$, and consequently,

$$\Pr(\mathcal{E}) = \frac{1}{2^{nR}}\sum \lambda_i = P_e^{(n)}. \tag{9.28}$$

Then

$$P_e^{(n)} = \Pr(\mathcal{E}) = \Pr(\mathcal{E}|W=1) \tag{9.29}$$

$$\leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \tag{9.30}$$

$$\leq \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \tag{9.31}$$

$$= 2\epsilon + \left(2^{nR}-1\right) 2^{-n(I(X;Y)-3\epsilon)} \tag{9.32}$$

$$\leq 2\epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \tag{9.33}$$

$$\leq 3\epsilon \tag{9.34}$$

for n sufficiently large and $R < I(X;Y) - 3\epsilon$. This proves the existence of a good $(2^{nR}, n)$ code.

想要 $P_e^{(n)} \to 0$, 则需要 $2^{-n(I(X;Y)-R-3\epsilon)} \to 0$, i.e. $R < I(X;Y) - 3\epsilon$. 这里所有提到的 $I(X;Y)$ 都是 $\max\limits_{p(x)} I(X;Y) = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$.

2. Converse proof:

证明 $\forall R > C = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$ 时, $P_e^{(n)} \not\to 0$ with $\mathbb{E}(X^2) \leq P$.
同样要用 Fano's Inequality, 对于信息传递的 Markov Chain $W \to X \to Y \to \hat{W}$, 有

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} \log 2^{nR} = 1 + P_e^{(n)} nR \triangleq n\epsilon_n$$

$$nR = H(W) = I(W;\hat{W}) + H(W|\hat{W}) \tag{9.38}$$
$$\le I(W;\hat{W}) + n\epsilon_n \tag{9.39}$$
$$\le I(X^n;Y^n) + n\epsilon_n \tag{9.40}$$
$$= h(Y^n) - h(Y^n|X^n) + n\epsilon_n \tag{9.41}$$
$$= h(Y^n) - h(Z^n) + n\epsilon_n \tag{9.42}$$
$$\le \sum_{i=1}^{n} h(Y_i) - h(Z^n) + n\epsilon_n \tag{9.43}$$
$$= \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n \tag{9.44}$$
$$= \sum_{i=1}^{n} I(X_i;Y_i) + n\epsilon_n. \tag{9.45}$$

Here $X_i = x_i(W)$, where W is drawn according to the uniform distribution on $\{1, 2, \ldots, 2^{nR}\}$. Now let P_i be the average power of the ith column of the codebook, that is,

$$P_i = \frac{1}{2^{nR}} \sum_{w} x_i^2(w). \tag{9.46}$$

Then, since $Y_i = X_i + Z_i$ and since X_i and Z_i are independent, the average power EY_i^2 of Y_i is $P_i + N$. Hence, since entropy is maximized by the normal distribution,

$$h(Y_i) \le \frac{1}{2}\log 2\pi e(P_i + N). \tag{9.47}$$

Continuing with the inequalities of the converse, we obtain

$$nR \le \sum (h(Y_i) - h(Z_i)) + n\epsilon_n \tag{9.48}$$
$$\le \sum \left(\frac{1}{2}\log(2\pi e(P_i + N)) - \frac{1}{2}\log 2\pi eN\right) + n\epsilon_n \tag{9.49}$$
$$= \sum \frac{1}{2}\log\left(1 + \frac{P_i}{N}\right) + n\epsilon_n. \tag{9.50}$$

Since each of the codewords satisfies the power constraint, so does their average, and hence

$$\frac{1}{n}\sum_i P_i \le P. \tag{9.51}$$

Since $f(x) = \frac{1}{2}\log(1 + x)$ is a concave function of x, we can apply Jensen's inequality to obtain

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\log\left(1 + \frac{P_i}{N}\right) \le \frac{1}{2}\log\left(1 + \frac{1}{n}\sum_{i=1}^{n}\frac{P_i}{N}\right) \tag{9.52}$$
$$\le \frac{1}{2}\log\left(1 + \frac{P}{N}\right). \tag{9.53}$$

Thus $R \le \frac{1}{2}\log(1 + \frac{P}{N}) + \epsilon_n$, $\epsilon_n \to 0$, and we have the required converse.

Note that the power constraint enters the standard proof in (9.46).

$$R \le \frac{1}{2}\log\left(1 + \frac{P}{N}\right) + \epsilon_n = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) + \frac{1}{n} + P_e^{(n)}R$$

当 $R > C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$ 时, 有 $\frac{1}{n} + P_e^{(n)}R > 0$, 由于 n, R 均为常数, 所以一定有 $P_e^{(n)} \not\to 0$.

8.2 Gaussian Channels with Bandlimited, Parallel, Colored Noise

1. AWGN

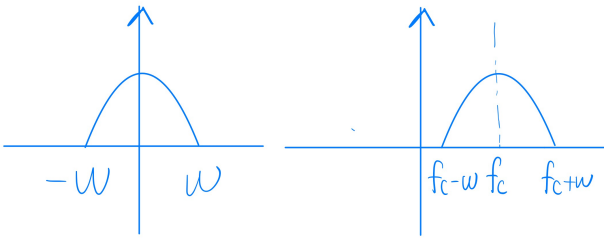
$$Y = X + Z, Z \sim \mathcal{N}(0, N), \mathbb{E}(X^2) \le P \Rightarrow C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

2. Bandwith Limited Channel

Bandwidth: 在正半轴的跨度. 由于实信号的频谱关于 y 轴对称, 所以只考

虑正半轴即可.

如在下图两个信号的频谱图中, 左边信号的 Bandwidth 为 W, 右边信号的 Bandwidth 为 $2W$.



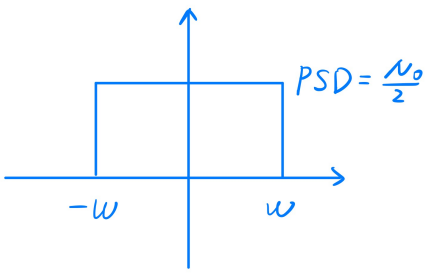
Bandwidth Limited Channel: 在信号传输时有 Bandwidth 限制, 信号的频谱在 W 范围内.

$$Y(t) = X(t) + Z(t)$$

其中信号的能量有限制: $\mathbb{E}(X(t)^2) \leq P$, $Z(t)$ 为 Guassian Noise with power spectral density(PSD) $\frac{N_0}{2}$.

在传输信号时, 对信号做量化: $x[n] = x(nT_s)$, 其中 T_s 为采样周期, $f_s = \frac{1}{T_s}$ 为采样频率. 由 Nyquist 采样定理, 有 $f_s \geq 2W$.

Take $T_s = \frac{1}{2W}$, 则在 T 时间内的采样点的个数为 $2WT$. T 时间内的总能量为 PT. 所以 $\mathbb{E}(X_i^2) = \frac{PT}{2WT} = \frac{P}{2W}$.



Noise 的 power spectral density 为 $\frac{N_0}{2}$ watts/hertz, bandwidth 为 W hertz, 所以噪声的功率为 $\frac{N_0}{2} \cdot 2W = N_0 W$ watts. 所以平均每个噪声 sample 的能量为 $\mathbb{E}(Z_i^2) = \frac{N_0 WT}{2WT} = \frac{N_0}{2}$.

由 1. AWGN 的结论, 我们可以得到每个 sample 的信道容量:

$$C = \frac{1}{2}\log\left(\frac{1+\frac{P}{2W}}{\frac{N_0}{2}}\right) = \frac{1}{2}\log\left(1+\frac{P}{N_0W}\right) \text{ bits per sample}$$

由于 $T = 1s$ 内有 $2W \cdot 1$ 个 sample, 所以信道容量为

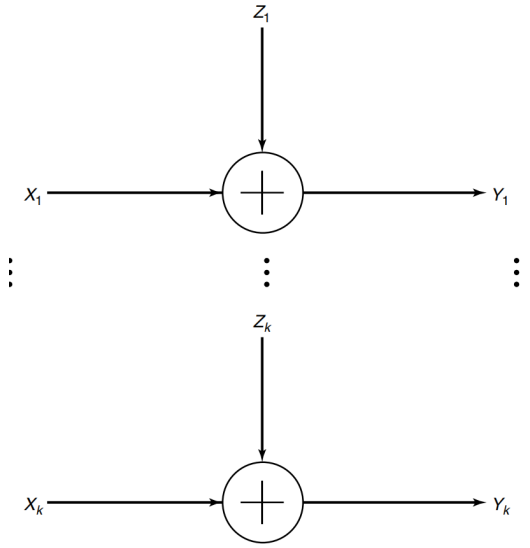
$$C = 2W \cdot \frac{1}{2}\log\left(1+\frac{P}{N_0W}\right) = W\log\left(1+\frac{P}{N_0W}\right) \text{ bits per second}$$

几个极端情况:

1. $P \to \infty$, $C \to +\infty$
2. $W \to \infty$, $C \to \dfrac{P}{N}\log e$

增加传输功率确实可以无限制增加信道容量, 但是增加带宽会达到瓶颈, 这个瓶颈也能和 $SNR = \dfrac{P}{N_0}$ 相对应.

3. Parallel AWGN



如上图, $Y_i = X_i + Z_i, i = 1, .., k$, $Z_i \sim \mathcal{N}(0, N_i), Z_i \perp Z_j, \mathbb{E}\left[\sum_{i=1}^{k} X_i^2\right] \leq P$.

则 Parallel AWGN 信道的容量为

$$C = \max_{f(x_1,\ldots,x_k):\sum_{i=1}^{k}\mathbb{E}(X_i^2)\leq P} I(X_1,\ldots,X_k;Y_1,\ldots,Y_k)$$

假设每个 X_i 的能量上限为 P_i, i.e. $\mathbb{E}(X_i^2) \leq P_i, \sum_{i=1}^{k} P_i \leq P$, 则有

$$\begin{aligned}
I(X_1,\ldots,X_k;Y_1,\ldots,Y_k) &= h(Y_1,\ldots,Y_k) - h(Y_1,\ldots,Y_k|X_1,\ldots,X_k) \\
&= h(Y_1,\ldots,Y_k) - h(Z_1,\ldots,Z_k) \\
&= h(Y_1,\ldots,Y_k) - \sum_{i=1}^{k} h(Z_i) \\
&\leq \sum_{i=1}^{k} (h(Y_i) - h(Z_i)) \qquad \text{(conditioning reduces entropy)} \\
&= \sum_{i=1}^{k} I(X_i;Y_i) \qquad (h(Z_i) = h(Z_i|X_i) = h(Y_i|X_i)) \\
&\leq \sum_{i=1}^{k} \frac{1}{2}\log\left(1+\frac{P_i}{N_i}\right) \qquad \text{(AWGN channel capacity)}
\end{aligned}$$

当且仅当 $Y_i \perp Y_j$, i.e. $X_i \perp X_j$, 且 $X_i \sim \mathcal{N}(0, P_i)$ 时取等. 所以接下来的问题就是如何分配 P_i, 使整体的信道容量最大. 我们可以转化成一下的优化问题:

$$\begin{aligned}
\underset{P_1,\ldots,P_k}{\text{minimize}} \quad & -\frac{1}{2}\sum_{i=1}^{k} \ln\left(1+\frac{P_i}{N_i}\right) \\
\text{subject to} \quad & \sum_{i=1}^{k} P_i = P \\
& -P_i \leq 0, i = 1,\ldots,k
\end{aligned}$$

So let $\lambda \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^k$ respectively be the multiplier of the equality constrain and inequality constrains, so $\boldsymbol{\mu} \succeq 0$. And the Lagrangian function is:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{p},\lambda,\boldsymbol{\mu}) &= -\frac{1}{2}\sum_{i=1}^{k} \ln\left(1+\frac{P_i}{N_i}\right) + \lambda\left(\sum_{i=1}^{k} P_i - P\right) - \boldsymbol{\mu}^T\boldsymbol{p} \\
\frac{\partial \mathcal{L}}{\partial P_i} &= \frac{1}{2}\cdot\frac{-1}{N_i+P_i} + \lambda - \mu_i, \quad i = 1,\ldots,k \\
\frac{\partial \mathcal{L}}{\partial P_i} = 0 \Rightarrow P_i^* &= \frac{1}{2}\cdot\frac{1}{\lambda-\mu_i} - N_i, \quad i = 1,\ldots,k
\end{aligned}$$

we can get the KKT conditions:

$$\begin{cases} \text{Primal feasibility:} & \begin{cases} \sum_{i=1}^{k} P_i^* = P \\ P_i^* \geq 0, \quad i = 1, \ldots, k \end{cases} \\ \text{Dual feasibility:} & \boldsymbol{\mu}^* \succeq 0 \\ \text{Complementary slackness:} & \mu_i^* P_i^* = 0, \quad i = 1, \ldots, k \\ \text{Stationarity:} & \frac{\partial \mathcal{L}}{\partial P_i} = 0 \Rightarrow P_i^* = \frac{1}{2} \cdot \frac{1}{\lambda^* - \mu_i^*} - N_i, \quad i = 1, \ldots, k \end{cases}$$

With KKT condition, we can get the optimal solution:

First of all, we define $\frac{1}{2\lambda^*} \triangleq \nu$. Then we have:

<1> If $P_i^* = 0$, then from the stationarity condition, we have $\mu_i^* = \frac{1}{2\nu} - \frac{1}{2N_i}$. And from the dual feasibility condition, we have $\mu_i^* \geq 0$, so we have $\nu \leq N_i$.

<2> If $P_i^* \neq 0$, then from the complementary slackness condition, i.e. $\mu_i^* = 0$. And from the stationarity condition, we have $P_i^* = \nu - N_i$.

And from the primal feasibility condition, we have $P_i^* \geq 0$, and since $P_i^* \neq 0$, so we have

$$P_i^* = \nu - N_i > 0 \Leftrightarrow \nu < N_i$$

So above all, we can get that

$$P_i^* = \begin{cases} 0, & \nu \leq N_i \\ \nu - N_i, & \nu > N_i \end{cases}$$

Which means that

$$P_i^* = \max\{0, \nu - N_i\} = (\nu - N_i)_+$$

And from the primal feasibility condition, we have

$$\sum_{i=1}^{k} P_i^* = P$$

So with some methods, such as the bisection method, we can get the value of ν with the above equation. And after calculating ν, we can get the optimal

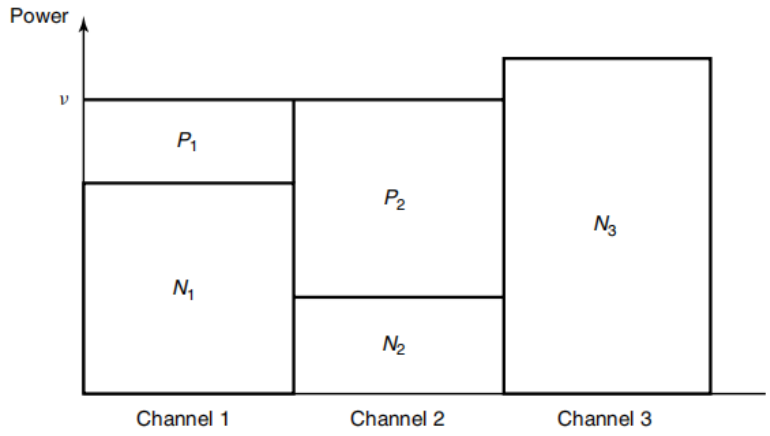
solution of P_i^*.

And the optimal value of the origin problem is

$$\begin{aligned}\text{obj} &= \frac{1}{2}\sum_{i=1}^{k}\ln\left(1+\frac{P_i^*}{N_i}\right)\\ &= \frac{1}{2}\sum_{i=1}^{k}\ln\left(1+\frac{\max\{0,\nu-N_i\}}{N_i}\right)\end{aligned}$$

这个方法也叫做 Water Filling Algorithm. 我们可以用二分或其他方法得到 ν, 然后由 $P_i=(\nu-N_i)_+$ 得到 P_i, 最后找到那个能够满足 $\sum_{i=1}^{k}P_i=P$ 的 ν 即可.

可以按照下图理解, 其中 ν 为水线, noise(N_i) 为石头, 水深为 P_i. N_i 越大, P_i 越小.



4. AWGN with correlated noise

Lemma: Hadamard's Inequality. 对于任意 n 阶半正定方阵 $A=(a_{ij})_{n\times n}$:

$$|A|\le\prod_{i=1}^{n}a_{ii}$$

当 A 为对角阵时取等.

之前噪声都是独立的, 若增加噪声直接的相关性: $Z^n=(Z_1,\dots,Z_n)\sim\mathcal{N}(0,\mathbf{K}_Z)$ 由于 Z_i 有相关性, 所以考虑 X^n 也有相关性, coviance matrix

为 $\mathbf{K}_X$. 也可以理解为**使 channels with memory**. 在 with memory 的情况下, 能量约束变为了 $\frac{1}{n}\mathbb{E}(X_i^2) \le P \Rightarrow \frac{1}{n}\mathrm{Tr}(\mathbf{K}_X) \le P$.

$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(X_i+Z_i, X_j+Z_j) = \mathrm{Cov}(X_i, X_j)+\mathrm{Cov}(Z_i, Z_j) \Rightarrow \mathbf{K}_Y = \mathbf{K}_X+\mathbf{K}_Z$

所以

$$\begin{aligned} I(X^n; Y^n) &= h(Y^n) - h(Y^n|X^n) \\ &= h(Y^n) - h(Z^n) \\ &\le \frac{1}{2}\log\left((2\pi e)^n|\mathbf{K}_Y|\right) - \frac{1}{2}\log\left((2\pi e)^n|\mathbf{K}_Z|\right) \end{aligned}$$

由于 $\mathbf{K}_Z$ 是 coviance matrix, 一定是对称阵, 所以可以正交对角化, 写作 $\mathbf{K}_Z = Q\Lambda Q^\top$, where $Q^\top Q = I$, so:

$$\begin{aligned} |\mathbf{K}_X + \mathbf{K}_Z| &= |\mathbf{K}_X + Q\Lambda Q^\top| \\ &= |Q(Q^\top\mathbf{K}_X Q + \Lambda)Q^\top| \\ &= |Q^\top\mathbf{K}_X Q + \Lambda| \end{aligned}$$

Let $A = Q^\top\mathbf{K}_X Q$, then:

$$\mathrm{Tr}(A) = \mathrm{Tr}(Q^\top\mathbf{K}_X Q) = \mathrm{Tr}(\mathbf{K}_X QQ^\top) = \mathrm{Tr}(\mathbf{K}_X) \le nP$$

由 Hadamard inequality:

$$|\mathbf{K}_X + \mathbf{K}_Z| = |A + \Lambda| \le \prod_{i=1}^{n}(\lambda_i + a_{ii})$$

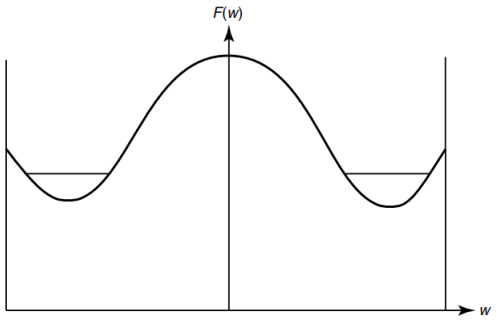
和往常一样, 应用 Water Filling Algorithm, $a_{ii} = (\nu - \lambda_i)_+$, ν 使得 $\mathrm{Tr}(A) = \sum_{i=1}^{n}(\nu - \lambda_i)_+ = nP$ 即可.
所以最终的信道容量为

$$C = \frac{1}{2}\log\left(\frac{\prod_{i=1}^{n}(\lambda_i + (\nu - \lambda_i)_+)}{|\mathbf{K}_Z|}\right)$$

其中 $\sum_{i=1}^{n}(\nu - \lambda_i)_+ = nP$, λ_i 为 $\mathbf{K}_Z$ 的特征值, $A = Q^\top\mathbf{K}_X Q$, Q 为 $\mathbf{K}_Z$ 正交对角化矩阵.

5. Gaussian Channels with Colored Noise

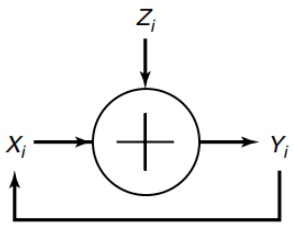
Channels in which the noise forms a stationary stochastic process, the input signal should be chosen to be a Gaussian process with a spectrum that is large at frequencies where the noise spectrum is small.



用频域上的 water-filling, ν 为水线, 满足 $\int(\nu - N(f))_+ df = P$ 即可.

$$C = \int_{-\pi}^{\pi} \frac{1}{2} \log \left(1 + \frac{(\nu - N(f))_+}{N(f)}\right) df$$

6. Feedback: Y_{i+1} 的产生有 Y_i 参与



无论是否 feedback, Channel capacity 均为 $C = \frac{1}{2} \log \left(\frac{\mathbf{K}_{X+Z}}{\mathbf{K}_Z}\right)$, 但是 without feedback 有 $\mathbf{K}_{X+Z} = \mathbf{K}_X + \mathbf{K}_Z$.

Memoryless Channel capacity: C_{noFB}, Channel without feedbeck C_{FB}, 满足 bound:

$$C_{FB} \leq C_{noFB} + \frac{1}{2}$$

$$C_{FB} \leq 2C_{noFB}$$

Without memory $p_e^{(n)}$ exponentially $\searrow 0$, with memory double exponentially $\searrow 0$, 可以更快的收敛到 0, more robust.

第九章　Rate Distortion Theory*

Book Chapter10. (P327)

率失真理论

9.1 Quantization

例 9.1.1. *1-bit Gaussian:* $X \sim \mathcal{N}(0, \sigma^2)$*,* 衡量指标*: MSE=*$\mathbb{E}\left[(X-\hat{X})^2\right]$*. Find* $\hat{X}$ *s.t.* $\hat{X}$ 有两个取值*,* 使得 *MSE* 最小*.*

假设 $\hat{X} \in \{a_1, a_2\}, \hat{X} = \begin{cases} a_1 & X \le b \\ a_2 & X > b \end{cases}$ 则

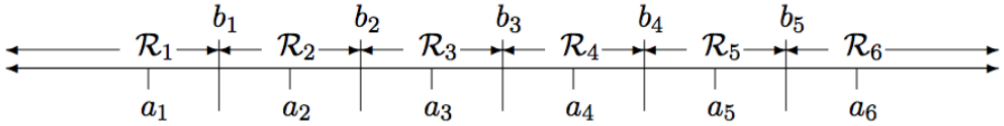
$$MSE = \mathbb{E}[(X-\hat{X})^2] = \int_{-\infty}^{b} f(x)(x-a_1)^2 dx + \int_{b}^{+\infty} f(x)(x-a_2)^2 dx$$

可以解得 $a_1 = -\sqrt{\frac{2}{\pi}}\sigma, a_2 = \sqrt{\frac{2}{\pi}}\sigma, b = \frac{a_1+a_2}{2} = 0.$

i.e.

$$\hat{X} = \begin{cases} -\sqrt{\frac{2}{\pi}\sigma} & X \le 0 \\ \sqrt{\frac{2}{\pi}\sigma} & X > 0 \end{cases}$$

例 9.1.2. *Quantization Region(1D): Aim: Given pdf* $f_U(u)$ *and alphabet size* M*, choose* $\{R_j, 1 \le j \le M\}$ *and* $\{a_j, 1 \le j \le M\}$ *to miminize MSE.*



如图所示, 我们要选择 a_i, b_i, 分成两步解决:

1. Given a_j, choose b_j such that $\mathbb{E}\left[(U-V)^2\right]$ is minimized.

$$\begin{aligned}\mathbb{E}\left[(U-V)^2\right] &= \sum_{j=1}^{M}\int_{\mathbf{R_J}} f_U(u)\left(u-a_j\right)^2 du \\ &= \sum_{j=1}^{M}\int_{b_{j-1}}^{b_j} f_U(u)\left(u-a_j\right)^2 du \\ &= \cdots + \int_{b_{j-1}}^{b_j} f_U(u)\left(u-a_j\right)^2 du + \int_{b_j}^{b_{j+1}} f_U(u)\left(u-a_{j+1}\right)^2 du + \cdots\end{aligned}$$

Let $\dfrac{\partial \mathbb{E}\left[(U-V)^2\right]}{\partial b_j} = 0$, we have $\left(\dfrac{\partial \int_{q(x)}^{g(x)} f(u)du}{\partial x} = f(g(x))\dfrac{\partial g(x)}{\partial x} - f(q(x))\dfrac{\partial q(x)}{\partial x}\right)$

$$\begin{aligned} f_U\left(b_j\right)\left(b_j-a_j\right)^2 - f_U\left(b_j\right)\left(b_j-a_{j+1}\right)^2 &= 0 \\ 2b_j\left(a_{j+1}-a_j\right) &= a_{j+1}^2 - a_j^2 \\ b_j &= \frac{a_j+a_{j+1}}{2}\end{aligned}$$

2. Choice of $\{a_j\}$ for given $\mathcal{R}_j$

$$\begin{aligned}\text{MSE} = \mathbb{E}\left[(U-V)^2\right] &= \int_{-\infty}^{\infty} f(u)(u-v)^2 du \\ &= \sum_{j=1}^{M}\int_{\mathcal{R}_j} f(u)\left(u-a_j\right)^2 du \\ &= \sum_{j=1}^{M}\int_{\mathcal{R}_j} f(u)\left(u^2-2a_ju+a_j^2\right) du\end{aligned}$$

Let $\frac{\partial \mathbb{E}\left[(U-V)^2\right]}{\partial a_j} = 0$, we have

$$-2\int_{\mathcal{R}_j} f(u)u du + 2\int_{\mathcal{R}_j} f(u)a_j du = 0$$
$$\Longrightarrow a_j = \frac{\int_{R_j} uf(u)du}{\int_{R_j} f(u)du}$$

总结步骤: 1. Choose $a_1 < a_2 < \cdots < a_M$

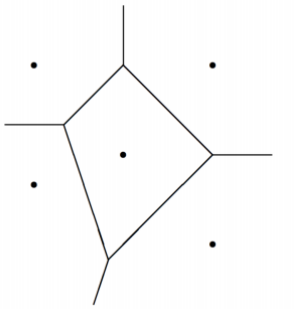
2. set $b_j = \frac{a_j + a_{j+1}}{2}$ for $1 \le j \le M-1$

3. Set $a_j = \frac{\int_{\mathcal{R}_j} uf(u)du}{\int_{\mathcal{R}_j} f(u)du}$, where $\mathcal{R}_j = (b_{j-1}, b_j]$ for $1 \le j \le M-1$

4. Iterate on 2 and 3 until improvement is negligible.

It find local min, not necessarily global min.

例 9.1.3. *Quantization Region(2D): 和 1D 类似.*
Given $\left\{\left(a_j, a_j'\right)\right\}$*, how to choose* $\{\mathcal{R}_j\}$*?*



Step1: 选择边界的划分方式:

1. The square error is $(u-a_j)^2 + \left(u'-a_j'\right)^2$, the point $\left\{a_j, a_j'\right\}$ which is the closest to (u, u') in Euclidean distance should be chosen.

2. $\{\mathcal{R}_j\}$ contains points that are closer to $\left(a_j, a_j'\right)$ than any other representation points, i.e., Voronoi regions.

Voronoi region 的划分: 在 CG 等领域已经有很多现成的算法, e.g. https://blog.csdn.net/yolon30
https://zhuanlan.zhihu.com/p/459884570

Step2: 选择中心点的位置:

Given a set of Voronoi region, how to find the $\{a_j, a'_j\}$?

Choose $\{a_j, a'_j\}$ to be the conditional means within those regions.

和 1D 一样, 反复迭代直至收敛.

The symbol + represents the updated point, and • the original point

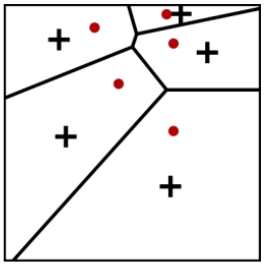


Figure: Iteration 1

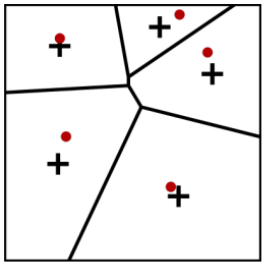


Figure: Iteration 2

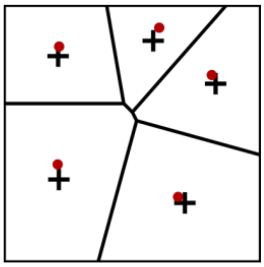


Figure: Iteration 3

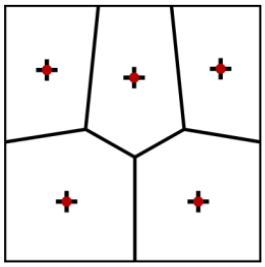
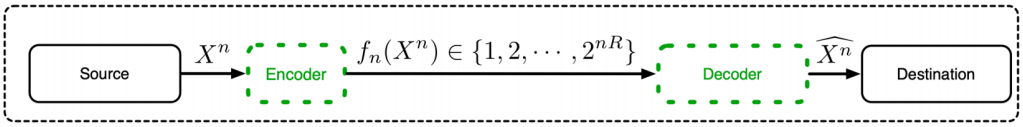


Figure: Iteration 4

9.2 Rate Distortion Theory

Rate Distortion Encoder and Decoder:



Rate-distortion theory describes the trade-off between lossy compression rate and the resulting distortion.

Lossless Source coding: Recover source data X without error

Lossy source coding: Recover source with some error and distortion

- A $(2^{nR}, n)$-*rate distortion code* consists of an encoding function

$$f_n : \mathcal{X}^n \to \{1, 2, \cdots, 2^{nR}\},$$

and a decoding (reproduction) function,

$$g_n : \{1, 2, \cdots, 2^{nR}\} \to \hat{\mathcal{X}}^n.$$

- The distortion associated with the $(2^{nR}, n)$ code is defined as

$$D = E[d(X^n, g_n(f_n(X^n))),$$

where the expectation is with respect to the probability distribution on $\mathcal{X}$,

$$D = \sum p(x^n) d(x^n, g_n(f_n(x^n))).$$

- The set of n-tuples $g_n(1), g_n(2), \cdots, g_n(2^{nR})$, denoted by $\hat{X}^n(1), \hat{X}^n(2), \cdots, \hat{X}^n(2^{nR})$ constitutes the *codebook* and $f_n^{-1}(1), \cdots f_n^{-1}(2^{nR})$ are the associated *assignment regions.*

其中 f 是 Encoder, g 是 Decoder, $g_n\left(f_n(X^n)\right)$ 即为 $\hat{X}^n$.

$f_n^{-1}(1), \ldots, f_n^{-1}(2^{nR})$: 一个区域量化成一个值.

D: distortion 上界, R: 压缩速率 (每个 symbol 用 R bit 表示).

$D = 0 \Rightarrow$ 无损压缩 (Lossless compression), $R = H(X) \Rightarrow$ 达到压缩极限.

D 越大, R 越小: 允许的误差越大, 就可以用更少的 bit 进行压缩, 所以压缩速率可以写作 $R(D)$.

Main Results

Theorem: The rate distortion function for an i.i.d. source X with distributed $p(x)$ and bounded distortion function $d(x,\hat{x})$ is equal to the associated information rate distortion function. Thus,

$$R(D) = R^{(I)}(D) = \min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x})\le D} I(X;\hat{X})$$

is the minimum achievable rate at distortion D.

Theorem: The rate distortion function for a Bernoulli(p) source with Hamming distortion is given by

$$R(D) = \begin{cases} H(p) - H(D), & 0 \le D \le \min\{p, 1-p\} \\ 0, & D > \min\{p, 1-p\} \end{cases}$$

Theorem: The rate distortion function for a $\mathcal{N}(0,\sigma^2)$ source with squared-error distortion is given by

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \le D \le \sigma^2 \\ 0, & D > \sigma^2\} \end{cases}$$

The key idea of the Proof:

- Converse: Find a lower bound on $I(X;\hat{X})$
- Achievability: Show that the lower bound is achievable

这里 W.L.O.G, $p \le \frac{1}{2}$.

$R(D) = 0$: 所有 symbol 都压缩成一个 constant $\Rightarrow$ D 太大了, 盲猜都能使得 distortion 小于 D.

"For a $\mathcal{N}(0,1)$ source": 假设的信源分布.

$X \sim Bern\,(p) \Rightarrow G(D) = H(X) = H(p)$.

$R(D) = \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right) \Rightarrow D = \sigma^2 2^{-2R}$.

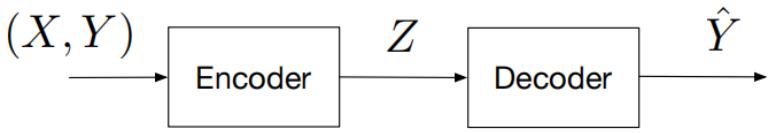
Some notes:

Given $p(x)$, $I(X;\hat{X})$ 关于 $p(\hat{x}|x)$ convcave.

$(\tilde{x},\tilde{y}) \sim p(x)p(y) \Rightarrow \Pr\left[(\tilde{x},\tilde{y}) \in A_\epsilon^{(n)}(P_{X,Y})\right] \to 2^{-nI(X;Y)}$.

Other details see in ppt.

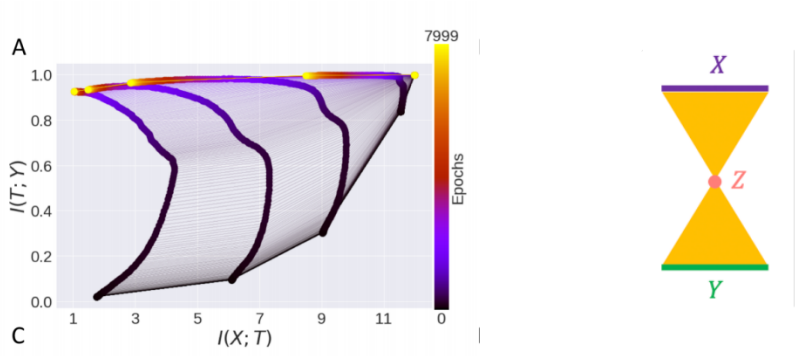
第十章　Information Bottleneck*



(X,Y): Source input, Z: extracted features, $\hat{Y}$: predicted label.

Information complexity: $I(X;Z)$, information utility: $I(Z;Y)$.

Trade-off: $L = \min_{p(z|x)} I(X;Z) - \beta I(Z;Y)$, 其中后半部分是为了防止过拟合.



若将 Encoder 和 Decoder 都换成 neural network, 则会出现上图的现象.

从左往右四条线是网络的深度, 在同一层中, 互信息会先增大后降低. 感性理解: 学习的过程中需要先把书学厚, 然后再学薄.

Deep Variational Information Bottleneck

$$\boxed{L = \min_{p(z|x)} I(X;Z) - \beta I(Z;Y)}$$

Rewrite: Need $p(x,y), r(z), p(z|x), q(y|z)$ to compute

$$I(Z;Y) \geq H(Y) + \int p(x,y)p(z|x)\log q(y|z)dxdydz$$

$$I(Z;X) \leq \int p(x,z)\log\frac{p(z|x)}{r(z)}dydz$$

Use Monte Carlo sampling: $p(x,y) = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}(x)\delta_{y_n}(y)$, we have

$$\min_{p(z|x)} I(X;Z) - \beta I(Z;Y) \leq \min_{p(z|x)} \int p(x,z)\log\frac{p(z|x)}{r(z)}dxdz - \beta E_{p(y,z)}\log q(y|z)$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\Big[\int p(z|x_n)\log\frac{p(z|x_n)}{r(z)}dz - \beta\int p(z|x_n)\log q(y_n|z)dz\Big]$$

$p(z|x)$ 为训练得到的 encoder, $p(y|z)$ 为训练得到的 decoder. $p(x,y)$ 无法直接得到, 用 Monte Carlo 从样本中采样估计.

VAE 中重参数化, ELBO 等 details in PPT.

第十一章　Appendix

11.1　Axiomatic definition of entropy

熵的公理定义

If we assume certain axioms for our measure of information, we will be forced to use a logarithmic measure such as entropy. Shannon used this to justify his initial definition of entropy.

If a sequence of symmetric functions $H_m(p_1, p_2, \ldots, p_m)$ satisfies the following properties:

1. Normalization: $H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$
2. Continuity: $H_2(p, 1-p)$ is a continuous function of p
3. Grouping: $H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$

prove that H_m must be of the form

$$H_m(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i, \quad m = 2, 3, \ldots$$

Solution

Notations:

$$f(m) \triangleq H_m\left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right)$$

$$S_k \triangleq \sum_{i=1}^{k} p_i \quad k = 1, 2, \ldots, m$$

From the grouping property, we can get its extension:

$$\begin{aligned}
&H_m\left(p_1, p_2, p_3, \ldots, p_m\right) \\
&= H_{m-1}\left(S_2, p_3, \ldots, p_m\right) + S_2 H_2\left(\frac{p_1}{S_2}, \frac{p_2}{S_2}\right) \\
&= H_{m-2}\left(S_3, p_4, \ldots, p_m\right) + S_3 H_2\left(\frac{p_1+p_2}{S_3}, \frac{p_3}{S_3}\right) + S_2 H_2\left(\frac{p_1}{S_2}, \frac{p_2}{S_2}\right) \\
&= \ldots \\
&= H_{m-(k-1)}(S_k, p_{k+1}, \ldots, p_m) + \sum_{i=2}^{k} S_i H_2\left(\frac{S_{i-1}}{S_i}, \frac{p_i}{S_i}\right) \\
&= H_{m-(k-1)}\left(S_k, p_{k+1}, \ldots, p_m\right) + S_k H_k\left(\frac{p_1}{S_k}, \frac{p_2}{S_k}, \ldots, \frac{p_k}{S_k}\right)
\end{aligned}$$

The last equality(blue part) can be obtained by expanding $H_k\left(\frac{p_1}{S_k}, \frac{p_2}{S_k}, \ldots, \frac{p_k}{S_k}\right)$.

And using this, we can get that

$$\begin{aligned}
f(mn) &= H_{mn}\left(\frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}\right) \\
&= H_{mn-(n-1)}\left(S_n, \frac{1}{mn}, \ldots, \frac{1}{mn}\right) + S_n H_n\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \\
&= H_{mn-2(n-1)}\left(S_n, S_n, \frac{1}{mn}, \ldots, \frac{1}{mn}\right) + 2 S_n H_n\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \\
&= \ldots \\
&= H_{mn-m(n-1)}\left(S_n, S_n, \ldots, S_n\right) + m S_n H_n\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \\
&= H_m\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) + m \frac{1}{m} H_n\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \\
&= f(m) + f(n)
\end{aligned}$$

And since we have the Continuity property, i.e. $H_2(p, 1-p)$ is a continuous function of p, from the property and provement of Cauchy function, we could get that

$$f(m) = \log_a m$$

And from the Normalization property, we have

$$f(2) = 1$$

So we can get that $a = 2$.

So above all, we have proved that

$$f(m) = H_m\left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right) = \log_2 m$$

And then prove $H_2(p, 1-p) = -p\log_2 p - (1-p)\log_2(1-p)$:

1. When p is rational, suppose $p = \frac{r}{s}$, where r, s are integers, $s > 1, 0 \leq r \leq s$, $\gcd(r, s) = 1$. Then from the Grouping property and its extension, we have:

$$\begin{aligned}
f(s) &= H_s\left(\frac{1}{s}, \ldots, \frac{1}{s}\right) = H_s\left(\underbrace{\frac{1}{s}, \ldots, \frac{1}{s}}_{r}, \underbrace{\frac{1}{s}, \ldots, \frac{1}{s}}_{s-r}\right) \\
&= H_{s-(r-1)}\left(\frac{r}{s}, \underbrace{\frac{1}{s}, \ldots, \frac{1}{s}}_{s-r}\right) + \frac{r}{s}H_r\left(\frac{1}{r}, \ldots, \frac{1}{r}\right) \\
&= H_{s-(r-1)-(s-1)}\left(\frac{r}{s}, \frac{s-r}{s}\right) + \frac{s-r}{s}H_{s-r}\left(\frac{1}{s-r}, \ldots, \frac{1}{s-r}\right) + \frac{r}{s}H_r\left(\frac{1}{r}, \ldots, \frac{1}{r}\right) \\
&= H_2(\frac{r}{s}, \frac{s-r}{s}) + \frac{s-r}{s}f(s-r) + \frac{r}{s}f(r)
\end{aligned}$$

And since $p = \frac{r}{s}$, so we can get that

$$\begin{aligned}
H_2(p, 1-p) &= H_2(\frac{r}{s}, \frac{s-r}{s}) \\
&= f(s) - \frac{s-r}{s}f(s-r) - \frac{r}{s}f(r) \\
&= \log_2 s - (1-p)\log_2\left(s(1-p)\right) - p\log_2(sp) \\
&= -p\log_2 p - (1-p)\log_2(1-p)
\end{aligned}$$

2. When p is irrational, since we have the Continuity property, so we can also get that

$$H_2(p, 1-p) = -p\log_2 p - (1-p)\log_2(1-p)$$

So above all, we have prove that $\forall p \in [0,1]$, we have

$$H_2(p, 1-p) = -p\log_2 p - (1-p)\log_2(1-p)$$

Then we use the induction to prove that, suppose that

$$H_k(p_1, p_2, \ldots, p_k) = -\sum_{i=1}^{k} p_i \log p_i, \forall k = 1, 2, \ldots, m$$

Then for $k = m+1$, we have

$$\begin{aligned}
H_k(p_1, p_2, \ldots, p_k) &= H_{m+1}(p_1, p_2, \ldots, p_{m+1}) \\
&= H_m(p_1 + p_2, p_3, \ldots, p_{m+1}) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) \\
&= \left(-(p_1+p_2)\log(p_1+p_2) - \sum_{i=3}^{m+1} p_i \log p_i\right) - (p_1+p_2)\left(\frac{p_1}{p_1+p_2}\log\frac{p_1}{p_1+p_2} + \right. \\
&= -(p_1+p_2)\log(p_1+p_2) - \sum_{i=3}^{m+1} p_i \log p_i - p_1\log p_1 - p_2 \log p_2 + (p_1+p_2)(\frac{p_1}{p_1+} \\
&= -\sum_{i=1}^{m+1} p_i \log p_i \\
&= -\sum_{i=1}^{k} p_i \log p_i
\end{aligned}$$

So we have proved that $\forall m$, we have

$$H_m(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i.$$

So above all, we have proved that [1]

$$H_m\left(p_1, p_2, \ldots, p_m\right) = -\sum_{i=1}^{m} p_i \log p_i, \quad m = 2, 3, \ldots$$

[1]The proof above has referenced from A Rényi.Wahrscheinlichkeitsrechnung, mit einem Anhang über Informationstheorie. Veb Deutscher Verlag der Wissenschaften, Berlin, 1962.