信息论与编码

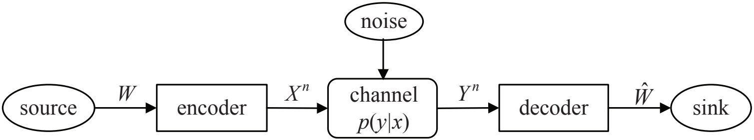
马啸
maxiao@mail.sysu.edu.cn

计算机学院
中山大学

2021 年春季学期

1. Channel coding
   - Mutual information
   - Channel capacity
   - Channel coding theorem

General Framework of Channel Coding



$$
\begin{aligned}
\text{Message Set} \quad & \mathcal{M} = \{1, \ldots, M_n\}, W \in \mathcal{M} \\
\text{Encoding} \quad & f : \mathcal{M} \to \mathcal{X}^n \\
& \qquad w \mapsto f(w) \\
\text{Decoding} \quad & g : \mathcal{Y}^n \to \mathcal{M} \\
& \qquad y^n \mapsto \hat{w} \\
\text{Probability of Error} \quad & \varepsilon^{(n)} = \frac{1}{M} \sum_{w=1}^{M} \lambda_w \\
& \lambda_w = \Pr(g(Y^n) \neq w | X^n = f(w)) \\
\text{Maximal Probability of Error} \quad & \lambda^{(n)} = \max_{w \in \mathcal{M}} \lambda_w \\
\text{Coding Rate} \quad & R = \frac{\log M}{n} \text{ bits per transmission}
\end{aligned}
$$

Mutual Information

$$\mathrm{I}(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \mathrm{E}_{P(x,y)}[\log \frac{P(X,Y)}{P(X)P(Y)}].$$

1. $\mathrm{I}(X;Y) = H(X) - H(X|Y)$; $\mathrm{I}(X;Y) = H(Y) - H(Y|X)$;
2. $\mathrm{I}(X;Y|Z) = H(X|Z) - H(X|Y,Z)$;
3. $\mathrm{I}(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} \mathrm{I}(X_i, Y|X_1, X_2, \ldots, X_{i-1})$;
4. $\mathrm{I}(X;Y) = D(p(x,y) \| p(x)p(y))$;
5. $\mathrm{I}(X;Y) \geq 0$ with equality iff X and Y are independent.
6. $\mathrm{I}(X;Y|Z) \geq 0$ with equality iff X and Y are conditionally independent given Z.
7. $\mathrm{I}(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.
8. If $X \to Y \to Z$ (i.e., X, Y, Z forms a Markov chain), then $\mathrm{I}(X;Y) \geq \mathrm{I}(X;Z)$.
9. If $X \to Y \to Z$, then $\mathrm{I}(X;Y|Z) \leq \mathrm{I}(X;Y)$.

Channel capacity

Definition 1

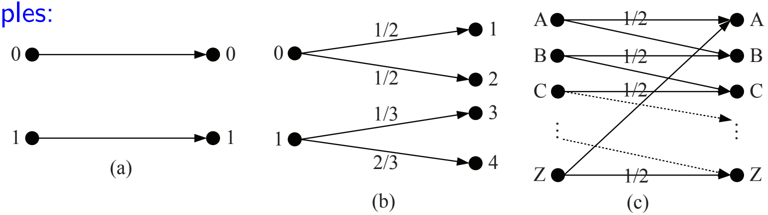
The information channel capacity of a DMC is

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

where the maximum is taken over all possible input distributions $p(x)$.

- $C \geq 0$
- $C \leq \log|\mathcal{X}|$, $C \leq \log|\mathcal{Y}|$
- $\mathrm{I}(X;Y)$ is a continuous function of $p(x)$.
- $\mathrm{I}(X;Y)$ is a concave function of $p(x)$.

Examples:



a. Noiseless Binary Channel:

$$C = 1 \text{ bits, achieved by } p(x) = (1/2, 1/2).$$

b. Noisy Channel with Non-overlapping Outputs:

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

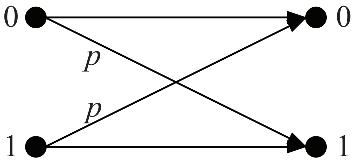
$$C = 1 \text{ bits, achieved by } p(x) = (1/2, 1/2).$$

c. Noisy Typewriter:

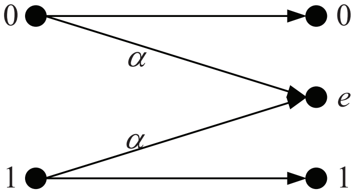
$$\mathrm{I}(X;Y) = H(Y) - H(Y|X) = H(Y) - 1,$$

$$C = \log 26 - 1 = \log 13 \text{ bits, achieved by uniform distribution.}$$

Examples:



(d)



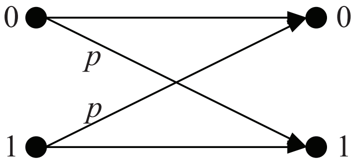
(e)

d. Binary Symmetric Channel (BSC):

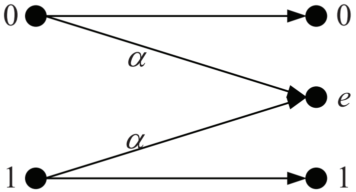
$$
\begin{aligned}
\mathrm{I}(X;Y) &= H(Y) - H(Y|X) = H(Y) - \sum p(x)H(Y|X=x) \\
&= H(Y) - \sum p(x)H(p) = H(Y) - H(p) \\
&\leq 1 - H(p)
\end{aligned}
$$

$$C = 1 - H(p) \text{ bits, achieved by } p(x) = (1/2, 1/2).$$

Examples:



(d)



(e)

e. Binary Erasure Channel (BEC):

$$\begin{aligned}
\mathrm{I}(X;Y) &= H(X) - H(X|Y) = H(X) - \sum p(y) H(X|Y=y) \\
&= H(X) - p(e) H(X|Y=e) = H(X) - \alpha H(X) \\
&= (1-\alpha) H(X)
\end{aligned}$$

$$C = 1 - \alpha \text{ bits}, \text{ achieved by } p(x) = (1/2, 1/2).$$

Channel coding theorem

The code characterized by the encoding f and decoding g is referred to as an (M, n) block code.

Definition 2

A rate R is said to be achievable if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the maximal probability of error $\lambda^{(n)}$ tends to 0 as $n \to \infty$.

Theorem 3 (The Channel Coding Theorem)

1. *For a DMC, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximal probability of error $\lambda^{(n)} \to 0$.*
2. *Conversely, any rate above capacity C cannot be achievable. Equivalently, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.*

Channel coding theorem

我们先以BEC为例，介绍一种可以逼近容量的编码方案。

我们已经知道，BEC信道的容量是 $1-\alpha$，其中 α 是删除概率。假设我们要传输 k 比特。

情形 1： 若有反馈，即，接收端的状态及时准确告知发送端。

设 u_i， $i \geq 0$ 是待传输的比特序列。

- 传输方案：传输 u_i，若接收到删除，则重传，直到正确接收为止。
- 码率：若共用 n 次信道，其中 k_n 次正确接收，则根据强大数定律，可以证明 $\frac{k_n}{n} \to 1-\alpha$， $n \to \infty$。

Channel coding theorem

情形 2： 无反馈，称之为前向纠错(FEC, forward error correction)。

- 输入：$(u_0, u_1, \cdots, u_{k-1}) \triangleq u$；
- 输出：$u \cdot G \triangleq c$，其中 G 是 $k \times n$ 的矩阵，称之为生成矩阵，收发两端都已知的。；

我们证明，只要 $k/n = R < 1 - \alpha$, 存在 G，使得正确恢复 u 的概率接近于1。

Channel coding theorem

为证明存在性，我们随机产生一个矩阵 G，其中每个元素都是独立同分布的二元均匀比特。我们由大数定律知道，当 C 在信道中传输时，有非常高的概率得知 $n(1-\alpha-\epsilon)$ 个比特可以正确接收。若 c_j 是正确接收的，则我们有方程：

$$\sum_{i=0}^{k-1} u_i g_{ij} = c_j$$

上述事实相当于说，我们在接收端可以看到 $n(1-\alpha-\epsilon)$ 个方程构成的线性方程组，其中 u 是未知的向量。

简记之，$u\tilde{G} = \tilde{c}$，其中 $\tilde{c}$ 表示正确接收的向量，长度 $\geq n(1-\alpha-\epsilon)$。而 $\tilde{G}$ 是 G 中对应列构成的子矩阵。

思考题：线性方程组有唯一解的条件是什么？

Channel coding theorem

$\tilde{G}$ 是否是行满秩的？为记号简单表现，不妨记

$$\tilde{G} = \begin{pmatrix} g_{00} & g_{01} & \cdots & g_{0,\tilde{n}-1} \\ g_{10} & g_{11} & \cdots & g_{1,\tilde{n}-1} \\ \cdots & & & \\ g_{k-1,0} & g_{k-1,1} & \cdots & g_{k-1,\tilde{n}-1} \end{pmatrix}$$

$\tilde{G}$ 行不满秩，等价于存在不全为0的 $x \in F_2^k$，使得 $x\tilde{G} = (0, 0, \cdots, 0)$。

对于此 x，上式成立的概率是 $2^{-\tilde{n}}$。

$$\begin{aligned} \Pr\{\mathrm{Rank}(\tilde{G} < k)\} &\le (2^k - 1)2^{-\tilde{n}}(\text{并集限}) \\ &\le 2^{-n(\frac{\tilde{n}}{n} - R)} \end{aligned}$$

由于 $\frac{\tilde{n}}{n} \ge (1 - \alpha - \epsilon)$，而 $R < 1 - \alpha$，所以可以选择 ϵ 使得上面的指数严格大于 0，因而概率 $\to 0$。

思考题：我们能否证明存在稀疏矩阵逼近BEC的容量？

Joint typical sequences

Definition 4

The set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$ with respect to the distribution $P(x, y)$ is the set of n-sequences with empirical entropies ϵ-close to the true entropies:

$$
\begin{aligned}
A_\epsilon^{(n)} \quad = \quad & \Big\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\
& \left| \frac{1}{n} \log P(x^n) - H(X) \right| \leq \epsilon, && (1) \\
& \left| \frac{1}{n} \log P(y^n) - H(Y) \right| \leq \epsilon, && (2) \\
& \left| \frac{1}{n} \log P(x^n, y^n) - H(X, Y) \right| \leq \epsilon \Big\}, && (3)
\end{aligned}
$$

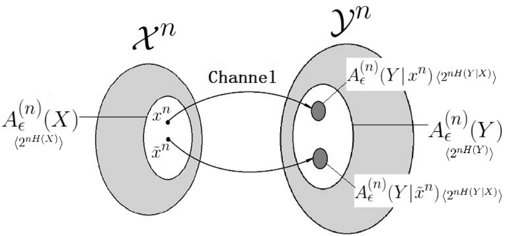
where $P(x^n, y^n) = \prod_{i=1}^{n} P(x_i, y_i)$.

Joint typical sequences

Let (X^n, Y^n) be drawn i.i.d. according to $p(x^n, y^n) = \prod_i^n p(x_i, y_i)$.

1. $\Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
2. $|A_\epsilon^{(n)}| \leq 2^{n[H(X,Y)+\epsilon]}$.
3. If $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same marginals as $p(x^n, y^n)$, i.e., $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then
   - $\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{-n[\mathrm{I}(X;Y)-3\epsilon]}$.
   - $\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \geq (1-\epsilon)2^{-n[\mathrm{I}(X;Y)+3\epsilon]}$, for sufficiently large n.

Joint typical sequences



There are about $2^{nH(X)}$ typical $\mathbf{X}$ sequences.
There are about $2^{nH(Y)}$ typical $\mathbf{Y}$ sequences.
There are about $2^{nH(X,Y)}$ jointly typical $(\mathbf{X}, \mathbf{Y})$ sequences.

$$2^{nH(Y)} \Big/ 2^{nH(Y|X)} = 2^{nI(X;Y)}.$$

Let the distribution on $\mathcal{X}$ be fixed, say $P(x)$.

(1). Code generation. Generate a $(2^{nR}, n)$ code at random according to $P(x)$. We exhibit the 2^{nR} codewords as the rows of a matrix:

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

Each entry in this matrix is generated i.i.d. according to $P(x)$. Thus the probability that we generate a particular code $\mathscr{C}$ is

$$\Pr(\mathscr{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} P(x_i(w)).$$

The code $\mathscr{C}$ is revealed to both sender and receiver.

(2). Encoding. A message W is chosen according to a uniform distribution

$$\Pr\{W = w\} = 2^{-nR}, w \in \mathcal{W} = \{1, 2, \ldots, 2^{nR}\}.$$

The chosen message w is encoded to the w-th row of the codeword matrix, i.e., $f(w) = x^n(w) = (x_1(w), x_2(w), \ldots, x_n(w))$. The codeword $f(w)$ is sent over the channel.

(3). Receiving. The receiver receives an n-sequence y^n with

$$P(y^n|f(w)) = \prod_{i=1}^{n} P(y_i|x_i(w)).$$

(4). Decoding. The receiver guesses which message was sent by using typical set decoding method. The receiver declares that the index $\hat{w}$ was sent if there exists a unique $\hat{w}$ such that $(f(w), y^n)$ is jointly typical. If no such $\hat{w}$ exists, then an error is declared.

(5). Analysis of error. The error event $\{\hat{W} \neq W\}$ is denoted by $\mathcal{E}$. Then the average probability of error is calculated as follows.

$$
\begin{aligned}
\Pr\{\mathcal{E}\} &= \Pr\{\hat{W} \neq W\} \\
&= \sum_{\mathscr{C}} \Pr\{\mathscr{C}\} \varepsilon^{(n)}(\mathscr{C}) \\
&= \sum_{\mathscr{C}} \Pr\{\mathscr{C}\} \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w\{\mathscr{C}\} \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathscr{C}} \Pr\{\mathscr{C}\} \lambda_w(\mathscr{C}) \\
&= \sum_{\mathscr{C}} \Pr\{\mathscr{C}\} \lambda_1(\mathscr{C}) \\
&= \Pr\{\mathcal{E} | W = 1\}.
\end{aligned}
$$

(5). Analysis of error. Let $\mathbf{Y}$ be the received sequence corresponding to the transmitted message $W = 1$. Define the following events:

$$E_i = \left\{(\mathbf{X}(i), \mathbf{Y}) \in A_\epsilon^{(n)}\right\}, \quad i \in \{1, 2, \ldots, 2^{nR}\}.$$

Then,

$$\Pr\{\mathcal{E}|W=1\} = \Pr\{E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}\} \leq \Pr\{E_1^c\} + \sum_{i=2}^{2^{nR}} \Pr\{E_i\}.$$

- By the joint AEP, $\Pr\{E_1^c\} \to 0$ as $n \to \infty$.
- By the independence of $\mathbf{X}(i)$ and $\mathbf{Y}$ for $i \neq 1$, we have

$$\Pr\{E_i\} \leq 2^{-n[I(X;Y)-3\epsilon]}.$$

Consequently,

$$\begin{aligned} \Pr\{\mathcal{E}\} &= \Pr\{\mathcal{E}|W=1\} \leqslant \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n[I(X;Y)-3\epsilon]} \\ &\leqslant \epsilon + 2^{nR} 2^{-n[I(X;Y)-3\epsilon)} \leqslant \epsilon + 2^{-n[I(X;Y)-R-3\epsilon]} \leqslant 2\epsilon \end{aligned}$$

if n is sufficiently large and $R < I(X;Y) - 3\epsilon$.

Hence, if $R < I(X;Y)$, we can choose ϵ and n so that the average probability of error, averaged over codebooks and codewords, is less than 2ϵ.

Finally,

- Choose $P^*(x)$, the distribution that achieves capacity. Then the condition is $R < C$.
- Get rid of the average over codebooks. There exists at least one codebook $\mathscr{C}^*$ such that $\varepsilon^{(n)}(\mathscr{C}^*) \leqslant 2\epsilon$
- Throw away the worst half of the codewords in the best codebook $\mathscr{C}^*$. There exist at least half codewords such that $\lambda_w \leqslant 4\epsilon$. If we reindex these codewords, we have 2^{nR-1} codewords, and the rate is $R - \frac{1}{n}$, where $\frac{1}{n}$ is negligible for large n.

This proves the achievability of any rate below capacity.

Fano's inequality

Theorem 5 (Fano's Inequality)

For any estimator $\hat{X}$ *such that* $X \to Y \to \hat{X}$ *with* $P_e = \Pr(X \neq \hat{X})$, *we have* $H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$.

作业

Exercise 1

考虑二元矩阵 $G_{2\times 4}$ 。若矩阵的每个元素都是均匀随机且独立产生的，计算 $\text{Rank}(G)=0, \text{Rank}(G)=1, \text{Rank}(G)=2$ 三个事件各自的概率，并检验

$$\Pr\{\text{Rank}(G)<2\}<\frac{1}{4}.$$

Exercise 2.

(X^n, Y^n) 联合典型可以推出 X^n 是典型的，Y^n 也是典型的，但反之未必成立。从典型序列的个数加以说明。

作业

Exercise 3.[田宝玉(2008)]

一离散无记忆信道的转移概率矩阵为

$$\begin{bmatrix} 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{bmatrix}$$

(1) 求该信道的信道容量。

(2) 求达到容量时的输入概率分布和输出概率分布。

作业

Exercise 4.[田宝玉(2008)]

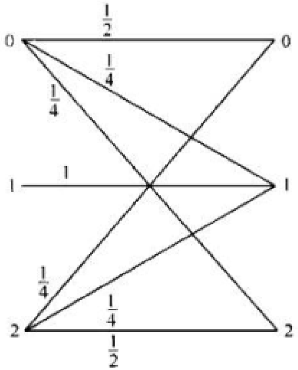
一离散无记忆信道如图所示

(1) 写出该信道的转移概率矩阵。

(2) 该信道是否为对称信道?

(3) 求该信道的信道容量。

(4) 求达到信道容量时的输出概率分布。



谢谢！