Elements of Information Theory

Chapter 5: Channel Capacity and Channel Coding Theorem

Bilingual course
(Chinese taught course)
Information and Communication Eng. Dept.
Deng Ke

Outline

- Sequence
- Channel Capacity
- Symmetric Channel
- Decoding Rule
- Joint Typical Set and Joint AEP
- Channel Coding Theorem

Sequence

- The channel input is X^N with length N, after N transmissions, the output is Y^N also with length N

$$X^N = (x_1, x_2, \ldots, x_N), x_i \in \{a_1, a_2, \ldots, a_K\} \qquad K^N$$
$$Y^N = (y_1, y_2, \ldots, y_N), y_i \in \{b_1, b_2, \ldots, b_J\} \qquad J^N$$

- This can be regarded as once transmission whose channel model with K^N input and J^N output
- Probability $p(y^N|x^N) = \prod_n p(y_n|x_n)$
- Condition entropy

Sequence

$$\begin{aligned} H(X^N|Y^N) &= \sum_{x^N,y^N} p(x^N,y^N)\log\frac{1}{p(x^N|y^N)} \\ &= \cdots = \sum_{x_1,y_1} p(x_1,y_1)\log\frac{1}{p(x_1|y_1)} + \cdots + \sum_{x_N,y_N} p(x_N,y_N)\log\frac{1}{p(x_N|y_N)} \\ &= H(X_1|Y_1) + \cdots + H(X_N|Y_N) \\ &= NH(X|Y) \end{aligned}$$

- Then

$$\begin{aligned} I(X^N;Y^N) &= H(X^N) - H(X^N|Y^N) \\ &= NH(X) - NH(X|Y) \\ &= N[H(X) - H(X|Y)] \\ &= NI(X;Y) \end{aligned}$$

Channel Capacity

- *Definition* We define the "*information*"*channel capacity* of a discrete memoryless channel as

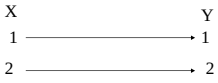
$$C = \max_{p(x)} I(X;Y) \qquad (7.1)$$

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

- Interpretation
  - Mutual information: A measure of the amount of information that one random variable contains about another random variable.
  - Original uncertainty (information): $H(X)$
  - Uncertainty at the receiver (Y observed): $H(X|Y)$
  - Uncertainty reduction: $H(X)$-$H(X|Y)$=$I(X;Y)$

Some Examples

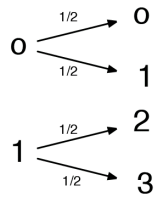
- Noiseless Channel



$$\begin{aligned} C &= \max I(X;Y) \\ &= \max\left[H(X) - H(X|Y)\right] \\ &= \max\left[H(X) - 0\right] \\ &= 1 \end{aligned}$$

Some Examples

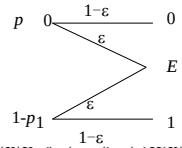
- 2-input-4-ouput channel



$$C = \max I(X;Y)$$
$$= \max\left[H(X) - H(X \mid Y)\right]$$
$$= \max\left[H(X) - 0\right]$$
$$= 1$$

Some Examples

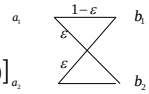
- Binary erasure channel



$$C=\max\{I(X;Y)\}=\max\{H(X)-H(X|Y)\}$$
$$=\max\{H(X)-\sum p(y)H(X|Y=y)\}$$
$$=\max\{H(X)-[p(1-\varepsilon)H(X|Y=0)+(1-p)(1-\varepsilon)H(X|Y=1)+(p\varepsilon+(1-p)\varepsilon)H(X|Y=E$$
$$=\max\{H(X)-[p(1-\varepsilon)0+(1-p)(1-\varepsilon)0+\varepsilon H(X)]$$
$$=\max\{H(X)(1-\varepsilon)\}$$
$$=(1-\varepsilon)$$

Some Examples

- BSC



$$C = \max I(X;Y)$$
$$= \max\left[H(Y) - H(Y \mid X)\right]$$
$$= \max\left[H(Y) - \Omega(\varepsilon)\right]$$
$$= 1 - \Omega(\varepsilon)$$

$I(X;Y) = \Omega(\varepsilon + p - 2\varepsilon p) - \Omega(\varepsilon)$ To achieve C

$\varepsilon + p - 2\varepsilon p = 1/2 \Rightarrow (1-2\varepsilon)p = (1-2\varepsilon)/2 \Rightarrow p = 1/2$

Channel Capacity

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

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

- (*Jensen's inequality*) *If f is a convex function and X is a random variable*, $Ef(X) \ge f(EX)$
- The sum of cap(cup) functions is also a cap(cup) fucntion.
- *The mutual information I(X;Y) is a cap function of p(x)*

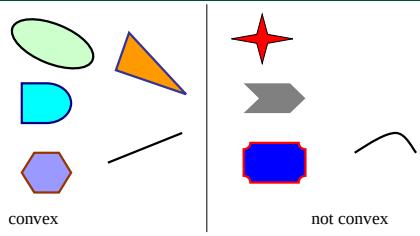
Convex Region

- A region R is convex if for each vector $\overline{\alpha}$ $\overline{\beta}$ in R, the vector $\theta\overline{\alpha} + (1-\theta)\overline{\beta}$ is in R for $0 \le \theta \le 1$

$$\theta\overline{\alpha} + (1-\theta)\overline{\beta} = \theta\overline{\alpha} + (1-\theta)\overline{\alpha} - (1-\theta)\overline{\alpha} + (1-\theta)\overline{\beta}$$
$$= \overline{\alpha} + (1-\theta)(\overline{\beta} - \overline{\alpha})$$

- For each pair of points in the region, the straight line between those points stays in the region
- A vector is defined to be a probability vector if its components are all nonnegative and sum to 1
- The region of probability vector is convex

Convex Region



Examples of convex regions for 2-dimensional vectors

Convex Region

- Let $\overline{\alpha}$ and $\overline{\beta}$ be probability vectors and let $\overline{\gamma} = \theta\overline{\alpha} + (1-\theta)\overline{\beta}$
- Nonnegative $\gamma_k = \theta\alpha_k + (1-\theta)\beta_k \geq 0$
- Sum $\sum_k \gamma_k = \sum_k [\theta\alpha_k + (1-\theta)\beta_k]$
  $= \theta \sum_k \alpha_k + (1-\theta)\sum_k \beta_k = \theta + (1-\theta) = 1$
- A real-valued function f of a vector is defined to be convex cap over a convex region R if for all $\overline{\alpha}$ $\overline{\beta}$ in R, and θ , $0<\theta<1$ the function satisfies
  $\theta f(\overline{\alpha}) + (1-\theta) f(\overline{\beta}) \leq f[\theta\overline{\alpha} + (1-\theta)\overline{\beta}]$

Maximize of a convex cap function

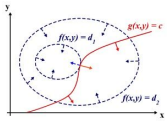
- Let $f(\overline{\alpha})$ be a convex cap function of $\overline{\alpha}$, where $\overline{\alpha}$ is a probablity vector. Assume that the partial derivatives, $\partial f(\overline{\alpha})/\partial\alpha_k$ are defined and continuous over the region R with possible exception that $\lim_{\alpha_k \to 0} \frac{\partial f(\overline{\alpha})}{\partial \alpha_k} = \infty$ Then the necessary and sufficient conditions on a probability vector $\overline{\alpha}$ to maximize f over the region R is

$$\frac{\partial f(\overline{\alpha})}{\partial \alpha_k} = \lambda \qquad \text{all } k \text{ such that } \alpha_k > 0$$

$$\frac{\partial f(\overline{\alpha})}{\partial \alpha_k} \leq \lambda \qquad \text{all } k \text{ such that } \alpha_k = 0$$

Lagrange Multiplier

$$\begin{cases} \max \quad f(x,y) \\ g(x,y) = c \end{cases}$$



$$L(x,y,\lambda) = f(x,y) - \lambda[g(x,y) - c]$$

$$\frac{\partial L(x,y,\lambda)}{\partial \lambda} = 0 \Rightarrow g(x,y) = c$$

$$\frac{\partial L(x,y,\lambda)}{\partial x} = 0 \Rightarrow \frac{\partial f(x,y)}{\partial x} = \lambda \frac{\partial g(x,y)}{\partial x}$$

$$\frac{\partial L(x,y,\lambda)}{\partial y} = 0 \Rightarrow \frac{\partial f(x,y)}{\partial y} = \lambda \frac{\partial g(x,y)}{\partial y}$$

Finding Channel Capacity

$$\frac{\partial I(X;Y)}{\partial Q(x_n)} = \lambda \text{ all } k \text{ with} Q(x_n) > 0; \quad \frac{\partial I(X;Y)}{\partial Q(x_n)} \leq \lambda \text{ all } k \text{ with} Q(x_n) = 0$$

$$\frac{\partial I(X;Y)}{\partial Q(x_n)} = \frac{\partial}{\partial Q(x_n)}\left[\sum_i \sum_j Q(x_i) p(y_j \mid x_i) \log \frac{p(y_j \mid x_i)}{p(y_j)}\right]$$

$$= \frac{\partial}{\partial Q(x_n)} \sum_i \sum_j \left[Q(x_i) p(y_j \mid x_i) \log \frac{p(y_j \mid x_i)}{\sum_i Q(x_i) p(y_j \mid x_i)} \right]$$

$$= \sum_j p(y_j \mid x_n) \log \frac{p(y_j \mid x_n)}{p(y_j)} + \sum_i \sum_j Q(x_i) p(y_j \mid x_i) \frac{\partial}{\partial Q(x_n)} \log \frac{p(y_j \mid x_i)}{p(y_j)}$$

$$\frac{\partial}{\partial Q(x_n)} \log \frac{p(y_j \mid x_i)}{p(y_j)} = \log e \frac{p(y_j)}{p(y_j \mid x_i)} \left[p(y_j \mid x_i) \left(-\frac{1}{p^2(y_j)} \right) \right] \frac{\partial}{\partial Q(x_n)} p(y_j)$$

$$= -\frac{\log e}{p(y_j)} \frac{\partial}{\partial Q(x_n)} \sum_i Q(x_i) p(y_j \mid x_i)$$

$$= -\frac{\log e}{p(y_j)} p(y_j \mid x_n)$$

Finding Channel Capacity

$$\sum_k \sum_j Q(x_k) p(y_j \mid x_k) - \frac{\log e}{p(y_j)} p(y_j \mid x_n)$$

$$= -\log e \sum_k \sum_j p(x_k, y_j) \frac{p(y_j \mid x_n)}{p(y_j)}$$

$$= -\log e \sum_j \frac{p(y_j \mid x_n)}{p(y_j)} \sum_k p(x_k, y_j)$$

$$= -\log e \sum_j \frac{p(y_j \mid x_n)}{p(y_j)} p(y_j)$$

$$= -\log e \sum_j p(y_j \mid x_n)$$

$$= -\log e$$

Finding Channel Capacity

$$\frac{\partial I(X;Y)}{\partial Q(x_n)} = \sum_j p(y_j \mid x_n) \log \frac{p(y_j \mid x_n)}{p(y_j)} - \log e$$

$$= I(X = x_n; Y) - \log e$$

- Then

$$\frac{\partial I(X;Y)}{\partial Q(x_n)} = I(X = x_n; Y) - \log e = \lambda \qquad Q(x_n) > 0,$$

$$\frac{\partial I(X;Y)}{\partial Q(x_n)} = I(X = x_n; Y) - \log e \leq \lambda \qquad Q(x_n) = 0.$$

- A set of necessary and sufficient conditions on an input probability vector $Q(x) = [Q(x_1), Q(x_2), \cdots, Q(x_k)]$ to achieve capacity on a DMC is

Symmetric Channel

$$I(X=x_n;Y)=C \qquad Q(x_n)>0,$$
$$I(X=x_n;Y)\le C \qquad Q(x_n)=0.$$

- inputs as rows and outputs as columns
- The rows of the probability transition matrix are permutations of each other and so are the columns. *symmetric I*
- The rows of the probability transition matrix are permutations of each other and all the column sums are equal *weakly symmetric*
- (The columns can be partitioned into subsets in such a way that in each subset, the rows are permutations of each other and so are the columns(if more than 1). *symmetric II*)

Symmetric Channel

$$\begin{bmatrix}0.3 & 0.2 & 0.5\\ 0.2 & 0.5 & 0.3\\ 0.5 & 0.3 & 0.2\end{bmatrix} \quad \begin{bmatrix}1/3 & 1/6 & 1/2\\ 1/2 & 1/2 & 1/6\end{bmatrix} \quad \begin{bmatrix}0.7 & 0.2 & 0.1\\ 0.1 & 0.2 & 0.7\end{bmatrix}$$

- For a symmetric DMC, capacity is achieved by using the inputs with equal probability

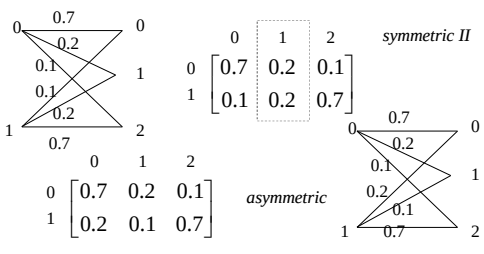
Symmetric Channel

$$I(x=k;Y)=I_1(x=k;Y)+I_2(x=k;Y)$$
$$I(x=1;Y)=\sum_{j=1}^{N}p(j|1)\log\frac{p(j|1)}{p(j)}+\sum_{j=N+1}^{J}p(j|1)\log\frac{p(j|1)}{p(j)}$$
$$=\sum_{j=1}^{N}p(j|1)\log\frac{p(j|1)}{\sum_i Q(i)p(j|i)}+\sum_{j=N+1}^{J}p(j|1)\log\frac{p(j|1)}{\sum_i Q(i)p(j|i)}$$
$$\underline{Q(i)=1/K}\sum_{j=1}^{N}p(j|1)\log\frac{p(j|1)}{\frac{1}{K}\sum_i p(j|i)}+\sum_{j=N+1}^{J}p(j|1)\log\frac{p(j|1)}{\frac{1}{K}\sum_i p(j|i)}$$
$$I(x=2;Y)=\sum_{j=1}^{N}p(j|2)\log\frac{p(j|2)}{\frac{1}{K}\sum_i p(j|i)}+\sum_{j=N+1}^{J}p(j|2)\log\frac{p(j|2)}{\frac{1}{K}\sum_i p(j|i)}$$
$$\vdots$$
$$I(x=K;Y)=\sum_{j=1}^{N}p(j|K)\log\frac{p(j|K)}{\frac{1}{K}\sum_i p(j|i)}+\sum_{j=N+1}^{J}p(j|K)\log\frac{p(j|K)}{\frac{1}{K}\sum_i p(j|i)}$$

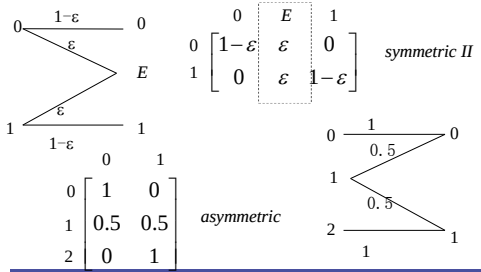
Symmetric Channel

- Consider the column permutation
$$\frac{1}{K}\sum_i p(j|i)=const$$
- Consider the row permutation
$$I_1(x=k;Y)=const$$
- And then $I_2(x=k;Y)=const$
- Hence
$$I(x=k;Y)=const \qquad k=1,2,\cdots,K$$
- Finally, we have
$$C=I(x=k;Y) \qquad k=1,2,\cdots,K$$

Symmetric Channel



Symmetric Channel



Symmetric Channel

- Example: Binary Erasure Channel

$$C = I(x=0;Y)\big|_{p(x=0)=0.5}$$
$$= \sum_y p(y\mid x=1)\log\frac{p(y\mid x=1)}{p(y)}$$
$$= (1-\varepsilon)\log\frac{1-\varepsilon}{\frac{1}{2}(1-\varepsilon)} + \varepsilon\log\frac{\varepsilon}{\varepsilon}$$
$$= 1-\varepsilon$$

$$\begin{array}{c} \\ 0 \\ 1 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & E & 1 \end{array} \\ \begin{bmatrix} 1-\varepsilon & \varepsilon & 0 \\ 0 & \varepsilon & 1-\varepsilon \end{bmatrix} \end{array}$$

Weakly Symmetric Channel

- Theorem 7.2.1 *For a weakly symmetric channel*

$$C = \log|Y| - H(\text{row of a transition matrix})$$

- *and this is achieved by a uniform distribution on the input alphabet*
- Proof: Letting r be a row of the transition matrix, we have $I(X;Y) = H(Y) - H(Y\mid X) = H(Y) - H(r)$
- From before we know, when $p(x) = 1/|\mathcal{X}|$, the capacity is achieved, and we have

$$p(y) = \sum p(y\mid x)p(x) = \frac{1}{|\mathcal{X}|}\sum p(y\mid x) = \frac{c}{|\mathcal{X}|} = \frac{1}{|Y|}$$

Decoding Rule

- Minimum error decoding rule
  - After receiveing y_j, for all *a posteriori* probability $P(x_1|y_j)$, $P(x_2|y_j)$, ..., $P(x_i\,|\,y_j)$, ...Find maximum $P(x^*|y_j)$
  - i.e. if $x_i \neq x^*$ we have $P(x^*|y_j) > P(x_i\,|\,y_j)$, then $g(y_j) = x^*$
  - This rule is also called MAP rule
- Example
  - Let the source probaliblity and channel matrix is

$$\begin{array}{cccc} x & x_1 & x_2 & x_3 \\ P(x) & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{array} \qquad \begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 1/6 & 1/2 & 1/3 \\ 1/3 & 1/6 & 1/2 \end{bmatrix}$$

Decoding Rule

- Find the decoding scheme to minimize P_E
- According to the MAP rule, first we get the MAP

$$P(x/y) = \frac{P(xy)}{P(y)}$$

$$P(xy) = \begin{bmatrix} 1/4 & 1/6 & 1/12 \\ 1/24 & 1/8 & 1/12 \\ 1/12 & 1/24 & 1/8 \end{bmatrix}$$

- and $P(y) = [3/8 \;\; 1/3 \;\; 7/24]$, $P(x/y) = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array}\begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} 2/3 & 1/2 & 2/7 \\ 1/9 & 3/8 & 2/7 \\ 2/9 & 1/8 & 3/7 \end{bmatrix} \end{array}$
- **hence**

Deconding Rule

Then, the decoding scheme is

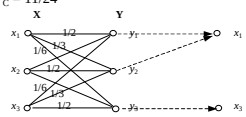
$y_1 \to x_1$, $y_2 \to x_1$, $y_3 \to x_3$

And the probability of correct transmission

$P_C = P(x_1)\cdot P(y_1|x_1) + P(x_1)\cdot P(y_2|x_1) + P(x_3)\cdot P(y_3|x_3)$
$= 1/2\times 1/2 + 1/2\times 1/3 + 1/4\times 1/2 = 13/24$

The error probability is

$P_E = 1 - P_C = 11/24$



Deconding Rule

- Maximum Likelihood Rule
  - MAP rule is the minimum error rule, but the computation is complex, we need other rule
  - When the source with M states is uniformly distributed
  - we have $P(x_i) = \frac{1}{M}$

$$P(x_i/y_j) = \frac{P(x_i y_j)}{P(y_j)} = \frac{1}{MP(y_j)}P(y_j/x_i)$$

  - which means, we can decode only from the $P(y|x)$
  - i.e. if $x_i \neq x^*$ we have $P(y_j|x^*) > P(y_j|x_i)$, then $g(y_j) = x^*$

Deconding Rule

- Example

  Let a BSC with ε=0.01, the source is uniformly distributed

  (1) Find minimum P_E (2) After the channel code "0" → "000", "1" → "111", find minimum P_E
- Solution

  (1) ε=$P(1|0)$=$P(0|1)$=0.01， $P(0)$=$P(1)$=1/2, then $g(0)$=0， $g(1)$=1, and

  $P_E = \frac{1}{2}(0.01 + 0.01) = 10^{-2}$

  (2) Let the channel input α_1=000 α_2=001, the output is β_0=000, β_1=001, β_2=010, β_3=100, β_4=011, β_5=101, β_6=110, β_7=111 and the channel matrix is

Deconding Rule

	000	001	010	100	011	101	110	111
000	$(1-\varepsilon)^3$	$(1-\varepsilon)^2\varepsilon$	$(1-\varepsilon)^2\varepsilon$	$(1-\varepsilon)^2\varepsilon$	$(1-\varepsilon)\varepsilon^2$	$(1-\varepsilon)\varepsilon^2$	$(1-\varepsilon)\varepsilon^2$	ε^3
111	ε^3	$(1-\varepsilon)\varepsilon^2$	$(1-\varepsilon)\varepsilon^2$	$(1-\varepsilon)\varepsilon^2$	$(1-\varepsilon)^2\varepsilon$	$(1-\varepsilon)^2\varepsilon$	$(1-\varepsilon)^2\varepsilon$	$(1-\varepsilon)^3$

$g(\beta_0)= g(\beta_1)= g(\beta_2)= g(\beta_3)=\alpha_1$=000 → 0

$g(\beta_4)= g(\beta_5)= g(\beta_6)= g(\beta_7)=\alpha_2$=111 → 1

and the error probability

$P_E = 3*(1-\varepsilon)\varepsilon^2 + \varepsilon^3 \approx 3\varepsilon^2 =$

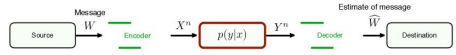
Channel Coding Theorem

- Information channel capacity:
  - The maximum of the mutual information
- Operational channel capacity:
  - Highest rate (bits/channel use) that can communicate at reliably
- Channel coding theorem says: information capacity = operational capacity
- Definition: the *rate* R of an (M, n) code is

$$R = \frac{\log M}{n} \qquad \text{bits per transmission}$$

Channel Coding Theorem

- *Definition: Achievablility*. A rate R is called *achievable* if there exists a sequence of ($\lceil 2^{nR} \rceil$, n) codes such that the maximal probability of error $\lambda^{(n)}$ (i.e. maximal Pr{Error}) tends to 0 as $n \to \infty$. Note (2^{nR}, n) codes means ($\lfloor 2^{nR} \rfloor$, n) codes
- *Definition: Capacity*. The *Capacity* of a channel is supremum of all ahievable rates.



Channel Coding Theorem

Definition: Channel code. An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following:

1. An index set $\{1, 2, \ldots, M\}$ over messages W.
2. An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2)$, (This set is called the *codebook* $\mathcal{C}$.)
   $x^n(W)$ passes through the channel and is received as a random sequence $Y^n \sim p(y^n \mid x^n)$.
3. A (deterministic) decoding function

$$g : \mathcal{Y}^n \to \{1, 2, ., ., .M\},$$

which is an estimator $\widehat{W} = g(Y^n)$ of $W \in \{1, 2, \ldots, M\}$. It declares an error if $\widehat{W} \neq W$.

Channel Coding Theorem

Definition: Conditional probability of error.

$$\text{Let } \lambda_i = \Pr\{g(Y^n) \neq i | X^n = x^n(i)\} = \sum_{y^n} p(y^n|x^n(i))\, I_{\{g(y^n)\neq i\}}(y^n)$$

be the *conditional probability of error* given that index i was sent.

Definition: Maximal probability of error. The *maximal probability of error* $\lambda^{(n)}$ for an (M, n) code is defined as

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i.$$

Definition: Average probability of error. The (*arithmetic*) *average probability of error* $P_e^{(n)}$ for an (M, n) code is:

$$P_e^{(n)} = \frac{1}{M} \sum_{i}^{M} \lambda_i.$$

- Note that $P_e^{(n)} = \Pr\{W \neq g(Y^n)\}$ if W is chosen uniformly.
- Also, $P_e^{(n)} \leq \lambda^{(n)}$; i.e., the average probability of error is less than the maximal probability of error.

Joint Typical Sequences

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

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left|-\frac{1}{n}\log p(x^n) - H(X)\right| < \epsilon, \left|-\frac{1}{n}\log p(y^n) - H(Y)\right| < \epsilon, \left|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| < \epsilon\},$$

where

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

Joint AEP

- Theroem 7.6.1 (*Joint AEP*) *Let* (X^n, Y^n) *be sequences of length n drawn i.i.d according to* $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$, *then*

1. $\Pr\left\{(X^n, Y^n) \in A_\epsilon^{(n)}\right\} \to 1$ as $n \to \infty$.
2. $\left|A_\epsilon^{(n)}\right| \le 2^{n(H(X,Y)+\epsilon)}$.

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

Joint AEP

- If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ so that $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same marginals as $p(x^n, y^n)$ then, we have

$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} \le 2^{-n(I(X;Y)-3\epsilon)}$$

$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} \ge (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

- for sufficient large n

$$\begin{aligned} \Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) &= \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n)p(y^n) \\ &\le 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\ &= 2^{-n(I(X;Y)-3\epsilon)}. \end{aligned}$$

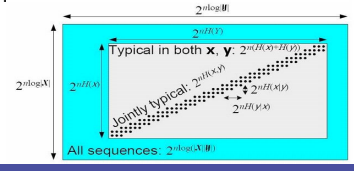
Joint AEP

For sufficiently large n, $\Pr(A_\epsilon^{(n)}) \ge 1-\epsilon$, and therefore

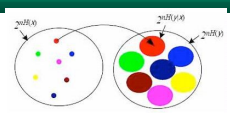
$$\begin{aligned} 1-\epsilon &\le \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n, y^n) \\ &\le |A_\epsilon^{(n)}| 2^{-n(H(X,Y)-\epsilon)} \qquad |A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X,Y)-\epsilon)} \\ \Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) &= \sum_{A_\epsilon^{(n)}} p(x^n)p(y^n) \\ &\ge (1-\epsilon)2^{n(H(X,Y)-\epsilon)} 2^{-n(H(X)+\epsilon)} 2^{-n(H(Y)+\epsilon)} \\ &= (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}. \quad \square \end{aligned}$$

Explanation

- There are about $2^{nH(X)}$ typical X in all
- Each typical Y is jointly typical with about $2^{nH(X|Y)}$ of those typlical X's
- The joint typical pairs are a fraction $2^{-nI(X;Y)}$ of inner rectangle



Explanation



- For each typical input sequence(how many?) there are about $2^{nH(Y|X)}$ possible output sequences, all equally likely.
- Want to ensure that no two typical input sequences produce the same output sequences.

Channel Coding Theorem

- Intuition
  - –Random choice of codeword
  - –Decoding rule: joint typicality decoding

The probability that any other codeword looks jointly typical with the received sequence is $2^{-nI(X;Y)}$ Hence, if we have fewer then $2^{nI(X;Y)}$ codewords in the input side, then with high probability there will be no other codewords that can be confused with the transmitted codeword, and the probability of error is small.

Channel Coding Theorem

- **Theorem 7.7.1** (*Channel coding theorem*) *For a DMC, all rates below capacity C are achivable. Specifically, for every rate R<C, there exists a sequence of* $(2^{nR}, n)$ *codes with maximum probability or error* $\lambda^{(n)} \to 0$.

  Conversely, any sequence of $(2^{nR}, n)$ *codes with* $\lambda^{(n)} \to 0$ *must have* $R \le C$
- Proof: All capacity proof can be divided into two parts:
  - i) Achievability: All $R \le C$ are achievable
  - ii) Converse: No rate $R > C$ is achievable
- Only Achievability

Channel Coding Theorem

- Generate a $(2^{nR}, n)$ code at random according to distribution $p(x)$

$$C = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix} \qquad \Pr(C) = \Pi_{w=1}^{2^{nR}} \Pi_{i=1}^{n} p(x_i(w))$$

- Each entry in this matrix is generated i.i.d.
- The code C is then revealed to both sender and receiver. Both sender and receiver are also assumed to know the channel transition matrix $p(y|x)$ for the channel.
- A message W is chosen according to a uniform distribution $\Pr(W = w) = 2^{-nR}, \quad w = 1,2,\ldots, 2^{nR}$

Channel Coding Theorem

- The codeword corresponding to the w-throw of C, is sent over the channel.
- The receiver receives a sequence Y^n according to the distribution $p(Y^n \mid X^n(w)) = \Pi_{i=1}^{n} p(y_i \mid x_i(w))$
- The receiver guesses which message was sent use jointly typical decoding
  - 1. $(X^n(\hat{w}), Y^n)$ is jointly typical
  - 2. There is NO other index $w' \neq \hat{w}$ such that $(X^n(w'), Y^n) \in A_\varepsilon^{(n)}$ Otherwise it declare an error
- Decoding error if $w \neq \hat{w}$, Let ε be the event $\{w \neq \hat{w}\}$

Channel Coding Theorem

- Instead of calculating the probability of error for a single code, we calculate the average over all codes generated at random accordin to the distribution.
- For a typical codeword, two kinds of error :
  - Either the output Y^n is not jointly typical with the transmitted codeword
  - Or there is some other codeword that is jointly typical with Y^n.
- For any rival codeword, the probability that it is jointly typical with the received sequence is approximately $2^{-nI(X;Y)}$

Channel Coding Theorem

$$\Pr(\varepsilon) = \sum_C \Pr(C) P_e^{(n)}(C) = \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_C \Pr(C) \lambda_w(C)$$

- By the symmetry of the code construction, the average probability of error averaged over all codes does not depend on the particular index that was sent.
- Thus, we can assume without loss of generality that the message W = 1 was sent

$$\Pr(\varepsilon) = \sum_C \Pr(C) \lambda_1(C) = \Pr(\varepsilon \mid W = 1)$$

Channel Coding Theorem

- Let $E_i = \{(X^n(i), Y^n) \text{ is in } A_\varepsilon^{(n)}\}, \quad i = 1,2,...,2^{nR}$
- Error scenario:
  - 1) When the transmitted codeword and the received sequence are not jointly typical
  - 2) When a wrong codeword is jointly typical with the received sequence

$$\Pr(\varepsilon \mid W = 1) = P(E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}} \mid W = 1)$$

- By joint AEP $P(E_1^c \mid W = 1) \to 0 \Rightarrow P(E_1^c \mid W = 1) \le \varepsilon$
- By the code generation process, $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$, Hence, the probability that $X^n(i)$ and Y^n are jointly typical is $\le 2^{-n(I(X;Y)-3\varepsilon)}$ by the joint AEP

Channel Coding Theorem

$$\begin{aligned}
\Pr(\mathcal{E}) = \Pr(\mathcal{E} \mid W = 1) &\le P(E_1^c \mid W = 1) + \sum_{i=2}^{2^{nR}} P(E_i \mid W = 1) \\
&\le \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\
&= \epsilon + (2^{nR} - 1)\, 2^{-n(I(X;Y)-3\epsilon)} \\
&\le \epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \le 2\epsilon \qquad R < I(X;Y) - 3\epsilon
\end{aligned}$$

$$\Pr(\mathcal{E}) \xrightarrow{n\to\infty} 0 \qquad \text{if} \quad R < I(X;Y)$$

- Choose $p(x) \to p^*(xI)$ that maximize $I(X;Y)$

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

Channel Coding Theorem

- Although the theorem shows that there exist good codes with arbitrarily small probability of error for long block lengths, it does not provide a way of constructing the best codes.
- However, without some structure in the code, it is very difficult to decode (the simple scheme of table lookup requires an exponentially large table). Hence the theorem does not provide a practical coding scheme.