Elements of Information Theory

Lecture 5
Channel Capacity and Channel Coding Theorem

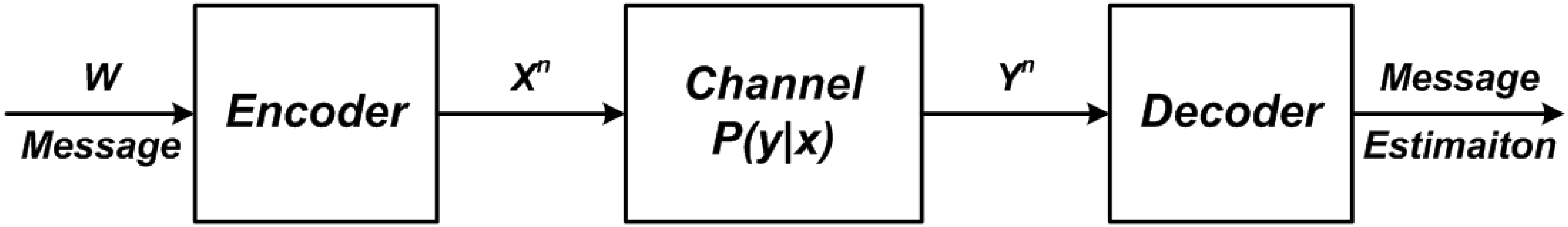
Instructor: Yichen Wang
Ph.D./Professor

School of Information and Communications Engineering
Division of Electronics and Information
Xi'an Jiaotong University

Outlines

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

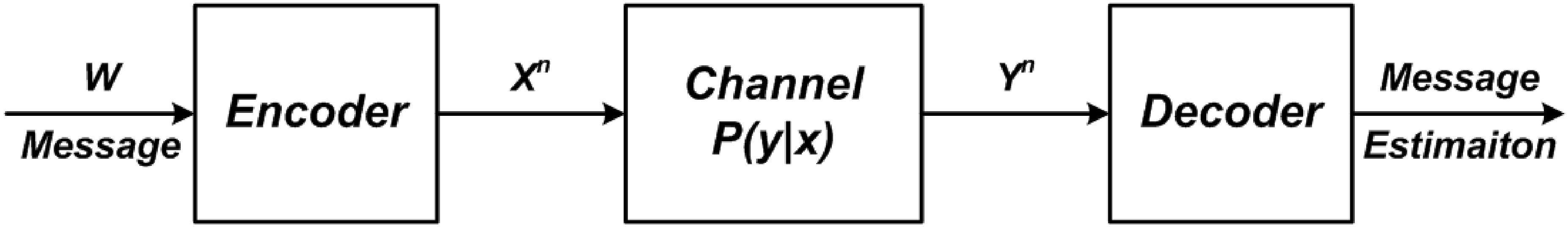
Channel Capacity



Definition

Define a discrete channel to be a system consisting of an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ and a probability transition matrix p(y|x) that expresses the probability of observing the output symbol y given that we send the symbol x.

Channel Capacity



The channel is said to be <u>memoryless</u> if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs.

Channel Capacity

Definition (Information Channel Capacity)

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

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

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

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

How to explain "information" channel capacity?

Channel Capacity

Operational Channel Capacity

The highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error.

Shannon's second theorem

Information channel capacity

||

Operational channel capacity

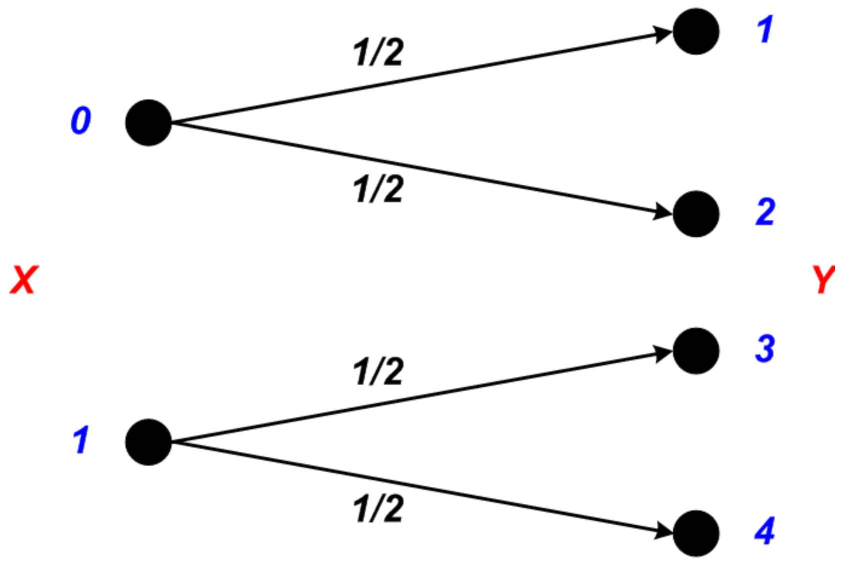
Channel Capacity

Example 1: Noiseless Binary Channel

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} \left\{ H(X) - H(X|Y) \right\} \\
&= \max_{p(x)} \left\{ H(X) - 0 \right\} \\
&= 1 \text{ bit}
\end{aligned}
$$

Channel Capacity

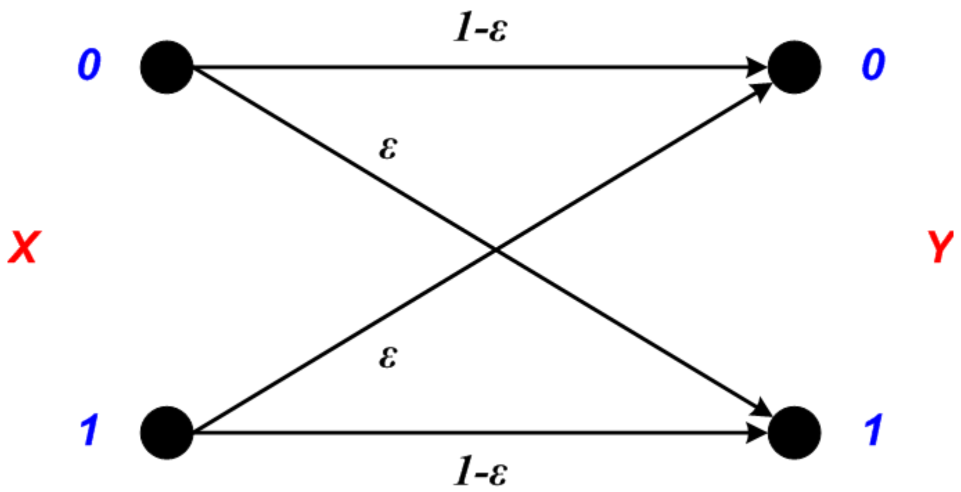
Example 2: Noisy Channel with Nonoverlapping Outputs



$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} \Big\{ H(X) - H(X|Y) \Big\} \\
&= \max_{p(x)} \Big\{ H(X) - 0 \Big\} \\
&= 1 \text{ bit}
\end{aligned}
$$

Channel Capacity

Example 3: Binary Symmetric Channel (BSC)



$$
\begin{aligned}
C &= \max I(X;Y) \\
&= \max \left\{ H(Y) - H(Y|X) \right\} \\
&= \max \left\{ H(Y) - \Omega(\epsilon) \right\} \\
&= 1 - \underline{\Omega(\epsilon)}
\end{aligned}
$$

$$\downarrow$$

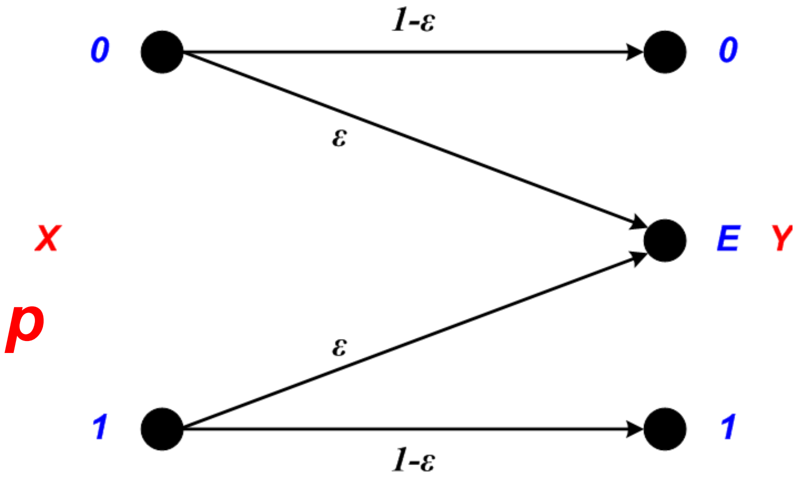
$$-\epsilon \log \epsilon - (1-\epsilon)\log(1-\epsilon)$$

When can such capacity be achieved?

Channel Capacity

Example 4: Binary Erasure Channel

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



Suppose that X=0 with probability p and X=1 with probability (1-p)

$$
\begin{aligned}
H(X|Y) &= \sum_y p(y) H(X|Y=y) \\
&= p(1-\epsilon)H(X|Y=0) + (1-p)(1-\epsilon)H(X|Y=1) \\
&\quad + \Big[p\epsilon + (1-p)\epsilon\Big] H(X|Y=E) \\
&= \epsilon H(X)
\end{aligned}
$$

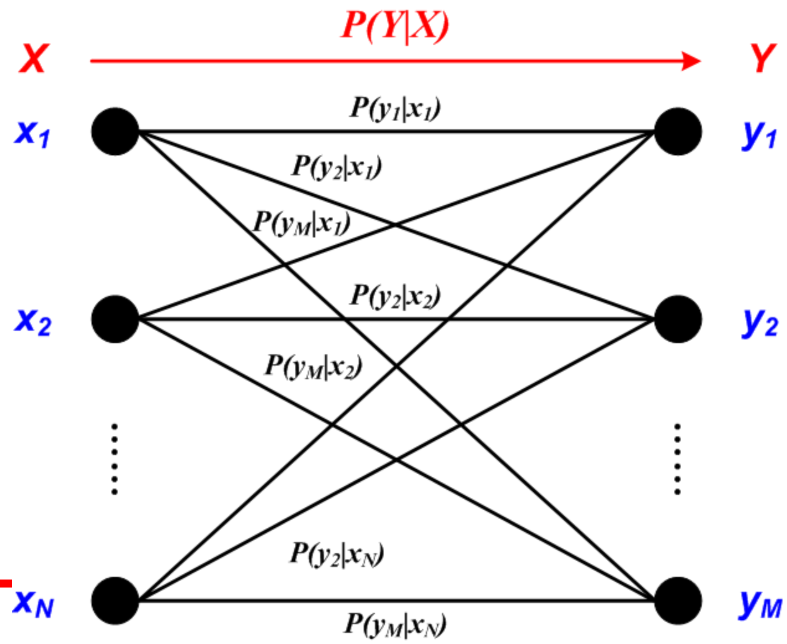
$$C = \max(1-\epsilon)H(X) = 1-\epsilon$$

Channel Capacity

Capacity Analysis for General DMC

Probability distribution of source X:

$$Q(\overline{\mathbf{x}}) = \left[Q(x_1), Q(x_2), \cdots, Q(x_N)\right]$$



Optimization Problem

$$\max_{Q(x_1),\cdots,Q(x_N)} \quad I(X;Y)$$

$$\text{s.t.} \quad \sum_{k=1}^{N} Q(x_k) = 1$$

$$Q(x_k) \geq 0, \ \ k = 1, 2, \cdots, N$$

Convex Optimization?

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*** $0 \le \lambda \le 1$***, we have***

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

Jensen's inequality

If f is a convex function and X is a random variable, then we have

$$\mathbb{E}\Big\{f(X)\Big\} \ge f\Big(\mathbb{E}\{X\}\Big)$$

Channel Capacity

Review

Theorem

Let (X,Y) ~ p(x,y)=p(x)p(y|x). Then, we can obtain that the mutual information I(X;Y) is a concave (cap convex) function of p(x) for fixed p(y|x).

$$\max_{Q(x_1),\cdots,Q(x_N)} \quad I(X;Y)$$

$$\text{s.t.} \quad \sum_{k=1}^{N} Q(x_k) = 1$$

$$Q(x_k) \geq 0,\ \ k = 1, 2, \cdots, N$$

Convex function

Is it a convex region?

Channel Capacity

Definition (Convex Region)

*A **region R** is defined to be **convex** if for each vector $\overline{\alpha}$ in R and each vector $\overline{\beta}$ in R, the vector $\theta\overline{\alpha} + (1-\theta)\overline{\beta}$ is in R for $0 \leq \theta \leq 1$.*

Definition (Probability Vector)

*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.

Channel Capacity

Definition

A real-valued function f of a vector is defined to be concave (cap function) over a convex region R of vector space, if for all $\overline{\alpha}$ in R, $\overline{\beta}$ in R, and θ ($0<\theta<1$), the function satisfies

$$\theta f\left(\overline{\alpha}\right)+(1-\theta)f\left(\overline{\beta}\right)\leq f\left(\theta\overline{\alpha}+(1-\theta)\overline{\beta}\right)$$

If the inequality is reversed for all such $\overline{\alpha}$, $\overline{\beta}$ and θ , f is convex (cup function).

If the inequality can be replaced with strict inequality, f is strictly concave or convex.

Channel Capacity

Theorem

Let f(   ) be a concave function of $\overline{\boldsymbol{\alpha}} = (\alpha_1, \cdots, \alpha_k)$ ***over the region R when*** $\overline{\boldsymbol{\alpha}}$ ***is a probability vector. Assume that the partial derivatives,*** $\partial f(\overline{\boldsymbol{\alpha}})/\partial\alpha_i$ ***are defined and continuous over the region R with the possible exception that*** $\lim_{\alpha_i \to 0} \partial f(\overline{\boldsymbol{\alpha}})/\partial\alpha_i$ ***may be +∞. Then, the sufficient and necessary conditions on a probability vector*** $\overline{\boldsymbol{\alpha}}$ ***to maximize f over the region R are***

$$\frac{\partial f(\overline{\boldsymbol{\alpha}})}{\partial\alpha_i} = \lambda; \quad \text{all } i \text{ such that } \alpha_i > 0$$

$$\frac{\partial f(\overline{\boldsymbol{\alpha}})}{\partial\alpha_i} \leq \lambda; \quad \text{all } i \text{ such that } \alpha_i = 0$$

Channel Capacity

Preliminaries on Convex Optimization

Convex optimization problem

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1, 2, \cdots, m \\ & h_i(x) = 0, i = 1, 2, \cdots, p \end{aligned}$$

$$\mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom} f_i \quad \cap \quad \bigcap_{i=1}^{p} \mathbf{dom} h_i$$

- ***The objective function must be convex;***
- ***The inequality constraint functions must be convex;***
- ***The equality constraint functions must be affine.***

Convex optimization: Lagrangian method

Channel Capacity

Preliminaries on Convex Optimization

Construct Lagrange function:

Lagrangian Multiplier

$\lambda_i \geq 0$

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)$$

- λ_i ***denotes the Lagrangian multiplier associated with the i-th inequality constraint $f_i(x) \leq 0$;***
- μ_i ***denotes the Lagrangian multiplier associated with the i-th equality constraint $h_i(x) = 0$.***

If original problem is convex, Lagrange function is also convex.

Channel Capacity

Preliminaries on Convex Optimization

Construct Lagrange dual function:

$$
\begin{aligned}
g(\lambda,\mu) &= \min_{x\in\mathcal{D}} L(x,\lambda,\mu) \\
&= \min_{x\in\mathcal{D}} f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)
\end{aligned}
$$

- ***Lagrange dual function yields lower bound on the optimal value of the original optimization problem***
- ***No matter the convexity of the original problem, Lagrange dual function is concave (cap function) over Lagrangian multipliers***

Channel Capacity

Preliminaries on Convex Optimization

Construct Lagrange dual problem:

$$\begin{aligned} &\max_{\lambda,\mu} \quad g(\lambda,\mu) \\ = \quad &\max_{\lambda,\mu} \quad \left\{ \min_{x\in\mathcal{D}} f_0(x) + \sum_{i=1}^{m}\lambda_i f_i(x) + \sum_{i=1}^{p}\mu_i h_i(x) \right\} \\ &\text{s.t.} \quad \lambda \succeq 0 \end{aligned}$$

- ***If original problem is not convex, we have***

$$g(\lambda^\star,\mu^\star) = \max\left\{g(\lambda,\mu)\right\} \le \min\left\{f_0(x)\right\} = f_0(x^\star)$$

- ***If original problem is convex, we have*** $g(\lambda^\star,\mu^\star) = f_0(x^\star)$

Channel Capacity

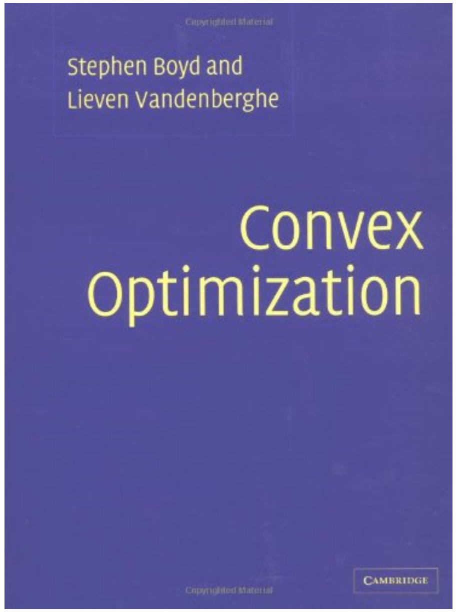
Preliminaries on Convex Optimization

Karush-Kuhn-Tucker (K.K.T.) Conditions:

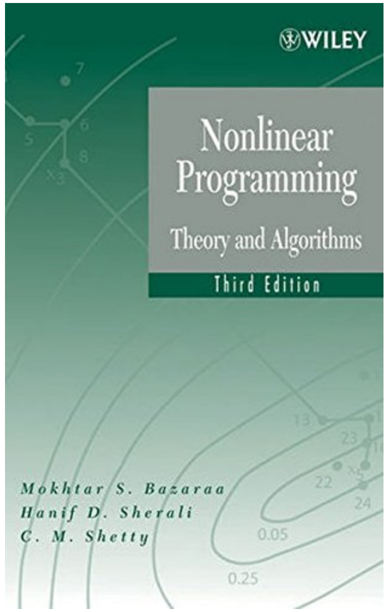
$$\begin{cases} f_i(x^\star) \leq 0, & i = 1, 2, \cdots, m \\ h_i(x^\star) = 0, & i = 1, 2, \cdots, p \\ \lambda_i^\star \geq 0, & i = 1, 2, \cdots, m \\ \lambda_i^\star f_i(x^\star) = 0, & i = 1, 2, \cdots, m \\ \nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^{p} \mu_i^\star \nabla h_i(x^\star) = 0 & \end{cases}$$

- $x^\star$ ***denotes the optimal solution for the original problem***
- $\lambda^\star = (\lambda_1^\star, \cdots, \lambda_m^\star)$ ***and*** $\mu^\star = (\mu_1^\star, \cdots, \mu_p^\star)$ ***denote the optimal solution for the Lagrange dual problem***

Preliminaries on Convex Optimization



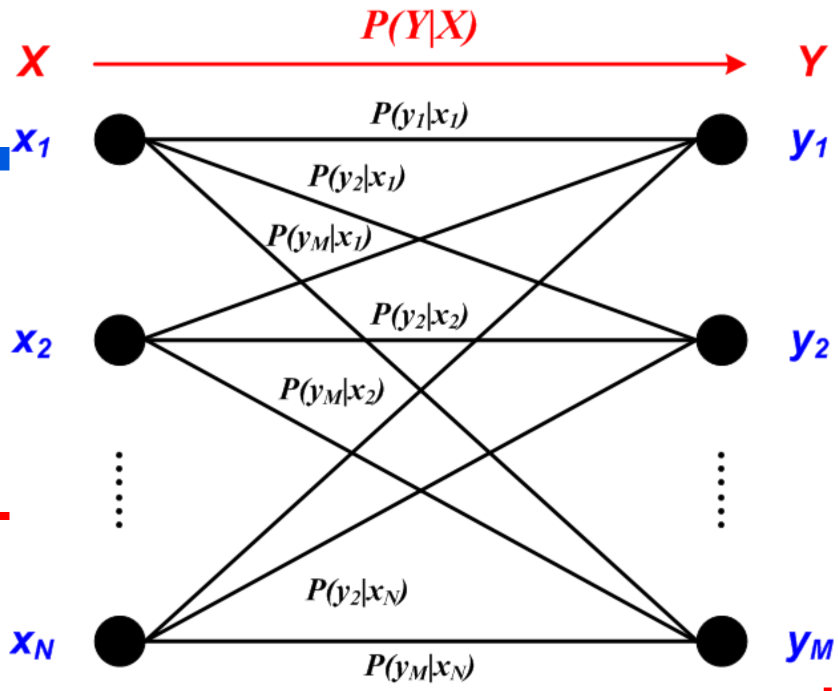
Stephen Boyd and Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2004.



Mokhtar S. Bazaraa, Hanif D. Sherali, and C. M. Shetty, Nonlinear Programming: Theory and Algorithms, Wiley-Interscience, 2006.

Channel Capacity

Capacity Analysis for General DMC



$$\max_{Q(x_1),\cdots,Q(x_N)} I(X;Y)$$

$$= \max_{Q(x_1),\cdots,Q(x_N)} \sum_{k=1}^{N}\sum_{j=1}^{M} Q(x_k)P(y_j|x_k)\log\frac{P(y_j|x_k)}{\sum_{i=1}^{N} Q(x_i)P(y_j|x_i)}$$

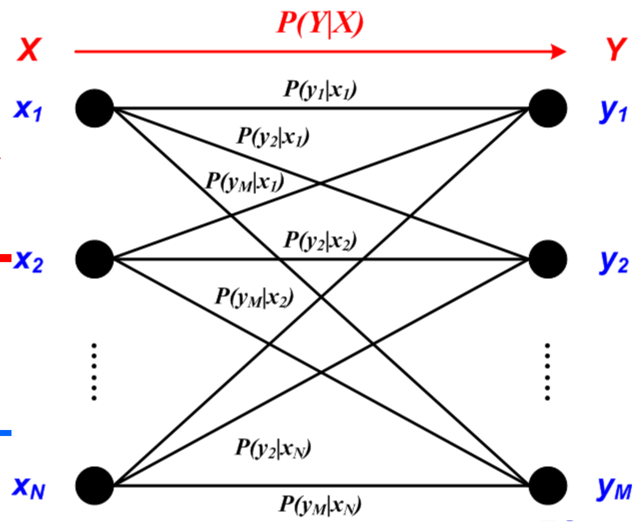
$$\text{s.t.} \quad \sum_{k=1}^{N} Q(x_k) = 1$$

$$Q(x_k) \geq 0,\ \ k = 1, 2, \cdots, N$$

Channel Capacity

Capacity Analysis for General DMC

$$
\begin{aligned}
& \max_{Q(x_1),\cdots,Q(x_N)} I(X;Y) \\
= & \max_{Q(x_1),\cdots,Q(x_N)} \sum_{k=1}^{N}\sum_{j=1}^{M} Q(x_k)P(y_j|x_k)\log\frac{P(y_j|x_k)}{\sum_{i=1}^{N} Q(x_i)P(y_j|x_i)} \\
\text{s.t.} & \sum_{k=1}^{N} Q(x_k) = 1 \\
& Q(x_k) \geq 0,\ k = 1,\cdots,N
\end{aligned}
$$



Convex optimization problem

Channel Capacity

Capacity Analysis for General DMC

Construct Lagrange function:

$$
\begin{aligned}
L\left(\{Q(x_k)\}_{k=1}^{N}, \lambda, \{\mu_k\}_{k=1}^{N}\right) &= I(X;Y) - \lambda\left\{\sum_{k=1}^{N} Q(x_k) - 1\right\} + \sum_{k=1}^{N} \mu_k Q(x_k) \\
&= \sum_{k=1}^{N}\sum_{j=1}^{M} Q(x_k)P(y_j|x_k)\log\frac{P(y_j|x_k)}{\sum_{i=1}^{N} Q(x_i)P(y_j|x_i)} \\
&\quad -\lambda\left\{\sum_{k=1}^{N} Q(x_k) - 1\right\} + \sum_{k=1}^{N} \mu_k Q(x_k)
\end{aligned}
$$

Channel Capacity

Capacity Analysis for General DMC

$$\frac{\partial L\left(\{Q(x_k)\}_{k=1}^{N}, \lambda, \{\mu_k\}_{k=1}^{N}\right)}{\partial Q(x_n)}$$

$$= \frac{\partial}{\partial Q(x_n)}\left\{\sum_{k=1}^{N}\sum_{j=1}^{M} Q(x_k)P(y_j|x_k)\log\frac{P(y_j|x_k)}{\sum_{i=1}^{N} Q(x_i)P(y_j|x_i)}\right\} - \lambda + \mu_n$$

$$= \boxed{\sum_{k=1}^{N}\sum_{j=1}^{M}\frac{\partial}{\partial Q(x_n)}\left\{Q(x_k)P(y_j|x_k)\log\frac{P(y_j|x_k)}{\sum_{i=1}^{N} Q(x_i)P(y_j|x_i)}\right\}} - \lambda + \mu_n$$

$$= \boxed{\sum_{j=1}^{M} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\sum_{i=1}^{N} Q(x_i)P(y_j|x_i)} - \log e} - \lambda + \mu_n$$

Channel Capacity

Capacity Analysis for General DMC

K.K.T. conditions:

$$
\begin{cases}
Q^\star(x_k) \geq 0, & k = 1, 2, \cdots, N \\
\sum_{k=1}^{N} Q^\star(x_k) - 1 = 0, & \\
\mu_k^\star \geq 0, & k = 1, 2, \cdots, N \\
\mu_k^\star Q^\star(x_k) = 0, & k = 1, 2, \cdots, N \\
\left.\dfrac{\partial L\left(\{Q(x_k)\}_{k=1}^{N}, \lambda, \{\mu_k\}_{k=1}^{N}\right)}{\partial Q(x_n)}\right|_{\{Q(x_k)=Q^\star(x_k)\}_{k=1}^{N}, \lambda=\lambda^\star, \left\{\mu_k=\mu_k^\star\right\}_{k=1}^{N}} = 0, & \\
& n = 1, 2, \cdots, N
\end{cases}
$$

Channel Capacity

Capacity Analysis for General DMC

Based on K.K.T. conditions, we can obtain the optimal probability distribution of the source:

$$Q^{\star}(\overline{\mathbf{x}}) = \Big[Q^{\star}(x_1), Q^{\star}(x_2), \cdots, Q^{\star}(x_N)\Big]$$

The optimal solution satisfies the following requirements:

$$\begin{cases} \sum_{j=1}^{M} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\sum_{i=1}^{N} Q^{\star}(x_i)P(y_j|x_i)} - \log e - \lambda^{\star} = 0, & \text{if } Q^{\star}(x_n) > 0 \\ \\ \sum_{j=1}^{M} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\sum_{i=1}^{N} Q^{\star}(x_i)P(y_j|x_i)} - \log e - \lambda^{\star} \le 0, & \text{if } Q^{\star}(x_n) = 0 \end{cases}$$

Channel Capacity

Capacity Analysis for General DMC

Define the following function:

$$I(x_n;Y)=\sum_{j=1}^{M}P(y_j|x_n)\log\frac{P(y_j|x_n)}{\sum_{i=1}^{N}Q(x_i)P(y_j|x_i)}$$

The mutual information for input x_n averaged over the outputs.

$$I(X;Y)=\sum_{n=1}^{N}Q(x_n)I(x_n;Y)$$

$$\begin{aligned}\max_{Q(x_1),\cdots,Q(x_N)} I(X;Y) &= \sum_{n=1}^{N}Q^\star(x_n)I(x_n;Y)\Big|_{\{Q(x_k)=Q^\star(x_k)\}_{k=1}^{N}}\\ &= \sum_{n=1}^{N}Q^\star(x_n)\left[\log e+\lambda^\star-\mu_n^\star\right]\\ &= \log e+\lambda^\star=C\end{aligned}$$

Channel Capacity

Theorem

A set of necessary and sufficient conditions on an input probability vector

$$Q(\overline{\mathbf{x}}) = \left[Q(x_1), Q(x_2), \cdots, Q(x_N)\right]$$

to achieve capacity on a discrete memoryless channel with transition probabilities $P(y_j|x_n)$ ***is that for some number C,***

$$I(x_n;Y) = C; \quad \text{all } n \text{ with } Q(x_n) > 0$$

$$I(x_n;Y) \le C; \quad \text{all } n \text{ with } Q(x_n) = 0$$

in which $I(x_n;Y)$ ***is the mutual information for input*** x_n ***averaged over the outputs.***

Furthermore, the number of C is the capacity of the channel.

Channel Capacity

Theorem

Let f(   ) be a concave function of $\overline{\boldsymbol{\alpha}} = (\alpha_1, \cdots, \alpha_k)$ ***over the region R when*** $\overline{\boldsymbol{\alpha}}$ ***is a probability vector. Assume that the partial derivatives,*** $\partial f(\overline{\boldsymbol{\alpha}})/\partial\alpha_i$ ***are defined and continuous over the region R with the possible exception that*** $\lim_{\alpha_i \to 0} \partial f(\overline{\boldsymbol{\alpha}})/\partial\alpha_i$ ***may be +∞. Then, the sufficient and necessary conditions on a probability vector*** $\overline{\boldsymbol{\alpha}}$ ***to maximize f over the region R are***

$$\frac{\partial f(\overline{\boldsymbol{\alpha}})}{\partial \alpha_i} = \lambda; \quad \text{all } i \text{ such that } \alpha_i > 0$$

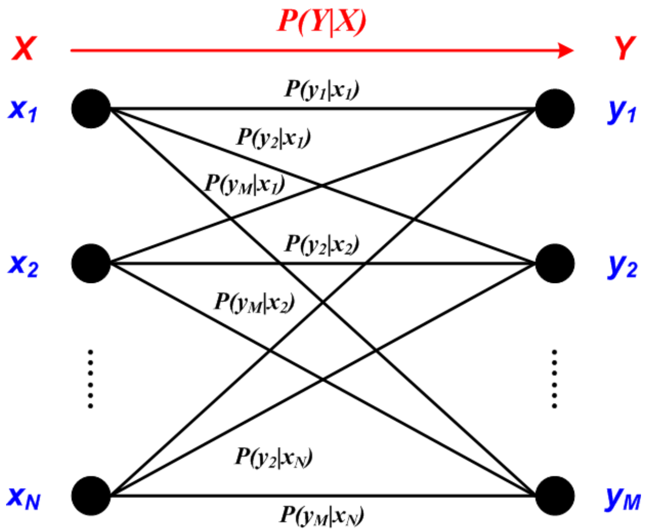
$$\frac{\partial f(\overline{\boldsymbol{\alpha}})}{\partial \alpha_i} \le \lambda; \quad \text{all } i \text{ such that } \alpha_i = 0$$

Outlines

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

Symmetric Channel

In information theory, channel can be represented by the <u>channel (probability) transition matrix</u>.



Channel Transition Matrix:

$$P(Y|X) = \begin{bmatrix} P(y_1|x_1) & P(y_2|x_1) & \cdots & P(y_M|x_1) \\ P(y_1|x_2) & P(y_2|x_2) & \cdots & P(y_M|x_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(y_1|x_N) & P(y_2|x_N) & \cdots & P(y_M|x_N) \end{bmatrix}_{N \times M}$$

Inputs as rows and outputs as columns

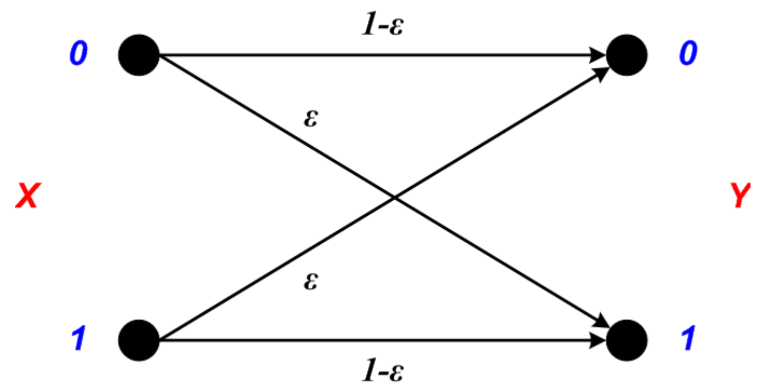
Symmetric Channel

Definition (Symmetric)

The channel is defined as symmetric if the rows of the channel transition matrix p(y|x) are permutations of each other and the columns are permutations of each other.

Symmetric I

$$p(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.2 & 0.5 & 0.3 \\ 0.5 & 0.3 & 0.2 \end{bmatrix}$$



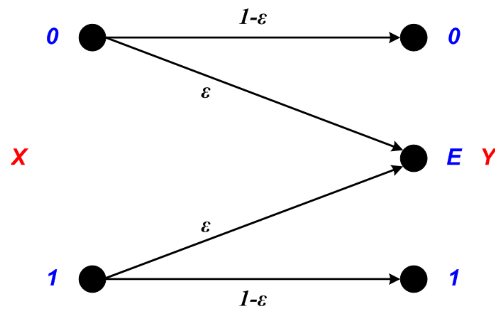
Symmetric Channel

Definition (Quasi-Symmetric)

The channel is defined as quasi-symmetric if the columns of the channel transition matrix p(y|x) 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

$$p(y|x) = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}$$



Symmetric Channel

Definition (Weakly Symmetric)

The channel is defined as weakly symmetric if every row of the channel transition matrix p(y|x) is a permutation of every other row and the column sums $\sum_x p(y|x)$ are equal.

$$p(y|x) = \begin{bmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}$$

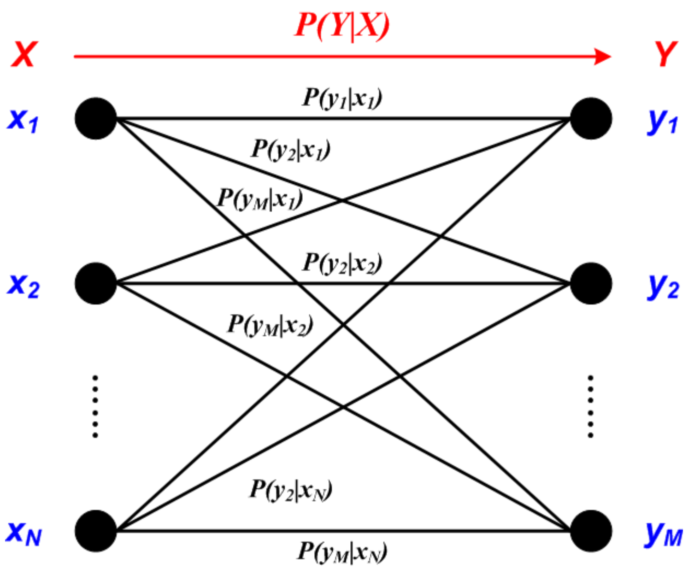
For the channel whose channel transition matrix does not meet the requirements of symmetric, quasi-symmetric, and weakly symmetric channels, the channel is viewed as asymmetric.

Symmetric Channel

Capacity Analysis for Quasi-Symmetric DMC

Theorem (Capacity of Quasi-Symmetric DMC)

For a quasi-symmetric discrete memoryless channel (DMC), capacity is achieved by using the inputs with equal probability.



$$Q^{\star}(x_1) = Q^{\star}(x_2) = \cdots = Q^{\star}(x_N) = \frac{1}{N}$$

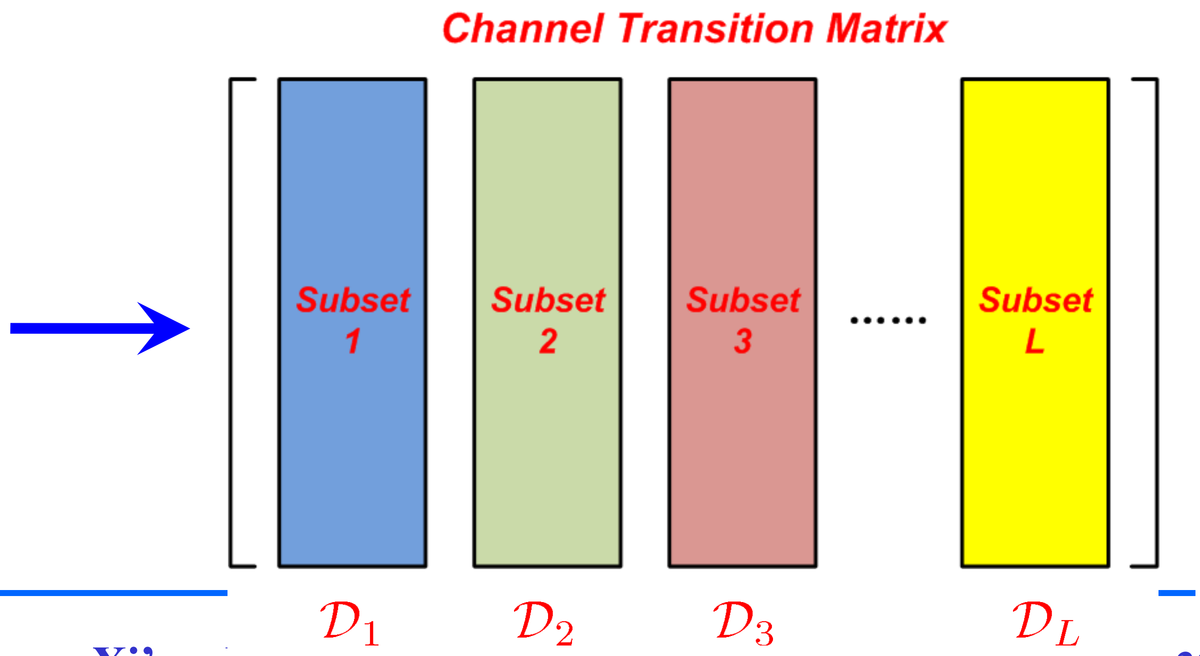
Symmetric Channel

<u>Capacity Analysis for Quasi-Symmetric DMC</u>

$$I(X;Y) = \sum_{n=1}^{N} Q(x_n) I(X = x_n; Y)$$

$$I(X = x_n; Y) = \sum_{j=1}^{M} P(y_j|x_n) \log \frac{P(y_j|x_n)}{\sum_{i=1}^{N} Q(x_i) P(y_j|x_i)}$$

As the channel is symmetric, we can partition the channel transition matrix based on the columns.



Symmetric Channel

Capacity Analysis for Quasi-Symmetric DMC

$$
\begin{aligned}
I(X=x_n;Y) &= \sum_{j\in\mathcal{D}_1} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\sum_{i=1}^{N}Q(x_i)P(y_j|x_i)} \\
&+ \sum_{j\in\mathcal{D}_2} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\sum_{i=1}^{N}Q(x_i)P(y_j|x_i)} \\
&+\cdots+ \sum_{j\in\mathcal{D}_L} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\sum_{i=1}^{N}Q(x_i)P(y_j|x_i)}
\end{aligned}
$$

$$Q(x_1)=Q(x_2)=\cdots=Q(x_N)=\frac{1}{N}$$

$$\boxed{\sum_{i=1}^{N}Q(x_i)P(y_j|x_i)=\frac{1}{N}\sum_{i=1}^{N}P(y_j|x_i)}$$

Symmetric Channel

Capacity Analysis for Quasi-Symmetric DMC

$$
\begin{aligned}
I(X=x_n;Y) &= \sum_{j\in\mathcal{D}_1} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\frac{1}{N}\sum_{i=1}^{N}P(y_j|x_i)} \\
&\quad + \sum_{j\in\mathcal{D}_2} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\frac{1}{N}\sum_{i=1}^{N}P(y_j|x_i)} \\
&\quad + \cdots + \sum_{j\in\mathcal{D}_L} P(y_j|x_n)\log\frac{P(y_j|x_n)}{\frac{1}{N}\sum_{i=1}^{N}P(y_j|x_i)}
\end{aligned}
$$

If we can obtain that $I(X=x_1;Y) = I(X=x_2;Y) = \ldots = I(X=x_N;Y)$, then channel capacity is achieved at

$$Q(x_1) = Q(x_2) = \cdots = Q(x_N) = \frac{1}{N}$$

Symmetric Channel

Capacity Analysis for Quasi-Symmetric DMC

$\forall j \in \mathcal{D}_l \ (l = 1, 2, \cdots, L)$, ***we have***

$$\frac{1}{N} \sum_{i=1}^{N} P(y_j|x_i) = \text{constant}$$

$\forall \mathcal{D}_l \ (l = 1, 2, \cdots, L)$, ***we can obtain***

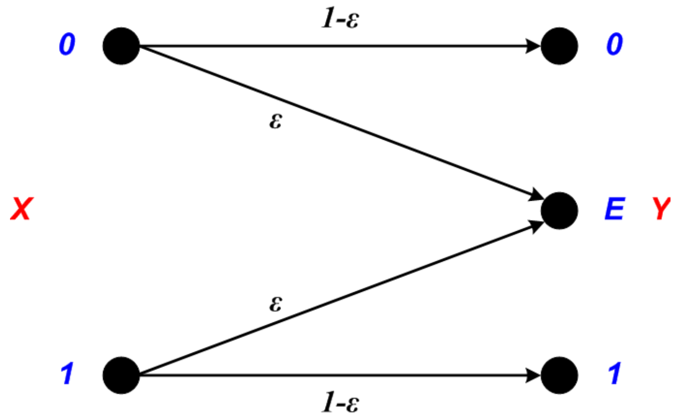
$$\sum_{j \in \mathcal{D}_l} P(y_j|x_n) \log \frac{P(y_j|x_n)}{\frac{1}{N} \sum_{i=1}^{N} P(y_j|x_i)} = \text{constant for all } x_n$$

$$I(X = x_1; Y) = \cdots = I(X = x_N; Y) = \text{constant}$$

The constant is only determined by the channel matrix. Consequently, the constant is CAPACITY!

Symmetric Channel

Example: Binary Erasure Channel



Channel Transition Matrix

$$\begin{bmatrix} 1-\epsilon & \epsilon & 0 \\ 0 & \epsilon & 1-\epsilon \end{bmatrix}$$

$$\begin{aligned} C &= I(X=0;Y)\Big|_{P(X=0)=0.5} \\ &= \sum_y P(y|X=0)\log\frac{P(y|X=0)}{0.5P(y|X=0)+0.5P(y|X=1)} \\ &= (1-\epsilon)\log\left(\frac{1-\epsilon}{\frac{1}{2}(1-\epsilon)}\right)+\epsilon\log\left(\frac{\epsilon}{\epsilon}\right)=1-\epsilon \end{aligned}$$

Symmetric Channel

Capacity Analysis for Symmetric DMC (Symmetric I)

As the symmetric DMC can be viewed as quasi-symmetric DMC, where the channel transition matrix p(y|x) is only partitioned into one set, capacity of symmetric DMC is achieved by using the inputs with equal probability.

Capacity Analysis for Weakly Symmetric DMC

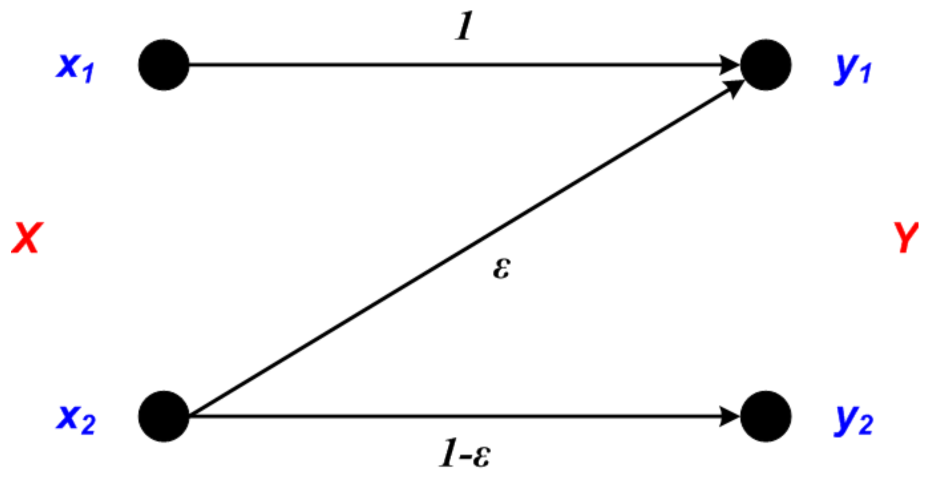
For a weakly symmetric channel, channel capacity is given by

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

and it is achieved by a uniform distribution on input alphabet.

Symmetric Channel

Example: Binary Asymmetric Channel (Z-Channel)



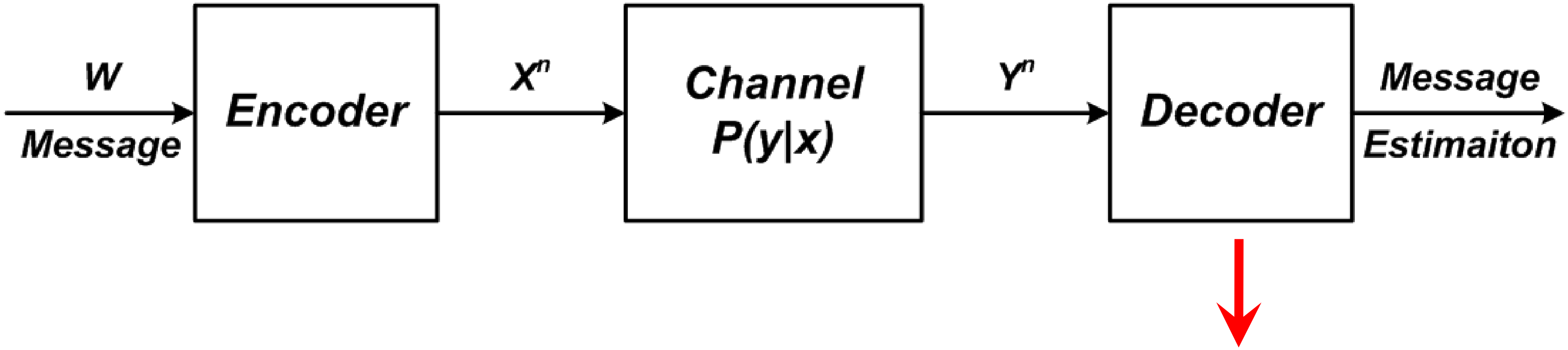
$Q(x_1)=p$ and $Q(x_2)=1-p$

Question:
What is the optimal value of p such that the channel capacity can be achieved?

Outlines

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

Decoding Rule



Objective: Guess X based on the received Y

Decoding rule:
The criteria following which X is viewed to be sent if Y is received.

$$\mathcal{G} : Y \longrightarrow X$$

Decoding Rule

Minimum Error Probability Decoding Rule

- ***Suppose that x_i is transmitted and y_j is received.***
- ***The decoding function is denoted by*** $\mathcal{G} : Y \longrightarrow X$
- ***The conditional error probability***

$$P(e|y_j) = 1 - P\big(\mathcal{G}(y_j) = x_i|y_j\big)$$

- ***The average error probability***

$$P_E = \mathbb{E}_Y\Big\{P(e|y_j)\Big\} = \sum_{y_j} P(y_j)P(e|y_j)$$

Decoding Rule

Minimum Error Probability Decoding Rule

The decoding rule is that we will decode y_j as x_i , i.e.,

$$\mathcal{G}(y_j) = x_i$$

such that the average error probability P_E is minimized.

⇕

Based on our obtained posteriori probabilities, we will decode y_j as x^*, i.e., $\mathcal{G}(y_i) = x^*$, if the requirement is satisfied:

$$\forall x_i \neq x^*, P(x^*|y_j) > P(x_i|y_j)$$

Maximum A Posteriori Probability (MAP) Rule

Decoding Rule

Example

Let the source probability and channel transition matrix are

$$\begin{bmatrix} X \\ P(x) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \\ 1/2 & 1/4 & 1/4 \end{bmatrix} \qquad P(Y|X) = \begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 1/6 & 1/2 & 1/3 \\ 1/3 & 1/6 & 1/2 \end{bmatrix}$$

Find the decoding scheme to minimize P_E.

Decoding Rule

$$P(X,Y) = P(X)P(Y|X) = \begin{bmatrix} 1/4 & 1/6 & 1/12 \\ 1/24 & 1/8 & 1/12 \\ 1/12 & 1/24 & 1/8 \end{bmatrix}$$

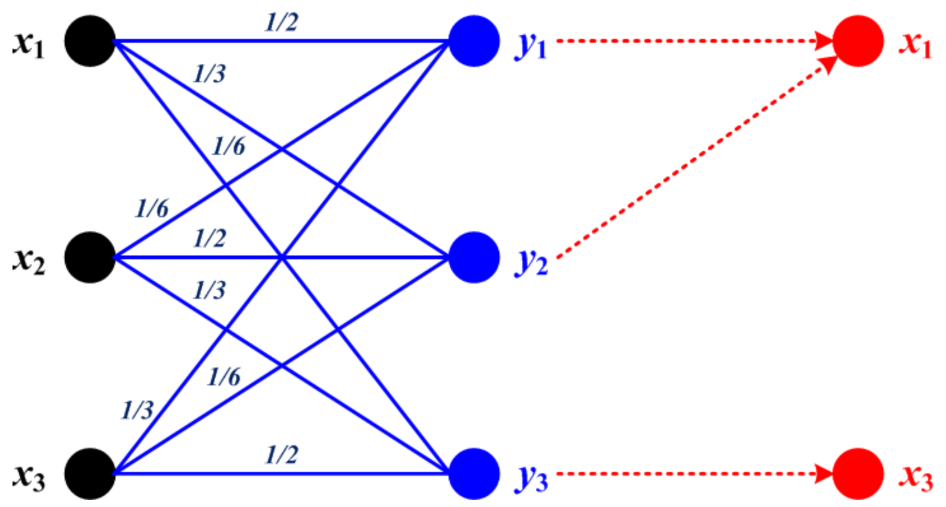
$$\begin{bmatrix} Y \\ P(y) \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & y_3 \\ 3/8 & 1/3 & 7/24 \end{bmatrix}$$

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} = \begin{bmatrix} 2/3 & 1/2 & 2/7 \\ 1/9 & 3/8 & 2/7 \\ 2/9 & 1/8 & 3/7 \end{bmatrix}$$

Decoding Rule

We can obtain the decoding rule:

$$y_1 \longrightarrow x_1;\ y_2 \longrightarrow x_1;\ y_3 \longrightarrow x_3$$



The average error probability:

$$\begin{aligned} P_E &= 1 - P_C \\ &= 1 - \Big[P(y_1)P(x_1|y_1) + P(y_2)P(x_1|y_2) + P(y_3)P(x_3|y_3)\Big] \\ &= 1 - \Big[P(x_1)P(y_1|x_1) + P(x_1)P(y_2|x_1) + P(y_3)P(y_3|x_3)\Big] = \frac{11}{24} \end{aligned}$$

Decoding Rule

Maximum Likelihood Rule

- ***The minimum error probability rule/maximum a posterior probability rule is complex.***

> ***Based on the channel transition probability matrix, we will decode y_j as x^*, i.e., $\mathcal{G}(y_i) = x^*$, if the requirement is satisfied:***
>
> $$\forall x_i \neq x^*, P(y_j|x^*) > P(y_j|x_i)$$

Question:

What is the relationship between minimum error probability rule and maximum likelihood rule?

Decoding 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" and "1" ➔ "111", find minimum P_E.***

Solution

1. ***Find minimum P_E***

$$\epsilon = P\big(Y=1|X=0\big) = P\big(Y=0|X=1\big) = 0.01$$

$$\mathcal{G}(0)=0 \quad \text{and} \quad \mathcal{G}(1)=1$$

$$P(X=0) = P(X=1) = \frac{1}{2}$$

$$P_E = \frac{1}{2}\Big[P\big(Y=1|X=0\big) + P\big(Y=0|X=1\big)\Big] = 10^{-2}$$

Decoding Rule

2. *After the channel code "0" ➔ "000" and "1" ➔ "111", find minimum P_E.*

Channel input: $\alpha_0 = 000$ ***and*** $\alpha_1 = 111$

Channel output: $\beta_0 = 000, \quad \beta_1 = 001, \quad \beta_2 = 010, \quad \beta_3 = 100$

$\beta_4 = 011, \quad \beta_5 = 101, \quad \beta_6 = 110, \quad \beta_7 = 111$

Channel transition matrix:

$$\begin{array}{c|cccccccc} & 000 & 001 & 010 & 100 & 011 & 101 & 110 & 111 \\ \hline 000 & (1-\epsilon)^3 & (1-\epsilon)^2\epsilon & (1-\epsilon)^2\epsilon & (1-\epsilon)^2\epsilon & (1-\epsilon)\epsilon^2 & (1-\epsilon)\epsilon^2 & (1-\epsilon)\epsilon^2 & \epsilon^3 \\ 111 & \epsilon^3 & (1-\epsilon)\epsilon^2 & (1-\epsilon)\epsilon^2 & (1-\epsilon)\epsilon^2 & (1-\epsilon)^2\epsilon & (1-\epsilon)^2\epsilon & (1-\epsilon)^2\epsilon & (1-\epsilon)^3 \end{array}$$

$$\mathcal{G}(\beta_0) = \mathcal{G}(\beta_1) = \mathcal{G}(\beta_2) = \mathcal{G}(\beta_3) = \alpha_0 = 000 \to 0$$
$$\mathcal{G}(\beta_4) = \mathcal{G}(\beta_5) = \mathcal{G}(\beta_6) = \mathcal{G}(\beta_7) = \alpha_1 = 111 \to 1$$

$$P_E = 3(1-\epsilon)\epsilon^2 + \epsilon^3 \approx 3\epsilon^2 = 3 \times 10^{-4}$$

Outlines

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

Joint Typical Set and Joint AEP

Definition (Joint Typical Set)

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:

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

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

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

where

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

Joint Typical Set and Joint AEP

Theorem (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. $\Pr\left\{(X^n, Y^n) \in A_\epsilon^{(n)}\right\} \to 1 \text{ as } n \to \infty$

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

3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ ***[i.e.,*** $\tilde{X}^n$ ***and*** $\tilde{Y}^n$ ***are independent with the same marginals as*** $p(x^n, y^n)$ ***], then***

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

Also, for sufficiently large n,

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

Joint Typical Set and Joint AEP

Proof for Property 1

Based on the weak law of large numbers, we have

$$\exists n_1, \quad \text{for all } n > n_1 \quad \Longrightarrow \quad \Pr\left\{ \left| -\frac{1}{n}\log p(X^n) - H(X) \right| \geq \epsilon \right\} < \frac{\epsilon}{3}$$

$$\exists n_2, \quad \text{for all } n > n_2 \quad \Longrightarrow \quad \Pr\left\{ \left| -\frac{1}{n}\log p(Y^n) - H(Y) \right| \geq \epsilon \right\} < \frac{\epsilon}{3}$$

$$\exists n_3, \quad \text{for all } n > n_3 \quad \Longrightarrow \quad \Pr\left\{ \left| -\frac{1}{n}\log p(X^n, Y^n) - H(X, Y) \right| \geq \epsilon \right\} < \frac{\epsilon}{3}$$

$$\text{For all } n > \max\{n_1, n_2, n_3\} \Longrightarrow \Pr\left\{ (X^n, Y^n) \notin A_\epsilon^{(n)} \right\} < \epsilon$$

For sufficient large n, the probability of the set $A_\epsilon^{(n)}$ is greater than 1-ε, establishing the first part of theorem.

Joint Typical Set and Joint AEP

Proof for Property 2

$$1 = \sum p(x^n, y^n) \geq \sum_{A_\epsilon^{(n)}} p(x^n, y^n) \geq \left|A_\epsilon^{(n)}\right| 2^{-n[H(X,Y)+\epsilon]} \rightarrow \left|A_\epsilon^{(n)}\right| \leq 2^{n[H(X,Y)+\epsilon]}$$

Proof for Property 3

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

$$\leq 2^{n[H(X,Y)+\epsilon]} 2^{-n[H(X)-\epsilon]} 2^{-n[H(Y)-\epsilon]} = 2^{-n[I(X;Y)-3\epsilon]}$$

For sufficient large n, we have $\Pr\left\{A_\epsilon^{(n)}\right\} \geq 1-\epsilon$ ***, and thereore***

$$1-\epsilon \leq \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n, y^n) \leq \left|A_\epsilon^{(n)}\right| 2^{-n[H(X,Y)-\epsilon]} \rightarrow \left|A_\epsilon^{(n)}\right| \geq (1-\epsilon) 2^{n[H(X,Y)-\epsilon]}$$

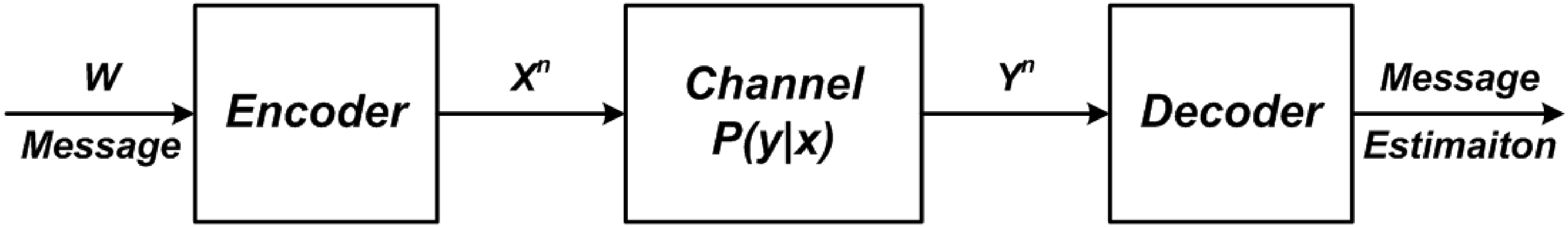
$$\Pr\left\{\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\epsilon^{(n)}\right\} = \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n) p(y^n)$$

$$\geq (1-\epsilon) 2^{n[H(X,Y)-\epsilon]} 2^{-n[H(X)+\epsilon]} 2^{-n[H(Y)+\epsilon]} = (1-\epsilon) 2^{-n[I(X;Y)+3\epsilon]}$$

Outlines

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

Channel Coding Theorem



- ***Message W drawn from the index set {1, 2, …, W} results in signal $X^n(W)$;***

- ***Signal $X^n(W)$ is received as a random sequence $Y^n \sim p(y^n|x^n)$;***

- ***Receiver guesses W by the decoding rule*** $\hat{W} = g\left(Y^n\right)$ ***;***

- ***If the guessed message is not equal to W, an error occurs.***

Channel Coding Theorem

<u>*Definition (Discrete Channel)*</u>

A discrete channel, denoted by $\left(\mathcal{X}, p(y|x), \mathcal{Y}\right)$, consists of two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a collection of probability mass functions p(y|x), one for each $x \in \mathcal{X}$, such that for every x and y, p(y|x)>0, and for every x, $\sum_y p(y|x) = 1$, with the interpretation that X is the input and Y is the output of the channel.

<u>*Definition (Extention)*</u>

The n-th extension of discrete memoryless channel (DMC) is the channel $\left(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n\right)$, where

$$p\left(y_k \middle| x^k, y^{k-1}\right) = p\left(y_k \middle| x_k\right), \quad k = 1, 2, \cdots, n$$

Channel Coding Theorem

Definition

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

1. An index set {1, 2, …, M}.

2. An encoding function $X^n : \{1, 2, \cdots, M\} \to \mathcal{X}^n$***, yielding codewords*** $x^n(1), x^n(2), \ldots, x^n(M)$***. The set of codewords is called the codebook.***

3. A decoding function

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

which is a deterministic rule that assigns a guess to each possible received vector.

Channel Coding Theorem

Definition (Conditional Probability of Error)

The conditional probability of error given that index i was sent is given by

$$\lambda_i = \Pr\left\{g\left(Y^n\right) \neq i \middle| X^n = x^n(i)\right\} = \sum_{y^n} p\left(y^n \middle| x^n(i)\right) I\left(g\left(y^n\right) \neq i\right)$$

where I(·) is the indicator function.

Definition (Maximal Probability of Error)

The maximal probability of error for an (M, n) code is defined as

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

Channel Coding Theorem

Definition (Average Probability of Error)

The (arithmetic) average probability of error $P_e^{(n)}$ for an (M, n) code is defined as

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

Definition (Rate)

The rate R of an (M, n) code is

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

Channel Coding Theorem

Definition (Achievable)

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

$\Downarrow$

The capacity of a channel is the supremum of all achievable rates.

$\Downarrow$

Rates less than the capacity yield arbitrarily small probability of error for sufficiently large block lengths.

Channel Coding Theorem

Theorem (Channel Coding Theorem)

For a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for every rate R < C, there exists a sequence of (2^{nR}, n) codes with maximum probability of error $\lambda^n \to 0$.

Conversely, any sequence of (2^{nR}, n) codes with $\lambda^n \to 0$ ***must have R ≤ C.***

Channel Coding Theorem

For the given p(x), we can generate a (2^{nR}, n) code at random according to the distribution p(x):

$$p\left(x^n\right) = \prod_{i=1}^{n} p\left(x_i\right)$$

Codeword matrix (codebook):

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

$$\Pr\{\mathcal{C}\} = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p\left(x_i(w)\right)$$

Channel Coding Theorem

1. A random code $\mathcal{C}$ is generated according to p(x).

2. The codebook is revealed to both sender and receiver. Both sender and receiver know channel transition matrix p(y|x).

3. A message W is chosen according to a uniform distribution:

$$\Pr\{W = w\} = 2^{-nR}, \quad w = 1, 2, \cdots, 2^{nR}$$

4. The wth codeword $X^n(w)$ is sent over the channel.

5. Receiver gets a sequence Y^n according to the distribution:

$$P\big(y^n \big| x^n(w)\big) = \prod_{i=1}^{n} p\big(y_i \big| x_i(w)\big)$$

Channel Coding Theorem

6. Jointly typical decoding

The receiver declares that the index $\hat{W}$ was sent if the following conditions are satisfied:

- ***If $\left(X^n(\hat{W}), Y^n\right)$ is jointly typical.***
- ***There is no other index $W' \neq \hat{W}$ such that***

$$\left(X^n(W'), Y^n\right) \in A_\epsilon^{(n)}$$

If no such $\hat{W}$ exists or if there is more than one such, an error is declared.

7. There is a decoding error if $\hat{W} \neq W$.

Let $\mathcal{E}$ be the event $\{\hat{W} \neq W\}$.

Channel Coding Theorem

Analysis of the probability of error

Average probability of error:

$$
\begin{aligned}
\Pr\{\mathcal{E}\} &= \sum_{\mathcal{C}} \Pr\{\mathcal{C}\} P_e^{(n)}(\mathcal{C}) \\
&= \sum_{\mathcal{C}} \Pr\{\mathcal{C}\} \boxed{\frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathcal{C})} \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \boxed{\sum_{\mathcal{C}} \Pr\{\mathcal{C}\} \lambda_w(\mathcal{C})}
\end{aligned}
$$

Error caused by jointly typical decoding

What does it mean?

Channel Coding Theorem

Analysis of the probability of error

We assume that the message W=1 was sent. Then, we have

$$\begin{aligned}\Pr\{\mathcal{E}\} &= \frac{1}{2^{nR}}\sum_{w=1}^{2^{nR}}\sum_{\mathcal{C}}\Pr\{\mathcal{C}\}\lambda_w(\mathcal{C}) \\ &= \sum_{\mathcal{C}}\Pr\{\mathcal{C}\}\lambda_1(\mathcal{C}) = \Pr\{\mathcal{E}\big|W=1\}\end{aligned}$$

Define the following events:

$$E_i = \left\{\left(X^n(i), Y^n\right) \text{ is in } A_\epsilon^{(n)}\right\}, \quad i \in \{1, 2, \cdots, 2^{nR}\}$$

Channel Coding Theorem

Analysis of the probability of error

The average probability of error becomes:

$$
\begin{aligned}
\Pr\{\mathcal{E}\} &= \Pr\left\{\mathcal{E}\middle|W=1\right\} \\
&= \Pr\left\{E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}\middle|W=1\right\} \\
&\le \Pr\left\{E_1^c\middle|W=1\right\} + \sum_{i=2}^{2^{nR}} \Pr\left\{E_i\middle|W=1\right\}
\end{aligned}
$$

Channel Coding Theorem

Analysis of the probability of error

For sufficiently large n and R < I(X;Y) - 3ε, we have

$$
\begin{aligned}
\Pr\{\mathcal{E}\} &= \Pr\left\{\mathcal{E} \middle| W=1\right\} \\
&\le \Pr\left\{E_1^c \middle| W=1\right\} + \sum_{i=2}^{2^{nR}} \Pr\left\{E_i \middle| W=1\right\} \\
&\le \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n[I(X;Y)-3\epsilon]} \\
&= \epsilon + \left(2^{nR}-1\right)2^{-n[I(X;Y)-3\epsilon]} \\
&\le \epsilon + 2^{3n\epsilon}2^{-n[I(X;Y)-R]} = \epsilon + 2^{-n[I(X;Y)-3\epsilon-R]} \le 2\epsilon
\end{aligned}
$$

The average probability of error goes to zero.

Channel Coding Theorem

Analysis of the probability of error

1. Choose p(x) to be $p^*(x)$ ***that achieves capacity, then we have***

$$R < I(X;Y) \Longrightarrow R < C$$

2. There must exist one codebook $\mathcal{C}^*$ ***such that*** $\Pr\{\mathcal{E}|\mathcal{C}^*\} \leq 2\epsilon$

$$\Pr\{\mathcal{E}|\mathcal{C}^*\} = \frac{1}{2^{nR}}\sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*)$$

Channel Coding Theorem

Analysis of the probability of error

3. At least half the indices i and their associated codewords $X^n(i)$ have conditional probability of error λ_i less than 4ε.

⇩

The best half of the codewords have a maximal probability of error less than 4ε.

⇩

Throw away the worst half of the codewords, we have 2^{nR-1} codes. Then, the rate changes from R to R-1/n.

⇩

The maximal probability of error $\lambda^{(n)} \leq 4\varepsilon$ for large n.

The achievability of any rate below capacity is proved.

Channel Coding Theorem

The converse to the coding theorem

The index W is uniformly distributed on the set $\mathcal{W} \in \{1, 2 \cdots, 2^{nR}\}$ ***and the sequence*** Y^n ***is related probabilistically to W.***

From Y^n***, we estimate the index W that was sent. For a fixed encoding rule*** $X^n(\cdot)$ ***and a fixed decoding rule*** $\hat{W} = g(Y^n)$ ***, we have*** $W \longrightarrow X^n(W) \longrightarrow Y^n \longrightarrow \hat{W}$ ***.***

Lemma (Fano's inequality)

For a DMC with a codebook $\mathcal{C}$ ***and the input message W uniformly distributed over*** 2^{nR}***, we have***

$$H\left(W|\hat{W}\right) \leq 1 + P_e^{(n)} nR$$

Channel Coding Theorem

The converse to the coding theorem

Lemma

Let Y^n be the result of passing X^n through a DMC of capacity C. Then

$$I(X^n;Y^n) \le C \quad \text{for all } p(x^n)$$

$$\begin{aligned} nR &= H(W) \\ &= H(W|\hat{W}) + I(W;\hat{W}) \\ &\le 1 + P_e^{(n)} nR + I(W;\hat{W}) \\ &\le 1 + P_e^{(n)} nR + I(X^n;Y^n) \\ &\le 1 + P_e^{(n)} nR + nC \end{aligned}$$

$$\Rightarrow \quad R \le P_e^{(n)} R + \frac{1}{n} + C$$

$$\Downarrow \; n \to \infty$$

$$R \le C$$

Review

1. Channel Capacity

<u>Definition (Information Channel Capacity)</u>

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

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

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

⇕ ***Channel Coding Theorem***

<u>Operational Channel Capacity</u>

The highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error.

Review

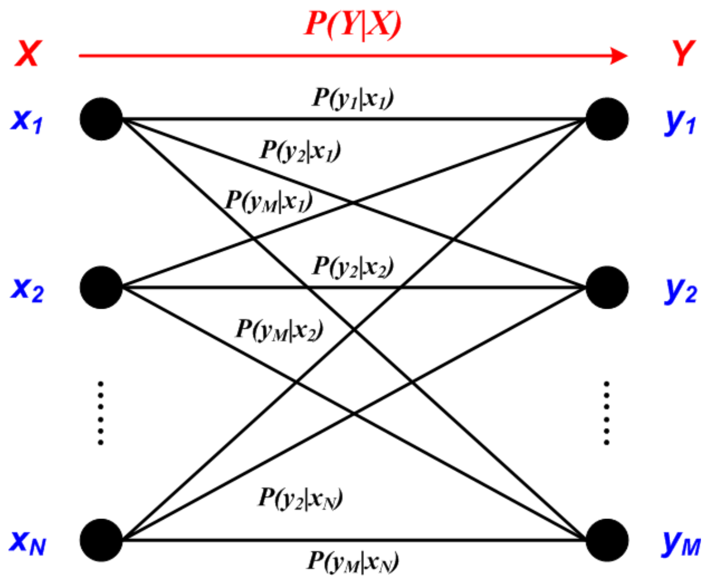
2. Capacity of General DMC

$$\max_{Q(x_1),\cdots,Q(x_N)} I(X;Y)$$

$$= \max_{Q(x_1),\cdots,Q(x_N)} \sum_{k=1}^{N}\sum_{j=1}^{M} Q(x_k)P(y_j|x_k)\log\frac{P(y_j|x_k)}{\sum_{i=1}^{N} Q(x_i)P(y_j|x_i)}$$

$$\text{s.t.} \quad \sum_{k=1}^{N} Q(x_k) = 1$$

$$Q(x_k) \geq 0,\ \ k = 1, 2, \cdots, N$$



Convex optimization problem

↓

Lagrange Method

Review

Theorem

A set of necessary and sufficient conditions on an input probability vector

$$Q(\overline{\mathbf{x}}) = \left[Q(x_1), Q(x_2), \cdots, Q(x_N)\right]$$

to achieve capacity on a discrete memoryless channel with transition probabilities $P(y_j|x_n)$ ***is that for some number C,***

$$I(x_n;Y) = C; \quad \text{all } n \text{ with } Q(x_n) > 0$$

$$I(x_n;Y) \le C; \quad \text{all } n \text{ with } Q(x_n) = 0$$

in which $I(x_n;Y)$ ***is the mutual information for input*** x_n ***averaged over the outputs.***

Furthermore, the number of C is the capacity of the channel.

Review

3. Capacity of Symmetric DMC

Definition (Symmetric)

The channel is defined as symmetric if the rows of the channel transition matrix p(y|x) are permutations of each other and the columns are permutations of each other.

Definition (Quasi-Symmetric)

The channel is defined as quasi-symmetric if the columns of the channel transition matrix p(y|x) 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).

Definition (Weakly Symmetric)

The channel is defined as weakly symmetric if every row of the channel transition matrix p(y|x) is a permutation of every other row and the column sums $\sum_x p(y|x)$ ***are equal.***

Review

Capacity of Quasi-Symmetric DMC

For a quasi-symmetric discrete memoryless channel (DMC), capacity is achieved by using the inputs with equal probability.

Capacity of Symmetric DMC

As the symmetric DMC can be viewed as quasi-symmetric DMC, where the channel transition matrix p(y|x) is only partitioned into one set, capacity of symmetric DMC is achieved by using the inputs with equal probability.

Capacity of Weakly Symmetric DMC

For a weakly symmetric channel, channel capacity is given by

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

and it is achieved by a uniform distribution on input alphabet.

Review

4. Decoding Rule

- ***Minimum Error Probability Decoding Rule/Maximum A Posteriori Probability (MAP) Rule***
- ***Maximum Likelihood Rule***

5. Joint Typical Set

6. Channel Coding Theorem