Chapter 4

Data Transmission and Channel Capacity

Po-Ning Chen, Professor

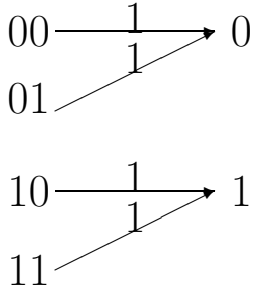
Department of Communications Engineering

National Chiao Tung University

Hsin Chu, Taiwan 30050, R.O.C.

Principle of Data Transmission

- Data transmission
  - To select codewords from the set of channel inputs so that a minimal ambiguity is induced at the channel output.
- **E.g.**, to transmit binary message through the following channel.

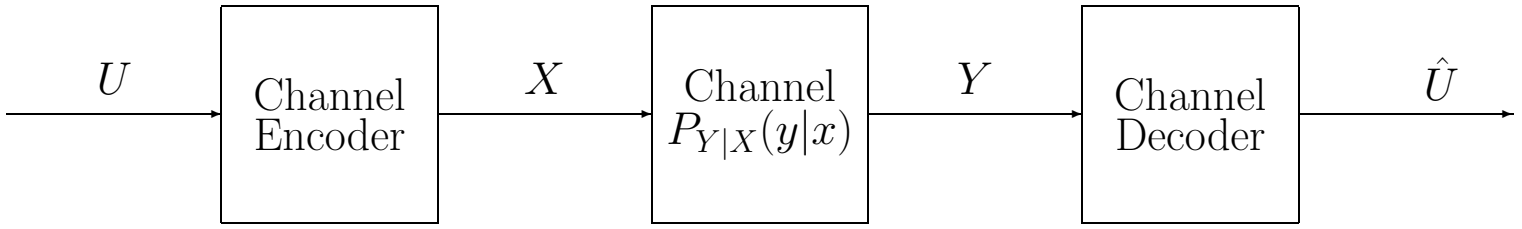


Code of (00 for event A, 10 for event B) obviously induces less ambiguity at the receiver than the code of (00 for event A, 01 for event B).

Reliable Transmission

- Definition of "reliable" transmission
  - The message can be transmitted with arbitrarily small error.
- Objective of data transmission
  - To transform a noisy channel into a reliable channel for the messages intended for transmission.
- How?
  - By taking advantage of the "common" parts (in statistics) between the sender and the receiver, "which are least affected by the noise." The "common" part is the "mutual information."

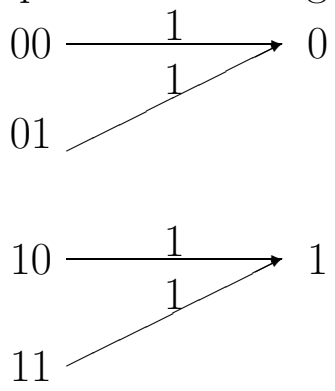
Notations



- A data transmission system, where
  - U represents the message for transmission,
  - X denotes the codeword corresponding to the channel input symbol U,
  - Y represents the received vector due to channel input X,
  - $\hat{U}$ denotes the reconstructed messages from Y.

Query?

- What is the maximum amount of information (per channel input) that can be reliably transmitted via a given noisy channel?
  - **E.g.** We can transmit 1 bit per channel usage by the following code.



    Code = (00 for event A, 10 for event B)

- Intuition
  - The *amount of information* that can be reliably transmitted for a *highly* noisy channel should be less than *that* for a *less* noisy channel.

Data Transmission Codes

- Categories
  - Fixed-length codes
    - Block codes (memoryless fixed-length codes)
      - We will focus on block codes in the data transmission coding theory.
    - Fixed-length tree codes (fixed-length codes with memory)
  - Variable-length codes
    - A **hard** problem.

Definition 4.1 (fixed-length data transmission code) An (n, M) fixed-length data transmission code for channel input alphabet $\mathcal{X}^n$ and output alphabet $\mathcal{Y}^n$ consists of

1. M informational messages intended for transmission;
2. an encoding function
$$f : \{1, 2, \ldots, M\} \to \mathcal{X}^n;$$
3. a decoding function
$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\},$$
which is (usually) a deterministic rule that assigns a guess to each possible received vector.

The channel inputs in $\{x^n \in \mathcal{X}^n \ : \ x^n = f(m) \text{ for some } 1 \leq m \leq M\}$ are the codewords of the data transmission code.

Definition 4.2 (average probability of error) The average probability of error for a $\mathscr{C}_n = (n, M)$ code with encoder $f(\cdot)$ and decoder $g(\cdot)$ transmitted over channel $Q_{Y^n|X^n}$ is defined as

$$P_e(\mathscr{C}_n) = \frac{1}{M} \sum_{i=1}^{M} \lambda_i,$$

where

$$\lambda_i \stackrel{\triangle}{=} \sum_{\left\{y^n \in \mathcal{Y}^n \,:\, g(y^n) \neq i\right\}} Q_{Y^n|X^n}(y^n|f(i)).$$

Under the criterion of average probability of error, all of the codewords are treated equally, namely the prior probability of the selected M codewords are uniformly distributed.

Definition 4.3 (discrete memoryless channel) A discrete memoryless channel (DMC) is a channel whose transition probability $Q_{Y^n|X^n}$ satisfies

$$Q_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} Q_{Y|X}(y_i|x_i).$$

Definition 4.4 (joint typical set) The set $\mathcal{F}_n(\delta)$ of joint δ-typical sequences (x^n, y^n) with respect to the memoryless distribution P_{X^n,Y^n} is defined by

$$\begin{aligned}\mathcal{F}_n(\delta) \triangleq \Bigg\{ & (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \\ & \left| -\frac{1}{n} \log P_{X^n}(x^n) - H(X) \right| < \delta, \ \left| -\frac{1}{n} \log P_{Y^n}(y^n) - H(Y) \right| < \delta, \\ & \text{and } \left| -\frac{1}{n} \log P_{X^n,Y^n}(x^n, y^n) - H(X,Y) \right| < \delta \Bigg\}.\end{aligned}$$

In short, it says that the empirical entropy is δ-close to the true entropy.

Theorem 4.5 (joint AEP) If (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), ... are i.i.d., then

$$-\frac{1}{n}\log P_{X^n}(X_1, X_2, \ldots, X_n) \to H(X) \quad \text{in probability;}$$

$$-\frac{1}{n}\log P_{Y^n}(Y_1, Y_2, \ldots, Y_n) \to H(Y) \quad \text{in probability;}$$

and

$$-\frac{1}{n}\log P_{X^n,Y^n}((X_1, Y_1), \ldots, (X_n, Y_n)) \to H(X, Y) \quad \text{in probability.}$$

Proof: By the weak law of large numbers, we have the desired result. □

Theorem 4.6 (Shannon-McMillan theorem for pairs) Given a dependent pair of DMSs with joint entropy $H(X,Y)$ and any δ greater than zero, we can choose n big enough so that the joint δ-typical set satisfies:

1. $P_{X^n,Y^n}\{\mathcal{F}_n^c(\delta)\} < \delta$ for sufficiently large n.
2. The number of elements in $\mathcal{F}_n(\delta)$ is at least $(1-\delta)e^{n[H(X,Y)-\delta]}$ for sufficiently large n, and at most $e^{n[H(X,Y)+\delta]}$ for every n.
3. If $(x^n,y^n) \in \mathcal{F}_n(\delta)$, its probability of occurrence satisfies

$$e^{-n[H(X,Y)+\delta]} < P_{X^n,Y^n}(x^n,y^n) < e^{-n[H(X,Y)-\delta]}.$$

Proof: The proof is similar to that of Shannon-McMillan theorem for a single memoryless source, and hence we omit it. □

Theorem 4.7 (Shannon's channel coding theorem) Consider a DMC with marginal transition probability $Q_{Y|X}(y|x)$. Define the channel capacity

$$C \stackrel{\triangle}{=} \max_{\{P_{X,Y} \,:\, P_{Y|X}=Q_{Y|X}\}} I(X;Y) = \max_{\{P_X\}} I(P_X, Q_{Y|X}).$$

and fix $\varepsilon > 0$ arbitrarily small. There exist $\gamma > 0$ and a sequence of data transmission block codes $\{\mathscr{C}_n = (n, M_n)\}_{n=1}^{\infty}$ with

$$\frac{1}{n} \log M_n > C - \gamma$$

such that

$$P_e(\mathscr{C}_n) < \varepsilon \quad \text{for sufficiently large } n.$$

Note that the mutual information is actually a function of the input statistics P_X and the channel statistics $Q_{Y|X}$. Hence, we may write it as

$$I(P_X, Q_{Y|X}) \stackrel{\triangle}{=} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_X(x) Q_{Y|X}(y|x) \log \frac{Q_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_X(x') Q_{Y|X}(y|x')}.$$

Such an expression is more suitable in calculating the channel capacity.

Block Codes for Data Transmission

Proof: The concept behind the **random-coding** proof:

- It suffices to prove the *existence* of a good block code sequence (satisfying the rate condition, i.e., $(1/n)\log M_n > C - \gamma$, for some $\gamma > 0$) whose average decoding error is ultimately less than ε.
- In the proof, the good block code sequence is not deterministically designed; instead, its existence is explicitly proven by showing that
  - for a class of block code sequences and a code-selecting distribution over these block code sequences, the expectation value of the average block decoding error, evaluated under the code-selecting distribution on these block code sequences, can be made smaller than ε for n sufficiently large.
- Hence, there must exist such a desired good code sequence among them.

The moral is: *If average midterm grade > 80, then there exists at least one student whose grade is higher than 80.*

Notations

- Fix some γ in $(0, 4\varepsilon)$.
- Observe that there exists N_0 such that for $n > N_0$, we can choose an integer M_n with
$$C - \frac{\gamma}{2} \geq \frac{1}{n} \log M_n > C - \gamma.$$
For $n \leq N_0$, let $M_n \stackrel{\triangle}{=} \left\lceil e^{n(C-\gamma)} \right\rceil + 1$.
- Define $\delta \stackrel{\triangle}{=} \gamma/8$.
- Let $P_{\hat{X}}$ be the probability distribution achieving the channel capacity, i.e.,
$$C \stackrel{\triangle}{=} \max_{\{P_X\}} I(P_X, Q_{Y|X}) = I(P_{\hat{X}}, Q_{Y|X}).$$
- Denote by $P_{\hat{Y}^n}$ the channel output distribution due to the channel input $P_{\hat{X}^n}$ (where $P_{\hat{X}^n}(x^n) = \prod_{i=1}^n P_{\hat{X}}(x_i)$) through channel $Q_{Y^n|X^n}$, i.e.,
$$P_{\hat{X}^n,\hat{Y}^n}(x^n, y^n) \stackrel{\triangle}{=} P_{\hat{X}^n}(x^n) Q_{Y^n|X^n}(y^n|x^n)$$
and
$$P_{\hat{Y}^n}(y^n) \stackrel{\triangle}{=} \sum_{x^n \in \mathcal{X}^n} P_{\hat{X}^n,\hat{Y}^n}(x^n, y^n).$$

We then present the proof in three steps.

Step 1: Code construction. For any blocklength n, independently select M_n channel inputs with replacement from $\mathcal{X}^n$ according to the distribution $P_{\hat{X}^n}(x^n)$, where

$$P_{\hat{X}^n}(x^n) \stackrel{\triangle}{=} \prod_{i=1}^{n} P_{\hat{X}}(x_i).$$

For the selected M_n channel inputs $\mathcal{C}_n \stackrel{\triangle}{=} \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{M_n}\}$, respectively define the encoder and decoder as:

$$f_n(m) = \boldsymbol{c}_m \quad \text{for } 1 \le m \le M_n,$$

and

$$g_n(y^n) = \begin{cases} m, & \text{if } \boldsymbol{c}_m \text{ is the only codeword in } \mathcal{C}_n \\ & \text{satisfying } (\boldsymbol{c}_m, y^n) \in \mathcal{F}_n(\delta); \\ \text{any one in } \{1, 2, \ldots, M_n\}, & \text{otherwise,} \end{cases}$$

where

$$\mathcal{F}_n(\delta) \stackrel{\triangle}{=} \Bigg\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \\ \left|-\frac{1}{n}\log P_{\hat{X}^n}(x^n) - H(\hat{X})\right| < \delta, \\ \left|-\frac{1}{n}\log P_{\hat{Y}^n}(y^n) - H(\hat{Y})\right| < \delta, \\ \text{and } \left|-\frac{1}{n}\log P_{\hat{X}^n,\hat{Y}^n}(x^n, y^n) - H(\hat{X},\hat{Y})\right| < \delta\Bigg\}.$$

Step 2: Error probability.

For the previously defined data transmission code, the conditional probability of error given that message m was sent, denoted by λ_m, can be upper bounded by:

$$\lambda_m \leq \sum_{\left\{y^n \in \mathcal{Y}^n \,:\, (\boldsymbol{c}_m, y^n) \notin \mathcal{F}_n(\delta)\right\}} Q_{Y^n|X^n}(y^n|\boldsymbol{c}_m) + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} \sum_{\left\{y^n \in \mathcal{Y}^n \,:\, (\boldsymbol{c}_{m'}, y^n) \in \mathcal{F}_n(\delta)\right\}} Q_{Y^n|X^n}(y^n|\boldsymbol{c}_m), \tag{4.3.1}$$

where

- the first term in (4.3.1) considers the case that the received channel output y^n is not weakly joint δ-typical with $\boldsymbol{c}_m$, (and hence, the decoding rule $g_n(\cdot)$ will possibly result in a wrong guess)
- and the second term in (4.3.1) reflects the situation when y^n is weakly joint δ-typical with not only the transmitted codeword $\boldsymbol{c}_m$ but also another codeword $\boldsymbol{c}_{m'}$ (which possibly causes error decision in decoding).

By taking the expectation with respect to the m^{th} codeword-selecting distribution $P_{\hat{X}^n}(\boldsymbol{c}_m)$, (4.3.1) can be written as:

$$\begin{aligned} E[\lambda_m] &= \sum_{\boldsymbol{c}_m \in \mathcal{X}^n} P_{\hat{X}^n}(\boldsymbol{c}_m)\lambda_m \\ &\leq \sum_{\boldsymbol{c}_m \in \mathcal{X}^n} \sum_{y^n \notin \mathcal{F}_n(\delta|\boldsymbol{c}_m)} P_{\hat{X}^n}(\boldsymbol{c}_m) Q_{Y^n|X^n}(y^n|\boldsymbol{c}_m) \\ &\quad + \sum_{\boldsymbol{c}_m \in \mathcal{X}^n} \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} \sum_{y^n \in \mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{\hat{X}^n}(\boldsymbol{c}_m) Q_{Y^n|X^n}(y^n|\boldsymbol{c}_m) \\ &= P_{\hat{X}^n,\hat{Y}^n}\left(\mathcal{F}_n^c(\delta)\right) + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} \sum_{\boldsymbol{c}_m \in \mathcal{X}^n} \sum_{y^n \in \mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m, y^n), \end{aligned} \tag{4.3.2}$$

where

$$\mathcal{F}_n(\delta|x^n) \stackrel{\triangle}{=} \{y^n \in \mathcal{Y}^n \ : \ (x^n, y^n) \in \mathcal{F}_n(\delta)\}.$$

Step 3: The expectation of average decoding error $P_e(\boldsymbol{\mathcal{C}}_n)$ (for the M_n selected codewords) with respect to the random selecting code $\boldsymbol{\mathcal{C}}_n$ can be expressed as:

$$
\begin{aligned}
E[P_e(\boldsymbol{\mathcal{C}}_n)] &= \sum_{\boldsymbol{c}_1 \in \mathcal{X}^n} \cdots \sum_{\boldsymbol{c}_{M_n} \in \mathcal{X}^n} P_{\hat{X}^n}(\boldsymbol{c}_1) \cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n}) \left(\frac{1}{M_n} \sum_{m=1}^{M_n} \lambda_m \right) \\
&= \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{\boldsymbol{c}_1 \in \mathcal{X}^n} \cdots \sum_{\boldsymbol{c}_{m-1} \in \mathcal{X}^n} \sum_{\boldsymbol{c}_{m+1} \in \mathcal{X}^n} \cdots \sum_{\boldsymbol{c}_{M_n} \in \mathcal{X}^n} \\
&\quad P_{\hat{X}^n}(\boldsymbol{c}_1) \cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1}) P_{\hat{X}^n}(\boldsymbol{c}_{m+1}) \cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n}) \\
&\quad \times \left(\sum_{\boldsymbol{c}_m \in \mathcal{X}^n} P_{\hat{X}^n}(\boldsymbol{c}_m) \lambda_m \right) \\
&= \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{\boldsymbol{c}_1 \in \mathcal{X}^n} \cdots \sum_{\boldsymbol{c}_{m-1} \in \mathcal{X}^n} \sum_{\boldsymbol{c}_{m+1} \in \mathcal{X}^n} \cdots \sum_{\boldsymbol{c}_{M_n} \in \mathcal{X}^n} \\
&\quad P_{\hat{X}^n}(\boldsymbol{c}_1) \cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1}) P_{\hat{X}^n}(\boldsymbol{c}_{m+1}) \cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n}) \times E[\lambda_m]
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & \frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\boldsymbol{c}_1\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{m-1}\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m+1}\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{M_n}\in\mathcal{X}^n} \\
& P_{\hat{X}^n}(\boldsymbol{c}_1)\cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1})P_{\hat{X}^n}(\boldsymbol{c}_{m+1})\cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n}) \\
& \times\left[P_{\hat{X}^n,\hat{Y}^n}\left(\mathcal{F}_n^c(\delta)\right)\right] \\
+\ & \frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\boldsymbol{c}_1\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{m-1}\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m+1}\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{M_n}\in\mathcal{X}^n} \\
& P_{\hat{X}^n}(\boldsymbol{c}_1)\cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1})P_{\hat{X}^n}(\boldsymbol{c}_{m+1})\cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n}) \\
& \times\left[\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n)\right]
\end{aligned}
\tag{4.3.3}
$$

$$
\begin{aligned}
&= P_{\hat{X}^n,\hat{Y}^n}\left(\mathcal{F}_n^c(\delta)\right) \\
&+ \frac{1}{M_n}\sum_{m=1}^{M_n}\left\{\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\left[\sum_{\boldsymbol{c}_1\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{m-1}\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m+1}\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{M_n}\in\mathcal{X}^n}\right.\right. \\
&\quad P_{\hat{X}^n}(\boldsymbol{c}_1)\cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1})P_{\hat{X}^n}(\boldsymbol{c}_{m+1})\cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n}) \\
&\quad \left.\left.\times\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n)\right]\right\},
\end{aligned}
$$

where (4.3.3) follows from (4.3.2), and the last step holds since

$$P_{\hat{X}^n,\hat{Y}^n}\left(\mathcal{F}_n^c(\delta)\right)$$

is a constant independent of $\boldsymbol{c}_1$, ..., $\boldsymbol{c}_{M_n}$ and m.

Observe that for $n > N_0$,

$$\begin{aligned}
&\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\Bigg[\sum_{\boldsymbol{c}_1\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{m-1}\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m+1}\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{M_n}\in\mathcal{X}^n}\\
&P_{\hat{X}^n}(\boldsymbol{c}_1)\cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1})P_{\hat{X}^n}(\boldsymbol{c}_{m+1})\cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n})\\
&\times\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n)\Bigg]\\
=&\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\Bigg[\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n}(\boldsymbol{c}_{m'})P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n)\Bigg]\\
=&\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\Bigg[\sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n}(\boldsymbol{c}_{m'})\left(\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n)\right)\Bigg]\\
=&\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\Bigg[\sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n}(\boldsymbol{c}_{m'})P_{\hat{Y}^n}(y^n)\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} \left[\sum_{(\boldsymbol{c}_{m'}, y^n) \in \mathcal{F}_n(\delta)} P_{\hat{X}^n}(\boldsymbol{c}_{m'}) P_{\hat{Y}^n}(y^n) \right] \\
&\leq \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} |\mathcal{F}_n(\delta)| e^{-n(H(\hat{X})-\delta)} e^{-n(H(\hat{Y})-\delta)} \\
&\leq \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} e^{n(H(\hat{X},\hat{Y})+\delta)} e^{-n(H(\hat{X})-\delta)} e^{-n(H(\hat{Y})-\delta)} \\
&= (M_n - 1) e^{n(H(\hat{X},\hat{Y})+\delta)} e^{-n(H(\hat{X})-\delta)} e^{-n(H(\hat{Y})-\delta)} \\
&\leq M_n \cdot e^{n(H(\hat{X},\hat{Y})+\delta)} e^{-n(H(\hat{X})-\delta)} e^{-n(H(\hat{Y})-\delta)} \\
&\leq e^{n(C-4\delta)} \cdot e^{-n(I(\hat{X};\hat{Y})-3\delta)} = e^{-n\delta},
\end{aligned}
$$

where the last step follows, since $C = I(\hat{X};\hat{Y})$ by definition of $\hat{X}$ and $\hat{Y}$, and $(1/n)\log M_n \leq C - (\gamma/2) = C - 4\delta$.

Consequently,

$$E[P_e(\boldsymbol{\mathcal{C}}_n)] \leq P_{\hat{X}^n,\hat{Y}^n}\left(\mathcal{F}_n^c(\delta)\right) + e^{-n\delta},$$

which for sufficiently large n (and $n > N_0$), can be made smaller than $2\delta = \gamma/4 < \varepsilon$ by Shannon-McMillan theorem for pairs. □

Data Compression Code Rate (For Stationary Ergodic Sources)

- *Ultimate data compression rate*

$$R = \limsup_{n\to\infty} \frac{1}{n} \log M_n \text{ nats/sourceword}$$

- *Shannon's source coding theorem*
  - *Arbitrary good performance can be achieved by extending the block-length*

$$\Big(\forall\ \varepsilon > 0 \text{ and } \delta > 0\Big)\ (\exists\ \mathscr{C}_n)$$

$$\text{such that} \quad \frac{1}{n} \log M_n < H(X) + \delta \quad \text{and} \quad P_e(\mathscr{C}_n) < \varepsilon.$$

- *How about $R < H(X)$? Answer:*

$$\left(\forall\ \{\mathscr{C}_n\}_{n\geq 1} \text{ with } \limsup_{n\to\infty} \frac{1}{n} \log |\mathscr{C}_n| < H(X)\right) \quad P_e(\mathscr{C}_n) \to 1.$$

Data Transmission Code Rate (For DMC)

- Ultimate data transmission code rate

$$R \stackrel{\triangle}{=} \liminf_{n\to\infty} \frac{1}{n} \log M_n \text{ nats per channel usage.}$$

- Shannon's channel coding theorem
  - Arbitrary good performance can be achieved by extending the blocklength.

$$\left(\forall\ \varepsilon > 0 \text{ and } \gamma > 0\right)\ (\exists\ \mathscr{C}_n)$$

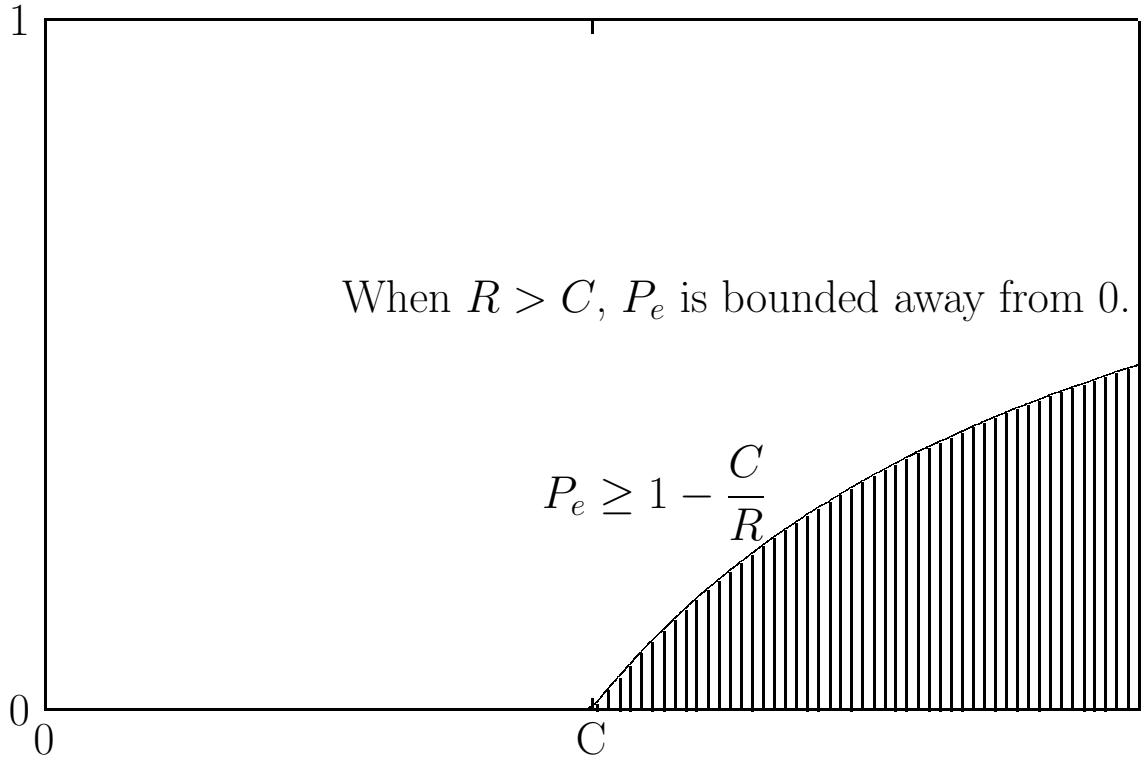
$$\text{such that} \quad \frac{1}{n} \log M_n > C - \gamma \quad \text{and} \quad P_e(\mathscr{C}_n) < \varepsilon.$$

- How about $R > C$? Answer:

$$\left(\forall\ \{\mathscr{C}_n\}_{n\geq 1} \text{ with } R = \liminf_{n\to\infty} \frac{1}{n} \log |\mathscr{C}_n| > C\right)$$

$$\liminf_{n\to\infty} P_e(\mathscr{C}_n) \geq 1 - \frac{C}{R} > 0.$$

Data Transmission Code Rate (For DMC)



Fano's Inequality

Lemma 4.8 (Fano's inequality) Let X and Y be two random variables, correlated in general, with values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, where $\mathcal{X}$ is finite but $\mathcal{Y}$ can be an infinite set. Let $\hat{x} \stackrel{\triangle}{=} g(y)$ be an estimate of x from observing y. Define the probability of estimating error as

$$P_e \stackrel{\triangle}{=} \Pr\{g(Y) \neq X\}.$$

Then for any estimating function $g(\cdot)$,

$$H_b(P_e) + P_e \cdot \log(|\mathcal{X}| - 1) \geq H(X|Y),$$

where $H_b(P_e)$ is the binary entropy function defined by

$$H_b(t) \stackrel{\triangle}{=} -t \cdot \log t - (1-t) \cdot \log(1-t).$$

Fano's Inequality

Proof: Define a new random variable,

$$E \stackrel{\triangle}{=} \begin{cases} 1, & \text{if } g(Y) \neq X \\ 0, & \text{if } g(Y) = X \end{cases} .$$

Then using the chain rule for conditional entropy, we obtain

$$H(E, X|Y) = H(X|Y) + H(E|X, Y) = H(E|Y) + H(X|E, Y). \qquad (4.3.4)$$

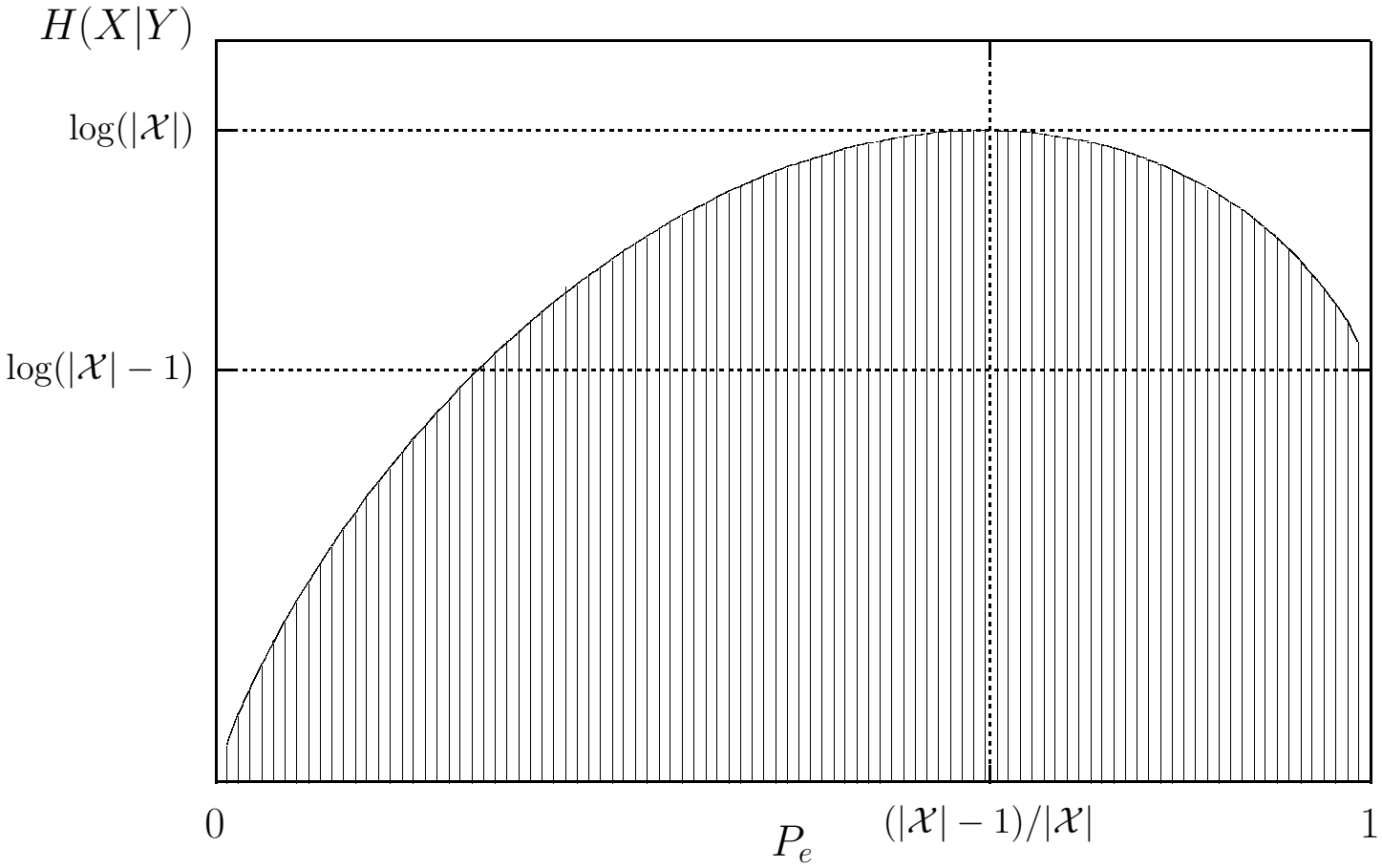
Observe that E is a function of X and Y; hence, $H(E|X, Y) = 0$. Since conditioning never increases entropy, $H(E|Y) \leq H(E) = H_b(P_e)$. The remaining term, $H(X|E, Y)$, can be bounded as follows:

$$\begin{aligned} H(X|E, Y) &= \Pr(E = 0)H(X|Y, E = 0) + \Pr(E = 1)H(X|Y, E = 1) \\ &\leq (1 - P_e) \cdot 0 + P_e \cdot \log(|\mathcal{X}| - 1), \end{aligned}$$

since $X = g(Y)$ for $E = 0$, and given $E = 1$, we can upper bound the conditional entropy by the log of the number of remaining outcomes, i.e., $(|\mathcal{X}| - 1)$. Combining these results, we obtain the Fano's inequality. □

Fano's Inequality

Permissible $(P_e, H(X|Y))$ region of the Fano's inequality.



Sharpness and Tightness

Definitions of sharpness and tightness

- A bound is said to be *sharp* if the bound is achievable for *some specific* cases.
- A bound is said to be *tight* if the bound is achievable for *all* cases.

Fano's inequality is *sharp*.

- There are cases where equality holds for Fano's inequality.

  Example 4.9 Suppose that X and Y are two independent random variables, which are both uniformly distributed over $\{0, 1, 2\}$.
  Let the estimator be $\hat{x} = g(y) = y$.
  Then

$$P_e = \Pr\{g(Y) \neq X\} = \Pr\{Y \neq X\} = 1 - \sum_{x=0}^{2} P_X(x) P_Y(y) = \frac{2}{3}.$$

  In this case, equality holds for the Fano's inequality, i.e.,

$$H_b\left(\frac{2}{3}\right) + \frac{2}{3} \cdot \log(3-1) = H(X|Y) = H(X) = \log(3).$$

Weak Converse for Shannon's Channel Coding Thm.

Theorem 4.10 (weak converse to Shannon's channel coding theorem) Fix a DMC with marginal transition probability $Q_{Y|X}$. For any data transmission code sequence $\{\mathscr{C}_n = (n, M_n)\}_{n=1}^{\infty}$, if

$$\liminf_{n\to\infty} \frac{1}{n} \log M_n > C,$$

the average probability of block decoding error is bounded away from zero for all n sufficiently large.

Proof: For an (n, M_n) block data transmission code, an encoding function is chosen as:

$$f_n : \{1, 2, \ldots, M_n\} \to \mathcal{X}^n,$$

and each index i is equally likely for the average probability of block decoding error criterion. Hence, we can assume that the information message $\{1, 2, \ldots, M_n\}$ is generated from a uniformly distributed random variable, and denote it by W.

As a result,

$$H(W) = \log M_n.$$

Since $W \to X^n \to Y^n$ forms a Markov chain because Y^n only depends on X^n, we obtain by the data processing lemma that $I(W;Y^n) \le I(X^n;Y^n)$. We can also bound $I(X^n;Y^n)$ by C as

$$\begin{aligned}
I(X^n;Y^n) &\le \max_{\{P_{X^n,Y^n} \,:\, P_{Y^n|X^n}=Q_{Y^n|X^n}\}} I(X^n;Y^n) \\
&\le \max_{\{P_{X^n,Y^n} \,:\, P_{Y^n|X^n}=Q_{Y^n|X^n}\}} \sum_{i=1}^{n} I(X_i;Y_i), \text{ (Theorem 2.20)} \\
&\le \sum_{i=1}^{n} \max_{\{P_{X^n,Y^n} \,:\, P_{Y^n|X^n}=Q_{Y^n|X^n}\}} I(X_i;Y_i) \\
&= \sum_{i=1}^{n} \max_{\{P_{X_iY_i} \,:\, P_{Y_i|X_i}=Q_{Y_i|X_i}\}} I(X_i;Y_i) \\
&= nC.
\end{aligned}$$

Weak Converse for Shannon's Channel Coding Thm.

Consequently, by defining $P_e(\mathscr{C}_n)$ as the error of guessing W by observing Y^n via a decoding function

$$g_n : \mathcal{Y}^n \to \{1, 2, \ldots, M_n\},$$

which is exactly the average block decoding failure, we get

$$\begin{aligned}
\log M_n &= H(W) \\
&= H(W|Y^n) + I(W; Y^n) \\
&\leq H(W|Y^n) + I(X^n; Y^n) \\
&\leq H_b(P_e(\mathscr{C}_n)) + P_e(\mathscr{C}_n) \cdot \log(|\mathcal{W}| - 1) + nC, \\
&\quad \text{(by Fano}'\text{s inequality)} \\
&\leq \log(2) + P_e(\mathscr{C}_n) \cdot \log(M_n - 1) + nC, \\
&\quad ((\forall\ t \in [0, 1])\ H_b(t) \leq \log(2)) \\
&\leq \log(2) + P_e(\mathscr{C}_n) \cdot \log M_n + nC,
\end{aligned}$$

which implies that

$$P_e(\mathscr{C}_n) \geq 1 - \frac{C}{(1/n) \log M_n} - \frac{\log(2)}{\log M_n} \quad \left(\stackrel{n \to \infty}{\longrightarrow}\ 1 - \frac{C}{R} \right).$$

Weak Converse for Shannon's Channel Coding Thm.

So if $\liminf_{n\to\infty}(1/n)\log M_n > C$, then there exists δ with $0 < \delta < 4\varepsilon$ and an integer N such that for $n \geq N$,

$$\frac{1}{n}\log M_n > C + \delta.$$

Hence, for $n \geq N_0 \stackrel{\triangle}{=} \max\{N, 2\log(2)/\delta\}$,

$$P_e(\mathscr{C}_n) \geq 1 - \frac{C}{C+\delta} - \frac{\log(2)}{n(C+\delta)} \geq \frac{\delta}{2(C+\delta)}.$$

□

Examples of DMC

- Identity channels

$$Q_{Y|X}(y|x) = \text{either 1 or 0} \;\Rightarrow\; H(Y|X) = 0.$$

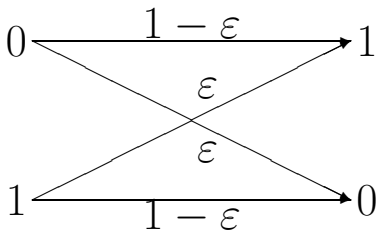
Then

$$\begin{aligned} C &= \max_X I(X;Y) \\ &= \max_X [H(Y) - H(Y|X)] \\ &= \max_X H(Y) \\ &= \log |\mathcal{Y}| \text{ nats/channel usage.} \end{aligned}$$

Examples of DMC

- Binary symmetric channel (BSC)

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x=0}^{1} P_X(x) H(Y|X=x) \\ &= H(Y) - \sum_{x=0}^{1} P_X(x) H_b(\varepsilon) \\ &= H(Y) - H_b(\varepsilon) \\ &\leq \log(2) - H_b(\varepsilon), \end{aligned}$$



where $H_b(u) \stackrel{\triangle}{=} -u \cdot \log u - (1-u) \cdot \log(1-u)$ is the binary entropy function, and the last inequality follows because Y is a binary random variable.
Equality is achieved when $H(Y) = \log(2)$, which is induced by uniform input distribution. Hence,

$$C \stackrel{\triangle}{=} \max_X I(X;Y) = [\log(2) - H_b(\varepsilon)] \text{ nats/channel usage.}$$

Examples of DMC

- Notes on BSC
  - An alternative way to derive the channel capacity for BSC is to first assume $P_X(0) = p = 1 - P_X(1)$, and to express $I(X;Y)$ as:

$$\begin{aligned} I(X;Y) \;=\; & (1-\varepsilon)\log(1-\varepsilon) + \varepsilon\log(\varepsilon) \\ & -[p(1-\varepsilon) + (1-p)\varepsilon]\log[p(1-\varepsilon) + (1-p)\varepsilon] \\ & -[p\varepsilon + (1-p)(1-\varepsilon)]\log[p\varepsilon + (1-p)(1-\varepsilon)]; \end{aligned}$$

    then to maximize the above quantity over $p \in [0,1]$ and yield that the maximizer is $p^* = 1/2$, which immediately gives $C = \log(2) - H_b(\varepsilon)$.

Weakly Symmetric and Symmetric Channels

- Definitions
  - Weakly symmetric channel
    - A channel is said to be *weakly symmetric* if the set

$$\mathcal{A}(x) \stackrel{\triangle}{=} \{P_{Y|X}(y_1|x), \ldots, P_{Y|X}(y_{|\mathcal{Y}|}|x)\}$$

      is identical for every x, and $\sum_{x \in \mathcal{X}} P_{Y|X}(y_i|x)$ equals constant for every $1 \leq i \leq |\mathcal{Y}|$
    - A channel is said to be *weakly symmetric* if every row of the matrix $\left[P_{Y|X}\right]$ is a permutation of the first row, and all the column sums are equal.
  - Symmetric channel
    - A channel is said to be *symmetric* if both

$$\mathcal{A}(x) \stackrel{\triangle}{=} \{P_{Y|X}(y_1|x), \ldots, P_{Y|X}(y_{|\mathcal{Y}|}|x)\} \text{ is identical for every } x,$$

      and

$$\mathcal{B}(y) \stackrel{\triangle}{=} \{P_{Y|X}(y|x_1), \ldots, P_{Y|X}(y|x_{|\mathcal{X}|})\} \text{ is identical for every } y.$$

    - A channel is said to be *symmetric* if every row of the matrix $\left[P_{Y|X}\right]$ is a permutation of the first row, and every column of $\left[P_{Y|X}\right]$ is also a permutation of the first column.

Weakly Symmetric and Symmetric Channels

- Examples of symmetric channels
  - BSC
  - A channel with ternary input and output alphabets and transition probability matrix

$$\begin{array}{lll} P_{Y|X}(0|0) = 0.4, & P_{Y|X}(1|0) = 0.1, & P_{Y|X}(2|0) = 0.5; \\ P_{Y|X}(0|1) = 0.5, & P_{Y|X}(1|1) = 0.4, & P_{Y|X}(2|1) = 0.1; \\ P_{Y|X}(0|2) = 0.1, & P_{Y|X}(1|2) = 0.5, & P_{Y|X}(2|2) = 0.4. \end{array}$$

  - mod-q channel, modelled as:

$$Y = (X + Z) \mod q,$$

    where the channel input X and noise Z are independent, and take value from the same alphabet $\{0, 1, \ldots, q-1\}$.

Weakly Symmetric and Symmetric Channels

- Capacity of symmetric channels

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x\in\mathcal{X}} P_X(x) H(Y|X=x) \\ &= H(Y) - \sum_{x\in\mathcal{X}} P_X(x) H(Z) \\ &= H(Y) - H(Z) \\ &\le \log|\mathcal{Y}| - H(Z), \end{aligned}$$

with equality holds if the output distribution is uniform, which is achieved by uniform input.
Therefore,

$$C \stackrel{\triangle}{=} \max_X I(X;Y) = [\log|\mathcal{Y}| - H(Z)] \text{ nats/channel usage.}$$

Weakly Symmetric and Symmetric Channels

- The capacity being achieved by uniform input is not restricted to *symmetric* and *weakly symmetric channels*, but can be extended to *quasi-symmetric* channel.
  - A quasi-symmetric channel is one that can be partitioned along the column of its channel transition matrix into *weakly symmetric* sub-arrays.
- Example of quasi-symmetric channels: Binary erasure channel with transition matrix

$$\begin{bmatrix} P_{Y|X}(0|0) & P_{Y|X}(1|0) & P_{Y|X}(e|0) \\ P_{Y|X}(0|1) & P_{Y|X}(1|1) & P_{Y|X}(e|1) \end{bmatrix} = \begin{bmatrix} 1-\varepsilon & 0 & \varepsilon \\ 0 & 1-\varepsilon & \varepsilon \end{bmatrix}.$$

  We can partition this transition matrix (along its column) into weakly symmetric sub-arrays as:

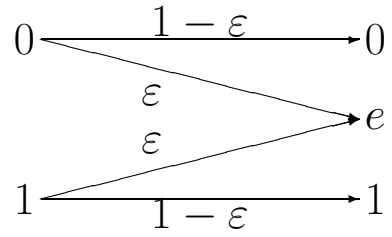
$$\left[\begin{array}{cc|c} 1-\varepsilon & 0 & \varepsilon \\ 0 & 1-\varepsilon & \varepsilon \end{array}\right]$$

- An intuitive interpretation for uniform input achieving the capacity of quasi-symmetric (hence, weakly symmetric and symmetric) channels is that *if the channel treats all input symbol equally, then all input symbols should be used equally often.*

Binary Erasure Channel

- Binary erasure channel (BEC)

$$\begin{aligned} C &= \max_X I(X;Y) \\ &= \max_X [H(Y) - H(Y|X)] \\ &= \max_X [H(Y)] - H_b(\varepsilon) \end{aligned}$$



- Note that $H(Y) \leq \log(3)$ because the size of the output alphabet $\{0, 1, e\}$ is 3, which is achieved by uniform channel output. (Question is "Can unform channel output be obtainable?")
- But since there is no input distribution yielding uniform channel output, we cannot take $\log(3)$ as an achievable maximum value.
- As stated previously, some specific approach needs to be developed for the calculation of its channel capacity.

Self-Mutual Information

Definition 4.11 (mutual information for specific input symbol) Define the mutual information for specific input symbol as:

$$I(x;Y) \stackrel{\triangle}{=} \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\log\frac{P_{Y|X}(y|x)}{P_Y(y)}.$$

- $I(x;Y) = \log\big(1/P_X(x)\big) - H(X=x|Y) = I(x) - H(X=x|Y)$; so $I(x;Y)$ can be conceptually interpreted as the self-information of $X=x$ minus the information that Y has about $X=x$.
- $I(x;Y)$ is the "mutual information" between self-information of $X=x$ and general receiver Y given x is transmitted.

Self-Mutual Information

Observation 4.12 An input distribution P_X achieves the channel capacity C if, and only if,

$$I(x;Y)\begin{cases} = C, & \text{for } P_X(x) > 0; \\ \leq C, & \text{for } P_X(x) = 0. \end{cases}$$

Proof: The *if* part holds straightforwardly; hence, we only provide proof for the *only-if* part.

Without loss of generality, we assume that $P_X(x) < 1$ for all $x \in \mathcal{X}$, since if $P_X(x) = 1$ for some $x \in \mathcal{X}$, $I(X;Y) = 0$ (Notably, $I(X;Y)$, which is the "common part" of $H(X)$ and $H(Y)$, equals zero simply because $H(X) = 0$.)

The problem of calculating the channel capacity is to maximize

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_X(x) Q_{Y|X}(y|x) \log \frac{Q_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_X(x') Q_{Y|X}(y|x')}, \qquad (4.4.2)$$

subject to the condition $\sum_{x\in\mathcal{X}} P_X(x) = 1$ for a given $Q_{Y|X}$.

Self-Mutual Information

We can solve the problem using the Lagrange multiplier argument. Define:

$$f(P_X) \stackrel{\triangle}{=} \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} P_X(x)Q_{Y|X}(y|x)\log\frac{Q_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_X(x')Q_{Y|X}(y|x')} + \lambda\left(\sum_{x\in\mathcal{X}} P_X(x) - 1\right).$$

We then take the derivative of the above quantity with respect to $P_X(x'')$, and obtain

$$\begin{aligned}
\frac{\partial f(P_X)}{\partial P_X(x'')} &= \frac{\partial}{\partial P_X(x'')}\left\{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)Q_{Y|X}(y|x)\log Q_{Y|X}(y|x)\right.\\
&\quad \left. -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)Q_{Y|X}(y|x)\log\left[\sum_{x'\in\mathcal{X}} P_X(x')Q_{Y|X}(y|x')\right] + \lambda\left(\sum_{x\in\mathcal{X}} P_X(x) - 1\right)\right\}\\
&= \sum_{y\in\mathcal{Y}} Q_{Y|X}(y|x'')\log Q_{Y|X}(y|x'') - \left(\sum_{y\in\mathcal{Y}} Q_{Y|X}(y|x'')\log\left[\sum_{x'\in\mathcal{X}} P_X(x')Q_{Y|X}(y|x')\right]\right.\\
&\quad \left. +\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)Q_{Y|X}(y|x)\frac{Q_{Y|X}(y|x'')}{\sum_{x'\in\mathcal{X}} P_X(x')Q_{Y|X}(y|x')}\right) + \lambda
\end{aligned}$$

$$
\begin{aligned}
&= I(x'';Y) - \sum_{y\in\mathcal{Y}} \left[\sum_{x\in\mathcal{X}} P_X(x)Q_{Y|X}(y|x)\right] \frac{Q_{Y|X}(y|x'')}{\sum_{x'\in\mathcal{X}} P_X(x')Q_{Y|X}(y|x')} + \lambda \\
&= I(x'';Y) - \sum_{y\in\mathcal{Y}} Q_{Y|X}(y|x'') + \lambda \\
&= I(x'';Y) - 1 + \lambda.
\end{aligned}
$$

Recall that $I(X;Y) = I(P_X, Q_{Y|X})$ is a concave function in P_X.

- Therefore, the maximum occurs at zero derivative when $P_X(x)$ does not locate in the boundary, namely $1 > P_X(x) > 0$.
- For those $P_X(x)$ locating on the boundary, i.e., $P_X(x) = 0$, the maximum occurs if, and only if, displacement from the boundary to interior decreases the quantity, which implies non-positive derivative, namely

$$I(x;Y) \leq -\lambda + 1, \quad \text{for those } x \text{ with } P_X(x) = 0.$$

Self-Mutual Information

To summarize, if an input distribution P_X achieves the channel capacity, then

$$I(x'';Y) \begin{cases} = -\lambda + 1, & \text{for } P_X(x'') > 0; \\ \leq -\lambda + 1, & \text{for } P_X(x'') = 0. \end{cases}$$

for some λ (With the above result, $C = -\lambda + 1$ trivially hold. Why?), which completes the proof of *only-if* part. □

Capacity of BEC

- From the previous observation, the capacity of BEC should satisfy one of the following three cases:

$$C = I(0;Y) = I(1;Y) \quad \text{for } P_X(0) > 0 \text{ and } P_X(1) > 0 \tag{4.4.4}$$

or

$$C = I(0;Y) \geq I(1;Y) \quad \text{for } P_X(0) = 1 \text{ and } P_X(1) = 0 \tag{4.4.5}$$

or

$$C = I(1;Y) \geq I(0;Y) \quad \text{for } P_X(0) = 0 \text{ and } P_X(1) = 1. \tag{4.4.6}$$

Since (4.4.5) and (4.4.6) only yield uninteresting zero capacity, it remains to verify whether or not (4.4.4) can give a positive capacity.

- By extending (4.4.4), we obtain

$$\begin{aligned} C &= I(0;Y) = -H_b(\varepsilon) - (1-\varepsilon) \cdot \log P_Y(0) - \varepsilon \cdot \log P_Y(e) \\ &= I(1;Y) = -H_b(\varepsilon) - (1-\varepsilon) \cdot \log P_Y(1) - \varepsilon \cdot \log P_Y(e), \end{aligned}$$

which implies $P_Y(0) = P_Y(1)$.

- Since $P_Y(e)$ is always equal to ε, the equality between $P_Y(0)$ and $P_Y(1)$ immediately gives that $P_Y(0) = P_Y(1) = (1-\varepsilon)/2$, and the uniform input process maximizes the channel mutual information.

Self-Mutual Information

- Finally, we obtain that the channel capacity of BEC is equal to

$$\begin{aligned} C &= -H_b(\varepsilon) - (1-\varepsilon)\cdot\log\frac{1-\varepsilon}{2} - \varepsilon\cdot\log(\varepsilon) \\ &= \log(2)\,(1-\varepsilon) \text{ nats/channel usage} \\ &= (1-\varepsilon) \text{ bits/channel usage} \end{aligned}$$

Final remark:

- When n independent bits with single bit error probability ε are transmitted, the expected number of correctly transmitted bits are $n - n\varepsilon = n(1-\varepsilon)$. So when averaging over the total number of bits, it becomes

$$\frac{n(1-\varepsilon)}{n} = (1-\varepsilon) \text{ average correctly transmitted bit per bit.}$$

- BEC is a peculiar case that its channel capacity meets the above number, which said that it is impossible to have reliable transmission efficiency exceeding $(1-\varepsilon)$.

Key Notes

- Definition of reliable transmission
- Discrete memoryless channels
- Data transmission code and its rate
- Joint typical set
- Shannon's channel coding theorem and its converse theorem
- Fano's inequality
- Calculation of channel capacity
  - Identity channels
  - BSC
  - Symmetric and weakly symmetric channels
  - BEC