Chapter 4

Data Transmission and Channel Capacity

Po-Ning Chen, Professor

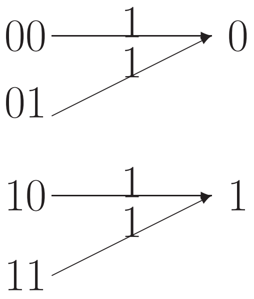
Institute of Communications Engineering

National Chiao Tung University

Hsin Chu, Taiwan 30010, R.O.C.

Principle of Data Transmission

- Data transmission
  - To carefully select *codewords* from the set of channel input words (of a given length) so that a minimal ambiguity is obtained at the channel receiver.
- **E.g.**, to transmit binary message through the following channel.

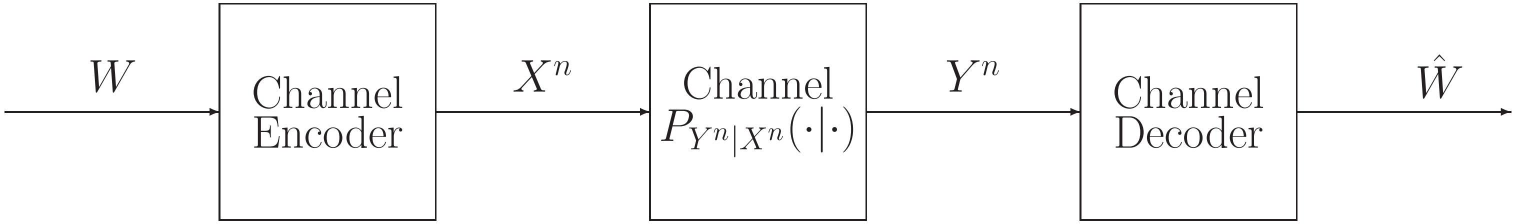


Code of (00 for event A, 10 for event B) obviously induces less ambiguity at the receiver than the code of (00 for event A, 01 for event B).

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 medium for sending messages and recovering them at the receiver.
- How?
  - By taking advantage of the **common parts** between the sender and the receiver sites that are least affected by the channel noise.
  - We will see that these **common parts** are probabilistically captured by the **mutual information** between the channel input and the channel output.

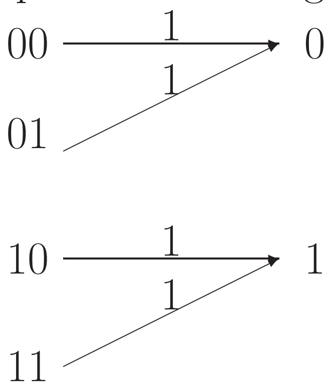
Notations



- A data transmission system, where
  - W represents the message for transmission,
  - $X^n = (X_1, \ldots, X_n)$ denotes the codeword corresponding to the channel input symbol W,
  - $Y^n = (Y_1, \ldots, Y_n)$ represents the received vector due to channel input X^n,
  - $\hat{W}$ denotes the reconstructed messages from Y^n.

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)

Discrete memoryless channels

Definition 4.1 (Discrete channel) A discrete communication channel is characterized by

- A finite input alphabet $\mathcal{X}$.
- A finite output alphabet $\mathcal{Y}$.
- A sequence of n-dimensional transition distributions

$$\{P_{Y^n|X^n}(y^n|x^n)\}_{n=1}^{\infty}$$

such that

$$\sum_{y^n \in \mathcal{Y}^n} P_{Y^n|X^n}(y^n|x^n) = 1$$

for every $x^n \in \mathcal{X}^n$, where $x^n = (x_1, \cdots, x_n) \in \mathcal{X}^n$ and $y^n = (y_1, \cdots, y_n) \in \mathcal{Y}^n$. We assume that the above sequence of n-dimensional distribution is **consistent**, i.e.,

$$P_{Y^i|X^i}(y^i|x^i) = \sum_{x_{i+1}\in\mathcal{X}} \sum_{y_{i+1}\in\mathcal{Y}} P_{X_{i+1}|X^i}(x_{i+1}|x^i) P_{Y^{i+1}|X^{i+1}}(y^{i+1}|x^{i+1})$$

for every x^i, y^i, $P_{X_{i+1}|X^i}$ and $i = 1, 2, \cdots$.

Discrete memoryless channels

Definition 4.2 (Discrete memoryless channel) A discrete memoryless channel (DMC) is a channel whose sequence of transition distributions $P_{Y^n|X^n}$ satisfies

$$P_{Y^n|X^n}(y^n|x^n) \;=\; \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) \qquad (4.2.1)$$

for every $n = 1, 2, \cdots$, $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$. In other words, a DMC is fully described by the channel's transition distribution matrix $\mathbb{Q} := [p_{x,y}]$ of size $|\mathcal{X}| \times |\mathcal{Y}|$, where

$$p_{x,y} := P_{Y|X}(y|x)$$

for $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Furthermore, the matrix $\mathbb{Q}$ is *stochastic*; i.e., the sum of the entries in each of its rows is equal to 1 $\big($since $\sum_{y\in\mathcal{Y}} p_{x,y} = 1$ for all $x \in \mathcal{X}\big)$.

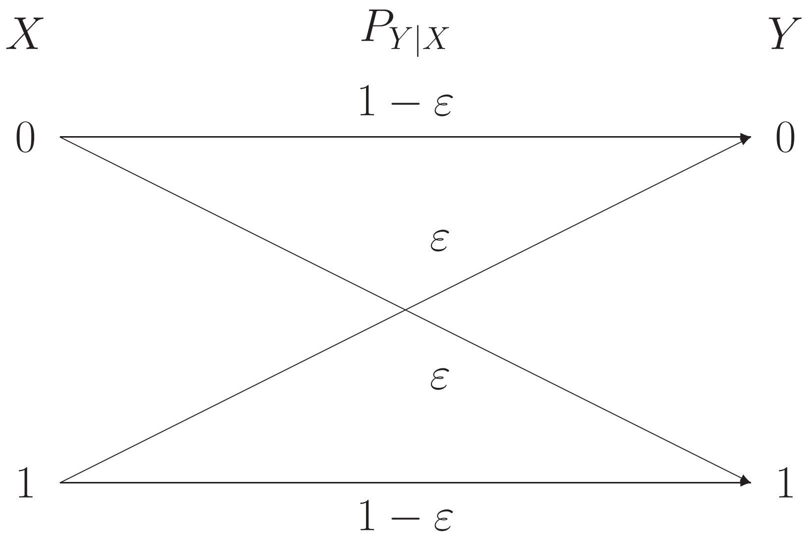
Frequently used channels

1. *Identity (noiseless) channels:* An identity channel has equal-size input and output alphabets ($|\mathcal{X}| = |\mathcal{Y}|$) and channel transition probability satisfying

$$P_{Y|X}(y|x) = \begin{cases} 1 & \text{if } y = x \\ 0 & \text{if } y \neq x. \end{cases}$$

   This is a noiseless or perfect channel as the channel input is received error-free at the channel output.

Frequently used channels



2. *Binary symmetric channels (BSC):*

- $\varepsilon \in [0,1]$ is called the channel's *crossover probability* or *bit error rate.*
- The channel's transition distribution matrix is given by

$$\begin{aligned} \mathbb{Q} &= [p_{x,y}] = \begin{bmatrix} p_{0,0} & p_{0,1} \\ p_{1,0} & p_{1,1} \end{bmatrix} \\ &= \begin{bmatrix} P_{Y|X}(0|0) & P_{Y|X}(1|0) \\ P_{Y|X}(0|1) & P_{Y|X}(1|1) \end{bmatrix} = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix} \end{aligned} \tag{4.2.4}$$

- $\varepsilon = 0$ reduces the BSC to the binary identity (noiseless) channel.

Frequently used channels

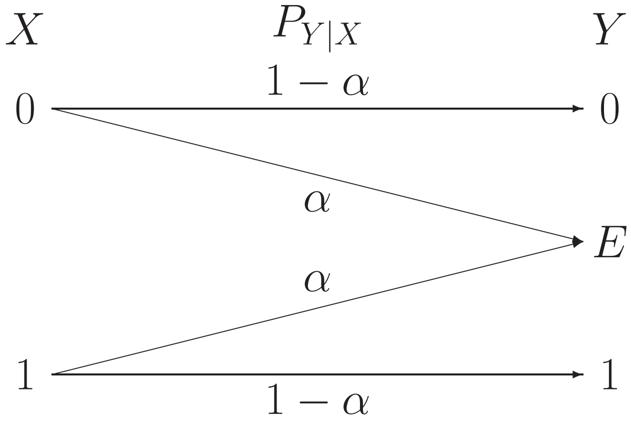
- BSC can be explicitly represented via a binary modulo-2 additive noise channel whose output at time i is the modulo-2 sum of its input and noise variables:

$$Y_i = X_i \oplus Z_i \qquad \text{for } i = 1, 2, \cdots$$

where

$$\begin{cases} \oplus \text{ denotes addition modulo-2,} \\ Y_i,\ X_i \text{ and } Z_i \text{ are the channel output, input and noise, respectively,} \\ \text{the alphabets } \mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\} \text{ are all binary,} \\ X_i \perp Z_j \text{ for any } i, j = 1, 2, \cdots, \text{ and} \\ \text{the noise process is a Bernoulli}(\varepsilon) \text{ process} \\ \qquad\qquad \text{– i.e., a binary i.i.d. process with } \Pr[Z = 1] = \varepsilon. \end{cases} .$$

Frequently used channels



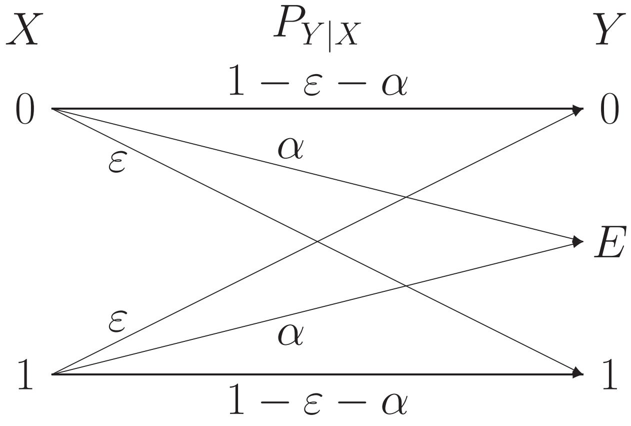
3. *Binary erasure channels (BEC):*

- In BEC, the receiver knows the exact location of the “error” bits in the received bitstream or codeword, but not their actual value.
- These “error” bits are then declared as “erased” during transmission and are called “erasures.”
- The channel transition matrix is given by

$$\begin{aligned}\mathbb{Q} &= [p_{x,y}] = \begin{bmatrix} p_{0,0} & p_{0,E} & p_{0,1} \\ p_{1,0} & p_{1,E} & p_{1,1} \end{bmatrix} \\ &= \begin{bmatrix} P_{Y|X}(0|0) & P_{Y|X}(E|0) & P_{Y|X}(1|0) \\ P_{Y|X}(0|1) & P_{Y|X}(E|1) & P_{Y|X}(1|1) \end{bmatrix} = \begin{bmatrix} 1-\alpha & \alpha & 0 \\ 0 & \alpha & 1-\alpha \end{bmatrix}\end{aligned}$$

where $0 \le \alpha \le 1$ is called the channel's *erasure probability.*

Frequently used channels



4. *Binary symmetric erasure channel (BSEC):*

- One can combine the BSC with the BEC to obtain a binary channel with both errors and erasures.
- The channel's transition matrix is given by

$$\mathbb{Q} = [p_{x,y}] = \begin{bmatrix} p_{0,0} & p_{0,E} & p_{0,1} \\ p_{1,0} & p_{1,E} & p_{1,1} \end{bmatrix} = \begin{bmatrix} 1-\varepsilon-\alpha & \alpha & \varepsilon \\ \varepsilon & \alpha & 1-\varepsilon-\alpha \end{bmatrix} \tag{4.2.8}$$

  where $\varepsilon, \alpha \in [0,1]$ are the channel's crossover and erasure probabilities, respectively.
- Clearly, setting $\alpha = 0$ reduces the BSEC to the BSC, and setting $\varepsilon = 0$ reduces the BSEC to the BEC.

Frequently used channels

- More generally, the channel needs not have a symmetric property in the sense of having identical transition distributions when inputs bits 0 or 1 are sent. For example, the channel's transition matrix can be given by

$$\mathbb{Q} = [p_{x,y}] = \begin{bmatrix} p_{0,0} & p_{0,E} & p_{0,1} \\ p_{1,0} & p_{1,E} & p_{1,1} \end{bmatrix} = \begin{bmatrix} 1-\varepsilon-\alpha & \alpha & \varepsilon \\ \varepsilon' & \alpha' & 1-\varepsilon'-\alpha' \end{bmatrix} \tag{4.2.10}$$

  where the probabilities $\varepsilon \neq \varepsilon'$ and $\alpha \neq \alpha'$ in general. We call such channel, an *asymmetric* channel with errors and erasures.

Frequently used channels

5. *q-ary symmetric channels:*

- Given an integer $q \geq 2$, the q-ary symmetric channel is a nonbinary extension of the BSC; it has alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1, \cdots, q-1\}$ of size q and channel transition matrix given by

$$\begin{aligned}\mathbb{Q} &= [p_{x,y}] \\ &= \begin{bmatrix} p_{0,0} & p_{0,1} & \cdots & p_{0,q-1} \\ p_{1,0} & p_{1,1} & \cdots & p_{1,q-1} \\ \vdots & \vdots & \vdots & \vdots \\ p_{q-1,0} & p_{q-1,1} & \cdots & p_{q-1,q-1} \end{bmatrix} \\ &= \begin{bmatrix} 1-\varepsilon & \frac{\varepsilon}{q-1} & \cdots & \frac{\varepsilon}{q-1} \\ \frac{\varepsilon}{q-1} & 1-\varepsilon & \cdots & \frac{\varepsilon}{q-1} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\varepsilon}{q-1} & \frac{\varepsilon}{q-1} & \cdots & 1-\varepsilon \end{bmatrix} \qquad (4.2.11)\end{aligned}$$

  where $0 \leq \varepsilon \leq 1$ is the channel's *symbol error rate (or probability).*

- When $q = 2$, the channel reduces to the BSC with bit error rate ε, as expected.

Frequently used channels

- Similar to the BSC, the q-ary symmetric channel can be expressed as a modulo-q additive noise channel with common input, output and noise alphabets $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1, \cdots, q-1\}$ and whose output Y_i at time i is given by

$$Y_i = X_i \oplus_q Z_i,$$

for $i = 1, 2, \cdots$, where $\oplus_q$ denotes addition modulo-q, and X_i and Z_i are the channel's input and noise variables, respectively, at time i.

- Here, the noise process $\{Z_n\}_{n=1}^{\infty}$ is assumed to be an i.i.d. process with distribution

$$\Pr[Z = 0] = 1 - \varepsilon \quad \text{and} \quad \Pr[Z = a] = \frac{\varepsilon}{q-1} \quad \forall a \in \{1, \cdots, q-1\}.$$

It is also assumed that the input and noise processes are independent from each other.

Frequently used channels

6. *q-ary erasure channels:*

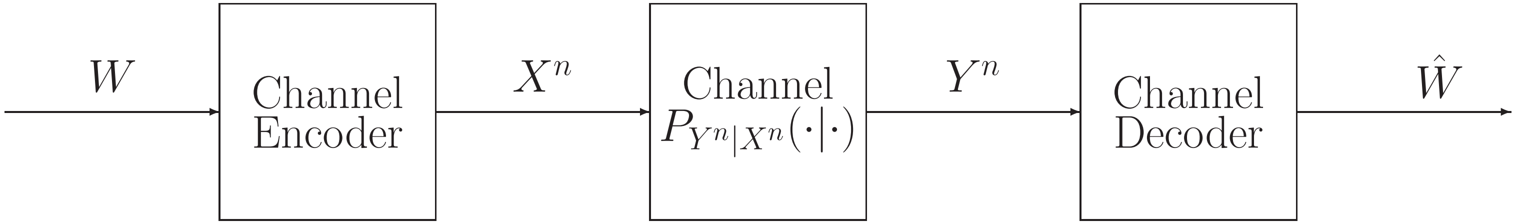
- Given an integer $q \geq 2$, one can also consider a non-binary extension of the BEC, yielding the so called q-ary erasure channel. Specifically, this channel has input and output alphabets given by $\mathcal{X} = \{0, 1, \cdots, q-1\}$ and $\mathcal{Y} = \{0, 1, \cdots, q-1, E\}$, respectively, where E denotes an erasure, and channel transition distribution given by

$$P_{Y|X}(y|x) = \begin{cases} 1-\alpha & \text{if } y = x,\, x \in \mathcal{X} \\ \alpha & \text{if } y = E,\, x \in \mathcal{X} \\ 0 & \text{if } y \neq x,\, x \in \mathcal{X} \end{cases} \tag{4.2.12}$$

  where $0 \leq \alpha \leq 1$ is the erasure probability.

- As expected, setting $q = 2$ reduces the channel to the BEC.

4.3 Block codes for data transmission over DMCs



Definition 4.4 (Fixed-length data transmission code) Given positive integers n and M, and a discrete channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, a fixed-length data transmission code (or block code) for this channel with blocklength n and rate $\frac{1}{n}\log_2 M$ message bits per channel symbol (or channel use) is denoted by $\mathscr{C}_n = (n, M)$ and consists of:

1. M information messages intended for transmission.
2. An encoding function
$$f : \{1, 2, \ldots, M\} \to \mathcal{X}^n$$
yielding codewords $f(1), f(2), \cdots, f(M) \in \mathcal{X}^n$, each of length n. The set of these M codewords is called the codebook and we also usually write $\mathscr{C}_n = \{f(1), f(2), \cdots, f(M)\}$ to list the codewords.
3. A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

4.3 Block codes for data transmission over DMCs

Definition 4.5 (Average probability of error) The average probability of error for a channel block code $\mathscr{C}_n = (n, M)$ ~~code~~ with encoder $f(\cdot)$ and decoder $g(\cdot)$ used over a channel with transition distribution $P_{Y^n|X^n}$ is defined as

$$P_e(\mathscr{C}_n) := \frac{1}{M}\sum_{w=1}^{M}\lambda_w(\mathscr{C}_n),$$

where

$$\begin{aligned}\lambda_w(\mathscr{C}_n) &:= \Pr[\hat{W} \neq W | W = w] = \Pr[g(Y^n) \neq w | X^n = f(w)] \\ &= \sum_{y^n \in \mathcal{Y}^n:\ g(y^n) \neq w} P_{Y^n|X^n}(y^n | f(w))\end{aligned}$$

is the code's conditional probability of decoding error given that message w is sent over the channel.

4.3 Block codes for data transmission over DMCs

Observation 4.6 Another more conservative error criterion is the so-called *maximal probability of error*

$$\lambda(\mathcal{C}_n) := \max_{w\in\{1,2,\cdots,M\}} \lambda_w(\mathcal{C}_n).$$

Clearly,

$$P_e(\mathcal{C}_n = (n, M)) \leq \lambda(\mathcal{C}_n = (n, M));$$

However,

$$2 \times P_e(\mathcal{C}_n = (n, M)) \geq \lambda(\mathcal{C}'_n = (n, M/2)),$$

where $\mathcal{C}'_n$ is constructed by throwing away from $\mathcal{C}_n$ half of its codewords with largest conditional probability of error $\lambda_w(\mathcal{C}_n)$.
So

$$\frac{1}{2}\lambda(\mathcal{C}'_n) \leq P_e(\mathcal{C}_n) \leq \lambda(\mathcal{C}_n)$$

with code rates

$$R = \frac{1}{n}\log_2(M) \quad \text{and} \quad R' = \frac{1}{n}\log_2(M/2) = R - \frac{1}{n}.$$

Consequently, a reliable transmission rate R under the average probability of error criterion is also a reliable transmission rate under the maximal probability of error criterion.

Definition 4.7 (Jointly typical set) The set $\mathcal{F}_n(\delta)$ of jointly δ-typical n-tuple pairs (x^n, y^n) with respect to the memoryless distribution

$$P_{X^n,Y^n}(x^n, y^n) = \prod_{i=1}^{n} P_{X,Y}(x_i, y_i)$$

is defined by

$$\begin{aligned}\mathcal{F}_n(\delta) \;:=\; \Bigg\{ & (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \;: \\ & \left| -\frac{1}{n} \log_2 P_{X^n}(x^n) - H(X) \right| < \delta, \\ & \left| -\frac{1}{n} \log_2 P_{Y^n}(y^n) - H(Y) \right| < \delta, \\ \text{and} \; & \left| -\frac{1}{n} \log_2 P_{X^n,Y^n}(x^n, y^n) - H(X,Y) \right| < \delta \Bigg\}.\end{aligned}$$

In short, a pair (x^n, y^n) generated by independently drawing n times under $P_{X,Y}$ is jointly δ-typical if its joint and marginal empirical entropies are respectively δ-close to the true joint and marginal entropies.

4.3 Block codes for data transmission over DMCs

Theorem 4.8 (Joint AEP) If (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n), … are i.i.d., i.e., $\{(X_i, Y_i)\}_{i=1}^{\infty}$ is a dependent pair of DMSs, then

$$-\frac{1}{n}\log_2 P_{X^n}(X_1, X_2, \ldots, X_n) \to H(X) \quad \text{in probability},$$

$$-\frac{1}{n}\log_2 P_{Y^n}(Y_1, Y_2, \ldots, Y_n) \to H(Y) \quad \text{in probability},$$

and

$$-\frac{1}{n}\log_2 P_{X^n,Y^n}((X_1, Y_1), \ldots, (X_n, Y_n)) \to H(X, Y) \quad \text{in probability}$$

as $n \to \infty$.

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

Theorem 4.9 (Shannon-McMillan-Breiman 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 jointly 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 $(1-\delta)2^{n(H(X,Y)-\delta)}$ for sufficiently large n, and at most $2^{n(H(X,Y)+\delta)}$ for every n.
3. If $(x^n, y^n) \in \mathcal{F}_n(\delta)$, its probability of occurrence satisfies
$$2^{-n(H(X,Y)+\delta)} < P_{X^n,Y^n}(x^n, y^n) < 2^{-n(H(X,Y)-\delta)}.$$

Proof: The proof is quite similar to that of the Shannon-McMillan-Breiman theorem for a single memoryless source presented in the previous chapter; we hence leave it as an exercise. □

4.3 Block codes for data transmission over DMCs

Definition 4.10 (Operational capacity) A rate R is said to be *achievable* for a discrete channel if there exists a sequence of (n, M_n) channel codes $\mathscr{C}_n$ with

$$\liminf_{n\to\infty} \frac{1}{n} \log_2 M_n \geq R \quad \text{and} \quad \lim_{n\to\infty} P_e(\mathscr{C}_n) = 0.$$

The channel's *operational capacity*, C_{op}, is the supremum of all achievable rates:

$$C_{op} = \sup\{R \colon R \text{ is achievable}\}.$$

- The next theorem shows $C_{op} = C$, i.e., the information capacity is equal to the operational capacity.

4.3 Block codes for data transmission over DMCs

Theorem 4.11 (Shannon's channel coding theorem) Consider a DMC with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$ and transition distribution probability $P_{Y|X}(y|x)$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Define the channel capacity (or information capacity)

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

where the maximum is taken over all input distributions P_X. Then the following hold.

- *Forward part (achievability):* For any $0 < \varepsilon < 1$, there exist $\gamma > 0$ and a sequence of data transmission block codes $\{\mathcal{C}_n = (n, M_n)\}_{n=1}^{\infty}$ with

  $$\big(C > \big) \quad \liminf_{n\to\infty} \frac{1}{n}\log_2 M_n \geq C - \gamma$$

  and

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

  where $P_e(\mathcal{C}_n)$ denotes the (average) probability of error for block code $\mathcal{C}_n$.

4.3 Block codes for data transmission over DMCs

- *Converse part:* For any $0 < \varepsilon < 1$, any sequence of data transmission block codes $\{\mathscr{C}_n = (n, M_n)\}_{n=1}^{\infty}$ with
$$\liminf_{n\to\infty} \frac{1}{n}\log_2 M_n > C$$
satisfies
$$P_e(\mathscr{C}_n) > (1-\epsilon)\mu \quad \text{for sufficiently large } n, \tag{4.3.1}$$
where
$$\mu = 1 - \frac{C}{\liminf_{n\to\infty}\frac{1}{n}\log_2 M_n} > 0,$$
i.e., the codes' probability of error is bounded away from zero for all n sufficiently large.

Notes:

- (4.3.1) actually implies that
$$\liminf_{n\to\infty} P_e(\mathscr{C}_n) \geq \lim_{\epsilon\downarrow 0}(1-\epsilon)\mu = \mu,$$
where the error probability lower bound is nothing to do with ϵ. Here we state the converse of Theorem 4.11 in a form in parallel to the converse statements in Theorems 3.6 and 3.15.

4.3 Block codes for data transmission over DMCs

- Also note that the mutual information $I(X;Y)$ is actually a function of the input statistics P_X and the channel statistics $P_{Y|X}$. Hence, we may write it as

  $$I(P_X, P_{Y|X}) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)P_{Y|X}(y|x)\log_2\frac{P_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_X(x')P_{Y|X}(y|x')}.$$

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

- Channel capacity C is well-defined
  - since for a fixed $P_{Y|X}$, $I(P_X, P_{Y|X})$ is **concave** and **continuous** in P_X with respect to both the variational distance and the Euclidean distance (i.e., $\mathcal{L}_2$-distance) [415, Chapter 2], and
  - since the set of all input distributions P_X is a compact (closed and bounded) subset of $\mathbb{R}^{|\mathcal{X}|}$ due to the finiteness of $\mathcal{X}$.

  For the above two reasons, there must exist a P_X that achieves the supremum of the mutual information and the maximum is attainable.

4.3 Block codes for data transmission over DMCs

Idea behind the proof of the forward part:

- It suffices to prove the *existence* of a good block code sequence, satisfying the rate condition,
$$\liminf_{n\to\infty} \frac{1}{n}\log_2 M_n \ge C - \gamma$$
for some $\gamma > 0$, whose average error probability is ultimately less than ε.

- **Random coding argument**:
  - The desired good block code sequence is not **deterministically constructed**;
  - instead, its existence is *implicitly* proven by showing that for a class (ensemble) of block code sequences $\{\mathcal{C}_n\}_{n=1}^{\infty}$ and a code-selecting distribution $\Pr[\mathcal{C}_n]$ over these block code sequences, the expectation value of the average error probability, evaluated under the code-selecting distribution on these block code sequences, can be made smaller than ε for n sufficiently large:
$$E_{\mathcal{C}_n}[P_e(\mathcal{C}_n)] = \sum_{\mathcal{C}_n} \Pr[\mathcal{C}_n] P_e(\mathcal{C}_n) \to 0 \quad \text{as } n \to \infty.$$
  - Hence, there must exist at least one such a desired good code sequence $\{\mathcal{C}_n^*\}_{n=1}^{\infty}$ among them (with $P_e(\mathcal{C}_n^*) \to 0$ as $n \to \infty$).

4.3 Block codes for data transmission over DMCs

Proof of the forward part:

- Since the forward part holds trivially when $C = 0$ by setting $M_n = 1$, we assume in the sequel that $C > 0$.
- Fix $\varepsilon \in (0, 1)$ and some γ with $0 < \gamma < \min\{4\varepsilon, C\}$.
- Observe that there exists N_0 such that for $n > N_0$, we can choose an integer M_n with
$$C - \frac{\gamma}{2} \geq \frac{1}{n} \log_2 M_n > C - \gamma. \tag{4.3.2}$$
(Since we are only concerned with the case of "sufficient large n," it suffices to consider only those n's satisfying $n > N_0$, and ignore those n's for $n \leq N_0$.)
- Define $\delta := \gamma/8$.

4.3 Block codes for data transmission over DMCs

- Let $P_{\hat{X}}$ be the probability distribution achieving the channel capacity:

$$C := \max_{P_X} I(P_X, P_{Y|X}) = I(P_{\hat{X}}, P_{Y|X}).$$

Denote by $P_{\hat{Y}^n}$ the channel output distribution due to channel input product distribution $P_{\hat{X}^n}$ with $P_{\hat{X}^n}(x^n) = \prod_{i=1}^n P_{\hat{X}}(x_i)$; in other words,

$$P_{\hat{Y}^n}(y^n) = \sum_{x^n \in \mathcal{X}^n} P_{\hat{X}^n,\hat{Y}^n}(x^n, y^n)$$

and

$$P_{\hat{X}^n,\hat{Y}^n}(x^n, y^n) := P_{\hat{X}^n}(x^n) P_{Y^n|X^n}(y^n|x^n)$$

for all $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$.

  – Note that since $P_{\hat{X}^n}(x^n) = \prod_{i=1}^n P_{\hat{X}}(x_i)$ and the channel is memoryless, the resulting joint input-output process $\{(\hat{X}_i, \hat{Y}_i)\}_{i=1}^{\infty}$ is also memoryless with

$$P_{\hat{X}^n,\hat{Y}^n}(x^n, y^n) = \prod_{i=1}^n P_{\hat{X},\hat{Y}}(x_i, y_i)$$

and

$$P_{\hat{X},\hat{Y}}(x, y) = P_{\hat{X}}(x) P_{Y|X}(y|x) \quad \text{for } x \in \mathcal{X}, y \in \mathcal{Y}.$$

We next present the proof in three steps.

4.3 Block codes for data transmission over DMCs

Step 1: Code construction.

- For any blocklength n, independently select M_n channel inputs **with replacement** from $\mathcal{X}^n$ according to the distribution $P_{\hat{X}^n}(x^n)$.
- For the selected M_n channel inputs yielding codebook

  $$\mathscr{C}_n := \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{M_n}\},$$

  define the encoder $f_n(\cdot)$ and decoder $g_n(\cdot)$, respectively, as follows:

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

  and

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

  where $\mathcal{F}_n(\delta)$ is defined in Definition 4.7 with respect to distribution $P_{\hat{X}^n,\hat{Y}^n}$. (We assume that the codebook $\mathscr{C}_n$ and the channel distribution $P_{Y|X}$ are known at both the encoder and the decoder.)

4.3 Block codes for data transmission over DMCs

$$\boxed{\begin{aligned}\mathcal{F}_n(\delta) \;:=\; \Big\{&(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \;: \\ &\left|-\frac{1}{n}\log_2 P_{X^n}(x^n) - H(X)\right| < \delta, \quad \left|-\frac{1}{n}\log_2 P_{Y^n}(y^n) - H(Y)\right| < \delta, \\ &\text{and } \left|-\frac{1}{n}\log_2 P_{X^n,Y^n}(x^n, y^n) - H(X,Y)\right| < \delta\Big\}.\end{aligned}}$$

- **Again, let me repeat the encoding and decoding process here!**
  - A message W is chosen according to the uniform distribution from the set of messages.
  - The encoder f_n then transmits the Wth codeword $\boldsymbol{c}_W$ in $\mathscr{C}_n$ over the channel.
  - Then Y^n is received at the channel output and the decoder guesses the sent message via $\hat{W} = g_n(Y^n)$.
  - Note that there is a total $|\mathcal{X}|^{nM_n}$ possible randomly generated codebooks $\mathscr{C}_n$ and the probability of selecting each codebook is given by

$$\Pr[\mathscr{C}_n] = \prod_{m=1}^{M_n} P_{\hat{X}^n}(\boldsymbol{c}_m).$$

4.3 Block codes for data transmission over DMCs

Step 2: Conditional error probability.

- For each (randomly generated) data transmission code $\mathscr{C}_n$, the conditional probability of error given that message m was sent, $\lambda_m(\mathscr{C}_n)$, can be upper bounded by:

$$\begin{aligned}\lambda_m(\mathscr{C}_n) \;\le\; & \sum_{y^n\in\mathcal{Y}^n:\ (\boldsymbol{c}_m,y^n)\notin\mathcal{F}_n(\delta)} P_{Y^n|X^n}(y^n|\boldsymbol{c}_m)\\ & + \sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\ \sum_{y^n\in\mathcal{Y}^n:\ (\boldsymbol{c}_{m'},y^n)\in\mathcal{F}_n(\delta)} P_{Y^n|X^n}(y^n|\boldsymbol{c}_m),\end{aligned} \tag{4.3.3}$$

  where

  – the first term in (4.3.3) considers the case that the received channel output y^n is not jointly δ-typical with $\boldsymbol{c}_m$, (and hence, the decoding rule $g_n(\cdot)$ would possibly result in a wrong guess), and

  – the second term in (4.3.3) reflects the situation when y^n is jointly δ-typical not only with the transmitted codeword $\boldsymbol{c}_m$, but also with another codeword $\boldsymbol{c}_{m'}$ (which may cause a decoding error).

- By taking expectation in (4.3.3) with respect to the m^{th} codeword-selecting distribution $P_{\hat{X}^n}(\boldsymbol{c}_m)$, we obtain

$$\begin{aligned}\sum_{\boldsymbol{c}_m\in\mathcal{X}^n} P_{\hat{X}^n}(\boldsymbol{c}_m)\lambda_m(\mathscr{C}_n) &\le \sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{y^n\notin\mathcal{F}_n(\delta|\boldsymbol{c}_m)} P_{\hat{X}^n}(\boldsymbol{c}_m)P_{Y^n|X^n}(y^n|\boldsymbol{c}_m)\\ &\quad+\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{\hat{X}^n}(\boldsymbol{c}_m)P_{Y^n|X^n}(y^n|\boldsymbol{c}_m)\\ &= P_{\hat{X}^n,\hat{Y}^n}\left(\mathcal{F}_n^c(\delta)\right)\\ &\quad+\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n),\end{aligned}\tag{4.3.4}$$

where

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

4.3 Block codes for data transmission over DMCs

Step 3: Average error probability.

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

$$
\begin{aligned}
\leq\ & \frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\boldsymbol{c}_1\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{m-1}\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m+1}\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{M_n}\in\mathcal{X}^n} \\
& P_{\hat{X}^n}(\boldsymbol{c}_1)\cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1})P_{\hat{X}^n}(\boldsymbol{c}_{m+1})\cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n}) \\
& \times P_{\hat{X}^n,\hat{Y}^n}\left(\mathcal{F}_n^c(\delta)\right) \\
+\ & \frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\boldsymbol{c}_1\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{m-1}\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m+1}\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{M_n}\in\mathcal{X}^n} \\
& P_{\hat{X}^n}(\boldsymbol{c}_1)\cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1})P_{\hat{X}^n}(\boldsymbol{c}_{m+1})\cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n}) \\
& \times\sum_{\substack{m'=1\\ m'\neq m}}^{M_n}\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n) \qquad (4.3.5)
\end{aligned}
$$

4.3 Block codes for data transmission over DMCs

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

where (4.3.5) follows from (4.3.4), 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.

(Then for $n > N_0$)

$$\sum_{\substack{m'=1\\m'\neq m}}^{M_n}\left[\sum_{\boldsymbol{c}_1\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{m-1}\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m+1}\in\mathcal{X}^n}\cdots\sum_{\boldsymbol{c}_{M_n}\in\mathcal{X}^n}\right.$$

$$P_{\hat{X}^n}(\boldsymbol{c}_1)\cdots P_{\hat{X}^n}(\boldsymbol{c}_{m-1})P_{\hat{X}^n}(\boldsymbol{c}_{m+1})\cdots P_{\hat{X}^n}(\boldsymbol{c}_{M_n})$$

$$\left.\times\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n)\right]$$

$$=\sum_{\substack{m'=1\\m'\neq m}}^{M_n}\left[\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n}(\boldsymbol{c}_{m'})P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n)\right]$$

$$=\sum_{\substack{m'=1\\m'\neq m}}^{M_n}\left[\sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n}(\boldsymbol{c}_{m'})\left(\sum_{\boldsymbol{c}_m\in\mathcal{X}^n}P_{\hat{X}^n,\hat{Y}^n}(\boldsymbol{c}_m,y^n)\right)\right]$$

$$=\sum_{\substack{m'=1\\m'\neq m}}^{M_n}\left[\sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})}P_{\hat{X}^n}(\boldsymbol{c}_{m'})P_{\hat{Y}^n}(y^n)\right]$$

4.3 Block codes for data transmission over DMCs

$$\begin{aligned}
&= \sum_{\substack{m'=1\\ m'\neq m}}^{M_n} \left[\sum_{(\boldsymbol{c}_{m'},y^n)\in\mathcal{F}_n(\delta)} P_{\hat{X}^n}(\boldsymbol{c}_{m'})P_{\hat{Y}^n}(y^n)\right] \\
&\le \sum_{\substack{m'=1\\ m'\neq m}}^{M_n} |\mathcal{F}_n(\delta)|2^{-n(H(\hat{X})-\delta)}2^{-n(H(\hat{Y})-\delta)} \\
&\le \sum_{\substack{m'=1\\ m'\neq m}}^{M_n} 2^{n(H(\hat{X},\hat{Y})+\delta)}2^{-n(H(\hat{X})-\delta)}2^{-n(H(\hat{Y})-\delta)} \\
&= (M_n-1)2^{n(H(\hat{X},\hat{Y})+\delta)}2^{-n(H(\hat{X})-\delta)}2^{-n(H(\hat{Y})-\delta)} \\
&< M_n\cdot 2^{n(H(\hat{X},\hat{Y})+\delta)}2^{-n(H(\hat{X})-\delta)}2^{-n(H(\hat{Y})-\delta)} \\
&\le 2^{n(C-4\delta)}\cdot 2^{-n(I(\hat{X};\hat{Y})-3\delta)} = 2^{-n\delta},
\end{aligned}$$

$$\text{where}\begin{cases}\text{the 1st inequality follows from the definition of the jointly typical set } \mathcal{F}_n(\delta),\\ \text{the 2nd inequality holds by the Shannon-McMillan-Breiman theorem for pairs (Theorem 4}\\ \text{the last inequality follows}\begin{cases}\text{since } C = I(\hat{X};\hat{Y}) \text{ by definition of } \hat{X} \text{ and } \hat{Y}\text{, and}\\ \text{since } (1/n)\log_2 M_n \le C-(\gamma/2) = C-4\delta.\end{cases}\end{cases}$$

4.3 Block codes for data transmission over DMCs

Consequently,

$$E_{\sim\mathcal{C}_n}[P_e(\sim\mathcal{C}_n)] \le P_{\hat{X}^n,\hat{Y}^n}(\mathcal{F}_n^c(\delta)) + 2^{-n\delta},$$

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

Fano's inequality

Relation between Fano's inequality and converse proof:

- Consider an (n, M_n) channel block code $\mathcal{C}_n$ with encoding and decoding functions given respectively by
$$f_n : \{1, 2, \cdots, M_n\} \to \mathcal{X}^n$$
and
$$g_n : \mathcal{Y}^n \to \{1, 2, \cdots, M_n\}.$$
- Let message W, which is uniformly distributed over the set of messages $\{1, 2, \cdots, M_n\}$, be sent via codeword $X^n(W) = f_n(W)$ over the DMC.
- Let Y^n be received at the channel output.
- At the receiver, the decoder estimates the sent message via $\hat{W} = g_n(Y^n)$.
- The probability of estimation error is given by the code's average error probability:
$$\Pr[W \neq \hat{W}] = P_e(\mathcal{C}_n).$$
- Then Fano's inequality yields
$$\begin{aligned} H(W|Y^n) &\leq 1 + P_e(\mathcal{C}_n) \log_2(M_n - 1) \\ &< 1 + P_e(\mathcal{C}_n) \log_2 M_n. \end{aligned} \tag{4.3.6}$$

4.3 Block codes for data transmission over DMCs

Proof of the converse part:

- For any (n, M_n) block channel code $\mathcal{C}_n$ as described above, we have that

$$W \to X^n \to Y^n$$

form a Markov chain; we thus obtain by the data processing inequality that

$$I(W;Y^n) \leq I(X^n;Y^n). \tag{4.3.7}$$

- We can also upper bound $I(X^n;Y^n)$ in terms of the channel capacity C as follows

$$\begin{aligned} I(X^n;Y^n) &\leq \max_{P_{X^n}} I(X^n;Y^n) \\ &\leq \max_{P_{X^n}} \sum_{i=1}^{n} I(X_i;Y_i) \quad \text{(by Theorem 2.21: Bounds on mutual information)} \\ &\leq \sum_{i=1}^{n} \max_{P_{X^n}} I(X_i;Y_i) \\ &= \sum_{i=1}^{n} \max_{P_{X_i}} I(X_i;Y_i) = nC. \end{aligned} \tag{4.3.8}$$

4.3 Block codes for data transmission over DMCs

- Consequently, code $\mathcal{C}_n$ satisfies the following:

$$\begin{aligned}\log_2 M_n &= H(W) && \text{(since } W \text{ is uniformly distributed)}\\ &= H(W|Y^n) + I(W;Y^n) &&\\ &\le H(W|Y^n) + I(X^n;Y^n) && \text{(by (4.3.7))}\\ &\le H(W|Y^n) + nC && \text{(by (4.3.8))}\\ &< 1 + P_e(\mathcal{C}_n)\cdot\log_2 M_n + nC. && \text{(by (4.3.6))}\end{aligned}$$

- This implies that

$$P_e(\mathcal{C}_n) > 1 - \frac{C}{(1/n)\log_2 M_n} - \frac{1}{\log_2 M_n} = 1 - \frac{C+1/n}{(1/n)\log_2 M_n}.$$

- So if

$$\liminf_{n\to\infty} \frac{1}{n}\log_2 M_n = \frac{C}{1-\mu},$$

then for any $0 < \varepsilon < 1$, there exists an integer N such that for $n \ge N$,

$$\frac{1}{n}\log_2 M_n \ge \frac{C+1/n}{1-(1-\varepsilon)\mu}, \tag{4.3.9}$$

because, otherwise, (4.3.9) would be violated for infinitely many n, implying a contradiction that

$$\liminf_{n\to\infty} \frac{1}{n}\log_2 M_n \le \liminf_{n\to\infty} \frac{C+1/n}{1-(1-\varepsilon)\mu} = \frac{C}{1-(1-\varepsilon)\mu}.$$

4.3 Block codes for data transmission over DMCs

- Hence, for $n \geq N$,

$$P_e(\mathscr{C}_n) > 1 - [1-(1-\varepsilon)\mu]\,\frac{C+1/n}{C+1/n} = (1-\epsilon)\mu > 0;$$

i.e., $P_e(\mathscr{C}_n)$ is bounded away from zero for n sufficiently large. □

- *Converse part:* For any $0 < \varepsilon < 1$, any sequence of data transmission block codes $\{\mathscr{C}_n = (n, M_n)\}_{n=1}^{\infty}$ with

$$R = \liminf_{n\to\infty} \frac{1}{n}\log_2 M_n > C$$

satisfies

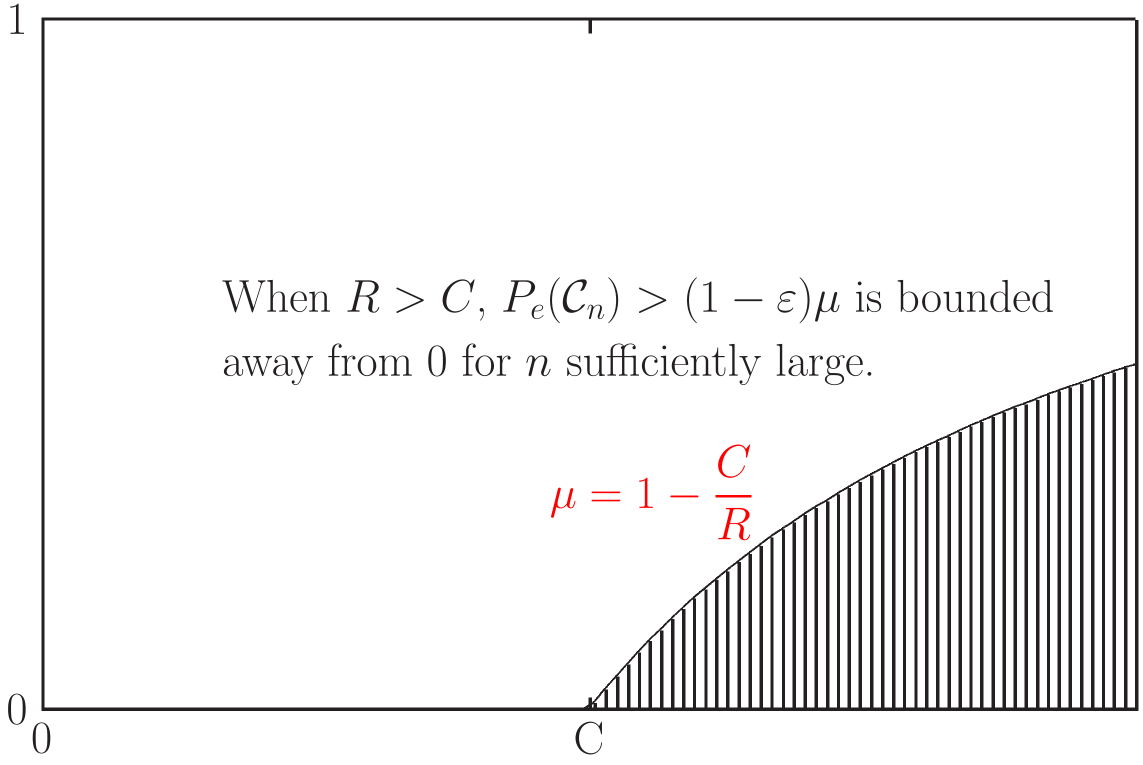
$$P_e(\mathscr{C}_n) > (1-\epsilon)\mu \quad \text{for sufficiently large } n,$$

where

$$\mu = 1 - \frac{C}{\liminf_{n\to\infty}\frac{1}{n}\log_2 M_n} = 1 - \frac{C}{R} > 0,$$

i.e., the codes' probability of error is bounded away from zero for all n sufficiently large.

4.3 Block codes for data transmission over DMCs



When $R > C$, $P_e(\mathcal{C}_n) > (1 - \varepsilon)\mu$ is bounded away from 0 for n sufficiently large.

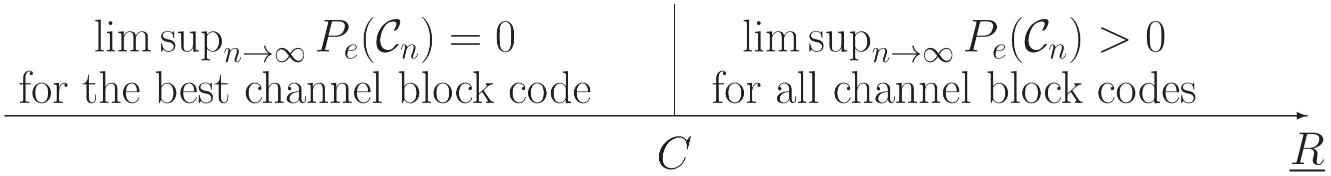
4.3 Block codes for data transmission over DMCs

Observation 4.12

- The results of the above channel coding theorem is illustrated in the figure below, where

$$\underline{R} = \liminf_{n\to\infty} R_n = \liminf_{n\to\infty} (1/n)\log_2 M_n \quad \text{message bits/channel use}$$

  is usually called the *asymptotic* coding rate of channel block codes, and R_n is the code rate for codes of blocklength n.



$\limsup_{n\to\infty} P_e(\mathcal{C}_n) = 0$ for the best channel block code

$\limsup_{n\to\infty} P_e(\mathcal{C}_n) > 0$ for all channel block codes

  - Note that Theorem 4.11 actually indicates

$$\begin{cases} \lim_{n\to\infty} P_e(\mathcal{C}_n) = 0, & \text{for } \underline{R} < C; \\ \liminf_{n\to\infty} P_e(\mathcal{C}_n) > 0, & \text{for } \underline{R} > C \end{cases}$$

  - Such a “two-region” behavior however only holds for a DMC.

4.3 Block codes for data transmission over DMCs

- For a more general channel, three partitions instead of two may result, i.e.,

$$(i)\ \underline{R} < C, \quad (ii)\ C < \underline{R} < \bar{C} \quad \text{and} \quad (iii)\ \underline{R} > \bar{C},$$

which respectively correspond to

$$\begin{cases} (i) \limsup_{n\to\infty} P_e(\mathcal{C}_n) = 0 \text{ for the best block code,} \\ (ii) \limsup_{n\to\infty} P_e(\mathcal{C}_n) > 0 \text{ but } \liminf_{n\to\infty} P_e = 0 \text{ for the best block code, and} \\ (iii) \liminf_{n\to\infty} P_e(\mathcal{C}_n) > 0 \text{ for all channel code codes,} \end{cases}$$

where $\bar{C}$ is named the optimistic channel capacity.

- Since $\bar{C} = C$ for DMCs, the three regions are thus reduced to two.

4.5 Calculating channel capacity

4.5.1 Symmetric, Weakly Symmetric, and Quasi-symmetric Channels

Definition 4.15

- A DMC with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$ and channel transition matrix $\mathbb{Q} = [p_{x,y}]$ of size $|\mathcal{X}| \times |\mathcal{Y}|$ is said to be *symmetric* if the rows of $\mathbb{Q}$ are permutations of each other and the columns of $\mathbb{Q}$ are permutations of each other.
- The channel is said to be *weakly-symmetric* if the rows of $\mathbb{Q}$ are permutations of each other and all the column sums in $\mathbb{Q}$ are equal.

Example of symmetric channel: A ternary DMC channel with $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$ and transition matrix

$$\mathbb{Q} = \begin{bmatrix} P_{Y|X}(0|0) & P_{Y|X}(1|0) & P_{Y|X}(2|0) \\ P_{Y|X}(0|1) & P_{Y|X}(1|1) & P_{Y|X}(2|1) \\ P_{Y|X}(0|2) & P_{Y|X}(1|2) & P_{Y|X}(2|2) \end{bmatrix} = \begin{bmatrix} 0.4 & 0.1 & 0.5 \\ 0.5 & 0.4 & 0.1 \\ 0.1 & 0.5 & 0.4 \end{bmatrix}.$$

Example of weakly symmetric but non-symmetric channel: A quadratry DMC with $|\mathcal{X}| = |\mathcal{Y}| = 4$ and

$$\mathbb{Q} = \begin{bmatrix} 0.5 & 0.25 & 0.25 & 0 \\ 0.5 & 0.25 & 0.25 & 0 \\ 0 & 0.25 & 0.25 & 0.5 \\ 0 & 0.25 & 0.25 & 0.5 \end{bmatrix} \tag{4.5.1}$$

is weakly-symmetric (but not symmetric).

Lemma 4.16 The capacity of a weakly-symmetric channel $\mathbb{Q}$ is achieved by a uniform input distribution and is given by

$$C = \log_2 |\mathcal{Y}| - H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) \tag{4.5.3}$$

where $(q_1, q_2, \cdots, q_{|\mathcal{Y}|})$ denotes any row of $\mathbb{Q}$ and

$$H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) := -\sum_{i=1}^{|\mathcal{Y}|} q_i \log_2 q_i$$

is the row entropy.

4.5 Calculating channel capacity

Proof:

- The mutual information between the channel's input and output is given by

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

  where

$$H(Y|X=x) = -\sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\log_2 P_{Y|X}(y|x) = -\sum_{y\in\mathcal{Y}} p_{x,y}\log_2 p_{x,y}.$$

- Noting that every row of $\mathbb{Q}$ is a permutation of every other row, we obtain that $H(Y|X=x)$ is independent of x and can be written as

$$H(Y|X=x) = H(q_1, q_2, \cdots, q_{|\mathcal{Y}|})$$

  where $(q_1, q_2, \cdots, q_{|\mathcal{Y}|})$ is any row of $\mathbb{Q}$.

4.5 Calculating channel capacity

- Thus

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} P_X(x) H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) \\ &= H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) \left(\sum_{x\in\mathcal{X}} P_X(x) \right) \\ &= H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}). \end{aligned}$$

This implies

$$\begin{aligned} I(X;Y) &= H(Y) - H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) \\ &\le \log_2 |\mathcal{Y}| - H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) \end{aligned}$$

with equality achieved iff Y is uniformly distributed over $\mathcal{Y}$.

- The proof is completed by confirming that for a weakly symmetric channel, the uniform input distribution induces the uniform output distribution (see the text). □

4.5 Calculating channel capacity

Example 4.18 (Capacity of the BSC) Since the BSC with crossover probability (or bit error rate) ε is symmetric, we directly obtain from Lemma 4.16 that its capacity is achieved by a uniform input distribution and is given by

$$C = \log_2(2) - H(1-\varepsilon, \varepsilon) = 1 - h_{\rm b}(\varepsilon) \tag{4.5.5}$$

where $h_{\rm b}(\cdot)$ is the binary entropy function.

Example 4.19 (Capacity of the q-ary symmetric channel) Similarly, the q-ary symmetric channel with symbol error rate ε described in (4.2.11) is symmetric; hence, by Lemma 4.16, its capacity is given by

$$\begin{aligned} C &= \log_2 q - H\left(1-\varepsilon, \frac{\varepsilon}{q-1}, \cdots, \frac{\varepsilon}{q-1}\right) \\ &= \log_2 q + \varepsilon \log_2 \frac{\varepsilon}{q-1} + (1-\varepsilon)\log_2(1-\varepsilon). \end{aligned}$$

Question: Does the uniform input achieve the channel capacity iff the channel is weakly symmetric? No.

4.5 Calculating channel capacity

Definition 4.20 (Quasi-symmetric channels) A DMC with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$ and channel transition matrix $\mathbb{Q} = [p_{x,y}]$ of size $|\mathcal{X}| \times |\mathcal{Y}|$ is said to be *quasi-symmetric* if $\mathbb{Q}$ can be partitioned along its columns into m **weakly-symmetric** sub-matrices $\mathbb{Q}_1, \mathbb{Q}_2, \cdots, \mathbb{Q}_m$ for some integer $m \geq 1$, where each $\mathbb{Q}_i$ sub-matrix has size $|\mathcal{X}| \times |\mathcal{Y}_i|$ for $i = 1, 2, \cdots, m$ with $\mathcal{Y}_1 \cup \cdots \cup \mathcal{Y}_m = \mathcal{Y}$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset \;\; \forall i \neq j,\ i, j = 1, 2, \cdots, m$.

Quasi- = "having some, but not all of the features of" such as quasi-scholar and quasi-official.

- The notion of "quasi-symmetry" we provide here is slightly more general than Gallager's notion [135, p. 94], as we herein allow each sub-matrix to be weakly-symmetric (instead of symmetric as in [135]).

Lemma 4.21 The capacity of a quasi-symmetric channel $\mathbb{Q}$ is achieved by a uniform input distribution and is given by

$$C = \sum_{i=1}^{m} a_i C_i \tag{4.5.6}$$

where

$$a_i := \sum_{y \in \mathcal{Y}_i} p_{x,y} = \text{sum of } \textit{any row} \text{ in } \mathbb{Q}_i, \qquad i = 1, \cdots, m,$$

and

$$C_i = \log_2 |\mathcal{Y}_i| - H\left(\text{any row in the matrix } \tfrac{1}{a_i}\mathbb{Q}_i\right), \qquad i = 1, \cdots, m$$

is the capacity of the ith weakly-symmetric "sub-channel" whose transition matrix is obtained by multiplying each entry of $\mathbb{Q}_i$ by $\frac{1}{a_i}$ (this normalization renders submatrix $\mathbb{Q}_i$ into a stochastic matrix and hence a channel transition matrix).

4.5 Calculating channel capacity

Example 4.22 (Capacity of the BEC) The BEC with erasure probability α and transition matrix

$$\mathbb{Q} = \begin{bmatrix} P_{Y|X}(0|0) & P_{Y|X}(E|0) & P_{Y|X}(1|0) \\ P_{Y|X}(0|1) & P_{Y|X}(E|1) & P_{Y|X}(1|1) \end{bmatrix} = \begin{bmatrix} 1-\alpha & \alpha & 0 \\ 0 & \alpha & 1-\alpha \end{bmatrix}$$

is quasi-symmetric (but neither weakly-symmetric nor symmetric).

- Its transition matrix $\mathbb{Q}$ can be partitioned along its columns into two symmetric (hence weakly-symmetric) sub-matrices

  $$\mathbb{Q}_1 = \begin{bmatrix} 1-\alpha & 0 \\ 0 & 1-\alpha \end{bmatrix}$$

  and

  $$\mathbb{Q}_2 = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}.$$

4.5 Calculating channel capacity

- Thus applying the capacity formula for quasi-symmetric channels of Lemma 4.21 yields that the capacity of the BEC is given by

$$C = a_1 C_1 + a_2 C_2$$

where $a_1 = 1 - \alpha$, $a_2 = \alpha$,

$$C_1 = \log_2(2) - H\left(\frac{1-\alpha}{1-\alpha}, \frac{0}{1-\alpha}\right) = 1 - H(1,0) = 1 - 0 = 1,$$

and

$$C_2 = \log_2(1) - H\left(\frac{\alpha}{\alpha}\right) = 0 - 0 = 0.$$

Therefore, the BEC capacity is given by

$$C = (1-\alpha)(1) + (\alpha)(0) = 1 - \alpha. \qquad (4.5.7)$$

4.5 Calculating channel capacity

Example 4.23 (Capacity of the BSEC) Similarly, the BSEC with crossover probability ε and erasure probability α and transition matrix

$$\mathbb{Q} = [p_{x,y}] = \begin{bmatrix} p_{0,0} & p_{0,E} & p_{0,1} \\ p_{1,0} & p_{1,E} & p_{1,1} \end{bmatrix} = \begin{bmatrix} 1-\varepsilon-\alpha & \alpha & \varepsilon \\ \varepsilon & \alpha & 1-\varepsilon-\alpha \end{bmatrix}$$

is quasi-symmetric; its transition matrix can be partitioned along its columns into two symmetric sub-matrices

$$\mathbb{Q}_1 = \begin{bmatrix} 1-\varepsilon-\alpha & \varepsilon \\ \varepsilon & 1-\varepsilon-\alpha \end{bmatrix}$$

and

$$\mathbb{Q}_2 = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}.$$

Hence by Lemma 4.21, the channel capacity is given by $C = a_1C_1 + a_2C_2$ where $a_1 = 1-\alpha$, $a_2 = \alpha$,

$$C_1 = \log_2(2) - H\left(\frac{1-\varepsilon-\alpha}{1-\alpha}, \frac{\varepsilon}{1-\alpha}\right) = 1 - h_{\mathrm{b}}\left(\frac{1-\varepsilon-\alpha}{1-\alpha}\right),$$

and

$$C_2 = \log_2(1) - H\left(\frac{\alpha}{\alpha}\right) = 0.$$

4.5 Calculating channel capacity

We thus obtain that

$$\begin{aligned} C &= (1-\alpha)\left[1-h_{\mathrm{b}}\left(\frac{1-\varepsilon-\alpha}{1-\alpha}\right)\right]+(\alpha)(0) \\ &= (1-\alpha)\left[1-h_{\mathrm{b}}\left(\frac{1-\varepsilon-\alpha}{1-\alpha}\right)\right]. \end{aligned} \tag{4.5.8}$$

4.5.2 Karuch-Kuhn-Tucker cond. for chan. capacity

Definition 4.24 (Mutual information for a specific input symbol) The mutual information for a specific input symbol is defined as:

$$I(x;Y) := \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)}.$$

From the above definition, the mutual information becomes:

$$\begin{aligned} I(X;Y) &= \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)} \\ &= \sum_{x \in \mathcal{X}} P_X(x) I(x;Y). \end{aligned}$$

4.5.2 Karuch-Kuhn-Tucker cond. for chan. capacity

Lemma 4.25 (KKT condition for channel capacity) For a given DMC, an input distribution P_X achieves its channel capacity iff there exists a constant C such that

$$\begin{cases} I(x:Y) = C & \forall x \in \mathcal{X} \text{ with } P_X(x) > 0; \\ I(x:Y) \le C & \forall x \in \mathcal{X} \text{ with } P_X(x) = 0. \end{cases} \tag{4.5.9}$$

Furthermore, the constant C is the channel capacity (justifying the choice of notation).

Proof: The forward (if) part holds directly; hence, we only prove the converse (only-if) part.

- Without loss of generality, we assume that $P_X(x) < 1$ for all $x \in \mathcal{X}$, since $P_X(x) = 1$ for some x implies that $I(X;Y) = 0$.

4.5.2 Karuch-Kuhn-Tucker cond. for chan. capacity

- The problem of calculating the channel capacity is to maximize

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_X(x') P_{Y|X}(y|x')}, \quad (4.5.10)$$

subject to the condition

$$\sum_{x\in\mathcal{X}} P_X(x) = 1 \quad (4.5.11)$$

for a given channel distribution $P_{Y|X}$.

- By using the Lagrange multiplier method (e.g., see Appendix B.10), maximizing (4.5.10) subject to (4.5.11) is equivalent to maximize:

$$f(P_X) := \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} P_X(x) P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_X(x') P_{Y|X}(y|x')} + \lambda \left(\sum_{x\in\mathcal{X}} P_X(x) - 1 \right).$$

4.5.2 Karuch-Kuhn-Tucker cond. for chan. capacity

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

$$\frac{\partial f(P_X)}{\partial P_X(x'')} = I(x'';Y) - \log_2(e) + \lambda.$$

The details for taking the derivative are as follows:

$$\begin{aligned}
& \frac{\partial}{\partial P_X(x'')} \left\{ \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 P_{Y|X}(y|x) \right. \\
& \left. - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 \left[\sum_{x'\in\mathcal{X}} P_X(x') P_{Y|X}(y|x') \right] + \lambda \left(\sum_{x\in\mathcal{X}} P_X(x) - 1 \right) \right\} \\
= \;& \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x'') \log_2 P_{Y|X}(y|x'') - \left(\sum_{y\in\mathcal{Y}} P_{Y|X}(y|x'') \log_2 \left[\sum_{x'\in\mathcal{X}} P_X(x') P_{Y|X}(y|x') \right] \right. \\
& \left. + \log_2(e) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_X(x) P_{Y|X}(y|x) \frac{P_{Y|X}(y|x'')}{\sum_{x'\in\mathcal{X}} P_X(x') P_{Y|X}(y|x')} \right) + \lambda \\
= \;& I(x'';Y) - \log_2(e) \sum_{y\in\mathcal{Y}} \left[\sum_{x\in\mathcal{X}} P_X(x) P_{Y|X}(y|x) \right] \frac{P_{Y|X}(y|x'')}{\sum_{x'\in\mathcal{X}} P_X(x') P_{Y|X}(y|x')} + \lambda \\
= \;& I(x'';Y) - \log_2(e) \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x'') + \lambda \\
= \;& I(x'';Y) - \log_2(e) + \lambda.
\end{aligned}$$

4.5.2 Karuch-Kuhn-Tucker cond. for chan. capacity

- By Property 2 of Lemma 2.46, $I(X;Y) = I(P_X, P_{Y|X})$ is a **concave** function in P_X (for a fixed $P_{Y|X}$). Therefore,

  1. the maximum of $I(P_X, P_{Y|X})$ occurs for a zero derivative when $P_X(x)$ does not lie on the boundary, namely $1 > P_X(x) > 0$.
  2. For those $P_X(x)$ lying on the boundary, i.e., $P_X(x) = 0$, the maximum occurs iff a displacement from the boundary to the interior decreases the quantity, which implies a non-positive derivative, namely

$$I(x;Y) \le -\lambda + \log_2(e), \quad \text{for those } x \text{ with } P_X(x) = 0.$$

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

$$\begin{cases} I(x'';Y) = -\lambda + \log_2(e), & \text{for } P_X(x'') > 0; \\ I(x'';Y) \le -\lambda + \log_2(e), & \text{for } P_X(x'') = 0. \end{cases}$$

  for some λ.

- With the above result, setting $C = -\lambda + 1$ yields (4.5.9).

- Finally, multiplying both sides of each equation in (4.5.9) by $P_X(x)$ and summing over x yields that $\max_{P_X} I(X;Y)$ on the left and the constant C on the right, thus proving that the constant C is indeed the channel's capacity. □

4.5.2 Karuch-Kuhn-Tucker cond. for chan. capacity

Question: Does the uniform input achieve the channel capacity iff the channel is quasi-symmetric? No.

Observation 4.28 (Capacity achieved by a uniform input distribution)

- T-symmetric channels [319, Section V, Definition 1]: A channel is T-symmetric if

$$T(x) := I(x;Y) - \log_2|\mathcal{X}| = \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\log_2\frac{P_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_{Y|X}(y|x')}$$

  is a constant function of x (i.e., functionally independent of x), where $I(x;Y)$ is the mutual information for input x under a uniform input distribution.

- An example of a T-symmetric channel that is not quasi-symmetric is the binary-input ternary-output channel with the following transition matrix

$$\mathbb{Q} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{bmatrix}.$$

  Its capacity is achieved by the uniform input distribution.

4.5.2 Karuch-Kuhn-Tucker cond. for chan. capacity

- Unlike quasi-symmetric channels, T-symmetric channels do not admit in general a simple closed-form expression for their capacity (such as the one given in (4.5.6)).

$$C = \sum_{i=1}^{m} a_i C_i \tag{4.5.6}$$

4.4 Example of Polar Codes for the BEC

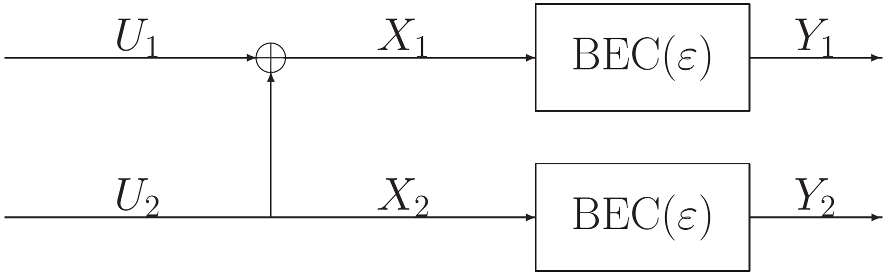
- Polar coding is a new channel coding method proposed by Arikan during 2008-2009, which can provably achieve the capacity of any binary-input memoryless channel $\mathbb{Q}$ whose capacity is realized by a uniform input distribution.

- The main idea behind polar codes is channel "polarization," which transforms n uses of BEC(ε) into extremal "polarized" channels; i.e., channels which are either perfect (noiseless) or completely noisy.

- It is shown that as $n \to \infty$, the number of unpolarized channels converges to 0 and the fraction of perfect channels converges to $I(X;Y) = 1 - \varepsilon$ under a uniform input, which is the capacity of the BEC (see Example 4.22 in Section 4.5).

- A polar code can then be naturally obtained by sending information bits directly through those perfect channels and sending known bits (usually called frozen bits) through the completely noisy channels.

4.4 Example of Polar Codes for the BEC



- We start with the simplest case (often named basic transformation) of $n = 2$.
- Under uniformly distributed X_1 and X_2, we have

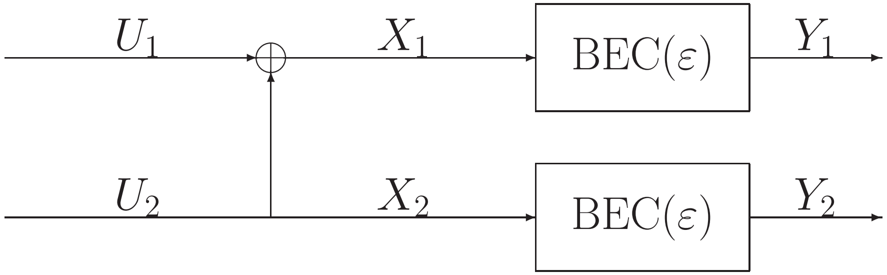
$$I(\mathbb{Q}):=I(X_1;Y_1) = I(X_2;Y_2) = 1 - \varepsilon.$$

- Now consider the following linear modulo-2 operation:

$$\begin{aligned} X_1 &= U_1 \oplus U_2, \\ X_2 &= U_2, \end{aligned}$$

where U_1 and U_2 represent uniformly distributed independent message bits.

4.4 Example of Polar Codes for the BEC

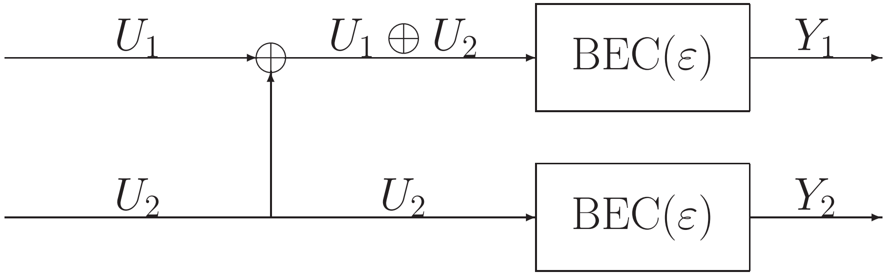


- The decoder performs successive cancellation decoding as follows.
  - It first decodes U_1 from the received (Y_1, Y_2),
  - and then decodes U_2 based on (Y_1, Y_2) and the previously decoded U_1 (assuming the decoding is done correctly).
- This will create two new channels; namely the "worse" channel $\mathbb{Q}^-$ and the "better" channel $\mathbb{Q}^+$ given by

$$\begin{aligned}\mathbb{Q}^- &: U_1 \to (Y_1, Y_2),\\ \mathbb{Q}^+ &: U_2 \to (Y_1, Y_2, U_1),\end{aligned}$$

respectively (the names of these channels will be justified shortly).

4.4 Example of Polar Codes for the BEC

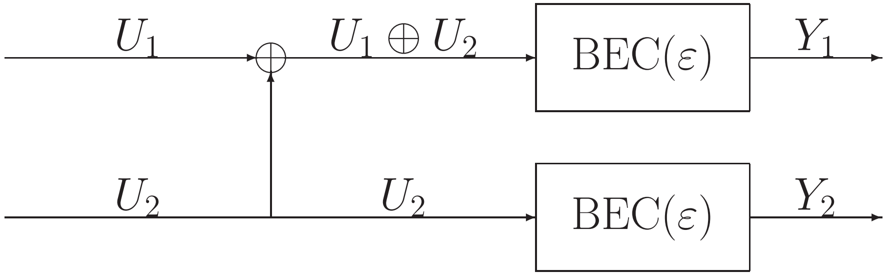


- $\mathbb{Q}^-$: $U_1 = \begin{cases} Y_1 \oplus Y_2, & \text{if } Y_1, Y_2 \in \{0,1\} \\ \;?\; \oplus Y_2, & \text{if } Y_1 = E, Y_2 \in \{0,1\} \\ Y_1 \oplus \;?\,, & \text{if } Y_1 \in \{0,1\}, Y_2 = E \\ \;?\; \oplus \;?\,, & \text{if } Y_1 = Y_2 = E \end{cases}$

  Noting that given output E for a BEC, the receiver knows “nothing” about the input.

- Thus, $\mathbb{Q}^-$ is a BEC with erasure probability $\varepsilon^- := 1 - (1-\varepsilon)^2$.

4.4 Example of Polar Codes for the BEC



- $\mathbb{Q}^+$: $U_2 = \begin{cases} Y_1 \oplus U_1, & \text{if } Y_1 \in \{0,1\} \\ Y_2, & \text{if } Y_2 \in \{0,1\} \\ \ ?\ , & \text{if } Y_1 = Y_2 = E \end{cases}$
- $\mathbb{Q}^+$ is a BEC with erasure probability $\varepsilon^+ := \varepsilon^2$.

Thus, let U_1 be the frozen bit and U_2 be the info bit. One can transform the system to a BEC(ε^2) with code rate 1/2 bits/channel usage.

4.4 Example of Polar Codes for the BEC

The channel capacity remains the same.

$$
\begin{aligned}
I(\mathbb{Q}^+) + I(\mathbb{Q}^-) &= I(U_2; Y_1, Y_2, U_1) + I(U_1; Y_1, Y_2) \\
&= (1 - \varepsilon^2) + [1 - (1 - (1 - \varepsilon)^2)] \\
&= 2(1 - \varepsilon) \\
&= 2I(\mathbb{Q}), \qquad (4.4.1)
\end{aligned}
$$

4.4 Example of Polar Codes for the BEC

- Now, let us consider the case of $n = 4$ and suppose we perform the basic transformation twice to send (i.i.d. uniform) message bits (U_1, U_2, U_3, U_4), yielding

$$\begin{array}{ll} \mathbb{Q}^- : V_1 \to (Y_1, Y_2), & \text{where } X_1 = V_1 \oplus V_2, \\ \mathbb{Q}^+ : V_2 \to (Y_1, Y_2, V_1), & \text{where } X_2 = V_2, \\ \mathbb{Q}^- : V_3 \to (Y_3, Y_4), & \text{where } X_3 = V_3 \oplus V_4, \\ \mathbb{Q}^+ : V_4 \to (Y_3, Y_4, V_3), & \text{where } X_4 = V_4, \end{array}$$

  where $V_1 = U_1 \oplus U_2$, $V_3 = U_2$, $V_2 = U_3 \oplus U_4$ and $V_4 = U_4$.

- $$\begin{cases} \mathbb{Q}^{--} : U_1 \to (Y_1, Y_2, Y_3, Y_4) & \text{with erasure probability } \varepsilon^{--} := 1 - (1 - \varepsilon^-)^2 \\ \mathbb{Q}^{+-} : U_3 \to (Y_1, Y_2, Y_3, Y_4, U_1, U_2) & \text{with erasure probability } \varepsilon^{+-} := 1 - (1 - \varepsilon^+)^2 \\ \mathbb{Q}^{-+} : U_2 \to (Y_1, Y_2, Y_3, Y_4, U_1) & \text{with erasure probability } \varepsilon^{-+} := (\varepsilon^-)^2 \\ \mathbb{Q}^{++} : U_4 \to (Y_1, Y_2, Y_3, Y_4, U_1, U_3, U_2) & \text{with erasure probability } \varepsilon^{++} := (\varepsilon^+)^2. \end{cases}$$

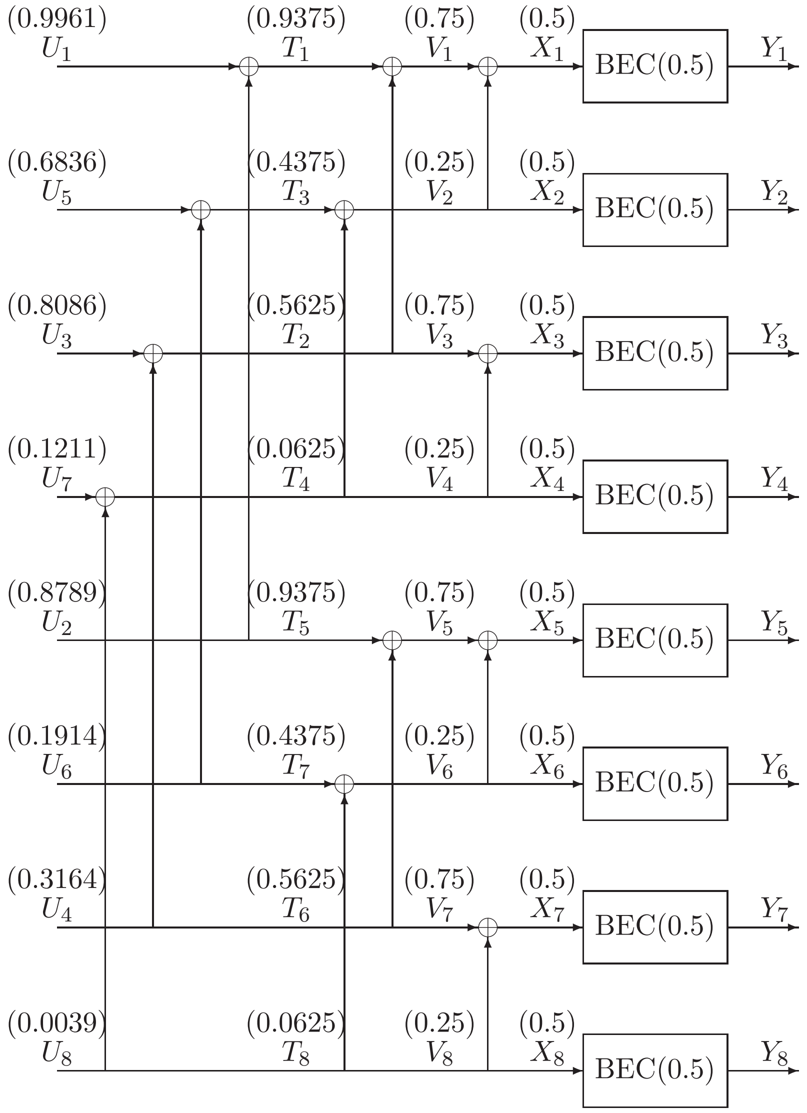
4.4 Example of Polar Codes for the BEC

In polar coding terminology,

- the process of using multiple basic transformations to get $X_1, \ldots, X_n$ from $U_1, \ldots, U_n$ (where the U_i's are i.i.d. uniform message random variables) is called channel "combining"
- and that of using $Y_1, \ldots, Y_n$ and $U_1, \ldots, U_{i-1}$ to obtain U_i for $i \in \{1, \ldots, n\}$ is called channel "splitting."
- Altogether, the phenomenon is called channel "polarization."

Example 4.14 Consider a BEC with erasure probability $\varepsilon = 0.5$ and let $n = 8$.

4.4 Example of Polar Codes for the BEC



4.4 Example of Polar Codes for the BEC

- A key reason for the prevalence of polar coding after its invention is that they form the first coding scheme that has an explicit low-complexity construction structure while being capable of achieving channel capacity as code length approaches infinity.

- More importantly, polar codes do not exhibit the error floor behavior, which Turbo and (to a lesser extent) LDPC codes are prone to.

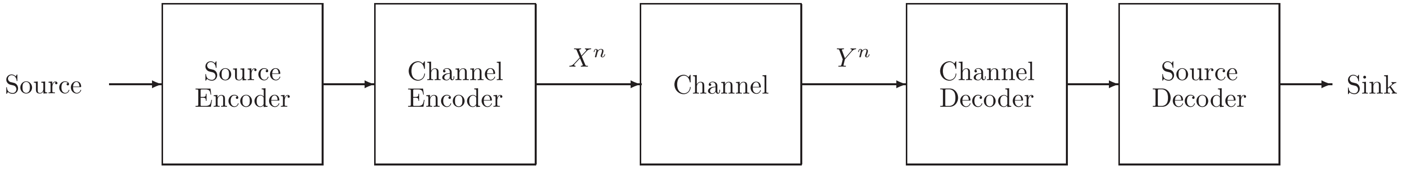
- Due to their attractive properties, polar codes were adopted in 2016 by the 3rd Generation Partnership Project (3GPP) as error correcting codes for the control channel of the 5th generation (5G) mobile communication standard.

4.6 Lossless joint source-channel coding and Shannon's separation principle

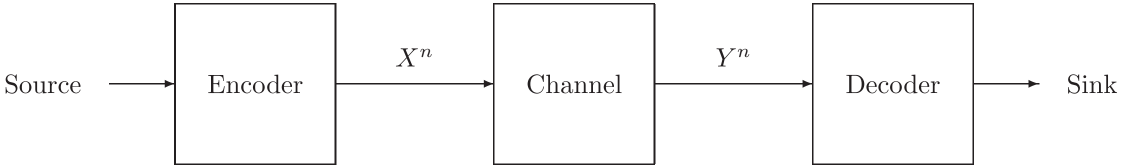
- We next establish *Shannon's lossless joint source-channel coding theorem* (or *lossless information transmission theorem*), which provides explicit (and directly verifiable) conditions for any communication system in terms of its source and channel information-theoretic quantities under which the source can be reliably transmitted (i.e., with asymptotically vanishing error probability).
- This key theorem is sometimes referred to as *Shannon's source-channel separation theorem or principle*.
  - Why it is named "separation principle"?
  - Answer: The theorem's necessary and sufficient conditions for reliable transmissibility are a function of entirely "separable" or "disentangled" information quantities, i.e., the source's minimal compression rate and the channel's capacity with no quantities that depends on both the source and the channel.

4.6 Lossless joint source-channel coding

- We will prove the theorem by assuming that the source is stationary ergodic in the forward part and just stationary in the converse part and that the channel is a DMC.
- Note that the theorem can be extended to more general sources and channels with memory (see Dobrushin 1963, Vembu & Verdu & Steinberg 1995, Chen & Alajaji 1999).



A separate (tandem) source-channel coding scheme.



A joint source-channel coding scheme.

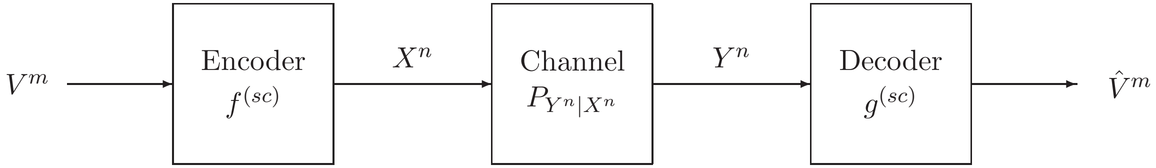
4.6 Lossless joint source-channel coding

Definition 4.29 (Source-channel block code) Given a discrete source $\{V_i\}_{i=1}^{\infty}$ with finite alphabet $\mathcal{V}$ and a discrete channel $\{P_{Y^n|X^n}\}_{n=1}^{\infty}$ with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, an m-to-n source-channel block code $\mathscr{C}_{m,n}$ with rate $\frac{m}{n}$ source symbol/channel symbol is a pair of mappings $(f^{(sc)}, g^{(sc)})$, where

$$f^{(sc)}\colon \mathcal{V}^m \to \mathcal{X}^n$$

and

$$g^{(sc)}\colon \mathcal{Y}^n \to \mathcal{V}^m.$$



An m-to-n block source-channel coding system.

The code's error probability is given by

$$P_e(\mathscr{C}_{m,n}) := \Pr[V^m \neq \hat{V}^m] = \sum_{v^m \in \mathcal{V}^m} \sum_{y^n \in \mathcal{Y}^n\colon\, g^{(sc)}(y^n) \neq v^m} P_{V^m}(v^m) P_{Y^n|X^n}(y^n|f^{(sc)}(v^m))$$

where P_{V^m} and $P_{Y^n|X^n}$ are the source and channel distributions, respectively.

4.6 Lossless joint source-channel coding

Theorem 4.30 (Lossless joint source-channel coding theorem for rate-one block codes) Consider a discrete source $\{V_i\}_{i=1}^{\infty}$ with finite alphabet $\mathcal{V}$ and entropy rate $H(\mathcal{V})$ and a DMC with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and capacity C, where both $H(\mathcal{V})$ and C are measured in the same units (i.e., they both use the same base of the logarithm). Then the following hold:

- *Forward part (achievability):* For any $0 < \epsilon < 1$ and given that the source is stationary ergodic, if

  $$H(\mathcal{V}) < C,$$

  then there exists a sequence of rate-one source-channel codes $\{\mathscr{C}_{m,m}\}_{m=1}^{\infty}$ such that

  $$P_e(\mathscr{C}_{m,m}) < \epsilon \qquad \text{for sufficiently large } m,$$

  where $P_e(\mathscr{C}_{m,m})$ is the error probability of the source-channel code $\mathscr{C}_{m,m}$.

4.6 Lossless joint source-channel coding

- *Converse part:* For any $0 < \epsilon < 1$ and given that the source is stationary, if

$$H(\mathcal{V}) > C,$$

then any sequence of rate-one source-channel codes $\{\mathscr{C}_{m,m}\}_{m=1}^{\infty}$ satisfies

$$P_e(\mathscr{C}_{m,m}) > (1-\epsilon)\mu \qquad \text{for sufficiently large } m, \tag{4.6.1}$$

where $\mu = H_D(\mathcal{V}) - C_D$ with $D = |\mathcal{V}|$, and $H_D(\mathcal{V})$ and C_D are entropy rate and channel capacity measured in D-ary digits, i.e., the codes' error probability is bounded away from zero and it is not possible to transmit the source over the channel via rate-one source-channel block codes with arbitrarily low error probability.

4.6 Lossless joint source-channel coding

Proof of the forward part:

- Without loss of generality, we assume throughout this proof that both the source entropy rate $H(\mathcal{V})$ and the channel capacity C are measured in nats (i.e., they are both expressed using the natural logarithm).

- **Key idea**: We will show the existence of the desired rate-one source-channel codes $\mathscr{C}_{m,m}$ via a separate (tandem or two-stage) source and channel coding scheme.

- Let $\gamma := C - H(\mathcal{V}) > 0$.

- Given any $0 < \epsilon < 1$, by the lossless source-coding theorem for stationary ergodic sources (Theorem 3.15), there exists a sequence of source codes of blocklength m and size M_m with

  $$\text{encoder } f_s\colon \mathcal{V}^m \to \{1, 2, \ldots, M_m\} \text{ and decoder } g_s\colon \{1, 2, \ldots, M_m\} \to \mathcal{V}^m$$

  such that

  $$\frac{1}{m} \log M_m < H(\mathcal{V}) + \gamma/2 \tag{4.6.2}$$

  and

  $$\Pr\left[g_s(f_s(V^m)) \neq V^m\right] < \epsilon/2$$

  for m sufficiently large.

4.6 Lossless joint source-channel coding

- Furthermore, by the channel coding theorem under the maximal probability of error criterion (see Observation 4.6 and Theorem 4.11), there exists a sequence of channel codes of blocklength m and size $\bar{M}_m$ with encoder

$$f_c \colon \{1, 2, \ldots, \bar{M}_m\} \to \mathcal{X}^m$$

and decoder

$$g_c \colon \mathcal{Y}^m \to \{1, 2, \ldots, \bar{M}_m\}$$

such that

$$\frac{1}{m} \log \bar{M}_m > C - \gamma/2 \ \big(= H(\mathcal{V}) + \gamma/2 > \frac{1}{m} \log M_m \big) \tag{4.6.5}$$

and

$$\lambda := \max_{w \in \{1, \ldots, \bar{M}_m\}} \Pr\left[g_c(Y^m) \neq w | X^m = f_c(w) \right] < \epsilon/2$$

for m sufficiently large.

4.6 Lossless joint source-channel coding

- Now we form our source-channel code by concatenating in tandem the above source and channel codes.
- Specifically, the m-to-m source-channel code $\mathscr{C}_{m,m}$ has the following encoder-decoder pair $(f^{(sc)}, g^{(sc)})$:

$$f^{(sc)}\colon \mathcal{V}^m \to \mathcal{X}^m \quad \text{with} \quad f^{(sc)}(v^m) = f_c(f_s(v^m)) \quad \forall v^m \in \mathcal{V}^m$$

and

$$g^{(sc)}\colon \mathcal{Y}^m \to \mathcal{V}^m$$

with

$$g^{(sc)}(y^m) = \begin{cases} g_s(g_c(y^m)), & \text{if } g_c(y^m) \in \{1, 2, \ldots, M_m\} \\ \text{arbitrary}, & \text{otherwise} \end{cases} \quad \forall y^m \in \mathcal{Y}^m.$$

- The above construction is possible since $\{1, 2, \ldots, M_m\}$ is a subset of $\{1,\ 2,\ \ldots,\ \bar{M}_m\}$.

4.6 Lossless joint source-channel coding

$$\begin{aligned}
P_e(\mathscr{C}_{m,m}) &= \Pr[g^{(sc)}(Y^m) \neq V^m] \\
&= \Pr[g^{(sc)}(Y^m) \neq V^m, g_c(Y^m) = f_s(V^m)] \\
&\qquad\qquad + \Pr[g^{(sc)}(Y^m) \neq V^m, g_c(Y^m) \neq f_s(V^m)] \\
&= \Pr[g_s(g_c(Y^m)) \neq V^m, g_c(Y^m) = f_s(V^m)] \\
&\qquad\qquad + \Pr[g^{(sc)}(Y^m) \neq V^m, g_c(Y^m) \neq f_s(V^m)] \\
&\leq \Pr[g_s(f_s(V^m)) \neq V^m] + \Pr[g_c(Y^m) \neq f_s(V^m)] \\
&= \Pr[g_s(f_s(V^m)) \neq V^m] \\
&\quad + \sum_{w \in \{1,2,\ldots,M_m\}} \Pr[f_s(V^m) = w] \Pr[g_c(Y^m) \neq w | f_s(V^m) = w] \\
&= \Pr[g_s(f_s(V^m)) \neq V^m] \\
&\quad + \sum_{w \in \{1,2,\ldots,M_m\}} \Pr[X^m = f_c(w)] \Pr[g_c(Y^m) \neq w | X^m = f_c(w)] \\
&\leq \Pr[g_s(f_s(V^m)) \neq V^m] + \lambda \\
&< \epsilon/2 + \epsilon/2 = \epsilon
\end{aligned}$$

for m sufficiently large. Thus the source can be reliably sent over the channel via rate-one block source-channel codes as long as $H(\mathcal{V}) < C$. □

4.6 Lossless joint source-channel coding

Proof of the converse part: For simplicity, we assume in this proof that $H(\mathcal{V})$ and C are measured in bits.

For any m-to-m source-channel code $\mathscr{C}_{m,m}$, we can write

$$\begin{aligned}
H(\mathcal{V}) &\le \frac{1}{m}H(V^m) && (4.6.6)\\
&= \frac{1}{m}H(V^m|\hat{V}^m) + \frac{1}{m}I(V^m;\hat{V}^m)\\
&\le \frac{1}{m}\left[P_e(\mathscr{C}_{m,m})\log_2(|\mathcal{V}|^m)+1\right] + \frac{1}{m}I(V^m;\hat{V}^m) && (4.6.7)\\
&\le P_e(\mathscr{C}_{m,m})\log_2|\mathcal{V}| + \frac{1}{m} + \frac{1}{m}I(X^m;Y^m) && (4.6.8)\\
&\le P_e(\mathscr{C}_{m,m})\log_2|\mathcal{V}| + \frac{1}{m} + C && (4.6.9)
\end{aligned}$$

where

- (4.6.6) is due to the fact that $(1/m)H(V^m)$ is non-increasing in m and converges to $H(\mathcal{V})$ as $m \to \infty$ since the source is stationary (see Observation 3.12),
- (4.6.7) follows from Fano's inequality,

  $H(V^m|\hat{V}^m) \le P_e(\mathscr{C}_{m,m})\log_2(|\mathcal{V}|^m)+h_b(P_e(\mathscr{C}_{m,m})) \le P_e(\mathscr{C}_{m,m})\log_2(|\mathcal{V}|^m)+1,$
- (4.6.8) is due to the data processing inequality since $V^m \to X^m \to Y^m \to \hat{V}^m$ form a Markov chain.

4.6 Lossless joint source-channel coding

Note that in the above derivation, the information measures are all measured in bits. This implies that for $m \geq \log_D(2)/(\varepsilon\mu)$,

$$P_e(\mathcal{C}_{m,m}) \geq \frac{H(\mathcal{V}) - C}{\log_2(|\mathcal{V}|)} - \frac{1}{m\log_2(|\mathcal{V}|)} = \underbrace{H_D(\mathcal{V}) - C_D}_{=\mu} - \underbrace{\frac{\log_D(2)}{m}}_{\leq \varepsilon\mu} \geq (1-\varepsilon)\mu.$$

□

4.6 Lossless joint source-channel coding

Theorem 4.32 (Lossless joint source-channel coding theorem for general rate block codes) Under the same notation as in Theorem 4.30, the following hold:

- *Forward part (achievability):* For any $0 < \epsilon < 1$ and given that the source is stationary ergodic, there exists a sequence of m-to-n_m source-channel codes $\{\mathscr{C}_{m,n_m}\}_{m=1}^{\infty}$ such that

  $$P_e(\mathscr{C}_{m,n_m}) < \epsilon \qquad \text{for sufficiently large } m$$

  if

  $$\limsup_{m\to\infty} \frac{m}{n_m} < \frac{C}{H(\mathcal{V})}.$$

- *Converse part:* For any $0 < \epsilon < 1$ and given that the source is stationary, any sequence of m-to-n_m source-channel codes $\{\mathscr{C}_{m,n_m}\}_{m=1}^{\infty}$ with

  $$\liminf_{m\to\infty} \frac{m}{n_m} > \frac{C}{H(\mathcal{V})},$$

  satisfies

  $$P_e(\mathscr{C}_{m,n_m}) > (1-\epsilon)\mu \qquad \text{for sufficiently large } m,$$

  for some positive constant μ that depends on $\liminf_{m\to\infty}(m/n_m)$, $H(\mathcal{V})$ and C.

4.6 Lossless joint source-channel coding

Discussion: separate vs joint source-channel coding

- Shannon's separation principle has provided the linchpin for most modern communication systems where source coding and channel coding schemes are separately constructed (with the source (resp., channel) code designed by only taking into account the source (resp., channel) characteristics) and applied in tandem without the risk of sacrificing optimality in terms of reliable transmissibility under unlimited coding delay and complexity.

- However, in practical implementations, there is a price to pay in delay and complexity for extremely long coding blocklengths (particularly when delay and complexity constraints are quite stringent such as in wireless communications systems).

- Under finite coding blocklengths and/or complexity, many studies have demonstrated that joint source-channel coding can provide better performance than separate coding.

4.6 Lossless joint source-channel coding

- Even in the infinite blocklength regime where separate coding is optimal in terms of reliable transmissibility, it can be shown that for a large class of systems, joint source-channel coding can achieve an *error exponent*that is as large as *double* the error exponent resulting from separate coding. This indicates that one can realize via joint source-channel coding the same performance as separate coding, while reducing the coding delay by *half* (this result translates into notable power savings of more than 2 dB when sending binary sources over channels with Gaussian noise, fading an output quantization).

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 the channel capacity
  - Symmetric, weakly symmetric, quasi-symmetric and T-symmetric channels
  - KKT condition
- Polar coding
- Joint source-channel coding theorem