Chapter 1

Winter School 2013: Basic Information Theory

Contents

1.1 Notation

- Upper case $X, Y, \ldots$ refer to random variables
- Script $\mathcal{X}, \mathcal{Y}, \ldots$ refer to discrete sets (alphabets)
- $|\mathcal{A}|$ is the cardinality of a discrete set $\mathcal{A}$
- $|A|$ is the determinant of the matrix A
- $X^n = (X_1, X_2, \ldots, X_n)$ is an n-sequence/vector of random variables
- $X_i^j = (X_i, X_{i+1}, \ldots, X_j), j \geq i$. By convention we take X_i^j to be the trivial random variable if $j < i$.
- $\mathrm{P}(A)$ denotes the probability of an event A

These notes are a modification of the lecture notes by Prof. Abbas El Gamal(Stanford) and Prof. Young-Han Kim(UCSD)

- $X^n \sim p(x^n)$: Probability mass function (pmf) of the random vector X^n is $p(x^n)$
  $p(x^n, y^n)$: Joint pmf of X^n and Y^n
  $p(y^n|x^n)$: Conditional pmf of Y^n given X^n
- Lower case $x, y, \ldots$ and $x^n, y^n, \ldots$ refer to scalars/vectors
- $\mathrm{E}_X(g(X))$, or $\mathrm{E}(g(X))$ in short, denotes the expected value of $g(X)$
- $X \to Y \to Z$ form a Markov chain if $p(x, y, z) = p(x)p(y|x)p(z|y)$
  $X_1 \to X_2 \to X_3 \to \cdots$ form a Markov chain if $p(x_i|x^{i-1}) = p(x_i|x_{i-1})$
- $X \sim \mathrm{Bern}(p)$ denotes that the binary random variable X is distributed according to the Bernoulli distribution with parameter p, i.e.,

$$X = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$

  $X^n \sim \mathrm{Bern}(p)$ denotes the binary random n-vector with X_i i.i.d. $\sim \mathrm{Bern}(p)$
- $[1:M]$ denotes the set $\{1, 2, \ldots, M\}$ for an integer M; more generally $[1:2^{nR}]$ denotes $\{1, 2, \ldots, \lfloor 2^{nR} \rfloor\}$ where $\lfloor 2^{nR} \rfloor$ denotes the integral part of the real number 2^{nR} (for channel coding problems, we use $\lceil \cdot \rceil$ instead of $\lfloor \cdot \rfloor$)
- $0 \cdot \log 0 = 0$ by convention
  (Recall: $\lim_{x \to 0} x \log x = 0$)

1.1.1 Convention of ϵ_n and $\delta(\epsilon)$

- We often use $\{\epsilon_n\}$ to denote a sequence of nonnegative numbers that approaches zero as $n \to \infty$
- When there are multiple sequences $\{\epsilon_{1n}\}, \{\epsilon_{2n}\}, \ldots, \{\epsilon_{kn}\} \to 0$, we denote them all by a generic $\{\epsilon_n\} \to 0$ with implicit understanding that $\epsilon_n = \max\{\epsilon_{1n}, \ldots, \epsilon_{kn}\}$
- Similarly, $\delta(\epsilon)$ denotes a generic function of ϵ such that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$
  (Example: $\delta(\epsilon) = \epsilon \log(\frac{1}{\epsilon})$)

1.2 Entropy and Mutual Information

1.2.1 Entropy

- Entropy of a discrete random variable $X \sim p(x)$:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\mathrm{E}_X(\log p(X))$$

  - $H(X)$ is nonnegative, continuous, and strictly concave function of $p(x)$
  - $H(X) \leq \log |\mathcal{X}|$
    This (as well as many other information theoretic inequalities) follows by Jensen's inequality:
    If g is a convex function, then

$$\mathrm{E}(g(X)) \geq g(\mathrm{E}(X))$$

  - Binary entropy function: For $0 \leq p \leq 1$

$$H(p) = -p \log p - (1-p) \log(1-p)$$
$$H(0) = H(1) = 0$$

- Conditional entropy: Let $(X, Y) \sim p(x, y)$

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = -\operatorname{E}_{X,Y}(\log p(Y|X))$$

  - $H(Y|X) \le H(Y)$, with equality iff X and Y are independent

- Joint entropy for random variables $(X, Y) \sim p(x, y)$:

$$\begin{aligned} H(X, Y) &= -\operatorname{E}(\log p(X, Y)) \\ &= -\operatorname{E}(\log p(X)) - \operatorname{E}(\log p(Y|X)) = H(X) + H(Y|X) \\ &= -\operatorname{E}(\log p(Y)) - \operatorname{E}(\log p(X|Y)) = H(Y) + H(X|Y) \end{aligned}$$

  - $H(X, Y) \le H(X) + H(Y)$, with equality iff X and Y are independent

- Let X be a discrete random variable and $g(X)$ be a function of X. Then

$$H(g(X)) \le H(X)$$

with equality iff g is one-to-one over the support of X, i.e., $\{x \in \mathcal{X} : p(x) > 0\}$
Proof:

$$\begin{aligned} H(X, g(X)) &= H(X) + H(g(X)|X) = H(X) + 0 = H(X) \\ H(X, g(X)) &= H(g(X)) + H(X|g(X)) \ge H(g(X)) \end{aligned}$$

with equality *iff* $H(X|g(X)) = 0$ or X can be determined from $g(X)$ (why?).

- Fano's inequality: If $(X, Y) \sim p(x, y)$ and $P_e = \operatorname{P}\{X \ne Y\}$, then

$$H(X|Y) \le H(P_e) + P_e \log(|\mathcal{X}| - 1) \le 1 + P_e \log(|\mathcal{X}| - 1)$$

Proof: Let the random variable E be defined as follows.

$$E = \begin{cases} 0 & X = Y \\ 1 & X \ne Y \end{cases}.$$

$$\begin{aligned} H(X|Y) &\le H(X, E|Y) = H(E|Y) + H(X|E, Y) \\ &\le H(E) + \operatorname{P}(E = 1) H(X|E = 1, Y) \text{ (why?)} \\ &\le 1 + P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

- Chain rule for entropies: Let X^n be a discrete random vector. Then

$$\begin{aligned} H(X^n) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \\ &= \sum_{i=1}^{n} H(X_i|X^{i-1}) \end{aligned}$$

1.2.2 Mutual Information

- For discrete random variables $(X, Y) \sim p(x, y)$:

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

A nonnegative function of $p(x, y)$, concave in $p(x)$ for fixed $p(y|x)$, and convex in $p(y|x)$ for fixed $p(x)$

- Conditional mutual information:

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

- Note that no general inequality relation exists between $I(X;Y|Z)$ and $I(X;Y)$

  Two important special cases:

  - If $Z \to X \to Y$ form a Markov chain, then $I(X;Y|Z) \leq I(X;Y)$
  - If $p(x,y,z) = p(z)p(x)p(y|x,z)$, then $I(X;Y|Z) \geq I(X;Y)$

- Chain rule:

$$I(X^n;Y) = \sum_{i=1}^{n} I(X_i;Y|X^{i-1})$$

- Data processing inequality: If $X \to Y \to Z$ form a Markov chain, then $I(X;Z) \leq I(Y;Z)$

  Proof: $I(X;Z) \leq I(X,Y;Z) = I(Y;Z)$.

1.3 Typical Sequences

- For a sequence $x^n \in \mathcal{X}^n$, we define its empirical distribution $\pi(\cdot|x^n)$ (often called its *type*) by

$$\pi(a|x^n) = \frac{|\{i : x_i = a\}|}{n} \qquad \text{for all } a \in \mathcal{X}$$

  $\mathbb{T}_n$ - number of *types* for x^n

  $\mathbb{T}_n \equiv$ number of ways you can have non-negative integers $a_1, ..., a_{|\mathcal{X}|}$ so that $\sum_i a_i = n$.

  Therefore $\mathbb{T}_n \leq (n+1)^{|\mathcal{X}|}$.

- Question: Suppose you have 2^{nR} sequences x^n, then prove that there is at least one type that has $2^{n(R-\epsilon)}$ of these sequences (for large n).?

  Solution: Let N be the maximum number of sequences of any one type. Then clearly,

$$N\mathbb{T}_n \geq 2^{nR} \Rightarrow N(n+1)^{|\mathcal{X}|} \geq 2^{nR}.$$

  Therefore $N \geq 2^{n(R - \frac{|\mathcal{X}| \log_2(n+1)}{n})} \geq 2^{n(R-\epsilon)}$ (for large n).

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim p_X(x)$. For each $a \in \mathcal{X}$ with $p_X(a) > 0$

$$\pi(a|X^n) \to p_X(a) \quad \text{in probability}$$

  This is a consequence of the (weak) law of large numbers (LLN)

  Thus most likely the random empirical distribution $\pi(\cdot|X^n)$ does not deviate much from the true distribution $p_X(\cdot)$

  Let $\{\epsilon_n\}$ be any sequence that satisfies: $\epsilon_n \to 0$, $\sqrt{n}\epsilon_n \to \infty$. (Example set $\epsilon_n = \frac{\log n}{\sqrt{n}}$.)

- A limit theorem (proof: follows from Chebyshev's ineq.)

  Let $X_1, X_2, \ldots$ be i.i.d. $\sim p_X(x)$. For each $a \in \mathcal{X}$ with $p_X(a) > 0$

$$\mathrm{P}\left(|\pi(a|X^n) - p_X(a)| > \epsilon_n p_X(a)\right) \to 0.$$

- The above theorem implies for any fixed $\epsilon > 0$ we have

$$\mathrm{P}\left(|\pi(a|X^n) - p_X(a)| > \epsilon p_X(a)\right) \to 0.$$

  Consider a sequence $\{\epsilon_n\}$ satisfying $\epsilon_n \to 0$ and $\sqrt{n}\epsilon_n \to \infty$.

- Typical set: For $X \sim p_X(x)$, define the set $T_\epsilon^{(n)}(X)$ of typical sequences x^n as

$$T_\epsilon^{(n)}(X) := \{x^n : |\pi(a|x^n) - p_X(a)| \le \epsilon_n \cdot p_X(a) \text{ for all } a \in \mathcal{X}\}$$

  When it is clear from the context, we will use $T_\epsilon^{(n)}$ instead of $T_\epsilon^{(n)}(X)$

- For each $x^n \in T_\epsilon^{(n)}$ (and n large enough)

$$2^{-n(1+\epsilon_n)H(X)} \le p(x^n) \le 2^{-n(1-\epsilon_n)H(X)}$$

  Notation: $p(x^n) \doteq 2^{-n(1\pm\epsilon_n)H(X)}$

  Proof: Note that $p(x^n) = \prod_a p_X(a)^{n\pi(a|x^n)}$.

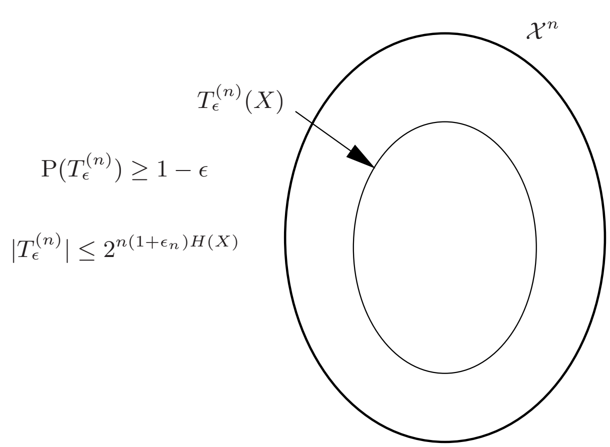
$$\begin{aligned} 2^{-n(1+\epsilon_n)H(X)} &= \prod_a p_X(a)^{np_X(a)(1+\epsilon_n)} \le \prod_a p_X(a)^{n\pi(a|x^n)} \\ &\le \prod_a p_X(a)^{np_X(a)(1-\epsilon_n)} = 2^{-n(1-\epsilon_n)H(X)}. \end{aligned}$$

- By summing the lower bound over the typical set, we have

$$\left|T_\epsilon^{(n)}\right| \le 2^{n(1+\epsilon_n)H(X)}$$

- If $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then by the LLN $\mathrm{P}\{X^n \in T_\epsilon^{(n)}\} \to 1$. Thus from the upper bound,

$$\left|T_\epsilon^{(n)}\right| \ge (1-\epsilon)2^{n(1-\epsilon_n)H(X)} \text{ for } n \text{ sufficiently large}$$



1.4 Jointly Typical Sequences

As before, consider a sequence $\{\epsilon_n\}$ such that $\epsilon_n \to 0$ and $\sqrt{n}\epsilon_n \to \infty$.

- Let $(X, Y) \sim p(x, y)$. The set $T_\epsilon^{(n)}(X, Y)$ (or $T_\epsilon^{(n)}$ in short) of *jointly typical* sequences (x^n, y^n) is defined as:

$$T_\epsilon^{(n)} := \{(x^n, y^n) : |\pi(a, b|x^n, y^n) - p(a, b)| \le \epsilon_n \cdot p(a, b) \text{ for all } a \in \mathcal{X}, b \in \mathcal{Y}\}$$

where

$$\pi(a, b|x^n, y^n) = \frac{|\{i : (x_i, y_i) = (a, b)\}|}{n}$$

is the empirical distribution of (x^n, y^n). In other words, $T_\epsilon^{(n)}(X, Y) = T_\epsilon^{(n)}((X, Y))$

- If $(x^n, y^n) \in T_\epsilon^{(n)}(X, Y)$, then

  1. $x^n \in T_\epsilon^{(n)}(X)$ and $y^n \in T_\epsilon^{(n)}(Y)$
  2. $p(x^n, y^n) \doteq 2^{-n(1\pm\epsilon_n)H(X,Y)}$
  3. $p(x^n) \doteq 2^{-n(1\pm\epsilon_n)H(X)}$ and $p(y^n) \doteq 2^{-n(1\pm\epsilon_n)H(Y)}$
  4. $p(x^n|y^n) \doteq 2^{-n(1\pm\epsilon)H(X|Y)}$ and $p(y^n|x^n) \doteq 2^{-n(1\pm\epsilon)H(Y|X)}$
     Proof:
     $p(x^n|y^n) = \frac{p(x^n,y^n)}{p(y^n)} = \frac{\prod_{(a,b)} p(a,b)^{n\pi(a,b|x^n,y^n)}}{\prod_{(b)} p(b)^{n(\sum_a \pi(a,b|x^n,y^n))}}$.
     Therefore
     $$\frac{2^{-n(1+\epsilon_n)H(X,Y)}}{2^{-n(1-\epsilon_n)H(Y)}} \le p(x^n|y^n) \le \frac{2^{-n(1-\epsilon_n)H(X,Y)}}{2^{-n(1+\epsilon_n)H(Y)}}.$$
     Thus, we obtain (for n large enough)
     $$2^{-n(1+\epsilon)H(X|Y)} \le p(x^n|y^n) \le 2^{-n(1-\epsilon)H(X|Y)}.$$
     (n should be large enough so that $\epsilon_n(H(X, Y) + H(Y)) < \epsilon H(X|Y)$ holds.)

- Remark: Check to see that everything is fine even when $H(X|Y) = 0$.

- As in the single random variable case,

  1. $|T_\epsilon^{(n)}(X, Y)| \le 2^{n(1+\epsilon_n)H(X,Y)}$
  2. $|T_\epsilon^{(n)}(X, Y)| \ge (1 - \epsilon)2^{n(1-\epsilon_n)H(X,Y)}$ for n sufficiently large

- Let $T_\epsilon^{(n)}(Y|x^n) := \{y^n : (x^n, y^n) \in T_\epsilon^{(n)}(X, Y)\}$. Then

$$|T_\epsilon^{(n)}(Y|x^n)| \le 2^{n(1+\epsilon)H(Y|X)} \qquad \text{for all } x^n \in T_\epsilon^{(n)}(X)$$

- Let $x^n \in T_\epsilon^{(n)}(X)$ and let Y^n be drawn according to $p(y^n|x^n) = \prod_{i=1}^n p(y_i|x_i)$. Then by the LLN

$$\mathrm{P}\{(x^n, Y^n) \in T_\epsilon^{(n)}(X, Y)\} \to 1 \quad \text{as } n \to \infty$$

This implies that

$$|T_\epsilon^{(n)}(Y|x^n)| \ge (1 - \epsilon)2^{n(1-\epsilon)H(Y|X)} \qquad \text{for all } x^n \in T_\epsilon^{(n)}(X)$$

- Observe that

$$(1 - \epsilon)2^{n(1-\epsilon)H(Y|X)} \le |T_\epsilon^{(n)}(Y|x^n)| \le 2^{n(1+\epsilon)H(Y|X)} \qquad \text{for all } x^n \in T_\epsilon^{(n)}(X)$$

- Given $(X,Y) \sim p(x,y)$, let $(\tilde{X}^n, \tilde{Y}^n)$ be drawn i.i.d. $\sim p(x)p(y)$; in other words, $\tilde{X}$ and $\tilde{Y}$ are from the product distribution with same marginals as X and Y respectively. Then, for n sufficiently large

  1. $\mathrm{P}\{(\tilde{X}^n, \tilde{Y}^n) \in T_\epsilon^{(n)}(X,Y)\} \leq \left(\frac{1}{1-\epsilon}\right) 2^{-n(I(X;Y)-\delta(\epsilon))}$
  2. $\mathrm{P}\{(\tilde{X}^n, \tilde{Y}^n) \in T_\epsilon^{(n)}(X,Y)\} \geq (1-\epsilon) 2^{-n(I(X;Y)+\delta(\epsilon))}$

  where $\delta(\epsilon) = \epsilon(H(X,Y) + H(X) + H(Y))$

- Intuition: We are determining the probability of picking one of $2^{nH(X,Y)}$ *jointly typical* pairs when we pick x^n uniformly from $2^{nH(X)}$ typical sequences and y^n independently from $2^{nH(Y)}$ typical sequences.

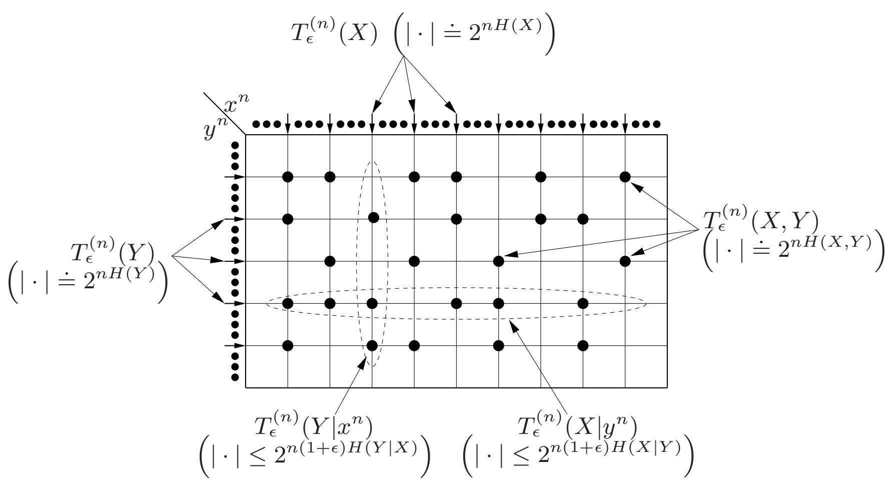
- For $\tilde{x}^n \in T_\epsilon^{(n)}(X)$ if $\tilde{Y}^n$ is drawn i.i.d. $p(y)$, then for n sufficiently large

  1. $\mathrm{P}\{(\tilde{x}^n, \tilde{Y}^n) \in T_\epsilon^{(n)}(X,Y)\} \leq \left(\frac{1}{1-\epsilon}\right) 2^{-n(I(X;Y)-\delta(\epsilon))}$
  2. $\mathrm{P}\{(\tilde{x}^n, \tilde{Y}^n) \in T_\epsilon^{(n)}(X,Y)\} \geq (1-\epsilon) 2^{-n(I(X;Y)+\delta(\epsilon))}$

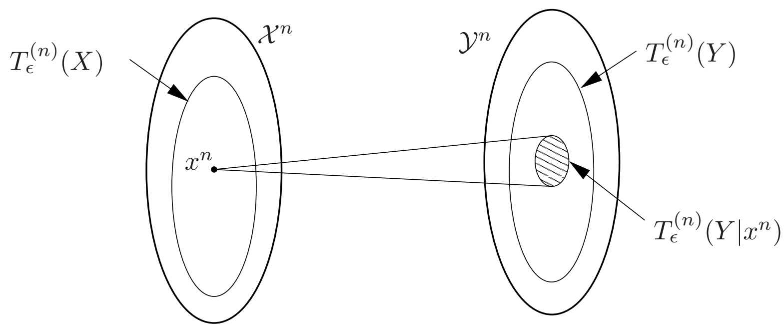
  where $\delta(\epsilon) = \epsilon(H(X,Y) + H(X) + H(Y))$

- Intuition: We are determining the probability of picking one of $2^{nH(Y|X)}$ sequences when we pick uniformly and randomly from $2^{nH(Y)}$ sequences.

1.4.1 Useful Picture



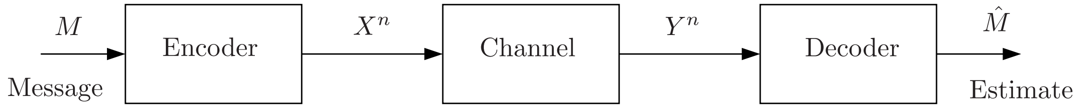
1.4.2 Another Useful Picture



1.5 Channel Coding Theorem

1.5.1 Channel Coding

- Point-to-point communication system model:



- We assume a *discrete memoryless channel* (DMC), denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$, consisting of two finite sets $\mathcal{X}$, $\mathcal{Y}$, and a collection of conditional pmfs $p(y|x)$

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

$$p(y_i|x^i, y^{i-1}) = p(y_i|x_i), \qquad i = 1, 2, \ldots, n$$

- For a channel with no feedback, i.e., $p(x_i|x^{i-1}, y^{i-1}) = p(x_i|x^{i-1})$, we have

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

  Proof:

$$\begin{aligned} p(x^n)p(y^n|x^n) = p(x^n, y^n) &= \prod_i p(x_i, y_i|x^{i-1}, y^{i-1}) \\ &= \prod_i p(x_i|x^{i-1}, y^{i-1}) p(y_i|x^i, y^{i-1}) = \prod_i p(x_i|x^{i-1}) p(y_i|x_i) \\ &= p(x^n) \prod_i p(y_i|x_i). \end{aligned}$$

- A $(2^{nR}, n)$ code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$, where R is the rate in bits/transmission, consists of the following:

  1. A message set $[2^{nR}] = \{1, 2, \ldots, \lceil 2^{nR} \rceil\}$
  2. An encoding function $x^n : [2^{nR}] \to \mathcal{X}^n$ that assigns a *codeword* $x^n(m)$ to each message $m \in [2^{nR}]$. The set $\{x^n(1), \ldots, x^n(2^{nR})\}$ is called the *codebook*
  3. A decoding function $\hat{m} : \mathcal{Y}^n \to [2^{nR}] \cup \{\mathrm{e}\}$ that assigns either an index $\hat{m} \in [2^{nR}]$ or an error index e to each received vector y^n

- Probability of error: Let $\lambda_m = \mathrm{P}\{\hat{M} \neq m | M = m\}$ be the conditional probability of error given that message m was sent

  The *average probability of error* $P_e^{(n)}$ for a $(2^{nR}, n)$ code is defined as

$$P_e^{(n)} = 2^{-nR} \sum_{m=1}^{2^{nR}} \lambda_m$$

  which corresponds to $\mathrm{P}\{\hat{M} \neq M\}$ when M is uniformly distributed over $[2^{nR}]$.

  Important: We assume throughout that the message M is a uniform random variable. (The assumption is quite general: If message is not uniform, then it does not have full entropy and we can compress the message sequence into another which is almost uniform.)

- A rate R is said to be *achievable* if there exists a sequence of $(2^{nR}, n)$ codes such that $P_e^{(n)} \to 0$ as $n \to \infty$

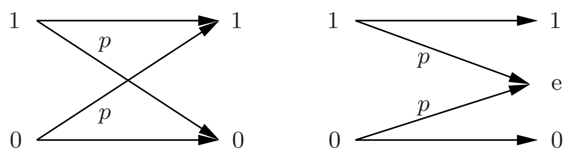
- The *capacity* C of a discrete memoryless channel is the supremum of all achievable rates

1.5.2 Channel Coding Theorem

- *Theorem* (Shannon [1]): The capacity of the DMC $(\mathcal{X}, p(y|x), \mathcal{Y})$ is given by

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

- Examples:

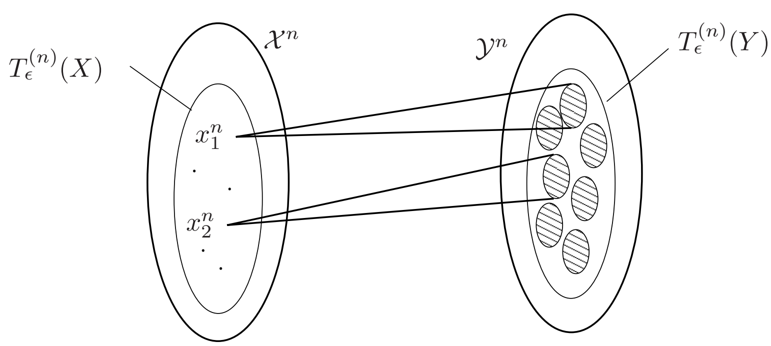


  - Binary symmetric channel (BSC) with crossover probability p: $C = 1 - H(p)$
  - Binary erasure channel (BEC) with erasure probability p: $C = 1 - p$

- To prove the theorem we need to prove:
  - Achievability: Any rate $R < C$ is achievable, i.e., there exists a sequence of $(2^{nR}, n)$ codes with average probability of error $P_e^{(n)} \to 0$
  - Weak converse: Given any sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$, $R \le C$

1.5.3 Sketch of Achievability Proof

- Let $p(x)$ be the optimal pmf. Consider a codebook of 2^{nR} randomly chosen ϵ-typical x^n codewords
- How many such codewords can be distiguished by the receiver?



  - There are $\approx 2^{nH(Y|X)}$ equally likely y^n sequences for each x^n sequence
  - The total number of likely y^n sequences is $\approx 2^{nH(Y)}$
  - Therefore, the maximum number of distinguishable x^n sequences is $\approx 2^{nH(Y)}/2^{nH(Y|X)} = 2^{nI(X,Y)} = 2^{nC}$

1.5.4 Proof of Achievability

- Random codebook generation (random coding): Fix $p(x)$. Generate a codebook $\mathcal{C}$ consisting of 2^{nR} i.i.d. x^n sequences according to $p(x^n) = \prod_{i=1}^n p(x_i)$. Label them $x^n(m)$, $m \in [1 : 2^{nR}]$. So

$$p(\mathcal{C}) = \prod_{m=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(m))$$

- The chosen codebook $\mathcal{C}$ is revealed to both sender and receiver before any transmission takes place
- Encoding: To send a message $m \in [2^{nR}]$, transmit $x^n(m)$
- Decoding: Let y^n be the received sequence

  The receiver declares that a message was sent if there exists one and only one index $\hat{m} \in [2^{nR}]$ such that $(x^n(\hat{m}), y^n) \in T_\epsilon^{(n)}$; otherwise an error is declared
- Probability of error: Assuming m is sent, there is a decoding error if $(x^n(m), y^n) \notin T_\epsilon^{(n)}$ or if there is an index $m' \neq m$ such that $(x^n(m'), y^n) \in T_\epsilon^{(n)}$
- Consider the probability of error averaged over M and over all codebooks

$$\begin{aligned}
\mathrm{P}(\mathcal{E}) &= \sum_{\mathcal{C}} p(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \\
&= \sum_{\mathcal{C}} p(\mathcal{C}) 2^{-nR} \sum_{m=1}^{2^{nR}} \lambda_m(\mathcal{C}) \\
&= 2^{-nR} \sum_{m=1}^{2^{nR}} \sum_{\mathcal{C}} p(\mathcal{C}) \lambda_m(\mathcal{C}) \\
&= \sum_{\mathcal{C}} p(\mathcal{C}) \lambda_1(\mathcal{C}) = \mathrm{P}(\mathcal{E}|M=1)
\end{aligned}$$

  Define the events

$$E_m = \{(X^n(m), Y^n) \in T_\epsilon^{(n)}\}, \quad m \in [2^{nR}]$$

  Hence

$$\begin{aligned}
\mathrm{P}(\mathcal{E}|M=1) &= \mathrm{P}\left(E_1^c \cup E_2 \cup E_3 \cup \ldots \cup E_{2^{nR}}\right) \\
&\leq \mathrm{P}(E_1^c) + \sum_{m=2}^{2^{nR}} \mathrm{P}(E_m)
\end{aligned}$$

  Since $(X^n(1), Y^n)$ is i.i.d. $\sim p(x,y)$, $\mathrm{P}(E_1^c) \leq \epsilon$, for n sufficiently large

  Since for $m \neq 1$ $X^n(m)$ is independent of $X^n(1)$, Y^n and $X^n(m)$ are independent

  Thus, the probability that $(X^n(m), Y^n)$ is jointly typical is $\leq 2^{-n(I(X;Y)-\delta(\epsilon))}$, where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, and

$$\begin{aligned}
\mathrm{P}(\mathcal{E}) &\leq \epsilon + \sum_{m=2}^{2^{nR}} 2^{-n(I(X;Y)-\delta(\epsilon))} \\
&= \epsilon + \left(2^{nR} - 1\right) 2^{-n(I(X;Y)-\delta(\epsilon))} \\
&\leq \epsilon + 2^{-n(I(X;Y)-R-\delta(\epsilon))} \\
&\leq 2\epsilon,
\end{aligned}$$

  provided that n is sufficiently large and $R < I(X;Y) - \delta(\epsilon)$
- To complete the proof, note that since the probability of error averaged over the codebooks $\mathrm{P}(\mathcal{E}) \leq 2\epsilon$, there must exist at least one codebook with $P_e^{(n)} \leq 2\epsilon$
- Probabilistic method. Simple and elegant
- Shannon's original arguments. Later made rigorous by Forney and Cover

- Alternative proofs
  - Feinstein's maximal coding theorem
  - Gallager's random coding exponent
- Remarks:
  - The capacity for the *maximal* probability of error $\lambda^* = \max_m \lambda_m$ is equal to that for the average probability of error $P_e^{(n)}$. This can be shown by throwing away the worst half of the codewords. In particular, the maximal probability of error for the remaining codewords should be $\leq 2P_e^{(n)}$. As we shall see, this is not always the case for multiple user channels
  - It can be shown (e.g., see [2]), that the probability of error decays exponentially in n. Close to tight bounds exist on the optimal error exponent (called the *reliability function*)

1.5.5 Proof of Weak Converse

- We need to show that for any sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$, $R \leq C$
- Each $(2^{nR}, n)$ code induces the joint pmf

$$(M, X^n, Y^n) \sim p(m, x^n, y^n) = 2^{-nR} p(x^n|m) \prod_{i=1}^{n} p(y_i|x_i)$$

- By Fano's inequality

$$H(M|\hat{M}) \leq 1 + P_e^{(n)} nR =: n\epsilon_n,$$

  where $\epsilon_n \to 0$ as $n \to \infty$ by the assumption that $P_e^{(n)} \to 0$
- From the data processing inequality,

$$H(M|Y^n) \leq H(M|\hat{M}) \leq n\epsilon_n$$

- Now consider

$$\begin{aligned}
nR &= H(M) \\
&= I(M; Y^n) + H(M|Y^n) \\
&\leq I(X^n; Y^n) + n\epsilon_n \\
&= H(Y^n) - H(Y^n|X^n) + n\epsilon_n \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\
&= \sum_{i=1}^{n} I(X_i; Y_i) + n\epsilon_n \\
&\leq nC + n\epsilon_n
\end{aligned}$$

  Dividing by n, we obtain $R \leq C + \epsilon_n$

  Now letting $n \to \infty$, we have $\epsilon_n \to 0$ and hence $R \leq C$

1.5.6 References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[2] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.