GI12/4C59: Information Theory

Lectures 13–15

Massimiliano Pontil

About these lectures

Theme of these lectures: We discuss the problem of data transmission through a noisy channel. We prove the key result of Information Theory which establishes that the fastest rate at which we can transmit a number of signals through the channel with arbitrarily small probability of error is bounded by the maximum of the mutual information of the channel.

Outline

1. Discrete channels

2. Typical sequences

3. Channel capacity

4. Channel coding theorem

5. Consequences of the theorem

Discrete channels

A channel is an input/output system where an input $x \in \mathcal{X}$ is transmitted and an output $y \in \mathcal{Y}$ is received with probability $p(y|x)$ (also called transition probability)

- x is called the sent symbol (or signal)

- y is called the received symbol (or signal)

If $\mathcal{X} = \mathcal{Y}$, the channel is said noiseless (or deterministic) if, for every $x \in \mathcal{X}$, $p(y|x) = 1$ for $y = x$ and zero otherwise. In this case it is always possible to infer the sent input from the received output.

Noisy channels

In practice the channel is noisy, that is, $p(y|x)$ is nonzero for more than one output.

Example 1 (Binary symmetric channel) $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, $p(1|1) = p(0, 0) = 1 - p$, $p(0|1) = p(1|0) = p$.

Example 2 (Binary erasure channel) $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, e, 1\}$, $p(0|0) = p(1, 1) = 1 - \alpha$, $p(e|1) = p(e|0) = \alpha$, $p(1|0) = p(0|1) = 0$.

Example 3 (Noisy typewriter) $\mathcal{X} = \mathcal{Y} = \{1, ..., 26\}$ (representing e.g. the 26 letter of the English alphabet), $p(y|x) = \frac{1}{2}$ if $y = x$ or $y = x + 1$ mod 26, and zero otherwise.

Can we still send messages through these channels with low probability of error? What does this mean?

Discrete memoryless channels

$$W \longrightarrow \boxed{\text{Encoder}} \rightarrow X^n \rightarrow \boxed{\textit{Channel}} \rightarrow Y^n \rightarrow \boxed{\text{Decoder}} \rightarrow \hat{W}$$

Suppose we have a set $\mathcal{W} = \{1, \ldots, M\}$ of M messages that we wish to send through a noisy channel. Each message w has a probability $p(w)$ of being selected for transmission.

Encoder: we code each message by a sequence of symbols from $\mathcal{X}$ of length n, that is

$$x^n(w), \quad w = 1, \ldots, M \qquad \text{called the codewords}$$

Discrete memoryless channels (cont.)

$$W \longrightarrow \boxed{\text{Encoder}} \rightarrow X^n \rightarrow \boxed{\textit{Channel}} \rightarrow Y^n \rightarrow \boxed{\text{Decoder}} \rightarrow \widehat{W}$$

Memoryless assumption: the received signal y^n has probability distribution

$$p(y^n|x^n) = p(y_1|x_1)p(y_2|x_2)\cdots p(y_n|x_n)$$

That is, the element y_i of the output sequence is only determined by the corresponding element x_i of x^n.

Discrete memoryless channels (cont.)

Decoder: based on y^n we produce a decoding rule $g : \mathcal{Y}^n \rightarrow \{1, \ldots, M\}$. $\widehat{w} = g(y^n)$ is our guess for the sent message w. An error occurs if $\widehat{w} \neq w$. In particular

$$\lambda_w(n) := P(\{g(Y^n) \neq w\}|\{X^n = x^n(w)\})$$

The map $x^n(w)$, $w \in \{1, \ldots, M\}$ coupled with a decoding function g is called an (M, n) *code* and we also denoted it by $\mathcal{C}^{(n)}$.

The probability of error of the code is defined by

$$\lambda(E|\mathcal{C}^{(n)}) := \max\{\lambda_w(n) : w \in \mathcal{W}\}$$

Channel capacity

Given an (M, n) code, the quantity $R = \frac{\log M}{n}$ is called the *transmission rate* of the code (log is the logarithm in base 2).

A rate R is said to be *achievable* if there exists a sequence of $\mathcal{C}^{(n)} = (\lceil 2^{nR} \rceil, n), n \in \mathbb{N}$ codes such that,

$$\lim_{n\to\infty} \lambda(E|\mathcal{C}^{(n)}) = 0$$

The **capacity** C of the channel is the supremum of all achievable rates.

Informally, $\mathcal{C}^{(n)}$ is a *"good code"* if it has small probability of error and its rate is close to C.

Note: the capacity does not depend on $p(x)$ but only on $p(y|x)$.

The channel coding theorem

Remember that the mutual information of a pair of the r.v X and Y is defined by

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Theorem (Shannon) The channel capacity is given by

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

that is, for every rate $R < C$, there exits a sequence $\mathcal{C}^{(n)} = (\lceil 2^{nR} \rceil, n), n \in \mathbb{N}$ of codes such that $\lambda(E|\mathcal{C}^{(n)}) \to 0$ as $n \to \infty$. Conversely, any sequence of codes for which $\lambda(E|\mathcal{C}^{(n)}) \to 0$ must have a rate $R \leq C$.

Application of the theorem

Before proving the theorem we use it to compute the capacity of the above channels.

Recall that the mutual information is a nonnegative concave function of $p(x)$ for fixed $p(y|x)$ (so the above maximization problem is well defined) and can be written as

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x \in \mathcal{X}} p(x) H(Y|X = x).$$

Using the properties of the entropy, we conclude that

- $0 \leq C \leq \min(\log |\mathcal{X}|, \log |\mathcal{Y}|)$.

Noiseless channel

In this case $C = 1$, because $H(Y|X = x) = 0$ and, so, $I(X;Y) = H(Y)$ which achieves its maximum when $p(x)$ (and $p(y)$) is uniform.

Any nonsingular code has zero probability of error and the identity code achieves capacity.

- In general, the computation of the capacity by the formula, $C = \max_{p(x)} I(X;Y)$, is not constructive, that is, this computation does not provide us with a sequence of codes whose rate is arbitrarily close to C.

Binary symmetric channel

We have $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, $p(1|1) = p(0|0) = 1 - p$, $p(0|1) = p(1|0) = p$.

In this case $C = 1 - H(p)$. In fact

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

Thus,

$$C = \max_{p(x)} \{I(X;Y)\} = 1 - H(p)$$

achieved for $p(x)$ uniform (in which case also $p(y)$ is uniform).

Noisy typewriter channel

Remember that $\mathcal{X} = \mathcal{Y} = \{1, ..., 26\}$ and $p(y|x) = \frac{1}{2}$ if $y = x$ or $y = x + 1 \bmod 26$, and zero otherwise.

We have $C = \log 13$. In fact

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = \sum_{x \in \mathcal{X}} p(x) 1 = 1$$

and, thus,

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \{H(Y) - 1\} = \log 26 - 1 = \log 13$$

again achieved when $p(y)$ (and, so, $p(x)$) is the uniform distribution.

Typewriter channel

For this channel it is also easy to choose a good code.

Simply take $n = 1$, $M = 13$ and $x(1) = a$, $x(2) = c$, $x(3) = e$, etc. This code has zero probability of error because each codeword is either transmitted at such or as the next symbol in $\mathcal{X}$.

This code also achieves capacity since its transmission rate is

$$R = \frac{\log M}{n} = \log 13$$

Binary erasure channel

Remember that $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, e, 1\}$, $p(0|0) = p(1, 1) = 1 - \alpha$, $p(e|1) = p(e|0) = \alpha$, $p(1|0) = p(0|1) = 0$.

Here $C = 1 - \alpha$. In fact

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = \sum_{x \in \mathcal{X}} p(x) H(\alpha) = H(\alpha).$$

If we let $p = p(X = 0)$ we have $p(Y = 0) = p(1 - \alpha)$, $p(Y = 1) = (1 - p)(1 - \alpha)$, $p(Y = e) = \alpha$. We then have

Binary erasure channel (cont.)

$$\begin{aligned} H(Y) &= -p(1-\alpha)\log p(1-\alpha) - (1-p)(1-\alpha)\log((1-p)(1-\alpha)) - \alpha\log\alpha \\ &= -(1-\alpha)\log(1-\alpha) - (1-\alpha)(p\log p + (1-p)\log(1-p)) - \alpha\log\alpha \\ &= (1-\alpha)H(p) + H(\alpha). \end{aligned}$$

We conclude that

$$C = \max_{p(x)} I(X;Y) = \max_{p}\{(1-\alpha)H(p)\} = 1-\alpha$$

achieved when $p(x)$ is the uniform distribution.

Symmetric channels

The above example of binary symmetric channel can be generalized as following.

We take $m = |\mathcal{X}|$, $\ell = |\mathcal{Y}|$ and let P be a $m \times \ell$ matrix whose rows are the numbers $p(y|x)$ for fixed x and columns are the numbers $p(y|x)$ for fixed y.

A channel is said weakly symmetric if the rows of the matrix P are permutations of each other and the columns all have the same sum.

Symmetric channels (cont.)

Since the rows are permutations of each other, we have $H(Y|X = x) = H(r)$ for every $x \in \mathcal{X}$, where r is, say, the first row of the transition matrix. Thus,

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = H(r)$$

and we conclude that

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \{H(Y) - H(r)\} = \log \ell - H(r)$$

achieved when $p(x)$ is the uniform distribution.

Note: for the binary symmetric channel we rediscover that $C = \log 2 - H(r) = 1 - H(p)$.

Jointly typical sequences

The proof of the theorem uses a simple decoding rule which is based on the idea of jointly typical sequences.

A sequence x^n is a called ϵ−typical if

$$\left| -\frac{\log p(x^n)}{n} - H(X) \right| < \epsilon \qquad \text{(likewise for } y^n\text{)}$$

and a pair of sequences (x^n, y^n) is said jointly ϵ−typical if x^n and y^n are ϵ−typical and

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

The set of ϵ−jointly typical sequences is denoted by $\mathcal{A}_\epsilon^{(n)}$.

Properties of jointly typical sequences

For every $\epsilon > 0$, we have that

1. $P\left(\left\{(X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\}\right) \to 1$ when $n \to \infty$.

2. $|\mathcal{A}_\epsilon^{(n)}| \in \left[(1-\epsilon)2^{n(H(X,Y)-\epsilon)}, 2^{n(H(X,Y)+\epsilon)}\right]$.

3. If S^n and T^n are independent with the same marginal distributions as X^n and Y^n respectively, then
$$P\left(\left\{(S^n, T^n) \in \mathcal{A}_\epsilon^{(n)}\right\}\right) \in \left[(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}, 2^{-n(I(X;Y)-3\epsilon)}\right]$$

Note: Remember that here X^n and (X^n, Y^n) are i.i.d..

Proof of property 1

The above properties follow by the *weak law of large numbers*, which says that if $X_i = X, i \in \mathbb{N}$ is a sequence of i.i.d. r.v., then
$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X] \qquad \text{in probability.}$$
that is, for every $\epsilon > 0$, $P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| > \epsilon\right) \to 0$ as $n \to \infty$.

To prove 1, note that
$$\frac{1}{n}\sum_{i=1}^{n} -\log p(X_i) \to -E[\log p(X)] = H(X)$$
and
$$-\frac{\log p(X^n, Y^n)}{n} = \frac{1}{n}\sum_{i=1}^{n} -\log p(X_i, Y_i) \to -E[\log p(X,Y)] = H(X,Y)$$

Proof of property 1 (cont.)

Then, consider the events

$$E_{1,\epsilon}^{(n)} = \left\{ \left| -\frac{\log p(X^n)}{n} - H(X) \right| > \epsilon \right\}, \quad E_{2,\epsilon}^{(n)} = \left\{ \left| -\frac{\log p(Y^n)}{n} - H(Y) \right| > \epsilon \right\}$$

and

$$E_{3,\epsilon}^{(n)} = \left\{ \left| -\frac{\log p(X^n, Y^n)}{n} - H(X,Y) \right| > \epsilon \right\}.$$

By the week law of large numbers there exist n_1, n_2, n_3 such that

$$P(E_{1,\epsilon}^{(n)}) < \frac{\epsilon}{3} \text{ if } n > n_1, \quad P(E_{2,\epsilon}(n)) < \frac{\epsilon}{3} \text{ if } n > n_2$$

and

$$P(E_{3,\epsilon}(n)) < \frac{\epsilon}{3} \text{ if } n > n_3.$$

Proof of property 1 (cont.)

Now set $E_{\epsilon}^{(n)} = E_{1,\epsilon}^{(n)} \cup E_{2,\epsilon}^{(n)} \cup E_{3,\epsilon}^{(n)}$ and note that $\mathcal{A}_{\epsilon}^{(n)} = \overline{E}_{\epsilon}^{(n)}$.

Using the union bound,

$$P(E_{\epsilon}^{n}) \leq \sum_{i=1}^{3} P(E_{i,\epsilon}^{(n)})$$

it follows that for $n > \max(n_1, n_2, n_3)$

$$P(\mathcal{A}_{\epsilon}^{(n)}) = 1 - P(\bar{\mathcal{A}}_{\epsilon}^{n}) \geq 1 - \sum_{i=1}^{3} P(E_{i,\epsilon}^{(n)}) > 1 - \epsilon.$$

- Properties 2 and 3 are proved similarly (see page 196-7 of Cover and Thomas).

Idea in proving the channel coding theorem

We focus on part 1 of the theorem: for every rate $R < C$, there exits a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the probability of error $\lambda(E|\mathcal{C}^{(n)}) \to 0$ for $n \to \infty$. The main steps are:

- Generate a code $\mathcal{C}$ at random (to simplify notation we drop the subscript (n) in $\mathcal{C}^{(n)}$).
- Use joint typical sequences to define a decoding rule.
- Compute the *average* probability of error w.r.t. a random choice of the sent codeword w *and* the generated code $\mathcal{C}$.
- Show that the above calculation guarantees that a good code exists.

Part 1: proof of $R < C$

We generate a $(\lceil 2^{nR} \rceil, n)$ code $\mathcal{C}$ at random according to $p(x)$. Each codeword $x^n(w), w = 1, \ldots, M := \lceil 2^{nR} \rceil$ is generated with probability

$$p(x^n(w)) = \prod_{i=1}^{n} p(x_i(w)) \quad \Rightarrow \quad P(C) = \prod_{w=1}^{M} \prod_{i=1}^{n} p(x_i(w))$$

Decoding function g: if there is only one $\widehat{w}$ such that $(x^n(\widehat{w}), y^n)$ is jointly $\epsilon-$typical, we set $g(y^n) = \widehat{w}$, otherwise we set $g(y^n) = 0$. Let $E = \{g(y^n) \neq w\}$ (we always commit an error in the second case).

Part 1 (cont.)

We compute the average probability of error $P(E)$ (with respect to the generated code $\mathcal{C}$ and uniformly sample codewords)

$$P(E) = \sum_{\mathcal{C}} P(E|\mathcal{C})P(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{M} \sum_{w=1}^{M} \lambda_w(\mathcal{C}) = \frac{1}{M} \sum_{w=1}^{M} \sum_{\mathcal{C}} P(\mathcal{C})\lambda_w(\mathcal{C})$$

Key observation: the inner sum does not depend on w because of the symmetric generation process of the code. Thus,

$$P(E) = \sum_{\mathcal{C}} P(\mathcal{C})\lambda_1(\mathcal{C}) = P(E|W=1)$$

Let $E_i = \{(x^n(i), y^n) : (x^n(i), y^n) \in \mathcal{A}_\epsilon^{(n)}\}$. Then

$$P(E|W=1) = P(\bar{E}_1 \cup E_2 \cup E_3 \cup \ldots \cup E_M) \leq P(\bar{E}_1) + \sum_{i=2}^{M} P(E_i)$$

Part 1 (cont.)

$$P(E|W=1) \leq P(\bar{E}_1) + \sum_{i=2}^{M} P(E_i)$$

now remember the properties of typical sequences (page 20).

- Property 1 $\Rightarrow P(\bar{E}_1) = 1 - P(E_1) \leq \epsilon$

$X^n(1)$ and $X^n(w)$ are independent if $w > 1$. This implies that Y^n is also independent of $X^n(w)$. Thus

- Property 3 $\Rightarrow P(E_i) \leq 2^{-n(I(X;Y)-3\epsilon)}$

Remember that $M = \lceil 2^{nR} \rceil$. If we chose $R \leq I(X;Y) - 3\epsilon$, we conclude that

$$P(E|W=1) \leq \epsilon + (\lceil 2^{nR} \rceil - 1)2^{-n(I(X;Y)-3\epsilon)} \leq \epsilon + 2^{nR}2^{-n(I(X;Y)-3\epsilon)} \leq 2\epsilon$$

where the last inequality holds provided that n is large enough.

Part 1 (cont.)

The above calculation show that if $R < I(X;Y)$, the average (w.r.t. $\mathcal{C}$ and W) probability of error goes to zero as n goes to infinity. To conclude the proof we observe that

- If we set $p(x)$ to be the probability which maximizes $I(X;Y)$, the above condition $R \leq I(X;Y)$ becomes $R < C$.
- There must exist at least one code $\mathcal{C}^*$ for which the average probability of error w.r.t. the codewords goes to zero as n goes to infinity.
- Since, above,

$$P(E|\mathcal{C}^*) = \frac{1}{2^{nR}} \sum_w \lambda_w(\mathcal{C}^*) \leq 2\epsilon$$

  at least half of the codewords of $\mathcal{C}^*$ must have probability of error less than 4ϵ. We keep such codewords to form a code which has 2^{nR-1} codewords. This code has maximal probability of error less than 4ϵ and a rate $R + \frac{1}{n}$. Thus, when $n \to \infty$ it achieves the rate R.

Part 1: some observation

We make some observations about the above proof technique.

- The symmetry of the above generation process greatly simplifies the calculation.
- The decoding rule also simplifies the calculation. We will see below that other decoding rules are possible.
- However, the proof technique is not constructive: it shows that a good code exists but it does not provide a procedure to find such a code.

Zero-error codes

Before proving the second part of the theorem, we analyze the case that our codes have zero probability of error for every n. In this case the output Y^n always determines the sent input index W and, so,

$$H(W|Y^n) = 0$$

Thus, assuming W has uniform distribution we have

$$nR = H(W) = H(W|Y^n) + I(W;Y^n) = I(W;Y^n)$$

Note: We have used the property $I(X_1;X_2) = H(X_1) - H(X_1|X_2)$

Zero-error codes (cont.)

Now recall the data processing inequality which says that, if $X \to Y \to Z$ forms a Markov chain (that is, $p(x,y,z) = p(x)p(y|x)p(z|y)$) then $I(X;Y) \geq I(X;Z)$.

Since $W \to X^n(W) \to Y^n$ forms a Markov chain, we have

$$I(W;Y^n) \leq I(X^n;Y^n)$$

Thus, so far we have

$$nR = I(W;Y^n) \leq I(X^n;Y^n)$$

Zero-error codes (cont.)

Now we observe that

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{i-1},\ldots,Y_1,X^n) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i;Y_i) \le nC \end{aligned}$$

where we used the property $H(Y^n) \le \sum_{i=1}^{n} H(Y_i)$ and the definition of capacity.

We conclude that if the a code $\mathcal{C}^{(n)}$ has zero probability of error then $R \le C$.

- Note that the inequality $I(X^n;Y^n) \le nC$ means that the capacity per transmission rate does not increas if we use the channel many times.

Fano inequality

Lemma: If $P^{(n)}$ is the average probability of error of a code $\mathcal{C}^{(n)}$ when $p(w)$ is uniform, then

$$H(X^n(W)|Y^n) \le 1 + nRP^{(n)} \qquad \text{(Fano inequality)}$$

Proof: By definition $P^{(n)} = P(g(Y^n) \ne W)$. If E is the binary r.v. defined by

$$E = \begin{cases} 0 & \text{if } g(Y^n) = W \\ 1 & \text{if } g(Y^n) \ne W \end{cases}$$

We have $P^{(n)} = P(E = 1)$ and, using the chain rule for entropy, we obtain

$$H(E,W|Y^n) = H(W|Y^n) + H(E|W,Y^n) = H(E|Y^n) + H(W|E,Y^n).$$

Fano inequality (cont.)

Since E is a function of W and $g(Y^n)$, we have $H(E|W,Y^n) = 0$ and, since E is binary $H(E|Y^n) \leq 1$. It follows that

$$H(W|Y^n) \leq 1 + H(W|E, Y^n).$$

We have

$$\begin{aligned} H(W|E,Y^n) &= P(E=0)H(W|Y^n, E=0) + P(E=1)H(W|Y^n, E=1) \\ &\leq (1-P^{(n)})0 + P^{(n)} \log(|\mathcal{W}|-1)) \leq P^{(n)}nR \end{aligned}$$

and, so,

$$H(W|Y^n) \leq 1 + H(W|E,Y^n) \leq 1 + P^{(n)}nR$$

Finally, note that, since X^n is a function of W, $H(X^n(W)|Y^n) \leq H(W|Y^n)$ and we conclude that

$$H(X^n|Y^n) \leq 1 + P^{(n)}nR$$

Note: this proof also tells us that $H(W|Y^n) \leq 1 + P^{(n)}nR$ (we will use this for channels with feedback next week)

Proof of part 2

We are now ready to prove part 2 of the theorem: any sequence of $(\lceil 2^{nR} \rceil, n)$ codes whose probability of error goes to zero as n goes to infinity has a rate $R \leq C$.

Since by hypothesis the maximal probability of the code $\mathcal{C}^{(n)}$ goes to zero as n grows, we also have that the average probability of error of that code goes to zero.

Again, we assume that W is drawn with the uniform distribution over $\mathcal{W} = \{1, \ldots, nR\}$ so that $P(g(Y^n) \neq W) = P^{(n)}$.

Proof of part 2 (cont.)

Using the previous results we have that

$$\begin{aligned} nR &= H(W) = H(W|Y^n) + I(W;Y^n) \\ &\leq H(W|Y^n) + I(X^n(W),Y^n) \quad \text{(Data processing ineq.)} \\ &\leq 1 + P^{(n)}nR + I(X^n(W),Y^n) \quad \text{(Fano inequality)} \\ &\leq 1 + P^{(n)}nR + nC \end{aligned}$$

which implies that

$$R \leq \left(C + \frac{1}{n}\right)\left(1 - P^{(n)}\right)^{-1} \to C \quad \text{for } n \to \infty$$

which proves the result.

An important remark

The above formula can be rewritten as

$$P^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}.$$

This shows than if $R > C$ and n is large enough, the average probability of error is bounded away from zero.

Indeed, this is also true for all n because if $P^{(n)} = 0$ for some $n = \bar{n}$, we could simply concatenate such code to have a code with large n and $P^{(n)} = 0$.

These observations confirm that we cannot achieve an arbitrarily low probability of error if $R > C$.

Bibliography

This lectures are based on Chapter 8 of Cover and Thomas's book.