Fundamentals of Mathematical Informatics

The Channel Capacity

Francesco Buscemi

Lecture Five

The information channel capacity: definition

- Consider a DMC $\mathcal{N}$ with input alphabet $\mathcal{X} = \{x_1, \cdots, x_m\}$, output alphabet $\mathcal{Y} = \{y_1, \cdots, y_n\}$, and channel matrix $[\![p_{ij}]\!]$ $(1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n)$.
- Let X be an input RV, with range equal to $\mathcal{X}$ and probability distribution π_i.
- Feeding X through the channel $\mathcal{N}$, we obtain a pair of dependent RVs (X, Y), with range $\mathcal{X} \times \mathcal{Y}$ and joint probability distribution $\Pr\{X = x_i, Y = y_j\} = \pi_i p_{ij}$.
- From $\pi_i p_{ij}$, we then compute the mutual information

$$I(X;Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} \Pr\{X = x_i, Y = y_j\} \log_2 \frac{\Pr\{X = x_i, Y = y_j\}}{\Pr\{X = x_i\} \Pr\{Y = y_j\}}.$$

  If the channel $\mathcal{N}$ is fixed, $[\![p_{ij}]\!]$ is fixed too, and $I(X;Y)$ is a function of the probability distribution π_i of X only.

The information channel capacity

The information capacity of the channel $\mathcal{N}$ is defined as

$$C(\mathcal{N}) \stackrel{\text{def}}{=} \max_{\{\pi_i\}} I(X;Y).$$

Example: the information capacity of the BSC

Consider a binary symmetric channel (BSC) with error probability γ. Then:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x=0,1} p(x) H(Y|X=x) \\ &= H(Y) - \sum_{x=0,1} p(x) \{ \underbrace{-\gamma \log_2 \gamma - (1-\gamma)\log_2(1-\gamma)}_{\stackrel{\text{def}}{=} H(\gamma)} \} \\ &= H(Y) - H(\gamma) \\ &\leqslant 1 - H(\gamma). \end{aligned}$$

On the other hand, choosing $p(0) = p(1) = 1/2$, we obtain $\Pr\{Y=0\} = \Pr\{Y=1\}$, i.e., $H(Y) = 1$.

Theorem: the capacity of the binary symmetric channel with error probability γ is equal to $C(\gamma) = 1 - H(\gamma)$.

Example: the information capacity of the BEC

Consider a binary erasure channel (BEC) with erasure probability γ. As for the binary symmetric channel, $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\gamma)$.
In order to compute $H(Y)$, we introduce the RV E, function of Y, defined as

$$E = \begin{cases} 0, & \text{if } Y \neq ☹, \\ 1, & \text{if } Y = ☹. \end{cases}$$

Since E is function of Y, $H(E|Y) = 0$. This implies that:

$$\begin{aligned} H(Y) &= H(Y,E) - H(E|Y) \\ &= H(Y,E) \\ &= H(E) + H(Y|E) \\ &= H(E) + \Pr\{E=0\} H(Y|E=0) + \Pr\{E=1\} H(Y|E=1) \\ &= H(\gamma) + (1-\gamma) H(X) + \gamma \cdot 0 \\ &= H(\gamma) + (1-\gamma) H(X). \end{aligned}$$

But then, $I(X;Y) = H(Y) - H(\gamma) = H(\gamma) + (1-\gamma)H(X) - H(\gamma) = (1-\gamma)H(X)$.
The maximum is achieved when $H(X) = 1$.

Theorem: the capacity of the binary erasure channel with erasure probability γ is equal to $C(\gamma) = 1 - \gamma$.

The operational channel capacity: definitions

Consider a DMC $\mathcal{N}$ with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$.

- an (M, n)-code $\mathscr{C}$ is given by an encoding $\boldsymbol{c} : \{1, 2, \cdots, M\} \to \mathcal{X}^{(n)}$ and a decoding $g : \mathcal{Y}^{(n)} \to \{1, 2, \cdots, M\}$.
- the rate of an (M, n)-code is $R \stackrel{\text{def}}{=} \frac{\log_2 M}{n}$, and is measured in 'bits per transmission.'
- (average) error probability: $\mathrm{e}(\mathscr{C}) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{i=1}^{M} \Pr\{g(Y^n) \neq i | X^n = \boldsymbol{c}_i\}$.
- maximum error probability: $\hat{\mathrm{e}}(\mathscr{C}) \stackrel{\text{def}}{=} \max_i \Pr\{g(Y^n) \neq i | X^n = \boldsymbol{c}_i\}$.
- a rate R is (asymptotically) achievable, if, for any $\epsilon > 0$, there exists a sequence of $(\lfloor 2^{nR} \rfloor, n)$-codes $\mathscr{C}_n$ and an integer $n_0(\epsilon)$ such that, for any $n \geqslant n_0(\epsilon)$, $\hat{\mathrm{e}}(\mathscr{C}_n) \leqslant \epsilon$. (That is, $\lim_{n\to\infty} \hat{\mathrm{e}}(\mathscr{C}_n) = 0$.)

The (asymptotic) operational channel capacity

The operational capacity of the channel $\mathcal{N}$ is defined as

$$C'(\mathcal{N}) \stackrel{\text{def}}{=} \sup_R \{R \text{ achievable rate}\}.$$

The noisy coding theorem for general DMCs

Information capacity $\equiv$ (asymptotic) operational capacity

For any DMC $\mathcal{N}$, any rate $R < C$ is asymptotically achievable, i.e.,

$$C(\mathcal{N}) = C'(\mathcal{N}).$$

- C.E. Shannon, *A mathematical theory of communication*. Bell Syst. Tech. J., **27**:379-423,623-656 (1948).
- A. Feinstein, *A new basic theorem of information theory*. IER Trans. Inf. Theory, **IT-4**:2-22 (1954).
- R.G. Gallager, *A simple derivation of the coding theorem and some applications*. IEEE Trans. Inf. Theory, **IT-11**:3-18 (1965).

Coding theorem for the BSC: direct part

We will only prove this particular statement:

Coding theorem: achievability (direct part)

Given a binary symmetric channel with bit-flip probability $0 \leqslant \gamma < \frac{1}{2}$, for any choice of parameters $0 < \delta \leqslant \frac{1}{2} - \gamma$ and $\eta > 0$, there exists a sequence of (M_n, n)-codes $\mathscr{C}_n$ such that

$$\lim_{n\to\infty} \hat{\mathrm{e}}(\mathscr{C}_n) = 0,$$

and

$$M_n = \left\lfloor 2^{n[C(\gamma+\delta)-\eta]} \right\rfloor,$$

i.e., any rate $R < C(\gamma)$ is asymptotically achievable.

Remark. The statement is restricted to the case $\gamma < 1/2$: the case $\gamma > 1/2$ is obtained by flipping all the bits received, while the case $\gamma = 1/2$ is obtained by continuity.

Useful facts required for the proof

Chebyshev's inequality (for coin tosses)

Consider a coin with $\Pr\{\text{head}\} = 1 - \Pr\{\text{tail}\} = \gamma$. The probability that, in a sequence of n tosses, the number of heads H is strictly greater than $n\gamma$ is bounded as

$$\Pr\{H \geqslant n\gamma + \Delta\} \leqslant \frac{n\gamma(1-\gamma)}{\Delta^2},$$

for any $\Delta > 0$.

Example: tossing 100 times a fair coin ($\gamma = 1/2$), the probability of obtaining 60 or more heads is at most 25%. For 70 heads, ⩽11%. For 90 heads, ⩽2%.

The tail inequality

For any $0 \leqslant \xi \leqslant 1/2$,

$$\sum_{k=0}^{\lfloor \xi n \rfloor} \binom{n}{k} \leqslant 2^{nH(\xi)}.$$

Reminder: the symbol $\binom{n}{k}$ denotes the Newton binomial coefficient $\frac{n!}{k!(n-k)!}$ (note that $0! \stackrel{\text{def}}{=} 1$): it gives the number of k-element subsets of an n-element set.

Proof: (random) construction of the code

- **Encoding**:
  1. Fix integers M (the size of the code) and n (the length of the code): the codebook is an M-element subset of V_n (the set of all 2^n binary strings of length n).
  2. All codewords $\boldsymbol{c}_i$ are drawn at random from V_n: $\Pr\{\boldsymbol{c}_i = \boldsymbol{x}\} = 2^{-n}$ for all $1 \leqslant i \leqslant M$ and for all $\boldsymbol{x} \in \mathsf{V}_n$. (For example, it could be $\boldsymbol{c}_i = \boldsymbol{c}_j$ for $i \neq j$; we do not care.)
- **Decoding**:
  1. Fix integer $r \geqslant 1$ and construct the sphere of Hamming radius r around each element $\boldsymbol{y} \in \mathsf{V}_n$: $S_r(\boldsymbol{y}) \overset{\text{def}}{=} \{\boldsymbol{z} : d(\boldsymbol{z}, \boldsymbol{y}) \leqslant r\}$.
  2. Upon receiving $\boldsymbol{y}$, if inside $S_r(\boldsymbol{y})$ is contained one and only one codeword $\boldsymbol{c}_j$, we decode $\boldsymbol{y}$ with j. Otherwise an error is declared.

Proof: error probability analysis (part 1 of 3)

Remember: $\gamma < 1/2$.

- Imagine that $\boldsymbol{Y}$ is received: a decoding error happens if more than r bit-flip errors occurred (event A) or if there are two (or more) codewords in $S_r(\boldsymbol{Y})$ (event B).
- Since $\Pr\{A \text{ or } B\} \leqslant \Pr\{A\} + \Pr\{B\}$, we independently consider events A and B.
- Let us begin with $\Pr\{A\} = \Pr\{\text{more than } r \text{ bit-flip errors}\}$.
- $\Pr\{A\}$ is equal to the probability of obtaining more than r 'heads' with n tosses of a coin with $\Pr\{\text{head}\} = \gamma$.
- Fix $\delta > 0$ such that $\gamma + \delta \leqslant 1/2$ and take $r = \lfloor n\gamma + n\delta \rfloor$.
- By Chebyshev's inequality, $\Pr\{A\} \leqslant \frac{\gamma(1-\gamma)}{n\delta^2}$.
- Let us move onto $\Pr\{B\}$.

Proof: error probability analysis (part 2 of 3)

Remember: $\gamma < 1/2$, $0 < \delta \leqslant 1/2 - \gamma$, and $r = \lfloor n\gamma + n\delta \rfloor$.

- How to evaluate $\Pr\{B\} = \Pr\{\text{two or more codewords in } S_r(\boldsymbol{Y})\}$?
- How many distinct elements are in $S_r(\boldsymbol{Y})$? There is $\boldsymbol{Y}$ itself... There are n distinct elements that differ from $\boldsymbol{Y}$ in one place... There are the $\frac{n(n-1)}{2}$ distinct elements that differ from $\boldsymbol{Y}$ in two places... In general, there are the $\binom{n}{k}$ distinct elements that differ from $\boldsymbol{Y}$ in k places. Therefore, for any $\boldsymbol{Y} \in \mathsf{V}_n$, $S_r(\boldsymbol{Y})$ contains exactly $\sum_{k=0}^{r}\binom{n}{k}$ distinct elements.
- Therefore, for each $\boldsymbol{Y} \in \mathsf{V}_n$, the probability that a codeword belongs to $S_r(\boldsymbol{Y})$ can be exactly computed as $2^{-n}\sum_{k=0}^{r}\binom{n}{k}$.
- Given that one codeword, say $\boldsymbol{c}_j$, is in $S_r(\boldsymbol{Y})$, then

$$\Pr\{\boldsymbol{c}_1 \in S_r(\boldsymbol{Y}) \text{ or } \cdots \text{ or } \boldsymbol{c}_{j-1} \in S_r(\boldsymbol{Y}) \text{ or } \boldsymbol{c}_{j+1} \in S_r(\boldsymbol{Y}) \text{ or } \cdots \text{ or } \boldsymbol{c}_M \in S_r(\boldsymbol{Y})\}$$
$$\leqslant \sum_{i \neq j} \Pr\{\boldsymbol{c}_i \in S_r(\boldsymbol{Y})\}$$
$$= (M-1)2^{-n}\sum_{k=0}^{r}\binom{n}{k} < M2^{-n}\sum_{k=0}^{r}\binom{n}{k} \leqslant M2^{-n}2^{nH(\gamma+\delta)} = M2^{-n(1-H(\gamma+\delta))}$$
$$= M2^{-nC(\gamma+\delta)}.$$

Proof: error probability analysis (part 3 of 3)

- Until now, we have evaluated the (average) error probability of a randomly constructed (M,n)-code $\mathscr{C}$ as follows:

$$\mathrm{e}(\mathscr{C}) \leqslant \frac{\gamma(1-\gamma)}{n\delta^2} + M2^{-nC(\gamma+\delta)},$$

  where n, M, and $0 < \delta \leqslant \frac{1}{2} - \gamma$ are free parameters.
- This means that, for any $0 < \delta \leqslant \frac{1}{2} - \gamma$, there always exists a sequence of random (M_n, n)-codes $\mathscr{C}_n$ such that $\mathrm{e}(\mathscr{C}_n) \to 0$, but... provided that $M_n 2^{-nC(\gamma+\delta)} \to 0$.
- For example, for any arbitrarily small $\eta > 0$, take $M_n = \lfloor 2^{n[C(\gamma+\delta)-\eta]} \rfloor$, so that $M_n 2^{-nC(\gamma+\delta)} = 2^{-n\eta} \to 0$.
- Then, for any $\delta > 0$, there exists a large enough n that achieves the rate $R_n = C(\gamma+\delta) - \eta$, for any arbitrarily small $\eta > 0$.
- We still need to evaluate the maximum error probability!

Proof: from average error probability to maximum error probability

- Assume that $\mathrm{e}(\mathscr{C}) = \frac{1}{M}\sum_{i=1}^{M} \Pr\{g(Y^n) \neq i | X^n = \boldsymbol{c}_i\} \leqslant \epsilon$.
- We can conclude that no more than $M/2$ codewords in $\mathscr{C}$ can be such that $\Pr\{g(Y^n) \neq i | X^n = \boldsymbol{c}\} > 2\epsilon$.
- This implies that there exist at least $M/2$ codewords in $\mathscr{C}$ such that $\Pr\{g(Y^n) \neq i | X^n = \boldsymbol{c}\} \leqslant 2\epsilon$.
- So, if we know that there exists a sequence of (M_n, n)-codes $\mathscr{C}_n$ with $\mathrm{e}(\mathscr{C}_n) \to 0$, we know that there exists a sequence of $(\frac{M_n}{2}, n)$-codes $\mathscr{C}'_n$ with $\hat{\mathrm{e}}(\mathscr{C}'_n) \to 0$.
- Computing the rate of $\mathscr{C}'_n$: $\frac{1}{n}\log_2(\frac{M_n}{2}) = \frac{1}{n}(\log_2 M_n - 1) \to \frac{1}{n}\log_2 M_n$.
- This implies that, without decreasing the asymptotic rate, we can make the maximum error probability go to zero.
- In other words, for any $\delta, \eta > 0$, the rate $R_n = C(\gamma + \delta) - \eta$ is asymptotically achievable.
- By taking the limits $\delta \to 0$ and $\eta \to 0$, any rate $R < C(\gamma)$ is asymptotically achievable.

Some remarks

- The proof shows that, for length n large enough, a good code can be constructed very easily, just by choosing the codewords at random.
- We pay this at the decoding stage: the receiver needs to use a table lookup scheme, i.e., a 'big book' where it's written what to do for each received $\boldsymbol{y}$, but the size of this book grows *exponentially* in n.
- Coding theory aims at constructing coding techniques that strike a good tradeoff between capacity and decoding efficiency.
- What happens if we try to transmit data at a rate $R > C$? **Weak converse**: the error probability cannot go to zero, i.e., for any sequence of (M_n, n)-codes with $\lim_n \frac{1}{n}\log_2 M_n > C$, there exists $\epsilon_0 > 0$ such that $\mathrm{e}(\mathscr{C}_n) > \epsilon_0$, for all n. **Strong converse**: for any sequence of (M_n, n)-codes with $\lim_n \frac{1}{n}\log_2 M_n > C$, $\mathrm{e}(\mathscr{C}_n) \to 1$.
- **Remark**: the theorem (and its converse) does not address the case $R = C$.

Summary of lecture five

- For any DMC channel, its information capacity is asymptotically achievable.
- The construction in the achievability proof involves a random coding argument.
- With random coding, coding is easy, decoding is hard.
- Actual codes try to balance rate and decoding efficiency.
- The capacity is a sharp transition point: error goes to zero for $R < C$, while it goes to one for $R > C$.

Keywords for lecture five

information channel capacity, operational channel capacity, the noisy coding theorem for DMCs, random coding argument