Lecture: Channel coding theorem

Chandra Nair

1 The communication problem

Alice wishes to communicate over a noisy channel to Bob. In many engineering problems like wireless networks, the errors are introduced in nature rather than an adversary. The other model where channel is an adversary is also very important, esp. in cryptography.

Instead of sending raw message bits, Alice sends a sequence with error correction capabilities so that Bob can counter the noisy behavior of the channel to recover the intended message with high probability. Thus the communication model we are studying is represented in Figure 1. Alice, the sender, first encodes the raw message m of nR bits, i.e. $m \in \{1, ..., 2^{nR}\}$, bits into a sequence of symbols $(x_1(m), \ldots, x_n(m))$. Here the channel is used n times. The noisy channel corrupts this sequence into another sequence y^n which is received by Bob. Bob then tries to estimate the message m. A rate R is said to be *achievable* if there are an encoding strategies and a decoding strategies (for each n) so that $\mathrm{P}(\hat{M} \neq M) \to 0$ as $n \to \infty$. The capacity of the channel is the supremum of all the achievable rates.

1.1 Channel model

The channel model that is assumed in Shannon's work is called the discrete memoryless model. Let the input alphabet be $\mathcal{X}$ and the output alphabet be $\mathcal{Y}$. In a discrete memoryless channel $|\mathcal{X}|, |\mathcal{Y}| < \infty$. Let x^i denote $\{x_1, \ldots, x_i\}$. Then for a memoryless channel $p(y_i|x^i, y^{i-1}) = p(y_i|x_i), \forall i$ i.e. the output at time i only depends on the input at time i and not on the behavior of the channel in the past time slots.

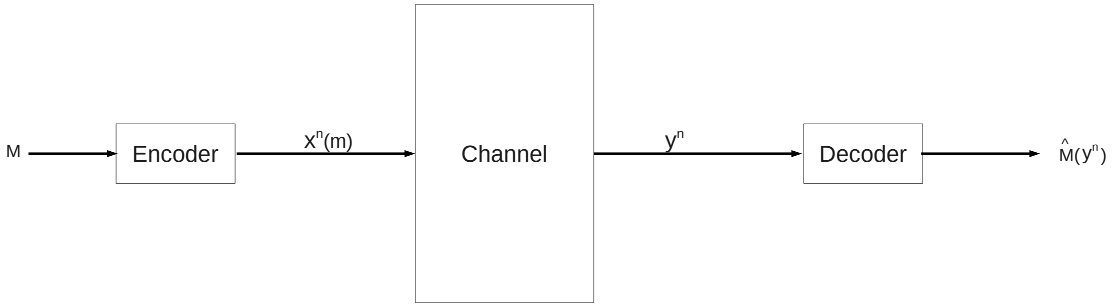

Figure 1: Communication problem

Common examples are:

- Binary symmetric channel $BSC(p)$: Here $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and
$$P(Y = y|X = x) = \begin{cases} p & y \neq x \\ 1 - p & y = x \end{cases} .$$

- Binary erasure channel $BEC(e)$: Here $\mathcal{X} = \{0, 1\}, \mathcal{Y} = \{0, E, 1\}$ and
$$P(Y = y|X = x) = \begin{cases} e & y = E \\ 1 - e & y = x \end{cases} .$$

1.2 Channel coding theorem

For an arbitrary discrete-memoryless channel Shannon [1948] proved this remarkable theorem.

Theorem 1. *The capacity, C, of a dicrete memoryless channel, characterized by $p(y|x)$, is given by*
$$C = \max_{p_X(x)} I(X; Y)$$

Remarks: To prove the channel coding theorem it suffices to assume that the messages are uniformly distributed in the set $\{1, .., 2^{nR}\}$. In particular, this is equivalent to saying that w.l.o.g. assume that the raw message bits are independent and uniformly distributed. (The justification for the assumption is that if this does not hold, then one can further compress the raw message bits into a smaller length sequence of independent and uniformly distributed bits, subject of source coding theorem, while involving only a negligible error in reconstruction.)

The proof of this theorem will consist of two arguments. The first part shows that for any $\epsilon > 0$, there exists a sequence of codebooks with rate $R = I(X; Y) - \epsilon$, so that the probability of decoding error goes to zero as $n \to \infty$. The second part, usually called the converse, shows that if there is a sequence of codebooks with rate R such that P(error) $\to 0$, then it must be that $R \leq \max_{p(x)} I(X; Y)$.

2 Mathematical Preliminaries

In this section we define some mathematical preliminaries that you encountered in the morning lecture:

- Entropy: The entropy of a random variable $X \sim p(x)$ is defined as
$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x).$$

  This is a measure of the uncertainty in X, or in other words, the number of uniformly random bits needed on average to describe X.

  Exercise: Use concavity of log (i.e. $\log(\sum_i p_i x_i) \geq \sum_i p_i \log(x_i)$ where p_i is a probability vector and $\{x_i\}$ is any collections of real numbers) to show that $0 \leq H(X) \leq \log |\mathcal{X}|$.

- Conditional entropy: The conditional of a random variable X given another random variable Y,m where $(X,Y) \sim p(x,y)$ is defined as

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

  This is a measure of the uncertainty in X conditioned on knowing Y.

- Mutual information: The mutual information between two random variables $(X,Y) \sim p(x,y)$ is defined as

  $$I(X;Y) = H(X) - H(X|Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}.$$

  This is a measure of the amount of information about X revealed by the knowledge of Y.

Exercises: Show the following:

1. $H(X|Y) \leq H(X)$(i.e. $I(X;Y) \geq 0$). (Hint: use concavity of the log function).
2. $H(X) - H(X|Y) = H(Y) - H(Y|X)$.

- ϵ-Typical sequences: Let x^n be distributed according to $\prod_{i=1}^n p_X(x_i)$, i.e. the components $x_1, .., x_n$ are independent and indentically distributed. Then

  $$\mathcal{T}_\epsilon^{(n)}(X) = \{x^n : |\{i : x(i) = a\}| \in [n(1-\epsilon)p_X(a), n(1+\epsilon)p_X(a)] \ \forall a \in \mathcal{X}\}.$$

  These are the sequences that have the correct proportion of each alphabet (up to an ϵ tolerance).

Exercises: Show the following:

1. $2^{-n(1+\epsilon)H(X)} \leq p(x^n) \leq 2^{-n(1-\epsilon)H(X)}, \forall x^n \in \mathcal{T}_\epsilon^{(n)}(X)$.
2. $|\mathcal{T}_\epsilon^{(n)}(X)| \leq 2^{n(1+\epsilon)H(X)}$.
3. Use Law of Large Numbers*(read the statement in Wiki) to conclude that for n large enough, we have $|\mathcal{T}_\epsilon^{(n)}(X)| \geq (1-\epsilon)2^{n(1-\epsilon)H(X)}$.

- ϵ-jointly typical sequences: Let (x^n, y^n) be distributed according to $\prod_{i=1}^n p_{X,Y}(x_i, y_i)$, i.e. the components $(x_1, y_1), .., (x_n, y_n)$ are independent and identically distributed. Then

  $$\mathcal{T}_\epsilon^{(n)}(X,Y) = \{(x^n, y^n) : |\{i : (x(i), y(i)) = (a,b)\}| \in [n(1-\epsilon)p_{X,Y}(a,b), n(1+\epsilon)p_{X,Y}(a,b)] \ \forall a, b \in \mathcal{X} \times \mathcal{Y}\}.$$

  These are the sequences that have the correct proportion of each alphabet pair (up to an ϵ tolerance).

- conditionally typical sequences: For every $y^n \in \mathcal{T}_\epsilon^{(n)}(Y)$ define

  $$\mathcal{T}_\epsilon^{(n)}(X|y^n) = \{(x^n, y^n) : |\{i : (x(i), y(i)) = (a,b)\}| \in [n(1-\epsilon)(a,b), n(1+\epsilon)p_{X,Y}(a,b)] \ \forall a, b \in \mathcal{X} \times \mathcal{Y}\}.$$

Exercises: Show the following for any ϵ:

1. $\mathcal{T}_\epsilon^{(n)}(X|y^n) \subseteq \mathcal{T}_\epsilon^{(n)}(X)$.
2. $\frac{2^{-n(1+\epsilon)H(X,Y)}}{2^{-n(1-\epsilon)H(Y)}} \leq p(x^n|y^n) \leq \frac{2^{-n(1-\epsilon)H(X,Y)}}{2^{-n(1+\epsilon)H(Y)}}, \forall x^n \in \mathcal{T}_\epsilon^{(n)}(X|y^n)$.
3. $|\mathcal{T}_\epsilon^{(n)}(X|y^n)| \leq \frac{2^{-n(1-\epsilon)H(Y)}}{2^{-n(1+\epsilon)H(X,Y)}}$
4. If $\tilde{x}^n$ is chosen according to $\prod_i p_X(\tilde{x}_i)$ independent of y^n , then

$$\mathrm{P}(\tilde{x}^n \in \mathcal{T}_\epsilon^{(n)}(X|y^n)) \leq 2^{-n(1-\epsilon)H(X)} \frac{2^{-n(1-\epsilon)H(Y)}}{2^{-n(1+\epsilon)H(X,Y)}} = 2^{-n(1-\epsilon)I(X;Y)} \times 2^{n2\epsilon H(X,Y)}$$

The result of this last exercise is crucial in the proof of the achievability. We have shown that if $\tilde{x}^n$ is chosen according to $\prod_i p_X(\tilde{x}_i)$ independent of y^n , then

$$\mathrm{P}(\tilde{x}^n \in \mathcal{T}_\epsilon^{(n)}(X|y^n)) \leq 2^{-n(1-\epsilon-\epsilon')I(X;Y)}$$

where $\epsilon' = 2\epsilon \frac{\log_2 |\mathcal{X}||\mathcal{Y}|}{I(X;Y)}$.

- Chain Rule: $H(X^n) = \sum_i H(X_i|X^{i-1}), \quad I(X;Y^n) = \sum_{i=1}^n I(X;Y_i|Y^{i-1})$.
- Data Processing inequality: If $X \to Y \to Z$ is a Markov chain (i.e. $p(z|y,x) = p(z|y)$) then $I(X;Y) \geq I(X;Z)$.
- Fano's inequality: If $\mathrm{P}(\hat{M} \neq M) \leq \epsilon$ then $H(M|\hat{M}) \leq 1 + \epsilon \log |M|$.

3 Proof of achievability

In this section we will show that any rate less than capacity is achievable. Let us fix a $p_X(x)$ and let $\delta > 0$ be arbitrary. Let $R = I(X;Y) - \delta$. We will use *random coding* and *jointly-typical* decoding to show the *existence* of a codebook with a small average probability of error.

Remark: How does one show the existence of a good codebook without actually explicitly demonstrating it? This is the beauty of the probabilistic method. We will construct a collection of codebooks, and average over this collection we will show that the probability of error is small. Therefore, there must exist a codebook with small probability of error.

Codebook generation

The ensemble(collection) of codebooks are generated as follows. For each $m \in \{1,..,2^{nR}\}$ we generate a sequence $x^n(m) \sim \prod_{i=1}^n p_X(x_i)$ independent of every other codeword. To transmit message m, the sender chooses the sequence $x^n(m)$.

Remark: Note that this will yield a collection of codebooks, each generated with certain probabilities. For instance, if $R = 0.5, n = 2, \mathcal{X} = \{0,1\}$ and $p_X(x=0) = p_X(x=1) = \frac{1}{2}$, then with probability $\frac{1}{16}$ we will have generated one of the 16 codebooks: $\{00,00\}, \{00,01\}, \{00,10\}, \{00,11\}$, $\{01,00\}, \{01,01\}, \{01,10\}, \{01,11\}, \{10,00\}, \{10,01\}, \{10,10\}, \{10,11\}, \{11,00\}, \{11,01\}, \{11,10\}, \{11,11\}$. Here there are 2 messages, and the first element of the set is the transmitted sequence corresponding to the first message, and the second element of the set is the transmitted sequence corresponding to the second message. If you are wondering, it is true that the first, sixth, eleventh, and sixteenth codebooks are really bad because we transmit the same sequence for the different messages. However for large n it will turn out (much deeper than what we require) that almost all of the codebooks will be good. We only require that at least one of the codebooks is good.

Decoding

The codebook is conveyed to the receiver. Note that this is a one-time event and if we use the same codebook for repeated communication, this can be done at a negligible cost.

Pick an ϵ, ϵ' small enough such that $\epsilon + \epsilon' < \delta$. (Note that this is always possible. We can assume $I(X;Y) > 0$, as otherwise we have nothing to prove. Why?)

The receiver, upon receiving y^n, declares that $\hat{m} = m$ if $x^n(m)$ is the only codeword such that $(x^n(m), y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)$. It declares that $\hat{m} = 1$ otherwise.

Probability of error analysis

Clearly, there is a decoding error if the transmitted message is not in $\mathcal{T}_\epsilon^{(n)}(X,Y)$ or there is another $m_1 \neq m$ such that $(x^n(m_1), y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)$. By the law of large numbers and the channel model, we have $\mathrm{P}\left((x^n(m), y^n) \notin \mathcal{T}_\epsilon^{(n)}(X,Y)\right) \to 0$ as $n \to \infty$.

Consider the event that $\mathrm{P}((x^n(m_1), y^n) \notin \mathcal{T}_\epsilon^{(n)}(X,Y)\big)$. When $m_1 \neq m$, averaged over the choice of codebooks, observe that $x^n(m_1)$ is chosen according to $\prod_i p_X(\tilde{x}_i)$ independent of y^n (this is not true if we fix a codebook). Hence for the given y^n we have that

$$\mathrm{P}((x^n(m_1), y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)) \leq 2^{-n(1-\epsilon-\epsilon')I(X;Y)}.$$

Hence the probability that there is at least one $m_1 \neq m$ such that $(x^n(m_1), y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)$ is bounded by (union bound)

$$\sum_{m_1 \neq m} \mathrm{P}((x^n(m_1), y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)) \leq 2^{nR} 2^{-n(1-\epsilon-\epsilon')I(X;Y)} = 2^{-n(\delta-\epsilon-\epsilon')} \overset{n\to\infty}{\longrightarrow} 0.$$

Thus the probability of decoding error (average over the choice of codebooks and over the messages) goes to zero. Hence there must be at least one codebook such that the probability of error averaged over the messages goes to zero. Thus we have shown that any $R < I(X;Y)$ is achievable for any $p(x)$.

This shows that $C \geq \max_{p(x)} I(X;Y)$.

4 Converse

Suppose we have a sequence of codebooks such that the average probability of error (here the average is over the choice of messages) goes to zero as $n \to \infty$. We assume that the messages are uniformly distributed.

Note that a codebook induces an empirical distribution in the space $\mathcal{X}^n$ with each codeword getting a probability $\frac{1}{2^{nR}}$. This in turn induces a probability distribution on the space $(M, \mathcal{X}^n, \mathcal{Y}^n, \hat{M})$ according to the channel and the decoding rule. In particular we have $M \to X^n \to Y^n \to \hat{M}$ is a Markov chain.

Observe that from Fano's inequality we have

$$H(M|\hat{M}) \leq 1 + \epsilon_n \log 2^{nR} = 1 + \epsilon_n nR$$

where $\epsilon_n \to 0$ is the probability of decoding error.

Since the message is uniformly distributed $H(M) = nR$. Hence

$$\begin{aligned}
nR &= H(M) \\
&= I(M;\hat{M}) + H(M|\hat{M}) \\
&\overset{(a)}{I} \ (M;Y^n) + 1 + \epsilon_n nR \\
&= 1 + \epsilon_n nR + \sum_{i=1}^{n} I(M;Y_i|Y^{i-1}) \\
&\leq 1 + \epsilon_n nR + \sum_{i=1}^{n} I(M,Y^{i-1};Y_i) \\
&\overset{(b)}{\leq} 1 + \epsilon_n nR + \sum_{i=1}^{n} I(X_i;Y_i) \\
&\leq 1 + \epsilon_n nR + \sum_{i=1}^{n} \max_{p(x)} I(X;Y) \\
&= 1 + \epsilon_n nR + n \max_{p(x)} I(X;Y).
\end{aligned}$$

Here (a) comes from data-processing inequality as well as Fano's inequality, and (b) follows from data-processing inequality and the fact that $M, Y^{i-1} \to X_i \to Y_i$ forms a Markov chain (due to the memoryless nature of the channel).

Thus any achievable rate (i.e. $\epsilon_n \to 0$) must satisfy

$$R \leq \max_{p(x)} I(X;Y).$$

This implies that $C \leq \max_{p(x)} I(X;Y)$ and this concludes the proof of the channel coding theorem.

***** Beware of typos in this document ****

4.1 Exercises

- Compute the channel capacity of the channel $BSC(p)$
- Compute the channel capacity of the channel $BEC(e)$.
- (For tomorrow) Prove that $I(X;Y)_{BEC} \geq I(X;Y)_{BSC}$ for every $p(x)$ if and only if $e \leq h(p)$ where $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function.

  In other words prove that

  $$(1-e)h(x) \geq h(p * x) - h(p), \quad \forall x \in [0,1]$$

  if and only if $e \leq h(p)$. Here $p * x = p(1-x) + x(1-p)$.