Lecture 16 Channel Coding Theorem

Textbook 7.6-7.11

November 28 December 3 and 5, 2024

Outline

1. Joint typical sequences
2. Channel coding theorem
3. The converse part of the channel coding theorem
4. Feedback Capacity
5. Source-channel separation theorem
6. examples

- Roughly speaking, we decode a channel output Y^n as the ith index if the codeword $X^n(i)$ is “jointly typical” with the received signal Y^n.
- We now define the important idea of joint typicality and find the probability of joint typicality when $X^n(i)$ is true cause of Y^N and when it is not.

Definition

The set $A_\epsilon^{(n)}$ of joint typical sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of n-sequences with empirical entropies ϵ-close to the true entropies:

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : | - \frac{1}{n} \log p(x^n) - H(X)| < \epsilon,$$
$$| - \frac{1}{n} \log p(y^n) - H(Y)| < \epsilon, \; | - \frac{1}{n} \log p(x^n, y^n) - H(X, Y)| < \epsilon\}$$

Theorem (Joint AEP)

Let (X^n, Y^n) *be sequences of length* n *drawn i.i.d. according to* $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$. *Then:*

1. *Pr*$((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ *as* $n \to \infty$.
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. *If* $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ [*i.e.,* $\tilde{X}^n$ *and* $\tilde{Y}^n$ *are independent with the same marginals as* $p(x^n, y^n)$], *then*

$$\text{Pr}((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

Also, for sufficient large n,

$$\text{Pr}((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}.$$

Proof.

1. We begin by showing that with high probability, the sequence is in the typical set. By the weak law of large numbers,

$$-\frac{1}{n}\log p(X^n) \to -\mathsf{E}[\log p(X)] = H(X) \text{ in probability.}$$

Hence, given $\varepsilon > 0$, there exists n_1, such that for all $n > n_1$,

$$\mathsf{Pr}\big(|-\frac{1}{n}\log p(X^n) - H(X)| \geq \varepsilon\big) < \frac{\varepsilon}{3}.$$

Proof.

Similarly, by the weak law of large numbers,

$$-\frac{1}{n}\log p(Y^n) \to -\mathsf{E}[\log p(Y)] = H(Y) \text{ in probability}$$

and

$$-\frac{1}{n}\log p(X^n, Y^n) \to -\mathsf{E}[\log p(X, Y)] = H(X, Y) \text{ in probability,}$$

and there exist n_2 and n_3, such that for all $n > n_2$,

$$\mathsf{Pr}\big(| -\frac{1}{n}\log p(Y^n) - H(Y)| \geq \varepsilon\big) < \frac{\varepsilon}{3}$$

and for all $n > n_3$,

$$\mathsf{Pr}\big(| -\frac{1}{n}\log p(X^n, Y^n) - H(X, Y)| \geq \varepsilon\big) < \frac{\varepsilon}{3}.$$

Proof.

Choosing $n > \max\{n_1, n_2, n_3\}$, we know that for n sufficiently large, the probability of the set $A_\varepsilon^{(n)}$ is greater than $1 - \varepsilon$, establishing the first part of the theorem.

Proof.

To prove the second part of the theorem, we have

$$1 = \sum p(x^n, y^n) \geq \sum_{A_\varepsilon^{(n)}} p(x^n, y^n) \geq |A_\varepsilon^{(n)}| 2^{-n(H(X,Y)+\varepsilon)},$$

and hence

$$|A_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)}.$$

Proof.

Now if$\tilde{X}^n$ and $\tilde{Y}^n$ are independent but have the same marginals as X^n and Y^n, then

$$\begin{aligned}\text{Pr}((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) &= \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n)p(y^n) \\ &\leq 2^{n(H(X,Y)+\varepsilon)} 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)} \\ &= 2^{-n(I(X;Y)-3\varepsilon)}.\end{aligned}$$

Proof.

For sufficiently large n, Pr$(A_\varepsilon^{(n)} \geq 1-\varepsilon$, and therefore

$$\begin{aligned} 1-\varepsilon &\leq \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n,y^n) \\ &\leq |A_\varepsilon^{(n)}| 2^{-n(H(X,Y)-\varepsilon)} \end{aligned}$$

and

$$|A_\varepsilon^{(n)}| \geq (1-\varepsilon)2^{n(H(X,Y)-\varepsilon)}.$$

Proof.

By similar arguments to the upper bound above, we can also show that for n sufficiently large,

$$\begin{aligned}
&\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) \\
=&\sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n)p(y^n) \\
\geq&(1-\varepsilon)2^{n(H(X,Y)+\varepsilon)}2^{-n(H(X)-\varepsilon)}2^{-n(H(Y)-\varepsilon)} \\
=&(1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)}.
\end{aligned}$$

□

Outline

1. Joint typical sequences
2. Channel coding theorem
3. The converse part of the channel coding theorem
4. Feedback Capacity
5. Source-channel separation theorem
6. examples

Ideas

Shannon used a number of new ideas to prove that information can be sent reliably over a channel at all rates up to the channel capacity. These ideas include:

- Allowing an arbitrarily small but nonzero probability of error.
- Using the channel many times in succession, so that the law of large numbers comes into effect.
- Calculating the average of the probability of error over a random choice of codebooks, which symmetrizes the probability, and which can then be used to show the existence of at least one good code.

Channel coding theorem

Theorem

For a discrete channel, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Achievability

Fix $p(x)$. Generate a $(2^{nR}, n)$ code at random according to the distribution $p(x)$. Specifically, we generate 2^{nR} codewords independently according to the distribution $p(x^n) = \pi_{i=1}^{n} p(x_i)$. We exhibit the 2^{nR} codewords as the rows of a matrix:

$$\mathcal{C} = \begin{pmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{pmatrix}$$

Each entry in this matrix is generated i.i.d. according to $p(x)$. Thus, the probability that we generate a particular code $\mathcal{C}$ is

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

1. A random code $\mathcal{C}$ is generated as described above according to $p(x)$.
2. The code $\mathcal{C}$ is then revealed to both sender and receiver. Both sender and receiver are also assumed to know the channel transition matrix $p(y|x)$ for the channel.
3. A message W is chosen according to a uniform distribution

$$P(W = w) = 2^{-nR},\ w = 1, 2, \cdots, 2^{nR}.$$

4. The wth codeword $X^n(w)$, corresponding to the wth row of $\mathcal{C}$, is sent over the channel.

5. The receiver receives a sequence Y^n according to the distribution

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

6. The receiver guesses which message was sent. We will use jointly typical decoding: the receiver declares that the index $\hat{W}$ was sent if the following conditions are satisfied:
   - $(X(\hat{W}), Y^n)$ is jointly typical.
   - There is no other index $W' \neq \hat{W}$ such that $(X^n(W'), Y^n) \in A_\epsilon^{(n)}$.

   If no such $\hat{W}$ exists or if there is more than one such, an error is declared.
7. There is a decoding error if $\hat{W} \neq W$. Let $\mathcal{E}$ be the event $\{\hat{W} \neq W\}$.

We let W be drawn according to a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$ and use jointly typical decoding $\hat{W}$ as described in step 6. Let $\mathcal{E} = \{\hat{W}(Y^n) \neq W\}$ be the error event. We will calculate the average probability of error. averaged over all codewords in the codebook, and averaged over all codebooks; that is, we calculate

$$\begin{aligned}
\Pr(\mathcal{E}) &= \sum_{\mathcal{C}} P(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \\
&= \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathcal{C}) \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_w(\mathcal{C}).
\end{aligned}$$

For every codebook $\mathcal{C}$, exchanging the 1st and wth row, we get a new codebook $\mathcal{C}'$. Note $P(\mathcal{C}) = P(\mathcal{C}')$, and $\lambda_1(\mathcal{C}) = \lambda_w(\mathcal{C}')$. And the operation that exchange the 1st and wth row is a bijection over the set of all codebooks. So

$$\sum_{\mathcal{C}} P(\mathcal{C})\lambda_1(\mathcal{C}) = \sum_{\mathcal{C}'} P(\mathcal{C}')\lambda_w(\mathcal{C}'),$$

and

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

Define the following events:

$$E_i = \{(X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)}\},\ i \in \{1, 2, \cdots, 2^{nR}\}.$$

Recall that Y^n is the result of sending the first codeword $X^n(1)$ over the channel.

Then an error occurs in the decoding scheme if and only if either E_1^c occurs (when the transmitted codeword and the received sequence are not jointly typical) or $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (when a wrong codeword is jointly typical with the received sequence).

Letting $P(\mathcal{E})$ denote $P(\mathcal{E}|W=1)$, we have

$$\begin{aligned} P(\mathcal{E}) &= P(\mathcal{E}|W=1) \\ &= P(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}|W=1) \\ &\leq P(E_1^c|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1). \end{aligned}$$

by the union of events bound for probabilities.

Now by the joint AEP, for n sufficiently large,

$$P(E_1^c|W=1) \leq \epsilon.$$

Since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$, so are Y^n and $X^n(i)$. Hence, the probability that $X^n(i)$ and Y^n are jointly typical is $\leq 2^{-nI(X;Y)-3\epsilon)}$ by the joint AEP. Consequently,

$$\begin{aligned} P(\mathcal{E}) &= P(\mathcal{E}|W=1) \leq P(E_1^c|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1) \\ &\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\ &\leq \epsilon + 2^{3n\epsilon}2^{-n(I(X;Y)-R)} \\ &\leq 2\epsilon, \end{aligned}$$

if n is sufficiently large and $R < I(X;Y) - 3\epsilon$. Hence, if $R < I(X;Y)$, we can choose ϵ and n so that the average probability of error, averaged over codebooks and codewords, is less than 2ϵ.

To finish the proof, we will strengthen the conclusion by a series of code selections.

1. Choose $p(x)$ in the proof to be $p^*(x)$, the distribution on X that achieves capacity. Then the condition $R < I(X;Y)$ can be replaced by the achievability condition $R < C$.
2. Get rid of the average over codebooks. Since the average probability of error over codebooks is small ($\leq 2\epsilon$), there exists at least one codebook $\mathcal{C}^*$ with a small average probability of error. Thus, $\Pr(\mathcal{E}|\mathcal{C}^*) \leq 2\epsilon$.

3. Throw away the worst half of the codewords in the best codebook $\mathcal{C}^*$. Since the arithmetic average probability of error $P_e^{(n)}(\rfloor^*)$ for this code is less than 2ϵ, we

$$P(\mathcal{E}|\mathcal{C}^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*).$$

   which implies that at least half the indices i and their associated codewords $X^n(i)$ must have conditional probability of error λ_i less than 4ϵ. Hence the best half of the codewords have a maximal probability of error less than 4ϵ. If we reindex these codewords, we have 2^{nR-1} codewords. Throwing out half the codewords has changed the rate from R to $R - \frac{1}{n}$, which is negligible for large n.

Combining all these improvements, we have constructed a code of rate $R' = R - \frac{1}{n}$, with maximal probability of error $\lambda^{(n)} \leq 4\epsilon$. This proves the achievability of any rate below capacity.

Outline

1. Joint typical sequences
2. Channel coding theorem
3. The converse part of the channel coding theorem
4. Feedback Capacity
5. Source-channel separation theorem
6. examples

Let us define the setup under consideration. The index W is uniformly distributed on the set $W = \{1, 2, \cdots, 2^{nR}\}$, and the sequence Y^n is realted probabilistically to W. From Y^n, we estimate the index W that was sent. Let the estimate be $\hat{W} = g(Y^n)$. Thus, $W \to X^n(W) \to Y^n \to \hat{W}$ forms a Markov chain. Note that the probability of error is

$$\mathrm{Pr}(\hat{W} \neq W) = \frac{1}{2^{nR}} \sum_i \lambda_i = P_e^{(n)}.$$

Lemma (Fano's inequality)

For a discrete memoryless channel with a codebook $\mathcal{C}$ the input message W uniformly distributed over 2^{nR}, we have

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} nR.$$

Proof.

Sine W is uniformly distributed, we have $P_e^{(n)} = \Pr(W \neq \hat{W})$. We apply Fano's inequality for W in an alphabet of size 2^{nR}. □

Lemma

Let Y^n be the result of passing X^n through a discrete memoryless channel of capacity C. Then for all $p(x^n)$,

$$I(X^n; Y^n) \leq nC.$$

Proof.

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

Converse part of the channel coding theorem

We have to show that any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$. Note that $P_e^{(n)} \to 0$.

For a fixed encoding rule $X^n(\cdot)$ and fixed decoding rule $\hat{W} = g(Y^n)$, we have $W \to X^n(W) \to Y^n \to \hat{W}$. For each n, let W drawn according to a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$. Since W has a uniform distribution,

$$\Pr(\hat{W} \neq W) = P_e^{(n)} = \frac{1}{2^{nR}} \sum_i \lambda_i.$$

Hence,

$$\begin{aligned} nR &= H(W) \\ &= H(W|\hat{W}) + I(W;\hat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(X^n;Y^n) \\ &\leq 1 + P_e^{(n)} nR + nC. \end{aligned}$$

Dividing by n, we obtain

$$R \leq P_e^{(n)} R + \frac{1}{n} + C.$$

Now letting $n \to \infty$, we see that the first two terms on the right-hand side tend to 0, and hence

$$R \leq C.$$

Note

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

This equation shows that if $R > C$, the probability of error is bounded away from 0 for sufficiently large n (and hence for all n). Hence, we cannot achieve an arbitrarily low probability of error at rates above capacity.

- We have proved the channel coding theorem and its converse.
- In essence, these theorems state that when $R < C$, it is possible to send information with an arbitrarily low probability of error, and when $R > C$, the probability of error is bounded away from zero.
- It is interesting and rewarding to examine the consequences of equality in the converse.
- Hopefully, it will give some ideas as to the kinds of codes that achieve capacity.

Repeating the steps of the converse in the case when $P_e = 0$, we have

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

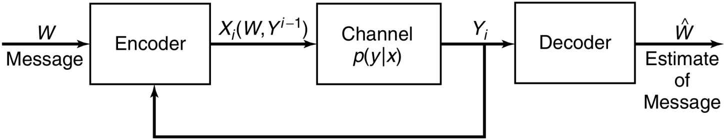
We have equality in the first inequality, the data-processing inequality, only if $I(Y^n; X^n(W)|W) = 0$ and $I(X^n : Y^n|\hat{W}) = 0$, which is true if all the codewords are distinct and if $\hat{W}$ is a sufficient statistic for decoding.

We have equality in the second inequality only if the Y_i's are independent, and equality in the third inequality only if the distribution of X_i is $p^*(x)$, the distribution on X that achieves capacity.

We have equality in the converse only if these conditions are satisfied.

Outline

1. Joint typical sequences
2. Channel coding theorem
3. The converse part of the channel coding theorem
4. Feedback Capacity
5. Source-channel separation theorem
6. examples


W
Message
Encoder
$X_i(W, Y^{i-1})$
Channel
$p(y|x)$
Y_i
Decoder
$\hat{W}$
Estimate
of
Message

- We assume that all the received symbols are sent back immediately and noiselessly to the transmitter, which can then use them to decide which symbol to send next.
- Can we do better with feedback?
- The surprising answer is no, which we shall now prove.
- We define a $(2^{nR}, n)$ **feedback code** as a sequence of mappings $x_i(W, Y^{i-1})$, where each x_i is a function only of the message $W \in 2^{nR}$ and the previous received values, Y_1, Y_2, $\cdots$, Y_{i-1}, and a sequence of decoding functions $g : \mathcal{Y}^n \to \{1, 2, \cdots, 2^{nR}\}$.
- Thus,

$$P_e^{(n)} = \mathrm{Pr}\{g(Y^n) \neq W\},$$

when W is uniformly distributed over $\{1, 2, \cdots, 2^{nR}\}$.

Definition

The **capacity with feedback**, C_{FB}, of a discrete memoryless channel is the supremum of all rates achievable by feedback codes.

Theorem

Feedback capacity

$$C_{FB} = C = \max_{p(x)} I(X : Y).$$

Proof.

- Since a nonfeedback code is a special case of a feedback code, any rate that can be achieved without feedback can be achieved with feedback, and hence

$$C_{\mathsf{FB}} \geq C.$$

- Proving the inequality the other way is slightly more tricky.
- We cannot use the same proof that we used for the converse to the coding theorem without feedback.
- Note X_i depends on the past received symbols, and it is no longer true that Y_i depends only on X_i and is conditionally independent of the future X's.

Proof.

- There is a simple change that will fix the problem with the proof.
- Instead of using X^n, we will use the index W and prove a similar series of inequalities.
- Let W be uniformly distributed over $\{1, 2, \cdots, 2^{nR}\}$.
- Then $\Pr(W \neq \hat{W}) = P_e^{(n)}$ and

$$\begin{aligned} nR = H(W) &= H(W|\hat{W}) + I(W;\hat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(W;Y^n), \end{aligned}$$

by Fano's inequality and the data-processing inequality.

Proof.

Now we can bound $I(W;Y^n)$ as follows:

$$\begin{aligned}
I(W;Y^n) &= H(Y^n) - H(Y^n|W) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, Y_2, \cdots, Y_{i-1}, W) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, Y_2, \cdots, Y_{i-1}, W, X_i) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i),
\end{aligned}$$

since X_i is a function of $Y_1, \cdots, Y_{i-1}$ and W; and conditional on X_i, Y_i is independent of W and past samples of Y.

Proof.

Continuing, we have

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

from the definition of capacity for a discrete memoryless channel.

Proof.

Putting these together, we obtain

$$nR \leq P_e^{(n)} nR + 1 + nC,$$

and dividing by n and letting $n \to \infty$, we conclude that $R \leq C$.

Thus, we cannot achieve any higher rates with feedback than we can without feedback, and

$$C_{\mathsf{FB}} = C.$$

□

- It is now time to combine the two main results that we have proved so far: data compression ($R > H$: Theorem 5.4.2) and data transmission ($R < C$: Theorem 7.7.1).
- Is the condition $H < C$ necessary and sufficient for sending a source over a channel?
- For example, consider sending digitized speech or music over a discrete memoryless channel.
- We could design a code to map the sequence of speech samples directly into the input of the channel, or we could compress the speech into its most efficient representation, then use the appropriate channel code to send it over the channel.
- It is not immediately clear that we are not losing something by using the two-stage method, since data compression does not depend on the channel and the channel coding does not depend on the source distribution.

- We will prove in this section that the two-stage method is as good as any other method of transmitting information over a noisy channel.
- This result has some important practical implications.
- It implies that we can consider the design of a communication system as a combination of two parts, source coding and channel coding.
- We can design source codes for the most efficient representation of the data.
- We can, separately and independently, design channel codes appropriate for the channel.
- The combination will be as efficient as anything we could design by considering both problems together.

- The result–that a two-stage process is as good as any one-stage process–seems so obvious that it may be appropriate to point out that it is not always true.
- There are examples of multiuser channels where the decomposition breaks down.
- We also consider two simple situations where the theorem appears to be misleading.
- A simple example is that of sending English text over an erasure channel.
- We can look for the most efficient binary representation of the text and send it over the channel.
- But the errors will be very difficult to decode.

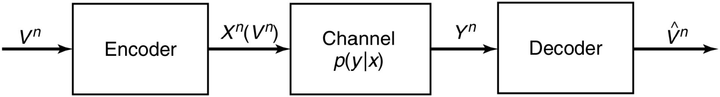
- If, however, we send the English text directly over the channel, we can lose up to about half the letters and yet be able to make sense out of the message.
- Similarly, the human ear has some unusual properties that enable it to distinguish speech under very high noise levels if the noise is white.
- In such cases, it may be appropriate to send the uncompressed speech over the noisy channel rather than the compressed version.
- Apparently, the redundancy in the source is suited to the channel.

- Let us define the setup under consideration.
- We have a source V that generates symbols from an alphabet $\mathcal{V}$.
- We will not make any assumptions about the kind of stochastic process produced by V other than that is it is from a finite alphabet and satisfies the AEP.
- Examples of such processes include a sequence of i.i.d. random variables and the sequence of states of a stationary irreducible Markov chain.

- We want to send the sequence of symbols $V^n = V_1, V_2, \cdots, V_n$ over the channel so that the receiver can reconstruct the sequence.
- To do this, we map the sequence onto a codeword $X^n(V^n)$ and send the codeword over the channel.
- The receiver looks at his received sequence Y^n aand makes an estimate $\hat{V}^n$ of the sequence V^n that was sent.
- The receiver makes an error if $V^n \neq \hat{V}^n$. We define the probability of error as

$$\Pr(V^n \neq \hat{V}^n) = \sum_{y^n} \sum_{v^n} p(v^n) p(y^n | x^n(v^n)) I(g(y^n) \neq v^n),$$

where I is the indicator function and $g(y^n)$ is the decoding function.


V^n
Encoder
$X^n(V^n)$
Channel
$p(y|x)$
Y^n
Decoder
$\hat{V}^n$

Theorem (Source-channel coding theorem)

If $V_1, V_2, \cdots, V^n$ *is a finite alphabet stochastic process that satisfies the AEP and* $H(\mathcal{V} < C$*, there exists a source-channel code with probability of error Pr*$(\hat{V}^n \neq V^n) \to 0.$ *Conversely, for any stationary stochastic process, if* $H(\mathcal{V}) > C$*, the probability of error is bounded away from zero, and it is not possible to send the process over the chnnel with arbitrarily low probability of error.*

Proof.

Achievability.

- The essence of the forward part of the proof is the two-stage encoding described earlier.
- Since we have assumed that the stochastic process satisfies the AEP, it implies that there exists a typical set $A_\epsilon^{(n)}$ of size $\leq 2^{n(H(\mathcal{V})+\varepsilon)}$ which contains most of the probability.
- We will encode only the source sequences belonging to the typical set; all other sequences will result in an error.
- This will contribute at most ε to the probability of error.

Proof.

- We index all the sequences belonging to $A_{\varepsilon}^{(n)}$.
- Since there are at most $2^{n(H+\varepsilon)}$ such sequences, $n(H+\varepsilon)$ bits suffice to index them.
- We can transmit the desired index to the receiver with probability of error less than ε if

$$H(\mathcal{V}) + \varepsilon = R < C.$$

Proof.

- The receiver can reconstruct V^n by enumerating the typical set $A_\varepsilon^{(n)}$ and choosing the sequence corresponding to the estimated index.
- This sequence will agree with the transmitted sequence with high probability.
- To be precise,

$$\begin{aligned} P(V^n \neq \hat{V}^n) &\leq P(V^n \notin A_\varepsilon^{(n)}) + P(g(Y^n) \neq V^n | V^n \notin A_\varepsilon^{(n)}) \\ &\leq \varepsilon + \varepsilon = 2\varepsilon \end{aligned}$$

  for n sufficiently large.
- Hence, we can reconstruct the sequence with low probability of error for n sufficiently large if

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

Proof.

- We wish to show that $\Pr(\hat{V}^n \neq V^n) \to 0$ implies that $H(\mathcal{V}) \leq C$ for any sequence of source-channel codes

$$X^n(V^n) : \mathcal{V}^n \to \mathcal{X}^n,$$
$$g_n(Y^n) : \mathcal{Y}^n \to \mathcal{V}^n.$$

- Thus $X^n(\cdot)$ is an arbitrary (perhaps rndom) assignment of codewords to data sequences V^n, and $g_n(\cdot)$ is any decoding function (assignment of estimates $\hat{V}^n$ to output sequences Y^n.
- By Fano's inequality, we must have

$$H(V^n|\hat{V}^n) \leq 1 + \Pr(\hat{V}^n \neq V^n) \log|\mathcal{V}^n| = 1 + \Pr(\hat{V}^n \neq V^n) n \log|\mathcal{V}|.$$

Proof.

Hence for the code,

$$\begin{aligned} H(\mathcal{V}) &\le \frac{H(V_1, V_2, \cdots, V_n)}{n} \\ &= \frac{H(V^n)}{n} \\ &= \frac{1}{n} H(V^n|\hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n) \\ &\le \frac{1}{n}(1 + \mathsf{Pr}(\hat{V} \ne V^n) n \log |\mathcal{V}|) + \frac{1}{n} I(V^n; \hat{V}^n) \\ &\le \frac{1}{n}(1 + \mathsf{Pr}(\hat{V} \ne V^n) n \log |\mathcal{V}|) + \frac{1}{n} I(X^n; Y^n) \\ &\le \frac{1}{n} + \mathsf{Pr}(\hat{V}^n \ne V^n) \log |\mathcal{V}| + C. \end{aligned}$$

Now letting $n \to \infty$, we have $\mathsf{Pr}(\hat{V}^n \ne V^n) \to 0$ and hence

$$H(\mathcal{V}) \le C.$$

□

- Hence, we can transmit a stationary ergodic source over a channel if and only if its entropy rate is less than the capacity of the channel.
- The joint source–channel separation theorem enables us to consider the problem of source coding separately from the problem of channel coding.
- The source coder tries to find the most efficient representation of the source, and the channel coder encodes the message to combat the noise and errors introduced by the channel.
- The separation theorem says that the separate encoders an achieve the same rates as the joint encoder.

- With this result, we have tied together the two basic theorems of information theory: data compression and data transmission.
- We will try to summarize the proofs of the two results in a few words.
- The compression theorem is a consequence of the AEP, which shows that there exists a "small" subset (of size $2^n H$) of all possible source sequences that contain most of the probability and that we can therefore represent the source with a small probability of error using H bits per symbol.

- The data transmission theorem is based on the joint AEP; it uses the fact that for long block lengths, the output sequence of the channel is very likely to be jointly typical with the input codeword, while any other codeword is jointly typical with probability $\approx 2^{-nI}$.
- Hence, we can use about 2^{nI} codewords and still have negligible probability of error.
- The source–channel separation theorem shows that we can design the source code and the channel code separately and combine the results to achieve optimal performance.

Example (2.10)

Let X_1 and X_2 be discrete random variables drawn according to probability mass functions $p_1(\cdot)$ and $p_2(\cdot)$ over the respective alphabets $\mathcal{X}_1 = \{1, 2, \ldots, m\}$ and $\mathcal{X}_2 = \{m+1, \ldots, n\}$. Let

$$X = \begin{cases} X_1, & \text{with probability } \alpha \\ X_2, & \text{with probability } 1-\alpha. \end{cases}$$

(a) *Find $H(X)$ in terms of $H(X_1)$ and $H(X_2)$ and α.*

(b) *Maximize over α to show that $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$ and interpret using the notion that $2^{H(X)}$ is the effective alphabet size.*

Example

A discrete memoryless source emits a sequence of statistically independent binary digits with probabilities $p(1) = 0.005$ *and* $p(0) = 0.995$. *The digits are taken* 100 *at time and a binary codeword is provided for every sequence of 100digits containing three or fewer* 1*'s.*

(1) *Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer* 1*'s.*

(2) *Calculate the probability of observing a source sequence for which no codeword has be assigned.*

(3) *Use Chebyshev's inequality to bound the probability of observing a source sequence for which no codeword has been assigned. Compare this bound with the actual probability computed in part (2).*

Example

Consider the following problem: m *binary signals* S_1, S_2, $\cdots$, S_m *are available at times* $T_1 \leq T_2 \leq \cdots T_m$, *and we would like to find their sum* $S_1 \oplus S_2 \oplus \cdots \oplus S_m$ *using two-input gates, each gate with one time unit delay, so that the final result is available as quickly as possible. A simple greedy algorithm is to combine the earliest two results, forming the partial result at time* $\max(T_1, T_2) + 1$. *We now have a new problem with* $S_1 \oplus S_2$, S_3, $\cdots$, S_m, *available at times* $\max(T_1, T_2) + 1$, T_3, $\cdots$, T_m. *We can now sort this list of* $T's$ *and apply the same merging step again, repeating this until we have the final result.*

Example (Continued)

(1) *Argue that the forgoing procedure is optimal, in that it constructs a circuit for which the final result is available as quickly as possible.*

(2) *Show that this procedure finds the tree that minimizes*

$$C(T) = \max_i (T_i + l_i),$$

where T_i is the time at which the result allotted to the ith leaf ia available and l_i is the length of the path from the i the leaf to the root.

Example (Continued)

(3) *Show that*

$$C(T) \geq \log_2(\sum_i 2^{T_i})$$

for any tree T*.*

(4) *Show that there exists a tree such that*

$$C(T) \leq \log_2(\sum_i 2^{T_i}) + 1.$$

Thus, $\log_2(\sum_i 2^{T_i})$ *is the analog of entropy for this problem.*

Example (2.30)

Find the probability mass function $p(x)$ that maximizes the entropy $H(X)$ of a nonnegative integer-valued random variable X subject to the constraint

$$EX = \sum_{n=0}^{\infty} np(n) = A$$

for a fixed value $A > 0$. Evaluate this maximum $H(X)$.

Example

Let $X_1 \to X_2 \to X_3 \to X_4$ *form a Markov chain. Show that*

$$I(X_1;X_3) + I(X_2;X_4) \leq I(X_1 : X_4) + I(X_2;X_3).$$

Proof.

By the chain rule of mutual information,

$$\begin{aligned} I(X_1; X_3, X_4) &= I(X_1; X_3) + I(X_1; X_4|X_3) \\ &= I(X_1; X_4) + I(X_1; X_3|X_4), \end{aligned}$$

and

$$\begin{aligned} I(X_2; X_3, X_4) &= I(X_2; X_3) + I(X_2; X_4|X_3) \\ &= I(X_2; X_4) + I(X_2; X_3|X_4). \end{aligned}$$

Since $X_1 \to X_2 \to X_3 \to X_4$ form a Markov chain, $I(X_1; X_4|X_3) = 0$ and $I(X_2; X_4|X_3) = 0$. And by data processing inequality, we have that

$$I(X_1; X_3|X_4) \leq I(X_2; X_3|X_4),$$

and the claim follows. □

Example (5.10)

A random variable X takes on m values and has entropy $H(X)$. An instantaneous ternary code is found for this source, with average length

$$L = \frac{H(X)}{\log_2 3} = H_3(X).$$

(1) *Show that each symbol of X has a probability of the form 3^{-i} for some i.*

(2) *Show that m is odd.*

Example (2.33)

Let Pr$(X = i) = p_i$, $i = 1, 2, \cdots, m$, *and let* $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_m$. *The minimal probability of error predictor of* X *is* $\hat{X} = 1$, *with resulting probability of error* $P_e = 1 - p_1$. *Maximize* $H(\mathbf{p})$ *subject to the constraint* $1 - p_1 = P_e$ *to find a bound on* P_e *in terms of* H.

Proof.

We have that

$$
\begin{aligned}
H(\mathbf{p}) &= -p_1 \log p_1 - \sum_{i=2}^{m} p_i \log p_i \\
&= -p_1 \log p_1 - \sum_{i=2}^{m} P_e \frac{p_i}{P_e} \log \frac{p_i}{P_e} - P_e \log P_e \\
&= H(P_e) + P_e H(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \cdots, \frac{p-m}{P_e}) \\
&\leq H(P_e) + P_e \log(m-1),
\end{aligned}
$$

since the maximum of $H(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \cdots, \frac{p_m}{P_e})$ is attained by a uniform distribution. Hence $H(X) \leq H(P_e) + P_e \log(m-1)$, which is the unconditional form of Fano's inequality. We can weaken the this inequality to obtain an explicit lower bound for P_e,

$$
P_e \geq \frac{H(X) - 1}{\log(m-1)}.
$$

□

Example (4.30)

Let X be the waiting time for the first heads to appear in successive flips of a fair coin. For example, Pr$\{X = 3\} = (\frac{1}{2})^3$. Let S_n be the waiting time for the nth head to appear. Thus,

$$S_0 = 0$$
$$S_{n+1} = S_n + X_{n+1},$$

where X_1, X_2, $X_3, \cdots$ are i.i.d. according to the distribution above.

(1) *Is the process $\{S_n\}$ stationary?*

(2) *Calculate $H(S_1, \cdots, S_n)$.*

(3) *Does the process $\{S_n\}$ have an entropy rate?*