Lecture 16 Channel Coding Theorem

Textbook 7.6-7.7, 7.9

October 28th and November 4th, 2022

Outline

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*

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

   Also, for sufficient large n,

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

Outline

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 $P_e^{(n)}([c]^*)$ 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

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

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

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

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

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.