Lecture 6, Oct. 7, 2003

Theorem: The Channel Coding Theorem:

All rates below capacity C are achievable, i.e., for every $R < C$ and any $\varepsilon > 0$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.

Theorem: Converse to the channel coding theorem

~~Converse by~~, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Proof of Coding ~~theorem~~ theorem (the achievability

Fix $p(x)$ and generate a $(2^{nR}, n)$ code at random according to the distribution, i

$$p(x^n) = \prod_{i=1}^{n} p(x_i)$$

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

Each entry in the matrix is generated i.i.d. according to $p(x)$.

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

This would be the sequence of events in the channel coding process:

1) A random code $\mathcal{C}$ is generated at random according to $p(x)$

2) The code $\mathcal{C}$ is then revealed to both the sender and receiver. Both sender and receiver know the transition matrix $p(y|x)$.

3) A message W is chosen according to a uniform distribution:

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

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

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

6) The receiver makes a decision about the message sent. In practice, the receiver uses maximum likelihood decoding, which is optimal. However, to make the analysis simple, we assume that the receiver uses typical set decoding. We describe this as:

The receiver declares that the index $\hat{W}$ was sent if the following conditions are satisfied:

- $(X^n(\hat{W}), Y^n)$ is jointly typical.
- There is no other index k such that $(X^n(k), Y^n) \in A_\varepsilon^{(n)}$

If no such $\hat{W}$ exist or if there are more than one such $\hat{W}$, then an error is declared.

7) There is a decoding error if $\hat{W} \neq W$. Denote by $\mathcal{E}$ the event $\hat{W} \neq W$.

Analysis of the probability of error:

$$P_r(\mathcal{E}) = \sum_C P(C) P_e^{(n)}(C)$$

$$= \sum_C P(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C)$$

$$= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_C P(C) \lambda_w(C)$$

where $P_e^{(n)}(C)$ is defined for typical set decoding.

The way we have constructed the code results in symmetry w.r.t. codewords, i.e., the average probability of error (averaged over all codes) does not depend on the particular index that was sent, i.e., $\sum_C P(C)\lambda_w(C)$ does not depend on w. Thus, we can assume that $W=1$ was sent. Then,

$$P_r(\mathcal{E}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_C P(C)\lambda_w(C)$$

$$= \sum_C P(C)\lambda_1(C)$$

$$= P_r(\mathcal{E}|W=1)$$

Define the even E_i as:

$$E_i = \{(x^n(i), y^n) \in A_\varepsilon^{(n)}\} \quad i \in \{1, 2, \ldots, 2^{nR}\}$$

E_i is the event that the ith codeword and y^n are jointly typical.

Then a decoding error occurs either when E_1^c occurs (when the transmitted and received sequences are not jointly typical) or $E_2 \cup E_3 \cdots \cup E_{2^{nR}}$ occurs, i.e., when another

codeword is jointly typical with y^n.

Hence, letting $P(\mathcal{E}) = P_r(\mathcal{E}|W=1)$, we have

$$P_r(\mathcal{E}|W=1) = P(E_1^c \cup E_2 \cup E_3 \cdots \cup E_{2^{nR}})$$

$$\leq P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i)$$

From the joint AEP, $P(E_1^c) \to 0$, hence

$$P(E_1^c) \leq \varepsilon \quad \text{for } n \text{ sufficiently large.}$$

Since by code construction $X^n(1)$ and $X^n(i)$ are independent so are Y^n and $X^n(i)$, $i \neq 1$.

Hence, the probability that $X^n(i)$ and Y^n are jointly typical is $\leq 2^{-n(I(X;Y)-3\varepsilon)}$

$$P(\mathcal{E}) = P(\mathcal{E}|W=1) \leq P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i)$$

$$\leq \varepsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\varepsilon)}$$

$$= \varepsilon + (2^{nR} - 1)\, 2^{-n(I(X;Y)-3\varepsilon)}$$

$$\leq \varepsilon + 2^{3n\varepsilon} \cdot 2^{-n(I(X;Y)-R)}$$

if n is sufficiently large and $R < I(X;Y) - 3\varepsilon$ then $P(\mathcal{E}) \leq 2\varepsilon$.

Hence if $R < I(X;Y)$, we can choose ε and n such that the average probability of error is

less than 2ε.

Now we finish the proof by:

1) Choosing $p(x)$ to be $p^*(x)$, i.e., the input probability distribution that maximizes $I(X;Y)$ (to get $C = I(X;Y)\big|_{p^*(x)}$). Then, we can replace $R < I(X;Y)$ by $R < C$.

2) Since probability of error averaged over codebooks is small, there should be one that results in small probability of error. "We can find" such a good code book C^* by exhaustive search over all $(2^{nR}, n)$ codes.

This way, we get rid of average over codebooks.

3) Take the best codebook (C^*). For this codebook C^*, we have

$$\frac{1}{2^{nR}} \sum_i \lambda_i(C^*) \leq 2\varepsilon$$

through throw away half of the codewords of C^* (the worst ones). The remaining half have conditional probability of error λ_i (associated to $X^n(i)$ less than 4ε.

If this was not true, these codewords themselves would contribute 2ε or more to the sum. If we re-index the remaining codewords, we get 2^{nR-1} codewords. So, we get a codebook of rate $R - \frac{1}{n} \xrightarrow[n\to\infty]{} R$ with maximal error probability $\lambda^{(n)} \leq 4\varepsilon$.

We have proved the achievability, i.e., the possibility of error free transmission for $R < C$. Now, we prove the error-free transmission is not possible if $R > C$. This is the converse to coding theorem

Theorem: Any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

In order to prove the converse channel coding theorem, we use the following two Lemmas.

Lemma: (Fano's inequality): For a discrete memory less channel with a codebook C and the input message uniformly distributed, let $P_e^{(n)} = P_r(W \neq g(y^n))$. Then:

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

Define the random variable E as

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

Then,

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

$H(E|W,Y^n) = 0$ since E is a function of W and $g(Y^n)$.

So,

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

$H(E|Y^n) \leq 1$ Since E is a binary r.v.

$$H(W|E,Y^n) = P(E=0)H(W|E=0,Y^n)$$
$$+ P(E=1)H(W|E=1,Y^n)$$
$$= (1-P_e^{(n)}) \times 0 + P_e^{(n)} H(W|E=1,Y^n)$$
$$\leq P_e^{(n)} \log(|W|-1)$$
$$\leq P_e^{(n)} nR$$

So,

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

Since for a given fixed code $X^n(W)$ is a

function of W.

$$H(X^n(W)|Y^n) \leq H(W|Y^n)$$

So,

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

Lemma: Let Y^n be the result of passing X^n through a DMC. Then

$$I(X^n;Y^n) \leq nC \quad \text{for all } p(x^n).$$

Interpretation: This means that by using a DMC several times, we cannot increase the capacity per use.

Proof:

$$I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n)$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\dots Y_{i-1},X^n)$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i)$$

{since the entropy of a collect of r.v.'s is less than the sum of their entropies} →

$$\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i)$$

$$= \sum_{i=1}^{n} I(X_i;Y_i) \leq nC.$$

Now, we proceed with the proof of the converse to Channel Coding Theorem, i.e., we show that any sequence $(2^{nR}, n)$ of codes with $\lambda^{(n)} \to 0$ must satisfy $R \leq C$.

Proof: if $\lambda^{(n)} \to 0$ then so does the average probability of error $P_e^{(n)} = \frac{1}{M}\sum_{i=1}^{M} \lambda_i \to 0$.

For each n, let W be drawn uniformly from $\{1, 2, \ldots, 2^{nR}\}$. Then $P_e^{(n)} = Pr(\hat{W} \neq W)$.

Hence,

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

$$\leq 1 + P_e^{(n)} nR + I(X^n(W); Y^n)$$

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

Dividing both sides by n,

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

letting $n \to \infty$ the first two terms $\to 0$ and we get $R \leq C$.

We can write (A) as

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

So, if $R > C$ then $P_e^{(n)}$ is bounded away from zero for sufficiently large n, and, hence for all n. The latter comes from the fact that if $P_e^{(n)} \to 0$ for small n, we could construct long codes by concatenating short codes.

X Lecture 7, Oct. 14, 2003

Continuous Sources

A continuous random variable X is characterised by a Cumulative Distribution Function $F_X(x)$,

$$F_X(x) = P_r(X \leq x)$$

the derivative of $F_X(x)$ w.r.t. x is called the probability density function (pdf),

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$

when there is no ambiguity, we drop the subscript and denote the pdf by $f(x)$.

It is clear that,

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

therefore

$$\int_S f(x)\,dx = 1$$