6.441 Transmission of Information March. 21, 2006

Lecture 12

Lecturer: Madhu Sudan *Scribe:* Costas Pelekanakis

1. Today's outline

a. Joint Typicality

b. Channel Coding Theorem for DMC

c. Achievability of $R < C = \max\limits_{p(x)} \{I(x;y)\}$

d. Nonachievability of $R > C$

2. Definitions

- DMC : $(\underbrace{X}_{\text{finite set}}, \underbrace{P_{y/x}}_{\text{transition probability matrix}}, \underbrace{Y}_{\text{finite set}})$
- Nth extension of DMC: $(X^n, P_{y^n/x^n}, Y^n)$. Check for properties of DMC:
    - memoryless $\Leftrightarrow X_{i-1} \to X_i \to Y_i$
    - no feedback $\Leftrightarrow Y_1,...Y_{i-1} \to X_1,...,X_{i-1} \to X_i$
    - memoryless + no feedback $\Leftrightarrow (X_1,...X_{i-1},Y_1,...,Y_{i-1}) \to X_i \to Y_i \Leftrightarrow P_{Y^n/X^n} = \prod_{i=1}^{n} P_{Y_i/X_i}$
- (M,n) code
    - message index set $\{1,...,M\}$
    - encoding function $f : W \to X^n$
    - codebook: $C = \begin{bmatrix} X_1(1) & X_2(1) \ldots\ldots\ldots X_n(1) \\ X_1(2) & X_2(2) \ldots\ldots\ldots X_n(2) \\ \ldots. & \\ X_1(M) & X_2(M) \ldots\ldots. X_n(M) \end{bmatrix}$
    - decoding function $g : Y^n \to W$
    - For M uniformly distributed messages, Rate = bits/channel uses = $\log_2(M)/n$. If we fix rate $R = \log_2(M)/n \Rightarrow M = 2^{nR}$.
- Probability of error conditioned on ith message sent $\lambda_i = \Pr[g(Y^n) \neq i / X^n = f(i)]$

- Maximum probability of error: $\lambda^{(n)} = \max_i \{\lambda_i\}$
- Arithmetic average probability of error: $P_e^{(n)} = \frac{1}{M}\sum_{i=1}^{M}\lambda_i$. Obviously, $P_e^{(n)} \le \lambda^{(n)}$
- A rate is said to be *achievable* if there exists a sequence of $\left(\lceil 2^{nR} \rceil, n\right)$ codes such that $\lambda^{(n)}$ tends to zero as n increases.
- The capacity of a DMC is the supremum of all the achievable rates.

3. Jointly typical sequences

Recall that a random vector X^n with i.i.d. components according to p_x is a typical sequence if $2^{-n(H(X)+\varepsilon)} \le p(X^n = x^n) \le 2^{-n(H(X)-\varepsilon)}$. We extend the notion of typical sequences to jointly typical sequences. We have two random vectors X^n and Y^n with i.i.d. components according to p_x and p_y respectively. In addition, $(X_i, Y_i) \sim p_{x,y}$. Hence, $p(X^n = x^n, Y^n = y^n) = \prod_{i=1}^{n} p(X_i = x_i, Y_i = y_i)$.

3.1. Definition: Jointly typical sequences

The set A_ε^n of jointly typical sequences (x^n, y^n) is defined as:

$$A_\varepsilon^{(n)} = \left\{ \begin{array}{l} (x^n, y^n): \\ \qquad 2^{-n(H(X)+\varepsilon)} \le p(X^n = x^n) \le 2^{-n(H(X)-\varepsilon)} \\ \qquad 2^{-n(H(Y)+\varepsilon)} \le p(Y^n = y^n) \le 2^{-n(H(Y)-\varepsilon)} \\ \qquad 2^{-n(H(X,Y)+\varepsilon)} \le p(X^n = x^n, Y^n = y^n) \le 2^{-n(H(X,Y)-\varepsilon)} \end{array} \right\}$$

where

$$p(X^n = x^n, Y^n = y^n) = \prod_{i=1}^{n} p(X_i = x_i, Y_i = y_i)$$

3.2. Theorem: Joint AEP

Let (X^n, Y^n) *be a random sequence with i.i.d. pairs* $(X_i, Y_i) \sim p_{x,y}$ *and*

$p(X^n = x^n, Y^n = y^n) = \prod_{i=1}^{n} p(X_i = x_i, Y_i = y_i)$. *Then the following are true:*

1. $\Pr\left(\left(X^n, Y^n\right) \in A_\varepsilon^n\right) \to 1$ as $n \to \infty$.
2. $\left|A_\varepsilon^n\right| \le 2^{n(H(X,Y)+\varepsilon)}$

3. *If* $\widetilde{X}^n$, $\widetilde{Y}^n$ *are independent with i.i.d. components and* $\widetilde{X}_i \sim p_x$ *and* $\widetilde{Y}_i \sim p_y$ *then for sufficient large n:* $\Pr\left(\left(\widetilde{X}^n,\widetilde{Y}^n\right)\in A_\varepsilon^n\right)\le 2^{-n(I(X;Y)-3\varepsilon)}$ and $\Pr\left(\left(\widetilde{X}^n,\widetilde{Y}^n\right)\in A_\varepsilon^n\right)\ge (1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)}$

Proof:

1 and 2 are derived from $2^{-n(H(X,Y)+\varepsilon)} \le p(X^n = x^n, Y^n = y^n) \le 2^{-n(H(X,Y)-\varepsilon)}$. We have shown a similar proof when we talked about typical sets. Here we give the proof for 3:

$$\Pr\left(\left(\widetilde{X}^n,\widetilde{Y}^n\right)\in A_\varepsilon^n\right)= \sum_{(x^n,y^n)\in A_\varepsilon^n} p(x^n)p(y^n) \le 2^{n(H(X/Y)+\varepsilon)}2^{-n(H(X)-\varepsilon)}2^{-n(H(Y)-\varepsilon)} \le 2^{-n(I(X;Y)-3\varepsilon)}$$

$$\Pr\left(\left(\widetilde{X}^n,\widetilde{Y}^n\right)\in A_\varepsilon^n\right)= \sum_{(x^n,y^n)\in A_\varepsilon^n} p(x^n)p(y^n) \ge (1-\varepsilon)2^{n(H(X/Y)-\varepsilon)}2^{-n(H(X)+\varepsilon)}2^{-n(H(Y)+\varepsilon)} \ge (1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)}$$

Comments: Not all pairs of typical X^n and typical Y^n are jointly typical since there exist approximately $2^{nH(X;Y)}$ typical pairs. Each typical X^n induces about $2^{nH(Y/X)}$ possible Y^n typical sequences, all of them equally likely. If we want to ensure that two different codewords will induce two disjoint sets of typical Y^n possible sequences then the total number of $2^{nH(Y)}$ typical Y^n must be divided into $2^{n(H(Y)-H(Y/X))} = 2^{nI(X;Y)}$ disjoint sets. Hence, we are allowed to send at most $2^{nI(X;Y)}$ different typical X^n sequences.

4. Channel Coding for the DMC

4.1. The Channel Coding Theorem

All rates below capacity C are achievable, namely, for every R<C there exists a sequence of $\left(2^{nR},n\right)$ *codes with* $\lambda^{(n)} \xrightarrow{n\to\infty} 0$. *Conversely, any sequence of* $\left(2^{nR},n\right)$ *codes with* $\lambda^{(n)} \xrightarrow{n\to\infty} 0$ must have $R \le C$.

Achievability:

Note that $\lambda^{(n)}$ is not easy to deal because it involves maximization which is a nonlinear operation while $P_e^{(n)}$ is something we can compute. But $P_e^{(n)} \le \lambda^{(n)}$, so let's try to generate a code $\aleph$ (*MxN* matrix of symbols) for which when $P_{e\aleph}^{(n)} \xrightarrow{n\to\infty} 0$ and show that there is a subcode $\aleph'$ of $\aleph$ (*TxN* submatrix of symbols, *T<M*) such that $\lambda_{\aleph'}^{(n)} \xrightarrow{n\to\infty} 0$. This is sufficient if $\lambda_{\aleph'}^{(n)} \le K \cdot P_{e\aleph}^{(n)}$. We do the following trick. We generate a code of *2M* codewords instead of *M*. Note that the new rate is $R_{2M} = \log_2(2M)/n = R_M + 1/n \xrightarrow{n\to\infty} R_M$. It is easy to show that $2P_{e2M}^{(n)} \ge \lambda_M^{(n)}$. The proof goes as follows:

assume that $\lambda_1 \le \lambda_2 \le ... \le \lambda_{2M}$. If this does not hold you can always swap the codewords in the codebook such that the first row has the smallest probability or error, the second row the second smallest probability of error, etc. Hence:

$$P_{e2M}^{(n)} = \frac{1}{2M}\sum_{i=1}^{2M}\lambda_i = \frac{1}{2M}\left(\sum_{i=1}^{M}\lambda_i + \sum_{i=M+1}^{2M}\lambda_i\right) \ge \frac{1}{2M}\sum_{i=M+1}^{2M}\lambda_i \ge \frac{1}{2M}M\lambda_M^{(n)} \ge \frac{\lambda_M^{(n)}}{2}.$$

Thus, using the code $\aleph$ and showing $P_{e\aleph}^{(n)} \overset{n\to\infty}{\rightarrow} 0$ it is equivalent as if we were using the subcode $\aleph'$ and showing $\lambda_{\aleph'}^{(n)} \overset{n\to\infty}{\rightarrow} 0$. In practice, we can throw away the worst M codewords and left with the rest M codewords (this is also called the expurgated code $\aleph'$ which has rate $R_{2M} - 1/n$).

Let's calculate the average probability of error averaged over all codewords in the codebook and averaged over all codebooks:

$$P\{error\} = \sum_{\aleph} \mathrm{P}\{\aleph\} P_e^{(n)}(\aleph) = \sum_{\aleph} \mathrm{P}\{\aleph\}\frac{1}{2^{nR}}\sum_{w=1}^{2^{nR}}\lambda_w(\aleph) = \frac{1}{2^{nR}}\sum_{w=1}^{2^{nR}}\sum_{\aleph}\mathrm{P}\{\aleph\}\lambda_w(\aleph)$$

by symmetry of the code construction, $P\{error\}$ does not depend on the particular message so:

$$P\{error\} = \sum_{\aleph}\mathrm{P}\{\aleph\}\lambda_1(\aleph) = P\{error/W=1\}$$

So we calculate the average probability of error based on the scenario that the first codeword was transmitted. An error occurs if the following events happen:

1. $E_i = \{(X^n(i), Y^n) \in A_\varepsilon^{(n)}\},\ i = 2,...,2^{nR}$
2. $E_1^c = \{(X^n(1), Y^n) \notin A_\varepsilon^{(n)}\}$

Thus,

$$P\{error/W=1\} = P\{E_1^C \cup E_2 \cup ... \cup E_{2^{nR}}\} \le P\{E_1^C\} + \sum_{i=2}^{2^{nR}} E_i \overset{n\to\infty}{\le} \varepsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X,Y)-3\varepsilon)}$$
$$\le \varepsilon + (2^{nR}-1)2^{-n(I(X,Y)-3\varepsilon)} \le \varepsilon + 2^{-n(I(X,Y)-R-3\varepsilon)} \underset{n\to\infty}{\overset{R<I(X;Y)}{\rightarrow}} 0$$

To finish the proof, we find the capacity achieving input distribution to generate the codebooks so $C = I(X;Y)$. Hence, we can drive the average probability of error as close to zero as desired as long as $R < C$ for sufficient large n.

Converse:

Let W to be uniformly distributed over the set $\{1,2,...,2^{nR}\}$. We have:

$$nR = H(W) = H(W/Y^n) + I(W;Y^n) \overset{W\to X^n(W)\to Y^n}{\le} H(W/Y^n) + I(X^n(W);Y^n)$$

If $\lambda^{(n)} \overset{n\to\infty}{\to} 0$ then $P_e^{(n)} \overset{n\to\infty}{\to} 0$. Since $P_e^{(n)} = P\big(W \neq g(Y^n)\big)$ by Fano's inequality we have:

$$H(W/Y^n) \leq 1 + P_\varepsilon^{(n)} nR$$

Moreover,

$$I(X^n(W);Y^n) \overset{DMC}{\leq} \sum_{i=1}^{n} I(X_i;Y_i) \leq nC$$

thus,

$$nR \leq H(W/Y^n) + I(X^n(W);Y^n) \leq 1 + P_\varepsilon^{(n)} nR + nC \Rightarrow R \leq P_\varepsilon^{(n)} R + \frac{1}{n} + C$$

so as $n \to \infty$ the first two terms go to 0 and we get the desired result: $R \leq C$

4.2. Example

When may we encode above capacity and have zero probability error?

Assume the BSC with ε = 0. Obviously C = 1 bit/channel use. Consider the channel code

1 → 0	p=1/4
2 → 1	p=1/4
3 → 00	p=1/4
4 → 01	p=1/4

then $R = \dfrac{\log_2 4}{2\cdot\frac{1}{2} + 1\cdot\frac{1}{2}} = \dfrac{4}{3} > 1$. Hence, we can have R > C and no errors at the receiver but note that the channel is not noisy anymore so the coding theorem does not hold anymore. Also note the code is not uniquely decodable.