EE/Stats 376A: Information theory **Winter 2017**

Lecture 10 — Feb 9

Lecturer: David Tse *Scribe: Daria R, Shiv K, Yuki N, Nipun A, Vivek B*

10.1 Outline

- Fano's Inequality
- Jointly typical sequences
- Getting to capacity: Sphere packing with random spheres

10.1.1 Readings

- CT: 7.6, 7.7, 13.7

10.2 Recap - Converse to the Coding Theorem

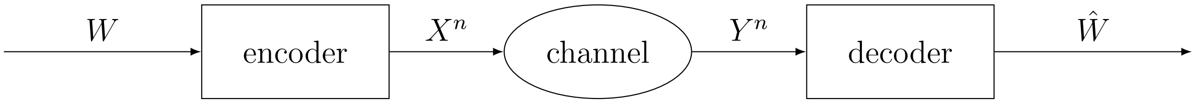


Figure 10.1: Channel Coding Model.

For the channel coding model (Figure 10.1), we proved

$$H(W|\hat{W}) \geq n(R - C). \tag{10.1}$$

Intuitively, nR is the number of bits of uncertainty at the receiver before receiving any signal, nC is the largest possibility number of bits of uncertainty that can be resolved through the channel, and the rest is the amount of leftover uncertainty. Therefore, for $R > C$, the original message W, conditioned on $\hat{W}$, has large uncertainty; which intuitively suggests that p_e must also be large. Now, we rigorously prove this claim using *Fano's Inequality.*

Theorem 1. (*Fano's Inequality*) For an **estimator** $\hat{U}$, and random variables U, V such that $U \to V \to \hat{U}$,

$$P(\hat{U} \neq U)\frac{\geq H(U|\hat{U}) - 1}{\log |\mathcal{U}|}.$$

Proof. Refer to CT. □

Applying Fano's inequality on Markov chain $W \to X^n \to Y^n \to \hat{W}$ (refer Figure 10.1), we obtain

$$\begin{aligned} P(\hat{W} \neq W) &\geq \frac{H(W|\hat{W}) - 1}{\log W} \\ &\geq \frac{n(R-C) - 1}{nR} \\ &= 1 - \frac{C}{R} - \frac{1}{nR} \\ &\approx 1 - \frac{C}{R} \end{aligned}$$

for n large.

This equation shows that if $R > C$, the probability of error is bounded away from 0 for sufficiently large n. Hence, we cannot achieve arbitrary low probability of error for $R > C$. In fact, there are tighter bounds, but even for this bound if $R > C$ the error probability is high.

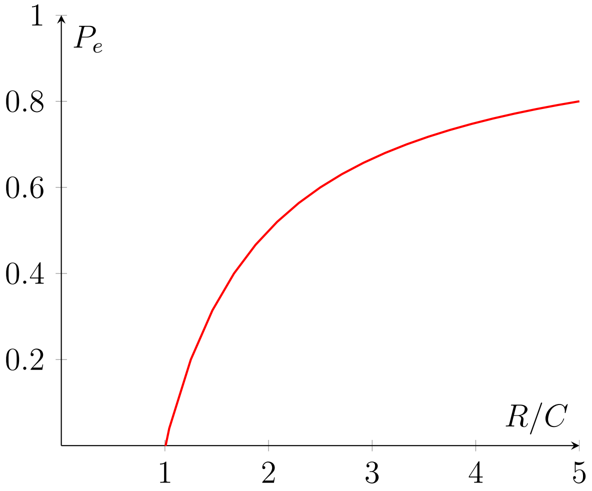


Figure 10.2: Error probability lower bound

10.2.1 Achieving capacity

To get to the capacity C with $P_e \approx 0$, we need:

- the data processing inequality $I(X^n;Y^n) \geq I(W;\hat{W})$ has to hold with equality, which means that the encoding and decoding should be information lossless;
- $I(X^n;Y^n) = nC$, which means that the Y_i's have to be independent

A natural way to make the channel outputs Y_i's independent would be to make the inputs X_i's independent, but redundancy in the code X^n is essential to reduce the effect of the noise, so X_i's are in general dependent. We will revisit this seeming paradox later on when we talk about the optimal communication scheme.

10.2.2 Data processing

In the last lecture, we used data processing theorem to derive the equation (10.1).

Theorem 2. Suppose random variables X, Y, and Z form a Markov Chain (i.e. $X \to Y \to Z$), then

$$I(X;Y) \geq I(X;Z).$$

Proof. Using the chain rule for mutual information, expand $I(X;Y,Z)$ in two following ways:

$$\begin{aligned} I(X;Y,Z) &= I(X;Y) + I(X;Z|Y) \\ I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \end{aligned}$$

Since X, Y, and Z form a Markov chain, $I(X;Z|Y) = 0$. From the non-negativity of mutual information we know that $I(X;Y|Z) \geq 0$. Thus

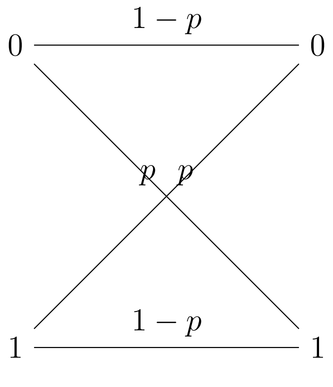
$$\begin{aligned} I(X;Y) &= I(X;Y,Z) \\ &= I(X;Z) + I(X;Y|Z) \\ &\geq I(X;Z) \end{aligned}$$

□

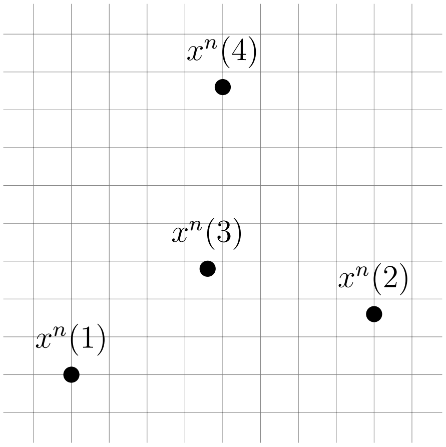
10.3 Sphere Packing

In this section, we will present a geometric view of the coding problem, in terms of sphere packing. Through this view, we will get another perspective of why one cannot communicate reliably at rate greater than the capacity. More importantly, it will lead us to a proof of the positive result, which is that we can reliably communicate at all rates up to the capacity.

Example 1. Consider a BSC with a flip probability p :



Code is a function: $\mathcal{C} : \{1 \ldots 2^{nR}\} \to \{0;1\}^n$, so each of the $M = 2^{nR}$ codewords can be represented as a binary vector $x^n(i), i \in \{1 \ldots M\}$:



The objective of an efficient code is to pack many points in the space while keeping them reasonably far to avoid errors caused by noise.

Typical noise results in $\sim np$ flips in the channel, so it is reasonable to keep points at a distance of $d(x^n(i), x^n(j)) \geq 2np$, which results in a combinatorial optimization problem (**packing problem**): find the largest number of codewords $x^n(i)$. such that every pair is at least $2np$ apart.

This is a hard combinatorial problem, and to this date remains unsolved. Shannon's idea was 1) to replace the hard constraint of $P_e = 0$ with a softer one $P_e \approx 0$ (but $P_e > 0$) so that it is ok that the spheres overlap a little; 2) only look at jointly typical sequences so that everything is approximately uniformly distributed.

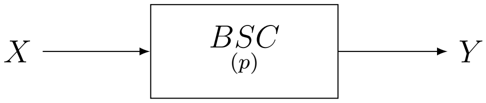
10.4 Jointly typical sequences

For ease of notation, let us denote the sequence $(x_1, x_2, \cdots, x_n)$ by x^n.

Definition 1. (x^n, y^n) is **jointly typical** with respect to an i.i.d. sequence of random variables (X^n, Y^n) if

- $p(x^n) \sim 2^{-nH(X)}$ (x^n is typical)
- $p(y^n) \sim 2^{-nH(Y)}$ (y^n is typical)
- $p(x^n, y^n) \sim 2^{-nH(X,Y)}$

Example 2. Let $(X, Y):\ X \sim \text{Ber}(1/2), Y$ is the output of BSC(p) with X as the input.



Let (x^n, y^n) be jointly typical, then

- $p(x^n) \sim 2^{-nH(X)}$
- $p(y^n) \sim 2^{-nH(Y)}$

- $p(x^n, y^n) \sim 2^{-n(H(p)+1)}$, where $H(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$

Also, the number of typical sequences:

- $\left|\left\{x^n : p(x^n) \sim 2^{-n}\right\}\right| \sim 2^n$.
- $\left|\left\{y^n : p(y^n) \sim 2^{-n}\right\}\right| \sim 2^n$.
- $\left|\left\{x^n, y^n : p(x^n, y^n) \sim 2^{-n(H(p)+1)}\right\}\right| \sim 2^{n(H(p)+1)} \left(< 2^{2n}\right)$.

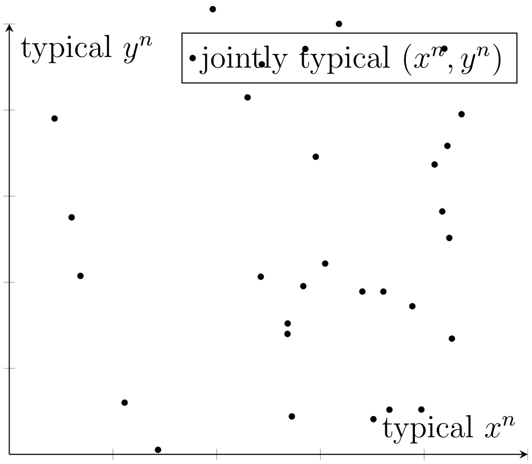


Figure 10.3: Typical sequences (schematic)

So, jointly typical sequences are sparse in the set of pairs of typical sequences.

For jointly typical sequences (x^n, y^n),

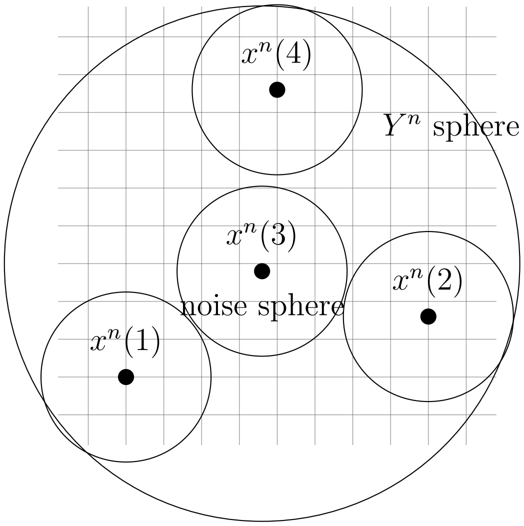
$$\begin{aligned} p(y^n|x^n) &= \frac{p(x^n, y^n)}{p(x^n)} \\ &\sim \frac{2^{-nH(X,Y)}}{2^{-H(X)}} \\ &= 2^{-n\left(H(X,Y)-H(X)\right)} \\ p(y^n|x^n) &\sim 2^{-nH(Y|X)}. \end{aligned} \tag{10.2}$$

In Example **??**, $p(y^n|x^n) = 2^{-nH(p)}$.

10.5 Geometric view of converse

Return to the BSC example. There are $2^{nH(Y|X)} = 2^{nH(p)}$ typical y^n sequences, jointly typical with x^n. Hence each noise sphere has roughly $2^{nH(Y|X)}$ sequences, while all in all there are $2^{nH(Y)}$ typical y^n's, so there can be no more than $\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X,Y)} \leq 2^{n(1-H(p))}$ noise spheres packed in the space if one does not want significant overlap among the noise spheres. Each noise sphere corresponds to a codeword (there are 2^{nR} of them), therefore for reliable communication,

$$2^{nR} \leq 2^{I(X,Y)} \Rightarrow R \leq I(X,Y) \leq C$$


$x^n(4)$
Y^n sphere
$x^n(3)$
$x^n(2)$
noise sphere
$x^n(1)$