ECE 7680
Lecture 7 – Channel Capacity

Objective: To define channel capacity and prove its fundamental properties.

Reading:

1. Read Chapter 6, which has some interesting things in it.
2. Read Chapter 8.

Channel capacity definition and examples

We are now ready to talk about the fundamental concept of the capacity of a channel. This is a measure of how much information *per channel usage* we can get through a channel.

Definition 1 The **information channel capacity** is defined as the maximum mutual information,

$$C = \max_{p(x)} I(X;Y),$$

where the maximum is taken over all possible input distributions $p(x)$. □

Example 1 Consider the noiseless channel, with error-free transmission. For every x we put in, we get out a y with no equivocation: $C = 1$ bit, which occurs when $p(x) = (1/2, 1/2)$ □

Example 2 Noisy channel with non-overlapping outputs. $C = 1$ bit, when $p(x) = (1/2, 1/2)$. □

Example 3 Noisy typewriter, with crossover of 1/2 and 26 input symbols. One way to think about this: we can use every other symbol, and get 13 bits through. Or we can go this way:

$$I(X;Y) = \max[H(Y) - H(Y|X)] = \max H(Y) - 1 = \log 26 - 1 = \log 13.$$

□

Example 4 The BSC, with crossover probability p.

$$\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - \sum_x p(x) H(Y|X = x) \\
&= H(Y) - \sum_x p(x) H(p) \\
&= H(Y) - H(p) \\
&\leq 1 - H(p)
\end{aligned}$$

So

$$\boxed{C = 1 - H(p)}.$$

More insight on the BSC from Lucky, p. 62. Suppose that we have a sequence of information that we send:

$$X^n = 10110100011011100101111001011000$$

and the received sequence

$$Y^n = 101100000011011100111111001111000$$

The error sequence is therefore

$$E^n = 000001000000000000010000000100000$$

If we knew this sequence, we could find X^n exactly. How much information does this sequence represent? We can think of it as being generated by a Bernoulli source with probability p (an error with probability p representing a crossover). Since we don't know this sequence, this deficiency represents the amount by which the information we receive drops: We thus have

$$1 - H(p)$$

bits of useful information left over for every bit of information received.

□

Example 5 Binary erasure channel: Probability $1 - \alpha$ of correct, α of erasure.

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\alpha)$$

To maximize this, we must find max $H(Y)$. There are three possibly Y outcomes, but we cannot achieve $H(Y) = \log 3$ for any input distribution. We have work a bit harder. Let $P(X = 1) = \pi$. Then

$$P(Y = 0) = P(X = 0)(1 - \alpha) = (1 - \pi)(1 - \alpha)$$

$$P(Y = e) = P(X = 0)\alpha + P(X = 1)\alpha = \alpha$$

$$P(Y = 1) = P(X = 1)(1 - \alpha) = \pi(1 - \alpha).$$

Then

$$H(Y) = -[(1-\pi)(1-\alpha)\log(1-\pi)(1-\alpha) + \alpha\log\alpha + \pi(1-\alpha)\log\pi(1-\alpha)] = H(\alpha) + (1-\alpha)H(\pi).$$

Then

$$I(X;Y) = H(\alpha) + (1-\alpha)H(\pi) - H(\alpha) = (1-\alpha)H(\pi)$$

This is maximized when $\pi = .5$, and $C = 1 - \alpha$. □

The channel capacity has the following properties:

1. $C \geq 0$. (why?)
2. $C \leq \log|\mathcal{X}|$. (why?)
3. $C \leq \log|\mathcal{Y}|$.
4. $I(X;Y)$ is continuous function of $p(x)$.
5. $I(X;Y)$ is a concave function of $p(x)$. (It has a maximum)

Symmetric channels

We now consider a specific class of channels for which the entropy is fairly easy to compute, the symmetric channels.

A channel can be characterized by a **transmission matrix** such as

$$p(y|x) = \begin{bmatrix} .3 & .2 & .5 \\ .5 & .3 & .2 \\ .2 & .5 & .3 \end{bmatrix} = P$$

The indexing is x for rows, y for columns: $P_{x,y} = p(y|x)$.

Definition 2 A channel is said to be **symmetric** if the rows of the channel transition matrix $p(y|x)$ re permutations of each other, and the columns are permutations of each other. A channel is said to be **weakly symmetric** if every row of the transition matrix $p(\cdot|x)$ is a permutation of every other row, and all the column sums $\sum_x p(y|x)$ are equal. □

Theorem 1 *For a weakly symmetric channel,*

$$C = \log|\mathcal{Y}| - H(\textit{row of transition matrix}).$$

This is achieved by a (discrete) uniform distribution over the input alphabet.

To see this, let $\mathbf{r}$ denote a row of the transition matrix. Then

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\mathbf{r}) \leq \log|\mathcal{Y}| - H(\mathbf{r}).$$

Equality holds if the output distribution is uniform.

A closer look at capacity

The rationale for the coding theorem: "for large block lengths, every channel looks like the noisy typewriter. **The channel has a subset of inputs that produce essentially disjoint sequences at the output**."

For each typical input n-sequence, there are approximately $2^{nH(Y|X)}$ possible Y sequences, each of them more or less equally likely by the AEP. In order to reliably detect them, we want to ensure that no two X sequences produce the same Y sequence. The total number of possible (typical) Y sequences is $\approx 2^{nH(Y)}$. This has to be divided into sets of size $2^{nH(Y|X)}$, corresponding to the different X sequences.

The total number of disjoint sets is therefore approximately $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$. Hence we can send at most $\approx 2^{nI(X;Y)}$ distinguishable sequences of length n.

Definition 3 A **discrete channel**, denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of an input set $\mathcal{X}$, and output set $\mathcal{Y}$, and a collection of probability mass functions $p(y|x)$, one for each $x \in \mathcal{X}$.

The **extension** of a discrete memoryless channel (DMC) is the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{X})$, where

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k).$$

In other words, the distribution of the channel given a sequence of inputs and prior outputs depends only on the current input (*memoryless* channel). □

When we talk about data transmission through a channel, the issue of coding arises. We define a code as follows:

Definition 4 An (M, n) code for the channel $(\mathcal{X}, p(x|y), \mathcal{Y})$ consists of:

1. An index set $\{1, \ldots, M\}$ (which represents the input alphabet W)
2. A function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$ yielding codewords $X^n(1), X^n(2), \ldots, X^n(M)$.
3. A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

□

In other words, the code takes symbols from W, and encodes them to produce a sequence of n symbols in $\mathcal{X}$.

The probability of error is defined as

$$\lambda_i = \Pr(g(Y^n) \neq i | X^n = X^n(i))$$

In other words, if the message symbol is i, but the output symbol is not i, then we have an error. This can be written using the indicator function $I(\cdot)$ as

$$\lambda_i = \Pr(g(Y^n) \neq i | X^n = X^n(i)) = \sum_{y^n} p(y^n|x^n(i)) I(g(y^n) \neq i)$$

In our development, it may be convenient to deal with the maximal probability of error. If it can be shown that the maximal probability of errors goes to zero, then clearly the other probabilities of error do also. The maximal probability of error is

$$\lambda^{(n)} = \max_i \lambda_i$$

The average probability of error is

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

Definition 5 The **rate** of an (M, n) code is

$$R = \frac{\log M}{n} \text{ bits per transmission.}$$

□

Example 6 Suppose $M = 4$, and codewords are $n = 4$ bits long. Then every input symbol supplies 2 bits of information, and takes 4 symbol transmissions to send. The rate is $R = 1/2$. □

Definition 6 A rate R is said to be *achievable* if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the maximal probability of error $\lambda^{(n)}$ tends to 0 as $n \to \infty$. □

Definition 7 The **capacity** of a DMC is the supremum of all achievable rates. □

This is a different definition than the "information" channel capacity of a DMC, presented above. What we will show (this is Shannon's theorem) is that the two definitions are equivalent.

The implication of this definition of capacity is that for an achievable rate, the probability of error tends to zero as the block length gets large. Since the capacity is the largest achievable rates, then **for rates less than the capacity, the probability of error goes to zero with the block length.**

Jointly typical sequences

Recall the definition of a typical sequence: it is the sequence we expect (probabilistically) to occur, given the statistics of the source. We had a theorem regarding the approximate number of typical sequences, the probability of a typical sequence, and so forth. In this section we generalize this.

All along this semester we have used the notion of representing sequences of random variables as vectors. For example, the r.v.'s (X_1, X_2) we could represent with a vector-valued random variable X. In a sense, this is all that we are doing with the jointly-typical sequences. We consider sequences (x^n, y^n) as if there were simply some sequence z^n, and ask: for the set of sequences z^n, what are the typical sequences:

Definition 8 The set $A_\epsilon^{(n)}$ of **jointly typical** sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of sequences with empirical (statistical averages) entropies close to the true entropy:

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

□

From this definition, we can conclude similar sorts of things about jointly typical sequences as we can about typical sequences:

Theorem 2 *(Joint AEP).*

1. *$Pr((X^n, Y^n) \in A_\epsilon^{(n)} \to 1$ as $n \to \infty$. (As the block length gets large, the probability that we observe a jointly typical sequence approaches 1.)*

2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$. *(The number of jointly typical sequences is close to the number of representations determined by the entropy of (X,Y).)*

3. *If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, that is, $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same marginals as $p(x^n, y^n)$ then*

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

For sufficiently large n,

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

Proof By WLLN,

$$-\frac{1}{n}\log p(X^n) \to -E\log p(X) = H(X)$$

so that, by the definition of convergence in probability, for any $\epsilon > 0$ there is an n_1 such that for $n > n_1$,

$$\Pr\left(|-\frac{1}{n}\log p(X^n) - H(X)| > \epsilon\right) < \frac{\epsilon}{3}.$$

Similar statements can be said about WLLN convergence in the Y data and the Z data: There is an n_2 and and n_3 such that for $n > n_2$

$$\Pr\left(|-\frac{1}{n}\log p(Y^n) - H(Y)| > \epsilon\right) < \frac{\epsilon}{3}.$$

and for $n > n_3$

$$\Pr\left(|-\frac{1}{n}\log p(X^n, Y^n) - H(X,Y)| > \epsilon\right) < \frac{\epsilon}{3}.$$

So for $n > \max(n_1, n_2, n_3)$, the probability of the union of the three sets must be $< \epsilon$. Hence the probability of the set $A_\epsilon^{(n)}$ must be $> 1 - \epsilon$.

The second part of the proof is proved exactly as for typical sequences:

$$\begin{aligned} 1 = \sum p(x^n, y^n) &\geq \sum_{A_\epsilon^{(n)}} p(x^n, y^n) \\ &\geq |A_\epsilon^{(n)}| e^{-n(H(X,Y)+\epsilon)} \end{aligned}$$

so that

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

For the third part, we note first that for n sufficiently large,

$$\Pr(A_\epsilon^{(n)}) \geq 1 - \epsilon$$

so we can write

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

from which

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

Now if $\tilde{X}$ and $\tilde{Y}$ are independent with the same marginals as X and Y,

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

(since there are approximately $2^{n(H(X,Y)+\epsilon)}$ terms in the sum). Similar arguments can be used to derive the last inequality. □

There are about $2^{nH(X)}$ typical X sequences, about $2^{nH(Y)}$ typical Y sequences, and only about $2^{nH(X,Y)}$ jointly typical sequences. This means that if we choose a typical X sequence and independently choose a typical Y sequence (without regard to the X sequence), in not all cases will the sequence (X^n, Y^n) so chosen be jointly typical. In fact, from the last part of the theorem, the probability that the sequences chosen independently will be jointly typical is about $2^{-nI(X;Y)}$. This means that we would have to try (at random) about $2^{nI(X;Y)}$ sequence pairs before we choose a jointly typical pair. Thinking now in terms of fixing Y and choosing X at random, this suggests that there are about $2^{nI(X;Y)}$ distinguishable sequences in X.

The channel coding theorem and its proof

We are now ready for the big theorem of the semester.

Theorem 3 *(The channel coding theorem) All rates below capacity C are achievable. That is, for every $\epsilon > 0$ and 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 $R \leq C$.

Note that by the statement of the theorem, we only state that "there exists" a sequence of codes. The proof of the theorem is *not* constructive, and hence does not tell us how to find the codes.

Also note that the theorem is asymptotic. To obtain an arbitrarily small probability of error, the block length may have to be infinitely long.

Being the big proof of the semester, it will take some time. As we grind through, don't forget to admire the genius of the guy who was able to conceive of such an intellectual *tour de force.* Aren't you glad this wasn't your homework assignment to prove this!

Proof We prove that rates $R < C$ are achievable. The converse will wait.

Fix $p(x)$. The first important step is to generate a $(2^{nR}, n)$ code **at random** using the distribution $p(x)$. (The idea of using a random code was one of Shannon's key contributions.) Specifically, we draw 2^{nR} codewords according to

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

Now we will list these codewords as the rows of a matrix

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

Each element in the matrix is independently generated according to $p(x)$. The probability that we generate any **particular** code is

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

The communication process is accomplished as follows:

1. A random code $\mathcal{C}$ is generated as described. The code is made available (prior to transmission) to both the sender and the receiver. Also, both sender and receiver are assumed to know the channel transition matrix $p(y|x)$.
2. A message W is chosen by the sender according to a uniform distribution,

   $$\Pr(W = w) = 2^{-nR}, \qquad w = 1, 2, \ldots, 2^{nR}.$$

   and the wth codeword $X^n(w)$ (corresponding to the wth row of the $\mathcal{C}$ matrix) is sent over the channel.
3. The receiver receives the sequence Y^n according to the distribution

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

   (the fact that the probabilities are multiplied together is a consequence of the memoryless nature of the channel).
4. Based on the received sequence, the receiver tries to determine which message was sent. In the interest of making the decision procedure tractable for this proof, a *typical set decoding* procedure is used. The receiver decides that message $\hat{W}$ was sent if
   - $(X^n(\hat{W}), Y^n)$ is jointly typical: in other words, we could expect this to happen, and
   - There is no other k such that $(X^n(k), Y^n) \in A_\epsilon^{(n)}$. If this happens, the receiver cannot tell which symbol should be decoded, and declares an error.
5. If $W \neq \hat{W}$, then there is a decoding error.

The next key step is we find the average probability of error, *averaged over all codewords in the codebook* **and** *averaged over all codebooks* $\mathcal{C}$. This is a key idea, because of the following observation: If the probability of error obtained by averaging over random codebooks can be shown to be made arbitrarily small (as we demonstrate below), then that means that there is *at least one* fixed codebook that has a small average probability of error. So the random codebook idea and the average over random codebooks is simply a mechanism to prove that there does exist a good code.

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

By the symmetry of the code construction, the average probability of error averaged over all codes does not depend on the particular index that was sent, so we can assume without loss of generality that $w = 1$.

Define the event

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

That is, E_i is the event that Y^n is jointly typical with the ith codeword. An error occurs if

- E_1^c occurs: Y^n is not jointly typical with the first codeword, or
- $E_2 \cup E_3 \cup \cdots \cup E_2^{2^{nR}}$ occurs (a wrong codeword is jointly typical with the received sequence).

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

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

(the union bound).

Finally we can use the typicality properties. By AEP,

$$P(E_1^c) \leq \epsilon$$

for n sufficiently large. And the probability that Y^b and $X^n(i)$ are jointly typical (for $i \neq 1$)is $\leq 2^{-nI(X;Y)-3\epsilon)}$. Hence

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

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.

We are not quite there yet, in light of all the averaging we had to do. Here are the final touch-ups:

1. In the proof, choose $p(x)$ to be that distribution that achieves capacity. Then $R < I(X;Y)$ can be replaced by the achievability conditions $R < C$.
2. Since the average probability of error, averaged over all codebooks is small, then there exists at least one codebook $\mathcal{C}^*$ with a small average probability of error. $P_e^n(\mathcal{C}^*) \leq e\epsilon$. (How to find it?!) (Is exhaustive search an option?)
3. Observe that the best half the codewords must have a probability of error less than 4ϵ (otherwise we would not be able to achieve the average less than 2ϵ. We can throw away the worst half the codewords, giving us a rate

$$R' = R - \frac{1}{n}$$

which asymptotically is negligible.

Hence we can conclude that for the best codebook, the **maximal** probability (not the average) has a probability of error $\leq 4\epsilon$. □

Whew!

The proof used a random-coding argument to make the math tractable. In practice, random codes are **never** used. Good, useful, codes are a matter of considerable interest and practicality.