EE/Stats 376A: Information theory	Winter 2017

Lecture 9 — Feb 7

Lecturer: David Tse | *Scribe: Rachel L, Jesse Z, Kevin C, Xin Z, Paris F*

9.1 Outline

- Properties of capacity
- Converse to the noisy channel coding theorem

9.1.1 Reading

- CT 2.8, 2.10, 7.8, 7.9

9.2 Recap

Last week, we looked at two important channel metrics: the probability of error p_e and the rate of transmission R. Typical tradeoff curves for repetition codes vs. optimal codes are shown in Figure 9.1 below.

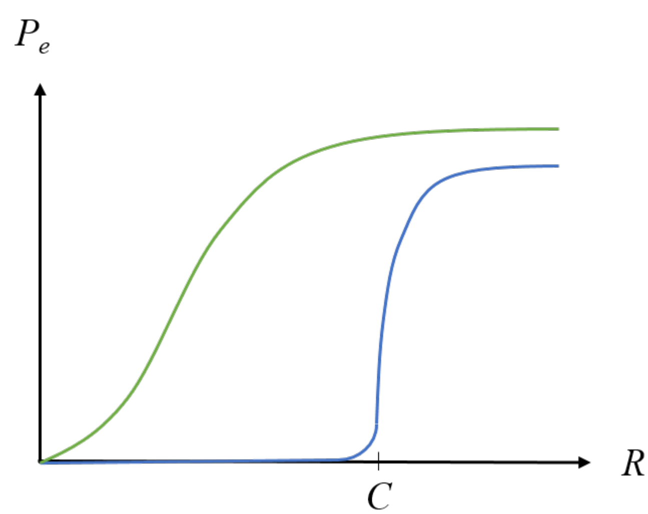

Figure 9.1: The tradeoff curve for repetition codes (green) and for optimal codes (blue).

This optimal tradeoff curve means that below the capacity C, we can get an arbitrarily small p_e, but above C, any code is bad. We defined the capacity as

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

and will justify the tradeoff picture in the next few lectures.

9.3 Properties of capacity

Recall that the binary symmetric channel (BSC), shown in Figure 9.2 below, is a channel model in which a transmitted bit is flipped with some crossover probability. We can keep this channel in mind for the following discussion - although our discussion will be general, it may be helpful to refer to something concrete.

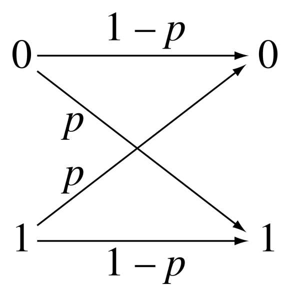

Figure 9.2: The binary symmetric channel (BSC) with crossover probability p.

We will now discuss two important properties of capacity:

1. $C \leq \log|\mathcal{X}|$, $C \leq \log|\mathcal{Y}|$
2. $C = \max_{p(x)} I(X;Y)$ is a convex optimization problem.

9.3.1 Upper bound on capacity

Let us take a closer look at the first property. Why is this true? Well, we can expand $I(X;Y)$ as follows:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &\leq H(X) \\ &\leq \log|\mathcal{X}| \end{aligned}$$

So $C \leq \log|\mathcal{X}|$. Similarly, we can expand $I(X;Y)$ as $H(Y) - H(Y|X)$ to show that $C \leq \log|\mathcal{Y}|$.

9.3.2 Capacity as a convex optimization problem

Now let us take a closer look at the second property. In order to prove that finding C is a convex optimization problem, we will first show the following fact:

Fact. $I(X;Y)$ is a concave function of $p(x)$.

Proof. Note that $I(X;Y)$ depends only on $p(y|x)$ and $p(x)$, and that $p(y|x)$ is fixed by the channel. Then, writing $I(X;Y) = H(Y) - H(Y|X)$ and expanding the second term, we get

$$\begin{aligned} H(Y|X) &= \sum_x p(x) H(Y|X = x) \\ &= \sum_x p(x) f(x) \end{aligned}$$

where we can express $H(Y|X = x)$ as $f(x)$ because it depends only on $p(y|x)$, which is fixed by the channel, as mentioned above. This last expression is a dot product between the vectors $p(x)$ and $f(x)$, so $H(Y|X)$ is a linear function of $p(x)$. A linear function of the input is always convex, so the second term of $I(X;Y)$ is concave!

What about the first term? Observe that $H(Y)$ is a concave function of $p(y)$, and $p(y)$ can be written as

$$p(y) = \sum_x p(y|x) p(x)$$

As before, $p(y|x)$ is fixed, so we can put the terms into transition matrix P and rewrite $p(y)$ as

$$p(y) = P \begin{bmatrix} p(1) \\ p(2) \\ \vdots \\ p(|\mathcal{X}|) \end{bmatrix}$$

Sine $H(Y)$ is a concave function of $p(y)$ and $p(y)$ is a linear transformation of $p(x)$, so $H(Y)$ is also a concave funcction of p(x).
$\implies I(X;Y) = H(Y) - H(Y|X)$ is concave. □

Now, thinking back to the BSC, note that by symmetry, an input distribution with probabilities p and $1-p$ yield the same mutual information as an input distribution with probabilities $1-p$ and p. If we take the average of the two distributions, we will achieve an even higher mutual information by concavity. Thus, if a channel has symmetry, the optimal distribution p^* must be uniform. Next we will generalize the symmetry of the BSC

Definition 1 (Symmetry)**.** . A channel is **symmetric** if for every permutation of the columns of transition matrix P, there exists a permutation of the rows that keeps P the same.

e.g. For the BSC, P is

$$P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

Let's take a look at another example - the binary erasure channel (BEC), which is shown in Figure 9.3.

This channel is also symmetric, so we expect that $p^*(0) = p^*(1) = 1/2$. Expanding C, we have

$$C = H(Y) - H(Y|X)$$

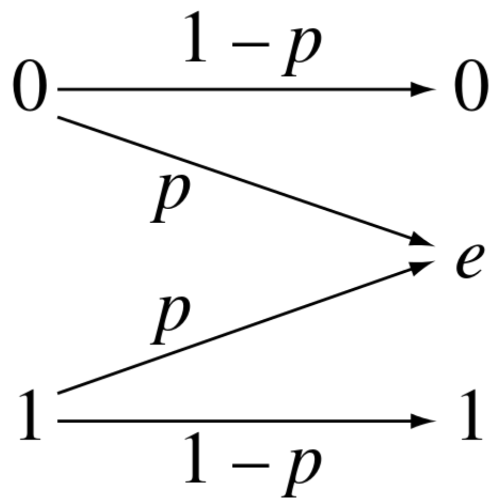

Figure 9.3: The binary erasure channel (BEC) with erasure probability p.

under the p^* distribution. We know that $H(Y|X) = H(p)$, and we can calculate $H(Y)$ as follows:

$$\begin{aligned}
H(Y) &= 2 * \frac{1}{2}(1-p)\log\frac{1}{\frac{1}{2}(1-p)} + p\log\frac{1}{p} \\
&= (1-p)\log\frac{1}{1-p} + p\log\frac{1}{p} + (1-p) \\
&= H(p) + (1-p)
\end{aligned}$$

Plugging these two terms back into the expression for C, we get

$$\begin{aligned}
C &= H(Y) - H(Y|X) \\
&= H(p) + (1-p) - H(p) \\
&= 1-p
\end{aligned}$$

9.4 Converse to the noisy channel coding theorem

Recall that the optimal tradeoff curve for p_e vs. R looks as shown in blue in Figure 9.1. The converse to the noisy channel coding theorem states that if $R > C$, then p_e will be bad for any code. (This is the counterpart to the fact that below C, we can get an arbitrarily good p_e.) To prove this, we will establish a lower bound for p_e. Our system is shown in Figure 9.4. As in Cover and Thomas, we use $W \in \{1, \ldots, 2^{nR}\}$ to represent the message. Recall that we assume W is uniformly distributed in its range.

$p_e = P(\hat{W} \neq W)$, and we want to show that if $R > C$, then p_e will be relatively high. We know that

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

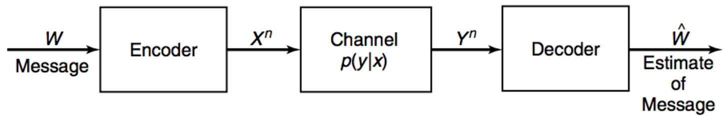

Figure 9.4: Full channel model.

and letting $X^n = (X_1, \ \dots \ X_n)$ and $Y^n = (Y_1 \ \dots \ Y_n)$, we can write

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X^n) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i; Y_i) \\ &\leq nC \end{aligned}$$

Claim: $I(W; \hat{W}) \leq I(X^n; Y^n)$

We can prove this claim by using the data-processing theorem.

Theorem 1 (Data-Processing Theorem)**.** If $U - V - Z$ forms a Markov chain, then $I(U; V) \geq I(U; Z)$.

We will prove this theorem next lecture.

Going back to our claim, we can see that $W - X^n - Y^n - \hat{W}$ forms a Markov chain, so the claim is true by the data-processing theorem. This implies that

$$I(W; \hat{W}) \leq I(X^n; Y^n) \leq nC$$

Now if we expand $I(W; \hat{W})$, we can write

$$\begin{aligned} I(W; \hat{W}) &= H(W) - H(W|\hat{W}) \\ &= nR - H(W|\hat{W}) \\ \implies H(W|\hat{W}) &\geq n(R - C) \end{aligned}$$

So if $R > C$, $H(W|\hat{W})$ is very large. Intuitively, the error probability will be large too, since there is so much uncertainty in W even given $\hat{W}$. We will make this precise in the next lecture.