10-704: Information Processing and Learning **Spring 2012**

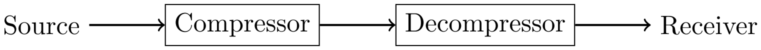
Lecture 15: Channel Capacity, Rate of Channel Code

Lecturer: Aarti Singh *Scribes: Martin Azizyan*

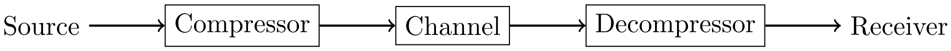
Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

15.1 The Information Theory Diagram

Until now we have studied the data compression (source coding) problem, schematically described by the following diagram:

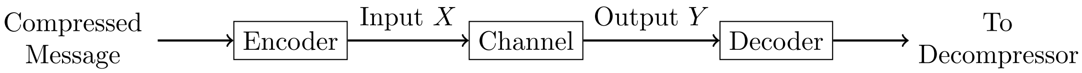


We now turn to the channel coding problem, where we introduce an intermediate step between the output of the compressor and the input to the decompressor:



A "channel" is any source of noise, for example a wireless link, an error prone storage device, an IP network with the potential for packet loss, etc.

In the presence of such noise we need to introduce some redundancy in the signal to ensure successful transmission, so we add two new steps in the diagram:



Essentially, a channel is a conditional distribution $p(Y|X)$. We will focus on **discrete memoryless channels** (DMC):

Discrete: Input and output alphabets $\mathcal{X}, \mathcal{Y}$ are finite; and

Memoryless: $p(Y|X)$ is independent of outputs in previous timesteps (later we will discuss more general feedback channels).

Definition 15.1 *The* ***information capacity*** *of a (DMC) channel is*

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

where the maximum is over all distributions on X.

Informally, the **operational capacity** of a channel is the highest rate in terms of bits/channel use (e.g. a single transmission over a wireless medium) at which information can be sent with arbitrarily low probability of error.

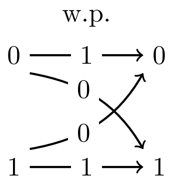
The separation of the source coding and channel coding problems (i.e. the compressor and encoder) is for convenience of analysis only. We will see later on that we don't pay any cost for this division.

15.2 Examples of Information Capacity

We now calculate the information capacity for some simple channel models.

Example: Noiseless binary channel

Consider the following transition model for the channel, where the arrows are annotated with values of $p(y|x)$:



Though we have not yet defined operational capacity exactly, intuitively we would expect that the most information that can be sent with a single use of this channel is exactly 1 bit.

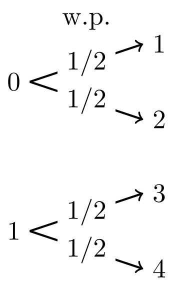
To calculate the information capacity, note that for any distribution over the inputs $p(x)$ we have

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= H(Y) && \text{because } Y \text{ is not random given } X \\ &= H(X) && \text{because distributions of } Y \text{ and } X \text{ are the same.} \end{aligned}$$

Since X is just a Bernoulli random variable, we know that its entropy is maximized when $p(0) = p(1) = 1/2$, and

$$\max_{p(x)} I(X,Y) = 1.$$

Example: Noisy channel with non-overlapping output distributions



As for the noiseless binary channel, we would expect the operational capacity of this channel to be exactly 1 bit/channel use.

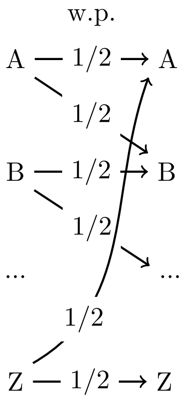
Also for the information capacity,

$$\begin{aligned}\max_{p(x)} I(X,Y) &= \max_{p(x)} H(X) - H(X|Y) \\ &= \max_{p(x)} H(X) \qquad\qquad \text{because knowing } Y \text{ fully determines } X \\ &= 1\end{aligned}$$

as before.

Example: Noisy typewriter

In this model, each of the 26 characters of the alphabet are either transmitted exactly with probability 0.5, or replaced by the next character in the alphabet with probability 0.5:



If all characters were transmitted exactly, we would expect a single transmission to carry $\log_2 26$ bits, which clearly can't be achieved for this channel. But if we only transmit every other letter (i.e. A, C, E...), then if we receive A or B we will know that A was transmitted, if we receive C or D we will know that C was transmitted, etc. Hence we can use the channel in such a way that it is essentially equivalent to the noisy channel with non-overlapping output distributions, and we might expect to transmit $\log_2 13$ bits/channel use.

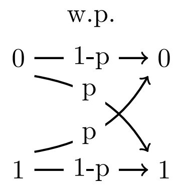
To calculate the information capacity, first note that for any input distribution $p(x)$, the marginal entropy of Y is at most $H(Y) \leq \log_2 26$, which can be achieved with equality if $p(x)$ is uniform. Also, the conditional entropy of Y given X is

$$\begin{aligned}H(Y|X) &= \sum_x p(x) H(Y|X=x) \\ &= \sum_x p(x) H(\text{Bernoulli}(1/2)) \\ &= \sum_x p(x) \\ &= 1.\end{aligned}$$

So

$$\begin{aligned}\max_{p(x)} I(X,Y) &= \max_{p(x)} H(Y) - H(Y|X) \\ &= \max_{p(x)} H(Y) - 1 \\ &= \log_2 26 - 1 \\ &= \log_2 13.\end{aligned}$$

Example: Binary symmetric channel (BSC)



We have

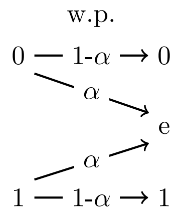
$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(\text{Bernoulli}(p)) \\ &\leq 1 - H(\text{Bernoulli}(p)) \end{aligned}$$

which is achieved with equality for $p(x)$ uniform, so

$$\max_{p(x)} I(X,Y) = 1 - H(\text{Bernoulli}(p)).$$

Note that in this case, we can interpret $H(\text{Bernoulli}(p))$ as the information lost per one bit of transmission.

Example: Binary erasure channel



We again begin by writing

$$I(X,Y) = H(X) - H(X|Y).$$

Note that if we observe $Y = 0$ or $Y = 1$, then X is known exactly, so $H(X|Y \neq e) = 0$. Furthermore $H(X|Y = e) = H(X)$, so

$$\begin{aligned} H(X|Y) &= 0 \cdot P(Y \neq e) + H(X) \cdot P(Y = e) \\ &= \alpha H(X) \end{aligned}$$

and

$$\begin{aligned} I(X,Y) &= (1-\alpha)H(X) \\ &\leq 1 - \alpha \end{aligned}$$

which is achieved with equality for $p(x)$ uniform, so

$$\max_{p(x)} I(X,Y) = 1 - \alpha.$$

Again, the interpretation is obvious. We lose α fraction of bits per transmission.

15.2.1 Weakly Symmetric Channels

Notice that in many of the above examples the maximizing distribution on X in the information capacity is uniform. Here we describe a class of channels that have this property.

Note that a DMC channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ can be described in terms of a transition probability matrix $P \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ where $P_{ij} = p(y_j|x_i)$. If both the rows and the columns of P are permutations of each other, then the channel is called **symmetric**. For such a channel, we have that $H(Y|X = x) = c$ is constant for all $x \in \mathcal{X}$ and so $H(Y|X) = c$ as well, and so for any $x \in \mathcal{X}$

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= H(Y) - c \\ &\le \log_2 |\mathcal{Y}| - c \end{aligned}$$

The inequality is achieved exactly when the distribution of Y is uniform. For a symmetric channel, Y is uniform whenever X is uniform, since then

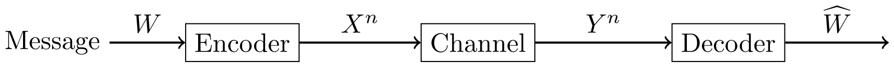
$$\begin{aligned} p(y) &= \sum_x p(y|x)p(x) \\ &= \frac{1}{|\mathcal{X}|}\sum_x p(y|x) \end{aligned}$$

which, by assumption, is constant for all y since the columns of P are permutations of each other.

Note that the above argument holds if we relax the assumption on the columns of P to summing to the same quantity (we only used the permutation property of the columns in the last step to argue precisely that their sum is constant). Channels of this type are called **weakly symmetric**.

15.3 Rate of a Channel Code

Consider the problem of transmitting sequences of length n over the channel.



Definition 15.2 *A* (M, n) ***code*** *for a channel* $(\mathcal{X}, \mathcal{Y}, p(Y|X))$ *consists of*

1. *an index set* $\{1, 2, ..., M\}$*;*
2. *an encoding function* $X^n : \{1, 2, ..., M\} \to \mathcal{X}^n$*;*
3. *and a decoding function* $g : \mathcal{Y}^n \to \{1, 2, ..., M\}$*.*

Definition 15.3 *The* ***rate*** *of an* (M, n) *code is*

$$R = \frac{\log_2 M}{n}$$

and is measured in terms of bits/transmission (i.e. channel use).

Definition 15.4 *A rate R is* ***achievable*** *if there exists a sequence of $(2^{nR}, n)$ codes such that the maximum probability of error in transmission for any message of length n converges to 0 as $n \to \infty$.*

We can now give a precise definition for operational capacity.

Definition 15.5 *The* ***operational capacity*** *of a channel is the supremum of all achievable rates on that channel.*

We will prove next time that the operational capacity of the channel is exactly equal to the information capacity $\max_{p(x)} I(X,Y)$. This is the Shannon Noisy Channel Coding theorem.