Electrical Engineering 229A Lecture 16 Notes

Daniel Raban

October 19, 2021

1 Discrete Memoryless Channels and Shannon's Channel Coding Theorem

1.1 Discrete memoryless channels

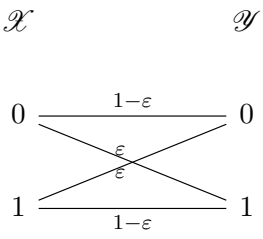
Shannon's **discrete memoryless channel** model of communication has 3 parts:

1. a finite set $\mathscr{X}$, called the **input alphabet**,
2. a finite set $\mathscr{Y}$, called the **output alphabet**,
3. a **channel matrix** of transition probabilities $[p(y \mid x)]_{x \in \mathscr{X}, y \in \mathscr{Y}}$ with $p(y \mid x) \geq 0$ and $\sum_y p(y \mid x) = 1$ for all x.

Using the channel n times with inputs $x_1, x_2, \ldots, x_n$ results in outputs $y_1, y_2, \ldots, y_n$ with

$$p(y_1^n \mid y_1^n) = \prod_{i=1}^n p(y_i \mid x_i).$$

Example 1.1. The binary symmetric channel with crossover probability ε has $\mathscr{X} = \mathscr{Y} = \{0, 1\}$ and $p(0 \mid 1) = \varepsilon = p(1 \mid 0)$.



Here is the physical background: Fix time $T > 0$ (a real number) called the **symbol interval**. Suppose $(g_1(t), t \in [0, T]), (g_2(t), t \in [0, T])$ are orthonormal functions:

$$\int_0^T g_1^2(t)\, dt = 1, \qquad \int_0^T g_2^2(t)\, dt = 1, \qquad \int_0^T g_1(T) g_2(t)\, dt = 0.$$

Example 1.2. For example, we could take

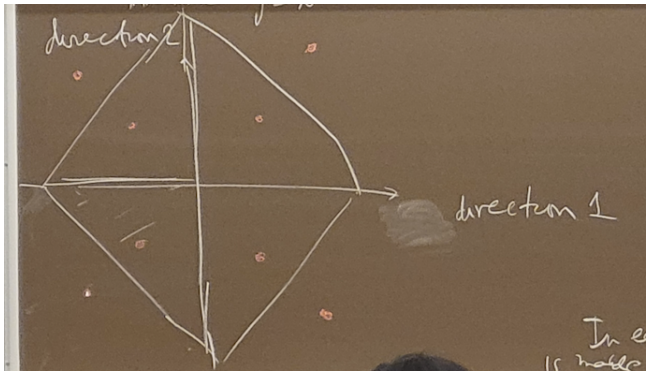
$$g_1(t) = \sqrt{\frac{2}{T}} \sin\left(\frac{2\pi t}{T}\right), \qquad g_2(t) = \sqrt{\frac{2}{T}} \cos\left(\frac{2\pi t}{T}.\right)$$

If we let $d = |\mathscr{X}|$ be the size of the input alphabet, then we have

$$\left\{ \begin{bmatrix} u_{1,1} \\ u_{1,2} \end{bmatrix}, \ldots, \begin{bmatrix} u_{d,1} \\ u_{d,2} \end{bmatrix} \right\},$$

where $g_i(t) = u_{i,1} g_1(t) + u_{i,2} g_2(t)$.

Now assume $\mathscr{Y} = \mathscr{X}$. The picture looks like this, called a **constellation**:



To send the sequence $i_1, \ldots, i_n$, what is physically sent is $\sum_{\ell=1}^{n} g_{i_\ell}(t - (\ell - 1)T)$. This is received in noise. In each interval, a decision is made as to what symbol was sent. For example, if the received output was in the bottom left triangle, we would make the decision that the dot in the center was the symbol sent.

1.2 Channel capacity and Shannon's channel coding theorem

Simple intuition suggests that in n uses of a DMC, we can hope to distinguish between a number of messages that is exponential in n. This motivates the Shannon formulation of "channel capacity."

Definition 1.1. Let

$$e_n : [M_n] \to \mathscr{X}^n, \qquad d_n : \mathscr{Y}^n \to [M_n].$$

We say that **communication is possible at rate** R if there is a sequence $((e_n, d_n), n \geq 1)$ such that $\mathbb{P}(d_n(e_n(W_n)) \neq W_n) \to 0$ as $n \to \infty$, where $W_n \sim \text{Unif}([M_n])$, and such that

$$\liminf_n \frac{1}{n} \log M_n \geq R.$$

Definition 1.2. Channel capacity is the supremum over rates at which communication is possible.

Theorem 1.1 (Shannon's channel coding theorem for a DMC)**.** *The channel capacity equals*

$$\sup_{(p(x),x\in\mathscr{X})} I(X;Y) = \sup_{(p(x),x\in\mathscr{X})} \sum_{x,y} p(x)p(y \mid x) \log \frac{p(y \mid x)}{\sum_{x'\in\mathscr{X}} p(y \mid x')p(x')}.$$

Remark 1.1. This is the maximum of a concave function. Often the maximizer is in the interior of the probability simplex.

Recall that $I(X;Y) = H(Y) - H(Y \mid X)$. Here, $H(Y \mid X) = \sum_x p(x)H(Y \mid X = x)$ is linear in $(p(x), x \in \mathscr{X})$ and $H(Y)$ is concave in $(p(y), y \in \mathscr{Y})$ and hence in $(p(y), x \in \mathscr{X})$.

Example 1.3 (Binary symmetric channel)**.** Suppose $p_X(1) = a = 1 - p_X(0)$. Then

$$\begin{aligned} I(X;Y) &= H(Y) - \underbrace{H(Y \mid X)}_{(1-a)H(Y|X=0)+aH(Y|X=1)} \\ &= h(a(1-\varepsilon) + (1-a)\varepsilon) - h(\varepsilon), \end{aligned}$$

to be optimized over a. This is maximized at $a = 1/2$. So the channel capacity is $1 - h(\varepsilon)$.

To get a feeling for why the theorem might be true, consider inputs to the channel $X_1, \ldots, X_n$ which are iid with $\mathbb{P}(X_1 = x) = p(x)$ for $x \in \mathscr{X}$. Then the outputs will be iid with marginals $(p(y), y \in \mathscr{Y})$, where $p(y) = \sum_x p(x)p(y \mid x)$. The inputs and outputs will be ε-jointly weakly typical with probability going to 1 as $n \to \infty$. The number of ε-weakly typical output sequences is $\geq (1-\varepsilon)2^{nH(Y|X)}2^{-n\varepsilon}$. The number of jointly ε-weakly typical output sequences with a specific ε-weakly typical input sequence is $\leq 2^{nH(Y|X)}2^{2n\varepsilon}$. Then, using

$$1 = \sum_{y_1^n} p(y_1^n \mid x_1^n) = \sum_{x_1^n, y_1^n} \frac{p(x_1^n, y_1^n)}{p(x_1^n)},$$

we get

$$\frac{(1-\varepsilon)2^{nH(Y)}2^{-n\varepsilon}}{2^{nH(Y|X)}2^{2n\varepsilon}} = (1-\varepsilon)2^{nI(X;Y)}2^{-3n\varepsilon}.$$

1.3 Proof of Shannon's channel coding theorem

Proof. For achievability, we need to show that for all rates $R < \max_{p(x),x\in\mathscr{X}} I(X;Y)$, we want to show that R is achievable. For the converse, we need to show that no $R > \max_{p(x),x\in\mathscr{X}} I(X;Y)$.

The achievability is given by a random coding argument.[1] We will take $M_n = \lceil 2^{nR} \rceil$, create random $e_n : [M_n] \to \mathscr{X}^n$ for each $n \geq 1$ and associated d_n and show that the error probability $\to 0$ as $n \to \infty$. Let

$$e_n(m) = (X_1(m), \ldots, X_n(m)), \qquad 1 \leq m \leq \lceil 2^{nR} \rceil = M_n$$

where $X_t(m) \sim (p(x), x \in \mathscr{X})$ is iid over $1 \leq t \leq n$ and $1 \leq m \leq M_n$. To define d_n, on receiving $y_1, \ldots, y_n$, find

$$\{1 \leq m \leq M_n \text{ s.t. } (x_1(m), \ldots, x_n(m)) \text{ is } \varepsilon\text{-jointly weakly typical with } (y_1, \ldots, y_n)\}.$$

If this set has exactly one member, return that member as $d_n(y_1, \ldots, y_n)$; otherwise, define $d_n(y_1, \ldots, y_n)$ arbitrarily.

Let $\mathcal{C}$ denote the (random) **codebook**

$$\begin{bmatrix} X)1(1) & \cdots & X_n(1) \\ \vdots & & \vdots \\ X_1(M_n) & \cdots & X_n(m) \end{bmatrix},$$

and let $P_e(\mathcal{C}) = \mathbb{P}(d_n(e_n(W_n) \neq W_n \mid \mathcal{C})$ denote the error probability conditioned on the codebook being $\mathcal{C}$. Then the expected error probability over the codebook is $\mathbb{P}(d_n(e_n(W_n)) \neq W_n)$. We have

$$\begin{aligned} \mathbb{P}(d_n(e_n(W_n)) \neq W_n) &= \sum_{\mathcal{C}} \mathbb{P}(\mathcal{C}) \mathbb{P}(d_n(e_n(W_n)) \neq W_n \mid \mathcal{C}) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) \sum_{m=1}^{M_n} \frac{1}{M_n} \lambda_m(\mathcal{C}), \end{aligned}$$

where $\lambda_m(c) = \mathbb{P}(d_n(e_n(m)) \neq m \mid \mathcal{C})$.

$$\begin{aligned} &= \sum_{m=1}^{M_n} \frac{1}{M_n} \sum_{\mathcal{C}} \mathbb{P}(\mathcal{C}) \lambda_m(\mathcal{C}) \\ &= \sum_{\mathcal{C}} \mathbb{P}(\mathcal{C}) \lambda_1(\mathcal{C}) \end{aligned}$$

by symmetry.

$$\begin{aligned} &\leq \mathbb{P}\left(E_0 \cup \left(\bigcup_{m=2}^{M_n} E_m \right) \right) \\ &\leq \mathbb{P}(E_0) + \sum_{m=2}^{M_n} \mathbb{P}(E_m). \end{aligned}$$

[1] This is one of the historically earliest uses of the probabilistic method. It predates Erdös' widespread usage of the method.

Note that E_0 is the event where $(Y_1, \ldots, Y_n)$ is not ε-jointly weakly typical with the sequence $(X_1(1), \ldots, X_n(1))$. For $m \geq 2$, E_m is the event where $(Y_1, \ldots, Y_n)$ is not ε-jointly weakly typical with $(X_1(m), \ldots, X_n(m))$. So

$$\mathbb{P}(d_n(e_n(W_N) \neq W_n) \leq \mathbb{P}(E_0^{(n)}) + (M_n - 1)\mathbb{P}(E_2^{(n)})$$

by symmetry. Now $\mathbb{P}(E_0^{(n)}) \to 0$ as $N \to \infty$, and $\mathbb{P}(E_2^{(n)}) \leq 2^{-nI(X;Y)}2^{3n\varepsilon}$ So if $R < I(X;Y) - 3\varepsilon$, this goes to 0 as $n \to \infty$. $\square$

Next time, we will prove the converse part of the theorem.