Information Theory — Foundations and Applications

Second Shannon theorem

Lucas Chibebe Céleri

Institute of Physics
Federal University of Goiás

2024 — Basque Center for Applied Mathematics
University of Basque Country

Preliminaries

From the definition of entropy, we can define the **joint entropy** of two random variables X and Y as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log\left(p_{XY}(x,y)\right),$$

where $p_{XY}(x,y)$ is the joint probability of the realizations x and y. This is the total uncertainty we have about both variables taken together

Now, we can ask by how much the uncertainty about one random variable changes when we learn the value of the other one. This is quantified by the **conditional entropy**

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log\left(p_{Y|X}(y|x)\right)$$

Preliminaries

Another important quantity is the measure of the amount of correlations shared by the random variables, which is called **mutual information**

$$I(X:Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right),$$

where $p_Y(y) = \sum_{x\in\mathcal{X}} p_{XY}(x,y)$ and $p_X(x) = \sum_{y\in\mathcal{Y}} p_{XY}(x,y)$ are the marginal probabilities.

The mutual information is related to the individual and joint entropies as

$$I(X:Y) = H(X) + H(Y) - H(X,Y)$$

Preliminaries

Let us now suppose that we have a random variable whose probability distribution is $q_X(x)$ but we think that it is $p_X(x)$. How inefficient is this? The answer is given by the relative entropy (Kullback-Liebler divergence)

$$D\left(p_X(x)||q_X(x)\right) = \sum_{x\in\mathcal{X}} p_X(x)\log\left(\frac{p_X(x)}{q_X(x)}\right)$$

Observe that, if supp$[q_X(x)] \not\subset$ supp$[p_X(x)]$, D diverges. This is a positive quantity that is zero if and only if $q_X(x) = p_X(x)$.

Now we can describe Shannon's second coding theorem.

The channel capacity — The problem

Let us now assume that the channel connecting Alice and Bob is **not ideal**. In other words, the information transmitted through the channel is not reliable.

The discrete channel is characterized by two random variables, X (the input) and Y (the output), and the conditional probability $p(x \in \mathcal{X} | y \in \mathcal{Y})$. The channel is also memoryless.

Since using the channel is expensive, Alice wants to minimize the uses of the channel while reliably communicate to Bob. It is natural to understand the **capacity** of the channel as

$$C = \max_{p_X(x)} I(X : Y)$$

The channel capacity — Error correction

Let us consider that the channel flips the input bit with probability p and leaves it unchanged with probability $1-p$. This is known as the bit-flip channel, that is assumed to be i.i.d.. Communication over such a channel works only if $p \to 0$. One solution for the noisy case is **error correction**. Let us consider redundant encoding of information

$$0 \to 000, \qquad 1 \to 111$$

Alice transmits the bit 0 using the codeword 000 that demands three uses of the channel. By majority vote, an error occurs when two or three flips are caused by the channel. For the code to work the error probability should satisfy

$$p_e = p^3 + 3p^2(1-p) = 3p^2 - 2p^3 < p \Rightarrow 0 < p < 1/2$$

The channel capacity — Error correction

We still have a significant probability of error. We can apply the majority vote again, using another code as an inner code, like $0 \to \overline{000}$ (similar for bit 1)

$$\overline{0} \to 000\,000\,000 \qquad \overline{1} \to 111\,111\,111$$

Such code reduces the error probability to $\mathcal{O}(p^4)$. Alice and Bob can continue this game until they reach a good probability of error. The problem is that the rate of the first coding is $1/3$, dropping to $1/9$ for the second and so one. Therefore, to achieve an arbitrarily small probability of error, the rate of communication vanishes.

Is there a way to code information into a noisy channel while keeping a good rate of communication?

The channel capacity — The problem

Now, Alice wants to achieve asymptotic reliability in communication. She selects messages from the set $[M] = \{1, ..., M\}$ with uniform probability. This implies that we do not care about the content of the message. The only thing that matters is her ability to send any information, reliably. The channel is modelled as a conditional probability

$$\mathcal{N} : p_{Y|X}(y|x)$$

Now, let $X^n = X_1X_2 \cdots X_n$ and $Y^n = Y_1Y_2 \cdots Y_n$ be the random variables associated with the sequences $x^n = x_1x_2 \cdots x_n$ and $y^n = y_1y_2 \cdots y_n$, respectively. Under i.i.d. assumption, we can write

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$$

The channel capacity — The problem

If Alice selects a message m, the code scheme translates it to the codeword $x^n(m)$ (n uses of the channel). Bob gets the corrupted codeword $y^n(\hat{m})$.

The rate of a given code scheme is

$$R = \frac{\#\text{ number of message bits}}{\#\text{ of channel uses}} = \frac{\log(M)}{n}.$$

Given a code $\mathcal{C}$, the average error probability is

$$\overline{p}_e(\mathcal{C}) = \frac{1}{M}\sum_{m=1}^{M} p_e(m, \mathcal{C})$$

The channel capacity — The problem

It is possible to prove that

$$\overline{p}_e(\mathcal{C}) \leq \epsilon \Rightarrow p_e(m, \mathcal{C}) \leq 2\epsilon$$

for at leas half of the messages in $[M]$.

Shannon's theorem states that there is a code that achieves the capacity of a given noisy channel with vanishingly small error probability.

The channel capacity — Proof (Bird's eye view)

To prove that there is a code that achives a vanishingly small error probability while reaching the capacity of the channel, Shannon considered the following

- Large number of uses of the channel (law of large numbers and probability theory)
- Typical sequences and the typical set
- The new ingredient was the notion of a *random code* (beyond the randomness of Alice's choice of the message and also the one coming from the channel)

The channel capacity — Proof (Bird's eye view)

The codewords themselves are chosen in a random way, accordingly with a random variable X. Each letter x_i of a given codeword x^n is selected according to the probability distribution $p_X(x_i)$. So, the codeword itself becomes a random variable $X^n(m)$. Given that this choice is i.i.d., we have

$$\text{Prob}\left[X^n(m) = x^n(m)\right] = p_{X_1 X_2 \ldots X_n}(x_1(m)x_2(m)\cdots x_n(m)) = \prod_{i=1}^{n} p_X(x_i(m))$$

So, the probability distribution does not depend explicitly on m (it is the same for all messages). The code itself becomes a random variable. The probability of choosing a particular code $\mathcal{C}_0$ is then

$$p_{\mathcal{C}}(\mathcal{C}_0) = \prod_{m=1}^{M} \prod_{i=1}^{n} p_X(x_i(m))$$

The channel capacity — Proof (Bird's eye view)

Shannon's insight was to study the expectation of the average error probability

$$\mathbb{E}_{\mathcal{C}}\left[\overline{p}_e(\mathcal{C})\right] = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{C}}\left[p_e(m,\mathcal{C})\right] = \mathbb{E}_{\mathcal{C}}\left[p_e(1,\mathcal{C})\right]$$

Since the probability is independent of the message, we could choose any number instead of $m = 1$. Shannon proceeded by computing an upper bound on this probability $\mathbb{E}_{\mathcal{C}}\left[\overline{p}_e(\mathcal{C})\right] \leq \epsilon$, which implies that there exists some deterministic code $\mathcal{C}_0$ for which

$$\overline{p}_e(\mathcal{C}_0) \leq \epsilon$$

Thus eliminating the randomness of the code!

The channel capacity — Proof (Bird's eye view)

Now, we come back a few slides and simply throw away half of the messages, the ones with the worst probability of error, thus reducing the number of messages from 2^{nR} to $2^{n(R-1/2)}$, causing the rate to change from R to $R - 1/n$, which is negligible in the large n limit. After this step, we have

$$\max[p_e(\mathcal{C}_0)] \leq 2\epsilon$$

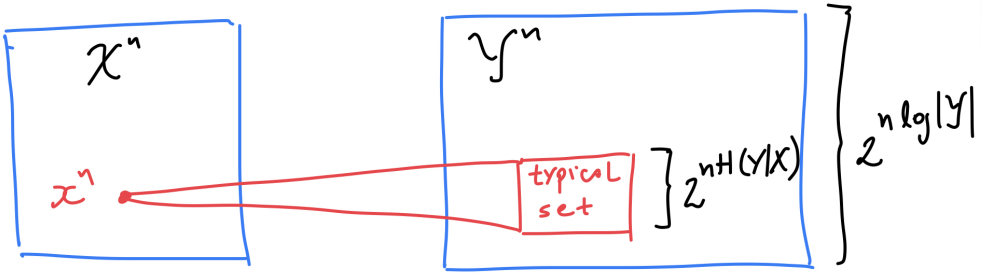
The only thing that we need to understand is the size of the code employed in the communication process given the size of the message set $M = 2^{nR}$.

The channel capacity — Proof (Bird's eye view)

We want to keep the rate constant while taking the asymptotic limit and ask about the maximum allowed rate, while maintaining a vanishingly small probability of error. To do this, we need to determine the number of distinguishable messages Alice can send to Bob.

This is done with the introduction of the conditional typicality, which is based on the the conditional entropy. The idea is that for a given codeword x^n we need to identify the set of possible output codewords y^n at Bob's side. Conditional typicality says that, for each x^n there is a corresponding conditional typical set at the output.

The channel capacity — Proof (Bird's eye view)



The channel capacity — Proof (Bird's eye view)

The conditional typical set has the following properties

Conditional AEP

- It has almost all the probability
- Its size is approximately $2^{nH(Y|Y)}$
- Uniform probability for the conditional sequences y^n

Let us now describe how the code works!

The channel capacity — Proof (Bird's eye view)

Steps of communication

- Alice generates 2^{nR} messages according to $p_X(x)$.
- Bob does not know the message, so the output sequence y^n must be generated according to the probability $p_Y(y)$.
- Bob verifies if y^n belongs to the typical set (whose size is $2^{nH(Y)}$).
- If yes, them he employs his knowledge of the code in order to determine to which conditional typical set y^n belongs. The size of these sets is $2^{nH(Y|X)}$.
- Now, Alice and Bob must structure the code in such a way to prevent the error in this last step.

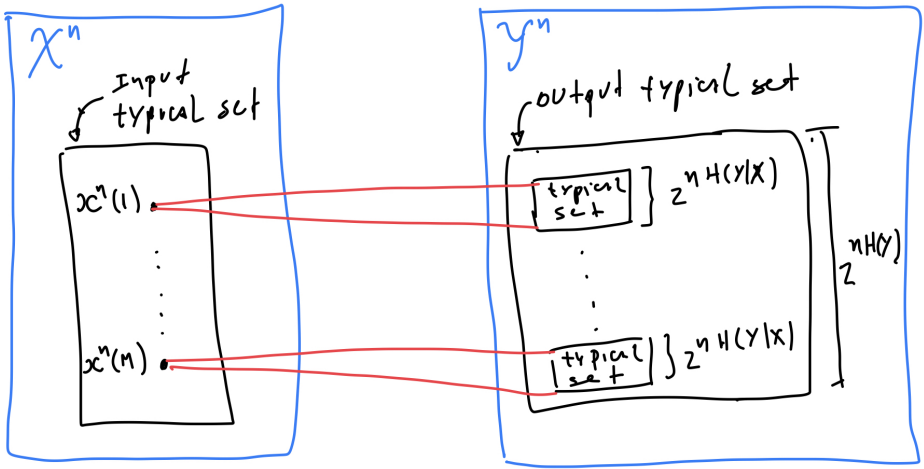
The channel capacity — Proof (Bird's eye view)

They must have no overlap between the output conditional sets! Then, Bob should be able to decode the output sequence to a unique input one. They must divide the set of the output typical sequences into M subsets of conditionally typical outputs, all of size $2^{nH(Y|X)}$. By setting

$$M = 2^{nR} = \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y)-H(Y|X))}$$

the job is done!

The channel capacity — Proof (Bird's eye view)



The channel capacity — Proof (Bird's eye view)

It turns out that Alice can reliably communicate to Bob with a rate

$$R \leq H(Y) - H(Y|X) = I(X:Y)$$

We finally achieve the channel capacity. Alice chooses her code according to $p_X(x)$. Since the mutual information is concave, there is a single distribution that maximizes it. We then define the **channel capacity** as

$$C(\mathcal{N}) = \max_{p_X(x)} I(X:Y)$$

Shannon theory

With this we finish the outlook at Shannon's theorems. In summary, we saw that if we compress information at the rate smaller than the entropy of the source, such information can be reliably decoded. Also, if we transmit information at a rate smaller than the channel capacity, this information can be reliably decoded at the output of the channel.

One thing that is missing is the proof that both of these rates are **optimal**. This is done by means of the so called **converse theorems**, that we will not present here. They can be found in the books mentioned earlier.

Thank you for your attention

lucas@qpequi.com

www.qpequi.com