Information Theory — Foundations and Applications

The first Shannon coding theorem

Lucas Chibebe Céleri

Institute of Physics
Federal University of Goiás

2024 — Basque Center for Applied Mathematics
University of Basque Country

Data compression — An example

Alice wants to communicate with Bob. Its information source, characterized by the random variable X, chooses among the symbols $\mathcal{X} = \{a, b, c, d\}$ according with the probabilities $p_X(x) = \{1/2, 1/8, 1/4, 1/8\}$.

The channel only takes bits as inputs. Therefore, we need a **code** that translates the information emitted by the source into something that can actually be transmitted over the channel. One possible choice is

$$a \to 00, \quad b \to 01, \quad c \to 10, \quad d \to 11$$

With this code we can translate the message into **codewords** that are accepted by the channel!

Data compression — An example

How can the efficiency of this code be characterized? The expected length of the codeword is a good measure

$$l = \sum_x p_X(x) l_x$$

Such a quantity, for our code, is just $l = 2$. This means that, on average, Alice has to send two bits over the channel (two uses of the channel).

Is this the best we can do?
No! We did not take into account the non-uniformity of the probability distribution. We can certainly do better!

Data compression — An example

Given the probability density $p_X(x) = \{1/2, 1/8, 1/4, 1/8\}$, let us then choose a code according to the following scheme

- Smallest codewords for the highest probability
- The code must be reliable

We can then define

$$a \to 0, \quad b \to 110, \quad c \to 10, \quad d \to 111$$

which results in $l = 7/4 < 2$.

Interesting $H(X) = 7/4$. But this is not a coincidence. Shannon's theorem says that such a code is optimal.

Data compression — Statement of the theorem

Theorem 1 (First Shannon coding theorem)
The entropy of an information source, $H(X)$, specified by a discrete random variable X, is the maximum achievable rate for data compression.

In other words, Shannon's first theorem states that there is an optimal code for which the rate of data compression is Shannon's entropy! This also provides an operational interpretation for $H(X)$.

Data compression — Shannon's insight

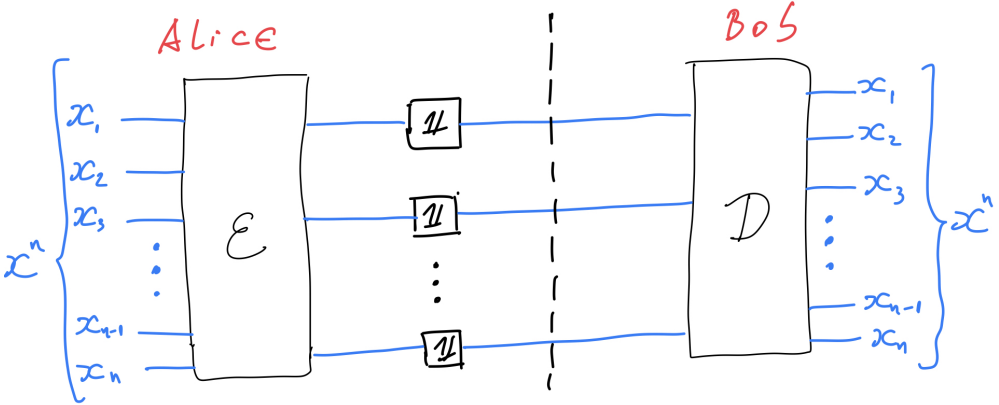
Shannon's idea was to let the source emit a large sequence of symbols

$$x^n = x_1 x_2 ... x_n \qquad \text{with } n \text{ very large}$$

x_i is the i-th emitted symbol as a realization of the random variable X_i. X^n is the random variable associated with the sequence x^n.

The next step is to code the sequence as a large block. But why this works? Shannon discovered that the space of all sequences can be split into two sets, the one containing the **typical sequences**, that contains all the probability, and the rest. Also, the cardinality of the typical set is exponentially smaller than that of the total set, in general.

Data compression — The problem



Data compression — The typical set

Let us assume that the information source is **i.i.d** (independently and identically distributed). This means that each symbol is independent of the previous ones and that $p_{X_i}(x) = p_X(x)\ \forall i$. Therefore

$$\begin{aligned} p_{X^n}(x^n) &= p_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \overset{\text{ind}}{=} p_{X_1}(x_1)p_{X_2}(x_2)...p_{X_n}(x_n) \\ &\overset{\text{id}}{=} p_X(x)p_X(x_2)...p_X(x_n) = \prod_{i=1}^{n} p_X(x_i). \end{aligned}$$

Let us now label the symbols in the alphabet $\mathcal{X}$ as a_i, with $i = 1, ..., |\mathcal{X}|$. Let $N(a_i|x^n)$ be the number of occurrences of the letter a_i in the sequence x^n. Then

$$p_{X^n}(x^n) = \prod_{i=1}^{n} p_X(x_i) = \prod_{i=1}^{|\mathcal{X}|} p_X(a_i)^{N(a_i|x^n)}.$$

Data compression — The typical set

This is much easier to compute since $|\mathcal{X}| \ll n$. Note that we used again the i.i.d. hypothesis, because we are interested only in the probabilities, which are invariant under permutation.

Now we have to investigate the probability of the random variable X^n. Let us consider the sample average of the information content of the random sequence X^n

$$\overline{H}(X^n) = -\frac{1}{n}\log\left(p_{X^n}(X^n)\right)$$

This is called **sample entropy** of the random sequence.

Data compression — The typical set

Let $N(a_i|X^n)$ be the number of appearances of a_i in the random sequence X^n. Then

$$
\begin{aligned}
-\frac{1}{n}\log\left(p_{X^n}(X^n)\right) &\overset{\text{i.i.d.}}{=} -\frac{1}{n}\log\left(\prod_{i=1}^{|\mathcal{X}|} p_X(a_i)^{N(a_i|X^n)}\right) \\
&= -\frac{1}{n}\sum_{i=1}^{|\mathcal{X}|}\log\left(p_X(a_i)^{N(a_i|X^n)}\right) \\
&= -\sum_{i=1}^{|\mathcal{X}|}\frac{N(a_i|X^n)}{n}\log\left(p_X(a_i)\right)
\end{aligned}
$$

Data compression — The typical set

Now we take the asymptotic limit

$$\lim_{n\to\infty} \frac{N(a_i|X^n)}{n} = p_X(a_i)$$

which implies that

$$\lim_{n\to\infty} \left[-\frac{1}{n} \log\left(p_{X^n}(X^n)\right) \right] = -\sum_{i=1}^{|\mathcal{X}|} p_X(a_i) \log\left(p_X(a_i)\right) = H(X).$$

It is highly likely that the random sequence X^n satisfies

$$\lim_{n\to\infty} \text{Prob} \left[\left| -\frac{1}{n} \log\left(p_{X^n}(X^n)\right) - H(X) \right| \leq \delta \right] = 1 \quad \forall\ \delta > 0.$$

Data compression — The typical set

It is highly likely that the information source emits a sequence whose sample entropy is close to the true entropy!

Typical sequence
A sequence x^n is called a typical sequence if its sample entropy is close to the true entropy $H(X)$.

Typical set
The set of all typical sequences is called the typical set.

Data compression — The typical set

In summary, we have shown that it is highly likely that a sequence in the typical set

$$T_\delta^{X^n} = \{x^n \in X^n \,|\, \left|\overline{H}(X^n) - H(X)\right| < \delta\}$$

will be emitted by the source in the asymptotic limit.

Shannon's idea
Alice just needs to code the sequences in $T_\delta^{X^n}$. If the source emits a non-typical sequence, an error is declared. Shannon showed that such error vanishes asymptotically.

This scheme works because of a very important set of properties of the typical set that we now discuss.

Data compression — Properties of $T_{\delta}^{X^n}$

The three properties of the typical set that justify Shannon's idea can be stated as follows

Asymptotic Equipartition Property (AEP)

- The typical set contains almost all the probability.
- The typical set is exponentially smaller than the set of all sequences.
- Each typical sequence has almost uniform probability.

Data compression — AEP

In order to prove the first property, we need to show that

$$\forall\, \epsilon > 0 \quad \text{Prob}\{x^n \in T_\delta^{X^n}\} = \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n) \geq 1 - \epsilon$$

for sufficiently large n.

We start by remembering that the weak law of large numbers states that the sample mean converges in probability to the expectation. So, let us consider the set $\{X_i\}_{i=1}^n$ of random variables. Its sample average is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Data compression — Proof of AEP

Denoting by μ the expectation of each random variable, the formal statement of the law of large numbers is

$$\forall \epsilon, \delta > 0 \quad \ni n_0 \, : \forall n > n_0 \quad \text{Prob}\{\left|\overline{X} - \mu\right| < \delta\} > 1 - \epsilon$$

To continue, consider the sequence of random variables $\{-\log\left(p_X(X_i)\right)\}_{i=1}^{n}$

$$-\frac{1}{n}\sum_{i=1}^{n}\log\left(p_X(X_i)\right) = -\frac{1}{n}\log\left(p_{X^n}(X^n)\right) = \overline{H}(X^n)$$

since $p_{X^n}(X^n) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} p_X(X_i)$.

Data compression — Proof of AEP

We know that the expectation of the random variable $-\log\left(p_X(X_i)\right)$ is Shannon entropy of X_i, which we denoted by μ in the last slide. Therefore, applying the law of large numbers to these random variables, we obtain

$$\forall\,\epsilon,\delta>0 \quad \ni n_0 \,:\, \forall\, n > n_0 \quad \text{Prob}\{\left|\overline{H}(X^n) - H(X)\right| < \delta\} > 1-\epsilon,$$

which is exactly the condition for the random sequence X^n to be in the typical set, and the probability of this event goes to one as n becomes very large.

The typical set contains all the probability in the asymptotic limit!

Data compression — Proof of AEP

Let us now prove the third property, the equipartition. We have shown that the sample entropy is δ-close to the entropy of the source, $H(X)$. Based on the definition of the sample entropy we can write

$$H(X) - \delta \leq -\frac{1}{n} \log \left(p_{X^n}(x^n) \right) \leq H(X) + \delta,$$

which implies that

$$2^{-n(H(X)+\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-\delta)}.$$

Since δ is arbitrary, we see that the probability distribution associated with the sequences in X^n is uniform, being equal to $2^{-nH(X)}$. This justify the name equipartition!

Data compression — Proof of AEP

Lastly, we need to show that $\left|T_\delta^{X^n}\right|$ is exponentially smaller than $|\mathcal{X}^n|$, which is the second property of the typical set. This can be done by considering that

$$\begin{aligned} 1 &= \sum_{x^n \in X^n} p_{X^n}(x^n) \geq \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n) \geq \sum_{x^n \in T_\delta^{X^n}} 2^{-n(H(X)+\delta)} \\ &= 2^{-n(H(X)+\delta)} \left|T_\delta^{X^n}\right| \end{aligned}$$

where we have used the equipartition property. Therefore, the cardinality of the typical set is bound from above as

$$\left|T_\delta^{X^n}\right| \leq 2^{n(H(X)+\delta)}$$

which should be compared with $|\mathcal{X}^n| = 2^{n \log(|\mathcal{X}|)}$.

Data compression — Optimal rate

We have introduced the typical set and the set of properties known as the Asymptotic Equipartition Property. We are now ready to show that the ultimate rate of data compression is Shannon entropy.

First, let us consider the code (n, R, ϵ) and let the encoding and the decoding maps be defined as

$$\mathcal{E} : \mathcal{X}^n \mapsto \{0,1\}^{nR} \quad \mathcal{D} : \{0,1\}^{nR} \mapsto \mathcal{X}^n$$

The error probability can then be computed as

$$p_e = \text{Prob}\left\{\left(\mathcal{D} \circ \mathcal{E}\right)\left(X^n\right) \neq X^n\right\} \leq \epsilon$$

The rate is defined as

$$R = \frac{\#\text{ number of channels uses}}{\text{length of the sequence}}$$

Data compression — Optimal rate

Based on the AEP, Shannon proposed to code only the sequence in $T_\delta^{X^n}$ and throw away the rest. The code will always work for large enough n since the probability of the source to emit a non-typical sequence vanishes asymptotically.

Due to AEP (exponentially small cardinality) we have that $nR = n(H(X) + \delta)$ (the number of bits in the sequence), while the length of the typical sequence is n. Therefore

$$R = \frac{n(H(X) + \delta)}{n} \approx H(X)$$

for sufficiently large n.

Data compression — Optimal rate

Therefore, we choose ϵ, $\delta > 0$ and and encoding function $f : x^n \mapsto \{0,1\}^{nR}$ with some error symbol e_0, that is, $f : x^n \mapsto e_0$ if $x^n \notin T_\delta^{X^n}$. Assuming AEP, $R = H(X) + \delta$. The decoding is the inverse function f^{-1} on $T_\delta^{X^n}$, such that $f^{-1} : e_0 \mapsto x_0^n$ signals the error. Then, due to the AEP, the probability of error in this code is less than ϵ and the code is achievable.

Note that it is not possible to compress a random variable whose probability distribution is uniform since $H(X) = \log |\mathcal{X}|$.

Thank you for your attention

lucas@qpequi.com

www.qpequi.com