Information Theory — Foundations and Applications

Classical Typicality

Lucas Chibebe Céleri

Institute of Physics
Federal University of Goiás

2024 — Basque Center for Applied Mathematics
University of Basque Country

Introduction

Typicality is a fundamental concept in the proofs of Shannon coding theorems. The first one, data compression, relies on the notion of typical set, while the channel capacity theorem is rooted into the conditional typicality.

The central object here is the **asymptotic equipartition property**, which is the application of the law of large numbers to a sequence drawn independently and identically from a distribution $p_X(x)$ for some random variable X.

This property shows that we can split the set of all possible sequences into two subsets: the **typical set** containing the sequences that are overwhelmingly likely to occur and the **atypical set** that contains all other sequences.

We here describe the definition and the properties of the classical typical set.

Example of typicality

Let us consider a random binary variable that outputs 1 with probability $3/4$ and 0 with probability $1/4$. We let such an information source to emit a sequence of n bits. The *sample entropy* of such a sequence is

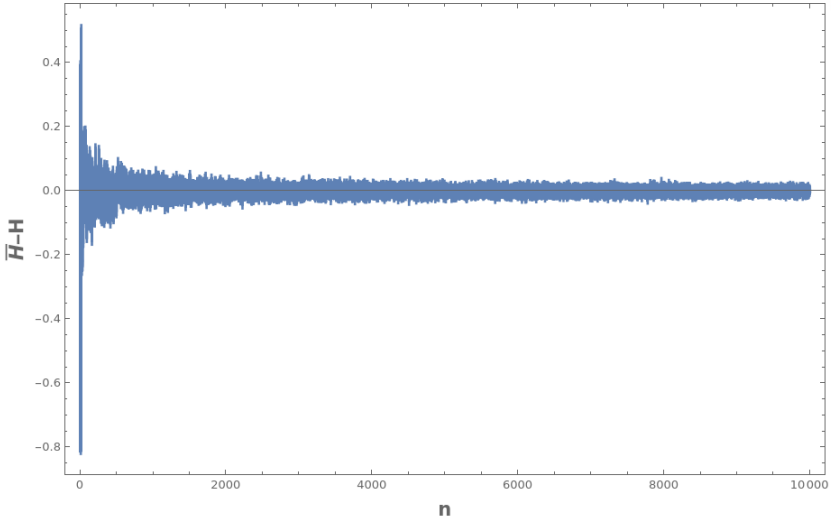
$$\overline{H}(X) = -\frac{1}{n}\log\left(p^{N(1|n)}(1-p)^{n-N(1|n)}\right)$$

where N is the number of times the bit 1 appeared in the sequence of length n.

The true entropy of the source is

$$H(X) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} \approx 0.81$$

Example of typicality



Weak typicality

Let X be a random variable $X = \{p_X(x), \mathcal{X}\}$ and let $\{a_i\}_{i=1}^{|\mathcal{X}|}$ labels the letters of the alphabet. This random variable describes the information source.

Now, assuming that the symbols are drawn i.i.d, let the source emits n symbols.

- $X^n = X_1 X_2 \cdots X_n$ is the random variable associated with the sequences
- $x^n = x_1 x_2 \cdots x_n$ is a particular realization of X^n

The probability of a particular string x^n is

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

Weak typicality

Intuitively, at large n we expect that

$$p_{X^n}(x^n) = p_X(x_1) \cdots p_X(x_n) \approx p_X(a_1)^{np_X(a_1)} \cdots p_X(a_{|\mathcal{X}|})^{np_X(a_{|\mathcal{X}|})}$$

The information of a particular string is then

$$-\frac{1}{n} \log \left(p_{X^n}(x^n)\right) \approx -\sum_{i=1}^{|\mathcal{X}|} p_X(a_i) \log p_X(a_i) = H(X)$$

Based on this, we define the **sample entropy** as

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

Weak typicality

From this, we can define the following.

Typical sequences
A sequence x^n is δ-typical if its sample entropy $\overline{H}(x^n)$ is δ-close to the entropy of the random variable X describing the information source, $H(X)$

Typical set
It is the set of all typical sequences

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

Properties of the typical set

- **Unity probability**: The typical set asymptotically has probability one

$$\Pr\left[x^n \in T_\delta^{X^n}\right] \geq 1 - \epsilon \qquad \forall\, \epsilon \in (0,1) \text{ and } \delta > 0$$

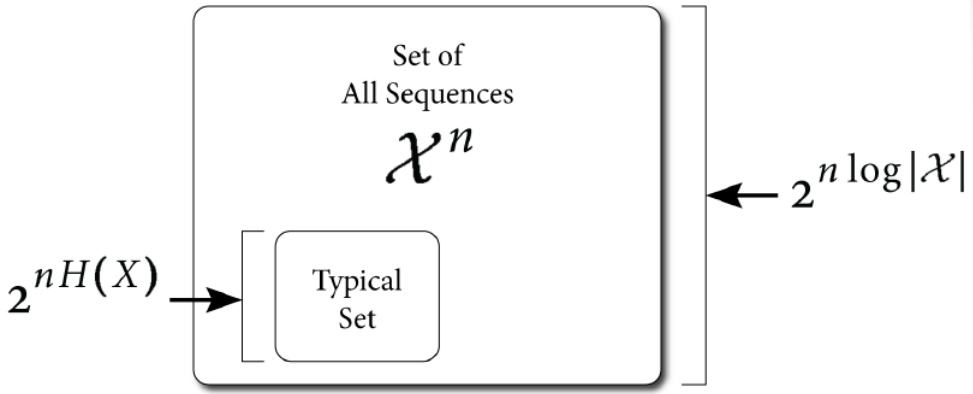
- **Exponentially smaller cardinality**: The total number of sequences $\left|T_\delta^{X^n}\right|$ in the typical set is exponentially smaller than the total number of sequences, $|\mathcal{X}|^n$, except when p_X is uniform

$$(1-\epsilon)2^{n(H(X)-\delta)} \leq \left|T_\delta^{X^n}\right| \leq 2^{n(H(X)+\delta)} \qquad \forall\, \epsilon \in (0,1) \text{ and } \delta > 0$$

- **Equipartition**: The probability of a particular δ-typical sequence x^n is approximately uniform

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

Typical set

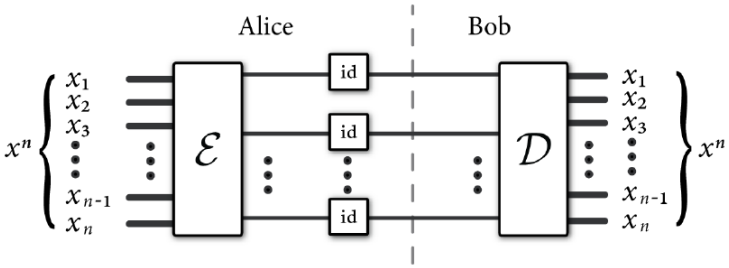

Set of
All Sequences
$\mathcal{X}^n$
$2^{nH(X)}$
Typical
Set
$2^{n \log |\mathcal{X}|}$

Data compression

Shannon compression

The entropy of an information source specified by a discrete random variable X is the smallest achievable rate for compression

$$\inf\{R : \text{R is achievable for } X\} = H(X)$$



Proof

The proof consists of two parts

1. **Direct coding**: the proof exhibits a coding scheme with an achievable rate and demonstrates that its rate converges to the entropy in the asymptotic limit

2. **Converse theorem**: It is a statement of optimality. It establishes that any coding scheme with rate below the entropy is not achievable.

Typical sequences and their properties are employed for proving a direct coding theorem, while the converse part resorts to entropy inequalities.

Weak jointly typicality

Let X and Y be random variables with $x^n = x_1 x_2 \cdots x_n$ and $y^n = y_1 y_2 \cdots y_n$ being two independent realizations of them. The sample joint entropy, under i.i.d assumption, is defined as

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

Jointly typical sequences

Two sequences x^n and x^n are δ-jointly typical if its sample joint entropy $\overline{H}(x^n, y^n)$ is δ-close to the joint entropy $H(X, Y)$

Jointly typical set

It is the set of all jointly typical sequences

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

Properties of the joint typical set

- **Unity probability**: The typical set asymptotically has probability one

$$\Pr\left[(x^n, y^n) \in T_\delta^{X^n,Y^n}\right] \geq 1 - \epsilon \quad \forall\, \epsilon \in (0,1) \text{ and } \delta > 0$$

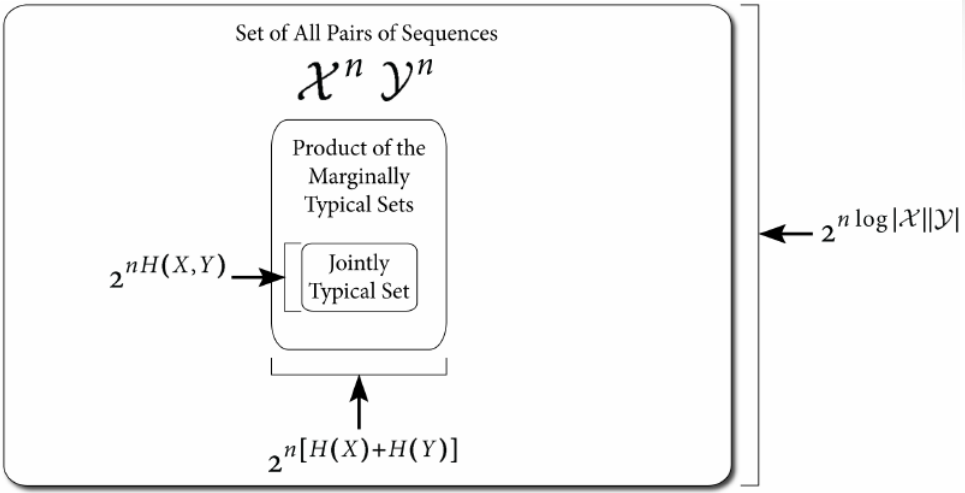
- **Exponentially smaller cardinality**: The total number of sequences $\left|T_\delta^{X^n,Y^n}\right|$ in the typical set is exponentially smaller than the total number of sequences, $(|\mathcal{X}||\mathcal{Y}|)^n$, except when the joint probability is uniform.

$$(1-\epsilon)2^{n(H(X,Y)-\delta)} \leq \left|T_\delta^{X^n,Y^n}\right| \leq 2^{n(H(X,Y)+\delta)} \quad \forall\, \epsilon \in (0,1) \text{ and } \delta > 0$$

- **Equipartition**: The probability of a particular δ-typical joint sequence (x^n, y^n) is approximately uniform

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

Joint typical set



Weak conditional typicality

Let us start by defining the **conditional sample entropy** of two sequences x^n and y^n with respect to $p_{X,Y}(x,y) = p_X(x)p_{Y|Y}(y|x)$ as

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

with

$$p_{Y^n|X^n}(y^n|x^n) = p_{Y|X}(y_1|x_1)\cdots p_{Y|X}(y_n|x_n)$$

Conditionally typical set
It is the set of all conditionally typical sequences

$$T_\delta^{Y^n,x^n} = \left\{y^n \,:\, \left|\overline{H}(y^n|x^n) - H(Y|Y)\right| \leq \delta\right\}$$

Properties of the conditionally typical set

Unity probability: On average with respect to a random sequence X^n, the set $T_\delta^{Y^n,x^n}$ has asymptotically unity probability. Therefore, it is highly likely that random sequences Y^n and X^n are such that Y^n is a conditionally typical sequence

$$\mathbb{E}_{X^n}\left\{\mathsf{Pr}_{Y^n|X^n}\left\{Y^n \in T_\delta^{Y^n|X^n}\right\}\right\} \geq 1-\epsilon$$

Exponentially small cardinality: The number $\left|T_\delta^{Y^n,x^n}\right|$ f conditional typical sequences is exponentially smaller than the total number $|\mathcal{Y}|^n$

$$\left|T_\delta^{Y^n,x^n}\right| \leq 2^{n(H(Y|X)+\delta)} \quad \text{and} \quad \mathbb{E}_{X^n}\left\{\left|T_\delta^{Y^n,X^n}\right|\right\} \geq (1-\epsilon)2^{n(H(Y|X)-\delta)}$$

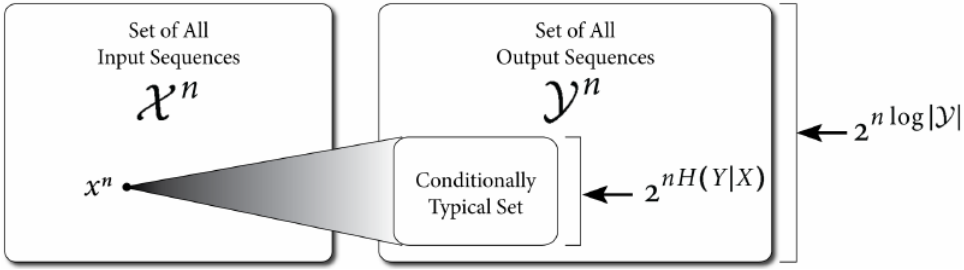
Properties of the conditionally typical set

Equipartition: The probability of a given δ-conditionally typical sequence y^n (corresponding to the sequence x^n) is approximately uniform

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

The proofs of these properties follow from the application of the law of large numbers.

Conditional typical set



Strong typicality

Instead of considering that the sample entropy approaches the entropy of the source, strong typicality assumes that the empirical frequencies converges to the true probability of the source.

$$\left| \frac{1}{n} N(x|x^n) - p_X(x) \right| \leq \delta$$

Strong typicality implies weak typicality. And all of them has the same properties.

Channel capacity theorem

The maximum mutual information $I(\mathcal{N})$ is equal to the capacity $C(\mathcal{N})$ of a channel $\mathcal{N} = p_{Y|X}(y|x)$

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

The proof also contains two parts. The first one, the direct coding, employs the joint and conditional notions of typicality in order to prove that there is a code for which the rate $C(\mathcal{N})$ is achievable. The converse theorem shows that such a rate is optimal.

Thank you for your attention

lucas@qpequi.com

www.qpequi.com