



Alberto Santini

Information theory notes

Information theory

Course held at Mathematics Department,
Copenhagen university.

Lecturer: Henrik Densing Petersen

Contents

1	Codes and entropy	2
1.1	Codes and Kraft's inequality	2
1.2	Appendix A: useful set theoretical notions	6
1.3	Appendix B: a non-set theoretical approach to prefix codes and Kraft's inequality	6
1.4	Appendix C: Extended Kraft's inequality	9
2	Entropy	11
2.1	Expected codelength, entropy, divergence	11
2.2	Mutual information and data-processing	19
2.3	Appendix A: Huffman codes	23
3	Asymptotic equipartition property	25
	Index	28
	Bibliography	29

1.1 Codes and Kraft's inequality

Definition 1.1. Given a random variable $X : \Omega \rightarrow X_0$ with both Ω, X_0 finite sets and given another finite set $A \neq \emptyset$ (usually called the alphabet A) we denote the set of all finite words over A as $A^{(\infty)}$, where as finite words we take all the finite tuples of elements of A . More formally, we can write:

$$A^{(\infty)} = \bigsqcup_{n \in \mathbb{N}} A^n = \{f : \mathbb{N} \rightarrow A \cup \{\infty\} \mid \exists N \in \mathbb{N} : f(n) = \infty \Leftrightarrow n > N\} \quad (1.1)$$

So, a **source code** for X_0 is a function $\kappa : X_0 \rightarrow A^{(\infty)}$. If this function is injective, the code is said to be **non-singular**. Given an element $x \in X_0$ we say that its **length** $l(x)$ is the number N associated with its code, as defined in equation 1.1; in other words $l(x) = \max\{n \in \mathbb{N} : \kappa(x)(n) \neq \infty\}$.

Observation 1.2. It easy to constructively produce a bijection between the sets $A^{(\infty)}$ and $(A^{(\infty)})^{(\infty)}$, in the sense that given a function in one of the two sets, we can give an explicit expression of a function in the other set, such that both produce

the same output string of letters of the alphabet. In particular, given $\bar{f} \in A^{(\infty)}$ we consider $f \in (A^{(\infty)})^{(\infty)}$ given by:

$$f(n) = \begin{cases} \bar{f} & \text{if } n = 1 \\ e & \text{if } n > 1 \end{cases} \quad (1.2)$$

Where $e : \mathbb{N} \rightarrow A \cup \{\infty\}$ is the empty codeword $e = \mathbb{N} \times \{\infty\}$. On the converse, given an $f \in (A^{(\infty)})^{(\infty)}$, we define $\bar{f} \in A^{(\infty)}$ as:

$$\bar{f}(n) = \begin{cases} f(i) \left(n - \sum_{j=1}^{i-1} l(f(j)) \right) & \text{if } \sum_{j=1}^{i-1} l(f(j)) < n \leq \sum_{j=1}^i l(f(j)) \\ \infty & \text{if } n > \sum_{j=1}^{l(f)} l(f(j)) \end{cases} \quad (1.3)$$

Definition 1.3. The **finite-length extension** of the code $\kappa : X_0 \rightarrow A^{(\infty)}$ is the function $\kappa^* : X_0^{(\infty)} \rightarrow A^{(\infty)}$ defined by:

$$\kappa^*(x_1, \dots, x_t)(n) = \begin{cases} \kappa(x_i) \left(n - \sum_{j=1}^{i-1} l(x_j) \right) & \text{if } \sum_{j=1}^{i-1} l(x_j) < i \leq \sum_{j=1}^i l(x_j) \\ \infty & \text{if } i > \sum_{j=1}^t l(x_j) \end{cases} \quad (1.4)$$

So basically, for each word made up by letters in the alphabet X_0 we produce a code that is the juxtaposition of the codes for each letter of the word, taken in order. If κ^* is injective, we say that the code κ is **uniquely decodable**. In other words, we want κ^* as a code over $X_0^{(\infty)}$ to be non-singular. We can of course use a length function for extension codes as well: $l^* : X_0 \rightarrow \mathbb{N}$ defined as

$$l^*(x_1, \dots, x_t) = \sum_{i=1}^t l(x_i) \quad (1.5)$$

Note 1.4. Warning : You may first want to read the appendix of this chapter before going on reading, if you're not familiar with some concept of set theory such as posets, chains, antichains and trees.

Definition 1.5. If we introduce on $A^{(\infty)}$ the following partial order:

$$f \leq g \Leftrightarrow \exists N \in \mathbb{N} : (f(n) = g(n) \forall n \leq N) \wedge (f(n) = \infty \forall n > N) \quad (1.6)$$

we say that the code $\kappa : X_0 \rightarrow A^{(\infty)}$ is a **prefix code** if $\kappa(X_0)$ is an antichain of $(A^{(\infty)}, \leq)$. This is to say that no codeword (with codeword we mean the image of an element of X_0 through κ) is the prefix of another codeword.

Observation 1.6. It's easy to notice that every prefix code is non-singular. In fact, if the function κ wasn't injective, we would have $x, y \in X_0$ with $x \neq y$ such that $\kappa(x) = \kappa(y)$. This, in particular, means that $\kappa(x)$ and $\kappa(y)$ are comparable, in contrast with the assumption that $\kappa(X_0)$ is an antichain.

Observation 1.7. It should now be clear that the shortest we want our codewords to be, the most unlikely will be for the code to be a prefix code. The following theorem sets a bound for this fact, but first let's make one more observation.

Observation 1.8. We now notice that if we include the empty word e in $A^{(\infty)}$ (call this new set $A_e^{(\infty)}$), we have that $(A_e^{(\infty)}, \leq)$ is a tree with root the empty word.

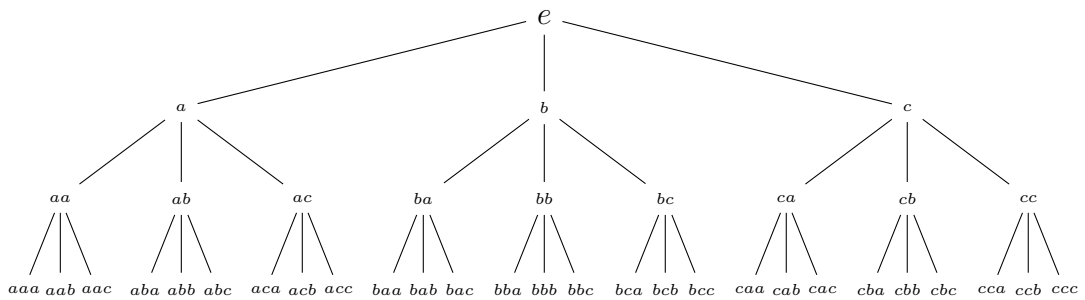


Figure 1.1: The first three level of $A_e^{(\infty)}$ for $A = \{a, b, c\}$, represented as a tree.

Theorem 1.9. **Kraft's inequality**. Given a prefix code $\kappa : X_0 \rightarrow A^{(\infty)}$ we have that:

$$\sum_{x \in X_0} |A|^{-l(x)} \leq 1 \quad (1.7)$$

Proof. First of all we notice that since κ is a prefix code, the tree $(A_e^{(\infty)}, \leq)$ has the property that no codeword has another codeword among its successors:

$$\forall x \in \kappa(X_0) \quad S_x \cap \kappa(X_0) = \emptyset \quad (1.8)$$

Let now be $l = \max_{x \in X_0} l(x)$, so the l -th level is the deepest level at which we can find a codeword in our tree. Each codeword $\kappa(x)$ at level $l(x)$ has exactly $|A|^{l-l(x)}$ successors at level l , while the total number of nodes at level l is $|A|^l$. Moreover, equation 1.8 ensures us that no two different codewords may have the same successors at the l -th level. Finally, we notice that there may be at level l some node which isn't among the successors of any codeword. All these facts tell us that if we sum over the codewords the number of their successors we are bounded by the number of elements at the l -th level:

$$\sum_{x \in X_0} |A|^{l-l(x)} \leq |A|^l \quad (1.9)$$

And so, dividing both sides by $|A|^l$:

$$\sum_{x \in X_0} |A|^{-l(x)} \leq 1 \quad (1.10)$$

□

Theorem 1.10. *Converse of theorem 1.9.* Given a function $l : X_0 \rightarrow \mathbb{N}$ which satisfies equation 1.7, it's always possible to find a code $\kappa : X_0 \rightarrow A^{(\infty)}$ for which l is the length function.

Proof. First we write $X_0 = \{x_1, \dots, x_t\}$ where we suppose, without loss of generality that $l(x_1) \leq \dots \leq l(x_t)$. We focus on the tree $(A_e^{(\infty)}, \leq)$. We fix an element y_1 on the $l(x_1)$ -th level of the tree and define $\kappa(x_1) = y_1$. Then we consider the tree $(A_e^{(\infty)} - (\{y_1\} \cup S_{y_1}), \leq)$ and we repeat the same reasoning now fixing an element y_2 on the $(l(x_1) - l(x_2))$ -th level and defining $\kappa(x_2) = y_2$. We can continue this way, considering at step i the tree $(A_e^{(\infty)} - (\{y_1, \dots, y_i\} \cup \bigcup_{j=1}^i S_{y_j}))$ and fixing an element on the $(\sum_{j=1}^i l(x_j))$ -th level of the tree. We will end at step $l = l(x_t)$. □

1.2 Appendix A: useful set theoretical notions

Definition 1.11. A **partially ordered set** (poset) is a couple (P, \leq) where:

- P is a non-empty set
- \leq is a reflexive, antisymmetric and transitive relation on P

The set P is said to be **total ordered** by \leq if every two elements of P are comparable by \leq . We can also define the **strict order** on P as $p < q \Leftrightarrow p \leq q \wedge p \neq q$. A subset $Q \subseteq P$ is said to be **well ordered** by $<$ if it is totally ordered and every non-empty $S \subseteq Q$ has a least element. We sometimes say that Q is a chain of P .

Definition 1.12. An **antichain** of the poset (P, \leq) is a subset $Q \subseteq P$ such that every two elements of Q are not comparable by \leq .

Definition 1.13. A **tree** is a poset (T, \leq) such that for every $t \in T$ the set $P_t = \{s \in T : s < t\}$ is well ordered by $<$. A **branch** of a tree is a maximal chain of T . For each element $t \in T$, the order type of the set P_t is called the **height** of t . Since we only work with finite-height trees we can just refer to the cardinality of P_t , instead to its order type. The n -th **level** of the tree is the set of all the elements of height n . The **root** of a tree is the only element of height 0. The elements of P_t are called the **predecessors** of t . The elements of $S_t = \{s \in T : t \in P_s\}$ are called the **successors** of t .

1.3 Appendix B: a non-set theoretical approach to prefix codes and Kraft's inequality

Definition 1.14. Let's first consider the set of words of length greater than a certain $M \in \mathbb{N}$:

$$A_{\geq M}^{(\infty)} = \{f : \mathbb{N} \rightarrow A \cup \{\infty\} \mid \exists N \geq M : f(n) = \infty \Leftrightarrow n > N\} \quad (1.11)$$

So, every $f \in A^{(\infty)}$ can be restricted to a function of $A_{\geq M}^{(\infty)}$, by an operator $P_M : A^{(\infty)} \rightarrow A_{\geq M}^{(\infty)}$:

$$(P_M f)(n) = \begin{cases} f(n) & \text{if } n \leq M \\ \infty & \text{if } n > m \end{cases} \quad (1.12)$$

With those notations, we say that a code $\kappa : X_0 \rightarrow A^{(\infty)}$ is a **prefix code** if

$$\forall x \in X_0 \nexists y \in X_0, y \neq x : \kappa(y) = P_{l(x)}\kappa(x) \quad (1.13)$$

Observation 1.15. In this approach is also very simple to find out that prefix codes are non-singular. In fact, if $\kappa(x) = \kappa(y)$ with $x, y \in X_0$ and $x \neq y$, we would have that $\kappa(y) = P_{l(y)}\kappa(x)$ and $\kappa(x) = P_{l(x)}\kappa(y)$ which contrasts the definition of prefix code.

Observation 1.16. It's also true that a prefix code κ is uniquely decodable. In fact, if

$$\kappa^*(x_{i_1}, \dots, x_{i_t}) = \kappa^*(x_{j_1}, \dots, x_{j_s}) \quad (1.14)$$

With $(x_{i_1}, \dots, x_{i_t}) \neq (x_{j_1}, \dots, x_{j_s})$ because e.g. $x_{i_1} \neq x_{j_1}$, then we could have two cases:

- If $l(x_{i_1}) \neq l(x_{j_1})$, then the shortest among $\kappa(x_{i_1})$ and $\kappa(x_{j_1})$ is a prefix of the longest one, which is absurd because κ is a prefix code.
- If $l(x_{i_1}) = l(x_{j_1})$, then $\kappa(x_{i_1}) = \kappa(x_{j_1})$ which is absurd because κ is non-singular, as noted in observation 1.15.

Lemma 1.17. *Let's now consider words of length exactly equal to $M \in \mathbb{N}$:*

$$A_M^{(\infty)} = \{f : \mathbb{N} \rightarrow A \cup \{\infty\} \mid f(n) = \infty \Leftrightarrow n > M\} \quad (1.15)$$

And consider also the set of all M -length extensions of a codeword $\kappa(x)$, where $M \geq l(x)$:

$$A_{M,x}^{(\infty)} = A_M^{(\infty)} \cap P_{l(x)}^{-1}\kappa(x) \quad (1.16)$$

Then κ is a prefix code if and only if for every $x, y \in X_0$ with $x \neq y$ and for all $M \geq \max\{l(x), l(y)\}$ we have that

$$A_{M,x}^{(\infty)} \cap A_{M,y}^{(\infty)} = \emptyset \quad (1.17)$$

Proof. The thesis is quite simple, if we notice that for every $x, y \in X_0$:

$$\kappa(x) \in A_{M,y}^{(\infty)} \Leftrightarrow \kappa(y) \in P_{l(y)}\kappa(x) \quad (1.18)$$

So the thesis follows from the very definition of prefix code. \square

Lemma 1.18. For every $x \in X_0$ and $M \geq l(x)$ we have that:

$$|A_{M,x}^{(\infty)}| = |A|^{M-l(x)} \quad (1.19)$$

Proof. Let's recall that in $A_{M,x}^{(\infty)}$ we have all the extensions of $\kappa(x)$ of length M . So, the first $l(x)$ letters of those words are all equal (and equal to the first $l(x)$ letters of $\kappa(x)$), so are the letters $M+1, \dots$ (because they are all ∞). It follows that we have only $M-l(x)$ letters that may vary and each of them may take every value from the alphabet. From this, the thesis follows trivially. \square

Theorem 1.19. *Kraft's inequality*. Given a prefix code $\kappa : X_0 \rightarrow A^{(\infty)}$ we have that:

$$\sum_{x \in X_0} |A|^{-l(x)} \leq 1 \quad (1.20)$$

Proof. Since all the M -length extensions of $\kappa(x)$ for every $x \in X_0$ and with $M = \max\{l(x), x \in X_0\}$ are of course less or equal than all the M -length words and since κ is a prefix code (so no two different words have common extensions), we have by

the preceding two alblemmata that:

$$\sum_{x \in X_0} |A_{M,x}| \leq |A|^M \Rightarrow \quad (1.21)$$

$$\Rightarrow \sum_{x \in X_0} |A|^{M-l(x)} \leq |A|^M \Rightarrow \quad (1.22)$$

$$\Rightarrow \sum_{x \in X_0} |A|^{-l(x)} \leq 1 \quad (1.23)$$

□

1.4 Appendix C: Extended Kraft's inequality

Theorem 1.20. *Extended Kraft's inequality.* Given an uniquely decodable code $\kappa : X_0 \rightarrow \{0, 1\}^{(\infty)}$ with length function $l : X_0 \rightarrow \mathbb{N}$, the following holds:

$$\sum_{x \in X_0} 2^{-l(x)} \leq 1 \quad (1.24)$$

Proof. Fix an integer $n \in \mathbb{N}$ and consider the code $\kappa^{(n)} : X_0^n \rightarrow \{0, 1\}^{(\infty)}$, naturally defined by:

$$\kappa^{(n)}(x_1, \dots, x_n) = \kappa^*(x_1, \dots, x_n) \quad (1.25)$$

And whose length function is $l^{(n)}(x_1, \dots, x_n) = \sum_{i=1}^n l(x_i)$. We have that:

$$\sum_{(x_1, \dots, x_n) \in X_0^n} 2^{-l^{(n)}(x_1, \dots, x_n)} = \sum_{(x_1, \dots, x_n) \in X_0^n} 2^{-\sum_{i=1}^n l(x_i)} = \quad (1.26)$$

$$= \sum_{(x_1, \dots, x_n) \in X_0^n} \prod_{i=1}^n 2^{-l(x_i)} = \quad (1.27)$$

$$= \left(\sum_{x \in X_0} 2^{-l(x)} \right)^n \quad (1.28)$$

But, on the other hand, for every $(x_1, \dots, x_n) \in X_0^n$ we have that:

$$n \leq l^{(n)}(x_1, \dots, x_k) \leq n \max_{x \in X_0} l(x) := n\bar{l} \quad (1.29)$$

We now rearrange the sum in 1.26 according to the codeword length, in ascending order:

$$\sum_{(x_1, \dots, x_n) \in X_0^n} 2^{-l^{(n)}(x_1, \dots, x_n)} = \sum_{i=1}^{\bar{n}l} |l^{(n)-1}(i)| 2^{-i} \leq \dots \quad (1.30)$$

Now consider that the total number of possible codewords with length i is 2^i and for the unique decodability each of those can be assigned to only one word of X_0^n . So:

$$\dots \leq \sum_{i=1}^{\bar{n}l} 2^i 2^{-i} = n\bar{l} \quad (1.31)$$

So, in conclusion:

$$\sum_{x \in X_0} 2^{-l(x)} \leq (n\bar{l})^{\frac{1}{n}} \quad (1.32)$$

And taking the limit for $n \rightarrow +\infty$ we get:

$$\sum_{x \in X_0} 2^{-l(x)} \leq 1 \quad (1.33)$$

□

2.1 Expected codelength, entropy, divergence

Definition 2.1. Given a finite set $X_0 \neq \emptyset$ we define **sub-probability distribution** on X_0 a function $p : X_0 \rightarrow [0, 1]$ such that

$$\sum_{x \in X_0} p(x) \leq 1 \tag{2.1}$$

So, this is quite similar to a probability distribution, except to the fact that it hasn't to sum to 1 over all the elements of X_0 , but just to some real number in $[0, 1]$. It is clear that probability distributions are a special case of sub-probability distributions.

Observation 2.2. We would like to establish some relation between a prefix code $\kappa : X_0 \rightarrow \{0, 1\}^{(\infty)}$ and a sub-probability distribution on X_0 (and vice-versa) by the mean of the length function of κ , $l : X_0 \rightarrow \mathbb{N}$. In fact, given the code κ , for Kraft's inequality we know that $\sum_{x \in X_0} 2^{-l(x)} \leq 1$ which gives immediately a sub-probability distribution $p(x) = 2^{-l(x)}$. If instead we have a sub-probability distribution p , we have

that:

$$\sum_{x \in X_0} p(x) \leq 1 \Rightarrow \quad (2.2)$$

$$\Rightarrow \sum_{x \in X_0} 2^{\log_2 p(x)} \leq 1 \Rightarrow \quad (2.3)$$

$$\Rightarrow \sum_{x \in X_0} 2^{\lfloor \log_2 p(x) \rfloor} \leq 1 \Rightarrow \quad (2.4)$$

$$\Rightarrow \sum_{x \in X_0} 2^{-\lceil -\log_2 p(x) \rceil} \leq 1 \quad (2.5)$$

And so, for theorem 1.10 there exists a prefix code κ which has length function $l(x) = \lceil -\log_2 p(x) \rceil$. Unfortunately, since here we are rounding we can't go back and forth from prefix codes to subprobability distribution: $\kappa \mapsto p$ is not a bijection, as we defined it. A work-around for this could be that we simply forget about the meaning of the length function in terms of letters and words and we allow it to take any value in \mathbb{R}^+ .

Definition 2.3. Given a function $f : \mathbb{N} \rightarrow [0, 1]$ we define its **support** as:

$$\text{supp}(f) = \{n \in \mathbb{N} : f(n) \neq 0\} \quad (2.6)$$

Note 2.4. Since we're only working with codes on finite discrete sets X_0 , for simplicity from now on, we'll identify them with subsets $\{1, \dots, t\}$ of the natural numbers.

Definition 2.5. Let's consider the **space of sub-probability distributions** on finite sets:

$$\underline{S} = \{p : \mathbb{N} \rightarrow [0, 1] : \text{supp}(p) < \aleph_0 \wedge \sum_{n \in \mathbb{N}} p(n) \leq 1\} \quad (2.7)$$

and its subset which contains only the probability distributions:

$$S = \{p : \mathbb{N} \rightarrow [0, 1] : \text{supp}(p) < \aleph_0 \wedge \sum_{n \in \mathbb{N}} p(n) = 1\} \quad (2.8)$$

Also, for every finite set $X_0 \subset \mathbb{N}$, we define:

$$\underline{S}_{X_0} = \{p : X_0 \rightarrow [0, 1] : \sum_{n \in X_0} p(n) \leq 1\} \quad (2.9)$$

Of course we can embed every $p \in \underline{S}_{X_0}$ into a function of \underline{S} , via the i_{X_0} operator:

$$i_{X_0}p(n) = \begin{cases} p(n) & \text{if } n \in X_0 \\ 0 & \text{if } n \notin X_0 \end{cases} \quad (2.10)$$

Let's equip $S_{X_0} \subset \mathbb{R}^{|X_0|}$ with the euclidean topology and \underline{S} with the topology induced by the maps $i_{X_0} : \underline{S}_{X_0} \rightarrow \underline{S}$ for every finite $X_0 \subset \mathbb{N}$, i.e. the strongest topology such that all the i_{X_0} are continuous maps.

Observation 2.6. Let's stop a moment and think about the real-case scenario in which usually we use code theory. We generally have a sender and a receiver, that communicate through a (noisy) channel. The sender has a set of possible inputs, encodes some of them with a code, sends them through the channel and at last they are decoded by the receiver (who expects to read exactly what the sender intended to communicate). So, it's clear that a lot of *probabilities* are involved: the probability for one input of the set to be sent by the receiver, the probability that noise will corrupt the input, the probability that the decoded output is what the sender has actually sent (this latter we should want to maximize, regardless of the former). Also, we've seen that there is a sub-probability associated with each code - or, to be more precise, with the length of the codewords. So, if we want an estimate of how long a generic codeword is, we have to take into account both the probability (whose distribution is e.g. $p \in S$) that each codeword has to be chosen by the sender to be put in the channel and the sub-probability (whose distribution is e.g. $q \in \underline{S}$) associated with the length of the codewords. This is exactly what we do in the following definition.

Definition 2.7. We define the **expected code-length function** $\Phi : S \times \underline{S} \rightarrow [0, 1]$

as:

$$\Phi(p, q) = \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n)(-\log_2 q(n)) = \quad (2.11)$$

$$= - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \log_2 q(n) = \quad (2.12)$$

$$:= \Phi(p||q) \quad (2.13)$$

Definition 2.8. Observe that we can write the expected code-length as:

$$\Phi(p||q) = - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \log_2 q(n) = \quad (2.14)$$

$$= - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \log_2 \left(p(n) \frac{q(n)}{p(n)} \right) = \quad (2.15)$$

$$= - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \log_2 p(n) - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \log_2 \frac{q(n)}{p(n)} \quad (2.16)$$

We call **entropy** of p the term:

$$H(p) = - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \log_2 p(n) \quad (2.17)$$

and **(Kullback-Leibler) divergence** - or relative entropy - between p and q the term:

$$D(p||q) = - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \log_2 \frac{q(n)}{p(n)} \quad (2.18)$$

If $X : \Omega \rightarrow X_0$ is the random variable described by p , we sometimes write $H(X)$ instead of $H(p)$.

Theorem 2.9. For all $p \in \mathcal{S}$ and $q \in \underline{\mathcal{S}}$ we have that $D(p||q) \geq 0$ with equality if and only if $p = q$.

Proof. Recall that $\forall x > 0 \log_2 x \leq x - 1$. So:

$$D(p||q) = - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \log_2 \frac{q(n)}{p(n)} \geq \quad (2.19)$$

$$\geq - \sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n) \left(\frac{q(n)}{p(n)} - 1 \right) = \quad (2.20)$$

$$= - \underbrace{\sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} q(n)}_{\leq 1} + \underbrace{\sum_{\substack{n \in \mathbb{N} \\ p(n) \neq 0}} p(n)}_{=1} \geq 0 \quad (2.21)$$

And in 2.19 the equality holds if and only if $p(n) = q(n)$. \square

Corollary 2.10. 2.12 From the previous theorem it follows immediately that for every $p \in S$ and $q \in \underline{S}$

$$\Phi(p||q) \geq H(p) \quad (2.22)$$

And the equality holds if and only if $p = q$.

Theorem 2.11. $H : S \rightarrow \mathbb{R}^+$ is continuous and concave.

Proof. For how we defined the topology on \underline{S} (and so on $S \subset \underline{S}$) we have that H is continuous if $H \circ i_{X_0}$ is continuous, for every finite X_0 , which is of course true, since

$$(H \circ i_{X_0})(p) = - \sum_{n \in X_0} p(n) \log_2 p(n) \quad (2.23)$$

And the function $x \mapsto x \log_2 x$ is continue and convex. \square

Theorem 2.12. Given $p \in S$, there exists a prefix code $\kappa : \text{supp}(f) \rightarrow \{0, 1\}^{(\infty)}$ with associated sub-probability distribution $q \in \underline{S}$, such that:

$$H(p) \leq \Phi(p||q) < H(p) + 1 \quad (2.24)$$

Proof. Let $l : \text{supp}(p) \rightarrow \mathbb{N}$ be defined by:

$$l(n) = \lceil -\log_2 p(n) \rceil \quad (2.25)$$

Then we have that:

$$\sum_{n \in \text{supp}(p)} 2^{-l(n)} = \sum_{n \in \text{supp}(p)} 2^{-\lceil -\log_2 p(n) \rceil} = \quad (2.26)$$

$$= \sum_{n \in \text{supp}(p)} 2^{\lfloor \log_2 p(n) \rfloor} \leq \quad (2.27)$$

$$= \sum_{n \in \text{supp}(p)} p(n) = \quad (2.28)$$

$$= 1 \quad (2.29)$$

So $l(n)$ satisfies Kraft's inequality and this mean, by theorem 1.10, that there is a prefix code $\kappa : \text{supp}(p) \rightarrow \{0, 1\}^{(\infty)}$ such that l is its length function. The sub-probability distribution associated with κ is $q(x) = 2^{-l(x)}$. Now, for corollary , we have that $H(p) \leq \Phi(p|q)$, so the first inequality is satisfied. Moreover:

$$\Phi(p|q) = \sum_{n \in \text{supp}(p)} p(n)(-\log_2 q(n)) = \quad (2.30)$$

$$= \sum_{n \in \text{supp}(p)} p(n)l(n) = \quad (2.31)$$

$$= \sum_{n \in \text{supp}(p)} p(n)\lceil -\log_2 p(n) \rceil = \quad (2.32)$$

$$= - \sum_{n \in \text{supp}(p)} p(n)\lfloor \log_2 p(n) \rfloor < \quad (2.33)$$

$$< - \sum_{n \in \text{supp}(p)} p(n)(\log_2 p(n) - 1) = \quad (2.34)$$

$$= - \sum_{n \in \text{supp}(p)} p(n) \log_2 p(n) + \sum_{n \in \text{supp}(p)} p(n) = \quad (2.35)$$

$$= H(p) + 1 \quad (2.36)$$

□

Theorem 2.13. *The entropy $H(p)$ has local maxima for $p = u$, where $u : X_0 \rightarrow [0, 1]$ is the uniform distribution $u(n) = |X_0|^{-1} \forall n \in X_0$; moreover $H(u) = \log_2 |X_0|$.*

Proof. Consider u as a sub-probability distribution. Then, for every $p \in S$ such that

$\text{supp}(p) \subseteq X_0$, we have that:

$$0 \leq D(p||u) = \Phi(p||q) - H(p) = \quad (2.37)$$

$$= \left(- \sum_{n \in \text{supp}(p)} p(n) \log_2 |X_0|^{-1} \right) - H(p) = \quad (2.38)$$

$$= \left(- \log_2 |X_0|^{-1} \sum_{n \in \text{supp}(p)} p(n) \right) - H(p) = \quad (2.39)$$

$$= - \log_2 |X_0|^{-1} - H(p) = \quad (2.40)$$

$$= \log_2 |X_0| - H(p) \Rightarrow \quad (2.41)$$

$$\Rightarrow H(p) \leq \log_2 |X_0| \quad (2.42)$$

□

Note 2.14. From now on, when we have more than one random variable - e.g. X_1, \dots, X_n , we'll use the letter p to denote all the probability distribution associated with the random variables and also the joint probability distribution. To allow unambiguity, when we write $p(x_i)$ or $p_{X_i}(x)$ we'll be referring to the probability distribution of X_i and when we write $p(x_1, \dots, x_n)$ we'll be referring to the joint probability distribution.

Definition 2.15. Given two random variables $X_1, X_2 : \Omega \rightarrow X_0$ with joint probability distribution $p : X_0^2 \rightarrow [0, 1]$, we define the **conditional entropy** of $X_2|X_1$ as follows:

$$H(X_2|X_1) = \sum_{x_1 \in X_0} p(x_1) H(X_2|X_1 = x_1) = \quad (2.43)$$

$$= \sum_{x_1 \in X_0} p(x_1) \left(\sum_{x_2 \in X_0} p(x_2|x_1) \log_2 p(x_2|x_1) \right) = \quad (2.44)$$

$$= \sum_{x_1 \in X_0} \sum_{x_2 \in X_0} p(x_1, x_2) \log_2 p(x_2|x_1) \quad (2.45)$$

Theorem 2.16. **Entropy chain rule.** Given the random variables $X_1, \dots, X_n : \Omega \rightarrow$

X_0 with joint probability distribution $p : X_0^n \rightarrow [0, 1]$, if we call $H(X_1, \dots, X_n) := H(p)$, we have that:

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \quad (2.46)$$

Proof. First of all, recall that the joint distribution can be written as:

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i | x_{i-1}, \dots, x_1) \quad (2.47)$$

Now, from the definition of entropy:

$$H(X_1, \dots, X_n) = - \sum_{(x_1, \dots, x_n) \in X_0^n} p(x_1, \dots, x_n) \log_2 p(x_1, \dots, x_n) = \quad (2.48)$$

$$= - \sum_{(x_1, \dots, x_n) \in X_0^n} p(x_1, \dots, x_n) \log_2 \left(\prod_{i=1}^n p(x_i | x_{i-1}, \dots, x_1) \right) = \quad (2.49)$$

$$= - \sum_{(x_1, \dots, x_n) \in X_0^n} \sum_{i=1}^n p(x_1, \dots, x_n) \log_2 p(x_i | x_{i-1}, \dots, x_1) = \quad (2.50)$$

$$= - \sum_{i=1}^n \sum_{(x_1, \dots, x_n) \in X_0^n} p(x_1, \dots, x_n) \log_2 p(x_i | x_{i-1}, \dots, x_1) = \quad (2.51)$$

$$= - \sum_{i=1}^n \sum_{(x_1, \dots, x_i) \in X_0^i} p(x_1, \dots, x_i) \log_2 p(x_i | x_{i-1}, \dots, x_1) = \quad (2.52)$$

$$= - \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \quad (2.53)$$

□

Corollary 2.17. Notice that if we only have two random variable, the chain rule assumes the form:

$$H(X_1, X_2) = H(X_1) + H(X_2 | X_1) \quad (2.54)$$

2.2 Mutual information and data-processing

Definition 2.18. Given two random variables $X_1, X_2 : \Omega \rightarrow X_0$ with joint probability distribution $p : X_0^2 \rightarrow [0, 1]$, we define the **mutual information** of X_1 and X_2 as:

$$I(X_1; X_2) = D(p \parallel p_{X_1} \otimes p_{X_2}) = \sum_{(x_1, x_2) \in X_0^2} p(x_1, x_2) \log_2 \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \quad (2.55)$$

Where $p_{X_1} \otimes p_{X_2}$ is the product distribution of X_1 and X_2 :

$$(p_{X_1} \otimes p_{X_2})(x_1, x_2) = p(x_1)p(x_2) \quad (2.56)$$

Observation 2.19. Intuitively, D measures the *distance* between two distributions. Here, we're measuring the distance between $p(x_1, x_2)$ and $p(x_1)p(x_2)$, that is how far they are from being independent. Indeed, if they *are* independent, by definition we have that $p(x_1, x_2) = p(x_1)p(x_2)$ and so $I(X_1; X_2) = 0$.

Definition 2.20. Given three random variables $X_1, X_2, X_3 : \Omega \rightarrow X_0$ with joint probability distribution $p : X_0^3 \rightarrow [0, 1]$, we say that they form a **Markov chain** if:

$$\frac{p(x_1, x_2, x_3)}{p(x_2)} = \frac{p(x_1, x_2)}{p(x_2)} \frac{p(x_2, x_3)}{p(x_2)} \quad \forall x_2 : p(x_2) \neq 0 \quad (2.57)$$

We say that X_1 and X_3 are independent given X_2 and we write $X_1 \rightarrow X_2 \rightarrow X_3$.

Observation 2.21. Given the previous definition, we expect that:

1. (X_2, X_3) should contain the same information about X_1 as the only X_2 .
2. (X_2, X_3) should contain more information about X_1 as the only X_3 .

For example, if X_1 represents a physical phenomenon, X_2 the data collected by an observer and X_3 the final result calculated from the data, we don't expect to have in the final result any more information than the one we have in the raw data. Summarizing, we expect that the mutual information between X_1 and X_3 is at most equal to the one between X_2 and X_3 , but never greater. We will prove this in the next theorem.

Theorem 2.22. *Data processing inequality.* Given a Markov chain $X_1 \rightarrow X_2 \rightarrow X_3$, then $I(X_1; X_2) \geq I(X_1; X_3)$.

Proof. Consider the quantity $I(X_1; (X_2, X_3))$. We have that:

$$I(X_1; (X_2, X_3)) = \sum_{(x_1, x_2, x_3) \in X_0^3} p(x_1, x_2, x_3) \log_2 \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2, x_3)} = \quad (2.58)$$

$$= \sum_{\substack{(x_1, x_2, x_3) \in X_0^3 \\ p(x_2) \neq 0}} p(x_1, x_2, x_3) \log_2 \left(\frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2)} \frac{p(x_2)}{p(x_2, x_3)} \right) = \quad (2.59)$$

$$= \sum_{\substack{(x_1, x_2, x_3) \in X_0^3 \\ p(x_2) \neq 0}} p(x_1, x_2, x_3) \log_2 \left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right) = \quad (2.60)$$

$$= I(X_1; X_2) \quad (2.61)$$

On the other hand:

$$I(X_1; (X_2, X_3)) = \sum_{(x_1, x_3) \in X_0^2} \sum_{x_2 \in X_0} p(x_1, x_2, x_3) \log_2 \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2, x_3)} \quad (2.62)$$

And, applying the well-known *log sum inequality*, we get:

$$I(X_1; (X_2, X_3)) = \sum_{(x_1, x_2, x_3) \in X_0^3} p(x_1, x_2, x_3) \log_2 \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2, x_3)} \geq \quad (2.63)$$

$$\geq \sum_{(x_1, x_3) \in X_0^2} \left(\left(\sum_{x_2 \in X_0} p(x_1, x_2, x_3) \right) \log_2 \frac{\sum_{x_2 \in X_0} p(x_1, x_2, x_3)}{\sum_{x_2 \in X_0} p(x_1)p(x_2, x_3)} \right) = \quad (2.64)$$

$$= \sum_{x_1, x_3 \in X_0} p(x_1, x_3) \log_2 \frac{p(x_1, x_3)}{p(x_1)p(x_3)} = \quad (2.65)$$

$$= I(X_1; X_3) \quad (2.66)$$

□

Corollary 2.23. Given two random variables $X_1, X_2 : \Omega \rightarrow X_0$, then $I(X_1; X_2) \leq H(X_1)$.

Proof. Consider the Markov chain $X_1 \rightarrow X_1 \rightarrow X_2$. Applying theorem 2.22 we get:

$$I({}_1X; X_1) \geq I(X_1; X_2) \quad (2.67)$$

And the thesis follows, since $I(X_1; X_1) = H(X_1)$. \square

Theorem 2.24. *Given two random variables $X_1, X_2 : \Omega \rightarrow X_0$, consider the diagonal of X_0^2 :*

$$\Delta = \{(x, x) \forall x \in X_0\} \quad (2.68)$$

And its characteristic probability distribution χ_Δ ; then we have that:

$$H(X_1, X_2) \leq H(\chi_\Delta) + \left(\sum_{(x_1, x_2) \in X_0 - \Delta} p(x_1, x_2) \right) \log_2 |X_0| \quad (2.69)$$

Proof. We will just use a long chain of equalities and inequalities; to simplify notations, let:

$$A = \{(x_1, x_2) \in X_0^2 - \Delta : p(x_2) \neq 0\} \quad (2.70)$$

$$B = \{(x_1, x_2) \in \Delta : p(x_2) \neq 0\} \quad (2.71)$$

$$C = \{(x_1, x_2) \in X_0^2 : p(x_2) \neq 0\} = A \cup B \quad (2.72)$$

Then we have that:

$$H(X_1|X_2) = - \sum_{(x_1, x_2) \in C} p(x_1, x_2) \log_2 \frac{p(x_1, x_2)}{p(x_2)} = \quad (2.73)$$

$$= - \sum_{(x_1, x_2) \in A} p(x_1, x_2) \log_2 \frac{p(x_1, x_2)}{p(x_2)} -$$

$$- \sum_{(x_1, x_2) \in B} p(x_1, x_2) \log_2 \frac{p(x_1, x_2)}{p(x_2)} \leq \quad (2.74)$$

$$\leq - \left(\sum_{(x_1, x_2) \in A} p(x_1, x_2) \right) \log_2 \underbrace{\frac{\sum_{(x_1, x_2) \in A} p(x_1, x_2)}{\sum_{(x_1, x_2) \in A} p(x_2)}}_{< 1} -$$

$$- \left(\sum_{(x_1, x_2) \in B} p(x_1, x_2) \right) \log_2 \frac{\sum_{(x_1, x_2) \in B} p(x_1, x_2)}{\underbrace{\sum_{(x_1, x_2) \in B} p(x_2)}_{<|B|=|X_0|}} \leq \quad (2.75)$$

$$\leq - \left(\sum_{(x_1, x_2) \in A} p(x_1, x_2) \right) \log_2 \sum_{(x_1, x_2) \in A} p(x_1, x_2) -$$

$$- \left(\sum_{(x_1, x_2) \in B} p(x_1, x_2) \right) \log_2 \frac{\sum_{(x_1, x_2) \in B} p(x_1, x_2)}{|X_0|} = \quad (2.76)$$

$$= - \left(\sum_{(x_1, x_2) \in A} p(x_1, x_2) \right) \log_2 \sum_{(x_1, x_2) \in A} p(x_1, x_2) -$$

$$- \left(\sum_{(x_1, x_2) \in B} p(x_1, x_2) \right) \log_2 \sum_{(x_1, x_2) \in B} p(x_1, x_2) +$$

$$+ \left(\sum_{(x_1, x_2) \in A} p(x_1, x_2) \right) \log_2 |X_0| = \quad (2.77)$$

$$= H(\chi_\Delta) + \left(\sum_{(x_1, x_2) \in A} p(x_1, x_2) \right) \log_2 |X_0| \quad (2.78)$$

□

Corollary 2.25. *Fano's inequality*. Given a Markov chain $X_1 \rightarrow X_2 \rightarrow X_3$, we call $P_e = P(X_1 \neq X_3)$. We will prove that:

$$H(X_1|X_2) \leq -P_e \log_2 P_e - (1 - P_e) \log_2(1 - P_e) + P_e \log_2 |X_0| \quad (2.79)$$

Proof. First notice that:

$$I(X; Y) = \sum_{(x_1, x_2) \in X_0^2} p(x_1, x_2) \log_2 \frac{p(x_1, x_2)}{p(x_1)p(x_2)} = \quad (2.80)$$

$$= \sum_{(x_1, x_2) \in X_0^2} p(x_1, x_2) \log_2 \frac{p(x_1|x_2)}{p(x_1)} = \quad (2.81)$$

$$= - \sum_{(x_1, x_2) \in X_0^2} p(x_1, x_2) \log_2 p(x_1) - \left(- \sum_{(x_1, x_2) \in X_0^2} p(x_1, x_2) \log_2 p(x_1 | x_2) \right) = \quad (2.82)$$

$$= - \sum_{x_1 \in X_0^2} p(x_1) \log_2 p(x_1) - \left(- \sum_{(x_1, x_2) \in X_0^2} p(x_1, x_2) \log_2 p(x_1 | x_2) \right) = \quad (2.83)$$

$$= H(X_1) - H(X_1 | X_2) \quad (2.84)$$

And so, using the *data processing inequality* and theorem 2.24:

$$H(X_1 | X_2) = H(X_1) - I(X_1; X_2) \leq \quad (2.85)$$

$$= H(X_1) - I(X_1; X_3) \leq \quad (2.86)$$

$$\leq -P_e \log_2 P_e - (1 - P_e) \log_2 (1 - P_e) + P_e \log_2 |X_0| \quad (2.87)$$

□

2.3 Appendix A: Huffman codes

Definition 2.26. Given a finite set X_0 and a probability distribution $p : X_0 \rightarrow [0, 1]$, we give the following algorithm to produce a code $\kappa : X_0 \rightarrow \{0, 1\}^{(\infty)}$.

- Let $p_0 = p$
- Given X_0, \dots, X_{n-1} and p_0, \dots, p_{n-1} we define X_n and p_n as follows:

– Let:

$$x = \operatorname{argmin}_{y \in X_{n-1}} p_{n-1}(y)$$

$$x' = \operatorname{argmin}_{y \in X_{n-1} - \{x\}} p_{n-1}(y)$$

– Let $X_n = X_{n-1} - \{x, x'\} \cup \{\{x, x'\}\}$.

– Let:

$$p_n(\bar{x}) = \begin{cases} p_{n-1}(\bar{x}) & \text{if } \bar{x} \in X_{n-1} \\ p_{n-1}(x) + p_{n-1}(x') & \text{if } \bar{x} = \{x, x'\} \end{cases}$$

So that p_n is still a probability distribution on X_n .

– If $|X_n| = 1$, we have finished.

• Let's say that we finished the above iterations at step s , then let:

$$N = \bigsqcup_{n=1}^s X_n = \bigcup_{n=1}^s \{(x, n), \forall x \in X_n\}$$

$$E = \{((x, n), (y, n+1)) : 0 \leq n < s \wedge ((x = y) \vee (x \in y))\}$$

- Let $T = (N, E)$ be the tree with nodes N and edges E .
- For each node (z, n) such that $|\{((x, n), (y, n+1)) \in E : y = z\}| = 2$ we label its two descending edges with 0 and 1. Notice that, by construction, fixed $n \in \{2, \dots, s\}$ there is exactly one such (z, n) .

Now, to every element x of X_0 we assign the codeword obtained by the juxtaposition of the labels encountered on the branches from the root of the tree to node x . The such created code is called an [Huffman code](#).

Theorem 2.27. *Huffman codes minimize the average code length, over all the prefix codes on some finite set X_0 .*

Asymptotic equipartition property

Note 3.1. In the very beginning of this chapter, we'll give some formal definition of basic probability theory structures that we have already used - and which we assumed the reader was familiar with. Then we'll state the weak law of large numbers in terms of our definitions and we'll talk about the asymptotic equipartition property (aep). Anyway, even in this chapter we won't start from the very beginning, but we'll assume that the reader has a basic knowledge of measure theory.

Definition 3.2. A triple $(\Omega, \mathcal{B}, \mu)$ is called a **probability space** if it is a measure space with base set Ω , σ -algebra of measurable sets \mathcal{B} and measure $\mu : \mathcal{B} \rightarrow \mathbb{R}^+$ such that $\mu(\Omega) = 1$. Given such a probability space and a measure space (X_0, \mathcal{X}, η) we call a **random variable** a function $X : \Omega \rightarrow X_0$ such that it is measurable, i.e. that:

$$\forall Y \in \mathcal{X} \quad X^{-1}(Y) \in \mathcal{B} \quad (3.1)$$

We say that a measure $p : X_0 \rightarrow \mathbb{R}^+$ over (X_0, \mathcal{X}) is the **probability distribution** associated with X if:

$$\forall Y \in \mathcal{X} \quad p(Y) = \mu(X^{-1}(Y)) \quad (3.2)$$

To say that p is the probability distribution associated with X we write $X \sim p$.

Theorem 3.3. Let $(\Omega_n, \mathcal{B}_n, \mu_n)$ for $n \in \mathbb{N}^+$ be a sequence of probability spaces. Consider their product $(\Omega_{\mathbb{N}}, \mathcal{B}_{\mathbb{N}}, \mu_{\mathbb{N}})$ where:

$$\Omega_{\mathbb{N}} = \prod_{n \in \mathbb{N}^+} \Omega_n \quad (3.3)$$

$$\mathcal{B}_{\mathbb{N}} = \bigotimes_{n \in \mathbb{N}^+} \mathcal{B}_n \quad (3.4)$$

We want to show that it is possible to chose a probability measure $\mu_{\mathbb{N}}$, that we'll denote also with $\bigotimes_{n \in \mathbb{N}^+} \mu_n$ in a way such that, if we call $\pi_n : \Omega_{\mathbb{N}} \rightarrow \Omega_n$ the canonical projections, we have that $\forall n \in \mathbb{N}^+$ and $\forall (A_1, \dots, A_n) \in \bigotimes_{i=1}^n \mathcal{B}_n$:

$$\mu_{\mathbb{N}}(\pi_1^{-1}(A_1) \cap \dots \cap \pi_n^{-1}(A_n)) = \mu_1(A_1) \cdot \dots \cdot \mu_n(A_n) \quad (3.5)$$

A proof of this theorem is due to Alexandra Ionescu Tulcea and can be found in [2].

Definition 3.4. Given a probability space $(\Omega, \mathcal{B}, \mu)$, consider the sequence of random variables:

$$\{X_n\}_{n \in \mathbb{N}^+} : X_n : \Omega \rightarrow X_0 \quad \forall n \in \mathbb{N}^+ \quad (3.6)$$

And the random variable $X : \Omega \rightarrow X_0$. We say that the sequence $\{X_n\}$ (that we may sometimes denote as (X_n)) **converges in measure** to X if, $\forall \varepsilon > 0$:

$$\lim_{n \rightarrow +\infty} \mu(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0 \quad (3.7)$$

We write in this case: $X_n \xrightarrow{m.} X$.

Theorem 3.5. Weak law of large numbers (wlln). Given a finite probability space $(\Omega_1, \mathcal{B}_1, \mu_1)$ consider the product probability space of \mathbb{N} copies of it, $(\Omega_{\mathbb{N}}, \mathcal{B}_{\mathbb{N}}, \mu_{\mathbb{N}})$, built as in theorem 3.3. Let $X : \Omega_1 \rightarrow X_0 \subset \mathbb{R}$ be a random variable and consider the random variables $X_n : \Omega_{\mathbb{N}} \rightarrow X_0$ for $n \in \mathbb{N}$, defined by:

$$X_n(\omega) = X(\pi_n(\omega)) \quad (3.8)$$

Where $\pi_n : \Omega_{\mathbb{N}} \rightarrow \Omega_1$ is the n -th canonical projection (i.e. the n -th coordinate of ω). Notice that this random variables are independent and identically distributed with distribution $p : X_0 \rightarrow [0, 1]$. Let e be the mean value of X :

$$e = \sum_{\omega \in \Omega} X(\omega) \mu_1(\omega) \quad (3.9)$$

And consider it as a constant random variable on $\Omega_{\mathbb{N}}$. Then, the wlln states that:

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{m.} e \quad (3.10)$$

Corollary 3.6. Asymptotic equipartition property (aep). In the same notations of theorem 3.5 we have that:

$$-\frac{1}{n} \log_2 (p(X_1) \dots p(X_n)) \xrightarrow{m.} H(p) \quad (3.11)$$

Definition 3.7. In the notations used until now, given $n \in \mathbb{N}$ and $\varepsilon > 0$ we define the (n, ε) -typical set :

$$A_\varepsilon^{(n)} = \left\{ (x_1, \dots, x_n) \in X_0^n : \left| H(p) + \frac{1}{n} \log_2 (p(x_1) \dots p(x_n)) \right| \leq \varepsilon \right\} = \quad (3.12)$$

$$= \left\{ (x_1, \dots, x_n) \in X_0^n : 2^{-n(H(p)+\varepsilon)} \leq p(x_1) \dots p(x_n) \leq 2^{-n(H(p)-\varepsilon)} \right\} \quad (3.13)$$

- convergence in measure, 26
- data processing inequality, 20
- divergence, 14
- entropy, 14
 - chain rule, 17
 - conditional entropy, 17
 - relative entropy, 14
- expected code-length function, 13
- Fano's inequality, 22
- Huffman codes, 23
- Kraft's inequality, 4, 8
 - extended, 9
- Kullback-Leibler divergence, 14
- Markov chain, 19
- mutual information, 19
- poset, 6
- antichain, 6
- probability distribution, 25
- probability space, 25
- random variable, 25
- source code, 2
 - finite-length extension, 3
 - prefix code, 3, 6
- sub-probability distribution, 11
 - as a topological space, 12
- support, 12
- tree, 6
- typical set, 27

Bibliography

- [1] Thomas Cover and Joy Thomas. *Elements of information theory*. Wiley, 2006.
- [2] A. Ionescu Tulcea. On pointwise convergence, compactness and equicontinuity in the lifting topology. i. *Probability Theory and Related Fields*, 26:197–205, 1973. 10.1007/BF00532722.