Lecture Notes in Mathematics

1851

Olivier Catoni

Statistical Learning Theory and Stochastic Optimization

Ecole d'Eté de Probabilités de Saint-Flour XXXI – 2001

Editor: J. Picard

Springer

Olivier Catoni

Statistical Learning Theory and Stochastic Optimization

Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001

Editor: Jean Picard

Springer

Author

Olivier Catoni

Laboratoire de Probabilités
et Modèles Aléatoires
UMR CNRS 7599, Case 188
Université Paris 6
4, place Jussieu
75252 Paris Cedex 05
France

e-mail: catoni@ccr.jussieu.fr

Editor

Jean Picard

Laboratoire de Mathématiques Appliquées
UMR CNRS 6620
Université Blaise Pascal Clermont-Ferrand
63177 Aubière Cedex, France

e-mail: Jean.Picard@math.univ-bpclermont.fr

The lectures of this volume are the second part of the St. Flour XXXI-2001 volume that has appeared as LNM 1837.

Cover picture: Blaise Pascal (1623-1662)

Library of Congress Control Number: 2004109143

Mathematics Subject Classification (2000):
62B10, 68T05, 62C05, 62E17, 62G05, 62G07, 62G08, 62H30, 62J02, 94A15, 94A17, 94A24, 68Q32, 60F10, 60J10, 60J20, 65C05, 68W20.

ISSN 0075-8434
ISBN 3-540-22572-2 Springer Berlin Heidelberg New York
DOI: 10.1007/b99352

Springer is a part of Springer Science + Business Media
www.springeronline.com

Printed in Germany

Typesetting: Camera-ready TEX output by the authors

41/3142/du - 543210 - Printed on acid-free paper

Preface

Three series of lectures were given at the 31st Probability Summer School in Saint-Flour (July 8–25, 2001), by the Professors Catoni, Tavaré and Zeitouni. In order to keep the size of the volume not too large, we have decided to split the publication of these courses into two parts. This volume contains the course of Professor Catoni. The courses of Professors Tavaré and Zeitouni have been published in the *Lecture Notes in Mathematics*. We thank all the authors warmly for their important contribution.

55 participants have attended this school. 22 of them have given a short lecture. The lists of participants and of short lectures are enclosed at the end of the volume.

Finally, we give the numbers of volumes of Springer *Lecture Notes* where previous schools were published.

Lecture Notes in Mathematics

1971: vol 307	1973: vol 390	1974: vol 480	1975: vol 539
1976: vol 598	1977: vol 678	1978: vol 774	1979: vol 876
1980: vol 929	1981: vol 976	1982: vol 1097	1983: vol 1117
1984: vol 1180	1985/86/87: vol 1362	1988: vol 1427	1989: vol 1464
1990: vol 1527	1991: vol 1541	1992: vol 1581	1993: vol 1608
1994: vol 1648	1995: vol 1690	1996: vol 1665	1997: vol 1717
1998: vol 1738	1999: vol 1781	2000: vol 1816	2001: vol 1837
2002: vol 1840			

Lecture Notes in Statistics

1986: vol 50

Jean Picard, Université Blaise Pascal
Chairman of the summer school

Contents

Introduction[1]

The main purpose of these lectures will be to estimate a probability distribution $P \in \mathcal{M}_+^1(\mathcal{Z})$ from an observed sample $(Z_1, \ldots, Z_N)$ distributed according to $P^{\otimes N}$. (The notation $\mathcal{M}_+^1(\mathcal{Z}, \mathcal{F})$ will stand throughout these notes for the set of probability distributions on the measurable space $(\mathcal{Z}, \mathcal{F})$ — the sigma-algebra $\mathcal{F}$ will be omitted when there is no ambiguity about its choice). In a regression estimation problem, $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ will be a set of two random variables, and the distribution to be estimated will rather be the conditional probability distribution $P(dY \mid X)$, or even only its mode (when $\mathcal{Y}$ is a finite set) or its mean (when $\mathcal{Y} = \mathbb{R}$ is the real line). A large number of pattern recognition problems could be formalized within this framework. In this case, the random variable Y_i takes a finite number of values, representing the different "labels" into which the "patterns" X_i are to be classified. The patterns may for instance be digital signals or images.

A major role will be played in our study by the risk function

$$R(Q) = \mathcal{K}(P, Q) \stackrel{\text{def}}{=} \begin{cases} \mathbb{E}_P \left(\log \frac{P}{Q} \right) & \text{if } P \ll Q \\ +\infty & \text{otherwise} \end{cases}, \qquad Q \in \mathcal{M}_+^1(\mathcal{Z}).$$

Let us remind that the function $\mathcal{K}$ is known as the Kullback divergence function, or relative entropy, that it is non negative and cancels only on the set $P = Q$. To see this, it is enough to remember that, whenever it is finite, the Kullback divergence can also be expressed as

$$\mathcal{K}(P, Q) = \mathbb{E}_Q \left[1 - \frac{P}{Q} + \frac{P}{Q} \log \left(\frac{P}{Q} \right) \right]$$

and that the map $r \mapsto 1 - r + r \log(r)$ is non negative, strictly convex on $\mathbb{R}_+$ and cancels only at point $r = 1$.

[1] I would like to thank the organizers of the Saint-Flour summer school for making possible this so welcoming and rewarding event year after year. I am also grateful to the participants for their kind interest and their useful comments.

In the case of regression estimation and pattern recognition, we will also use risk functions of the type

$$R(f) = \mathbb{E}\big[d\big(f(X), Y\big)\big], \qquad f : \mathcal{X} \to \mathcal{Y},$$

where d is a non negative function measuring the discrepancy between Y and its estimate $f(X)$ by a function of X. We will more specifically focuss on two loss functions : the quadratic risk $d\big(f(X), Y\big) = (f(X) - Y)^2$ in the case when $\mathcal{Y} = \mathbb{R}$, and the error indicator function $d\big(f(X), Y\big) = \mathbb{1}\big(f(X) \neq Y\big)$ in the case of pattern recognition.

Our aim will be to prove, for well chosen estimators $\hat{P}(Z_1, \ldots, Z_N) \in \mathcal{M}_+^1(\mathcal{Z})$ [resp. $\hat{f}(Z_1, \ldots, Z_N) \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$], non asymptotic oracle inequalities. Oracle inequalities is a point of view on statistical inference introduced by David Donoho and Iain Johnstone. It consists in making no (or few) restrictive assumptions on the nature of the distribution P of the observed sample, and to restrict instead the choice of an estimator $\hat{P}$ to a subset $\{P_\theta : \theta \in \Theta\}$ of the set $\mathcal{M}_+^1(\mathcal{Z})$ of all probability distributions defined on $\mathcal{Z}$ [resp. to restrict the choice of a regression function $\hat{f}$ to a subset $\{f_\theta : \theta \in \Theta\}$ of all the possible measurable functions from $\mathcal{X}$ to $\mathcal{Y}$]. The estimator $\hat{P}$ is then required to approximate P almost as well as the best distribution in the estimator set $\{P_\theta : \theta \in \Theta\}$ [resp. The regression function $\hat{f}$ is required to minimize as much as possible the risk $R(f_\theta)$, within the regression model $\{f_\theta : \theta \in \Theta\}$]. This point of view is well suited to "complex" data analysis (such as speach recognition, DNA sequence modeling, digital image processing, ...) where it is crucial to get quantitative estimates of the performance of approximate and simplified models of the observations.

Another key idea of this set of studies is to adopt a "pseudo-Bayesian" point of view, in which $\hat{P}$ is not required to belong to the reference model $\{P_\theta : \theta \in \Theta\}$ [resp. $\hat{f}$ is not required to belong to $\{f_\theta : \theta \in \Theta\}$]. Instead $\hat{P}$ is allowed to be of the form $\hat{P}(Z_1, \ldots, Z_N) = \mathbb{E}_{\hat{\rho}(Z_1, \ldots, Z_N)(d\theta)}(P_\theta)$, [resp. $\hat{f}$ is allowed to be of the form $\hat{f} = \mathbb{E}_{\hat{\rho}(d\theta)}(f_\theta)$], where $\hat{\rho}(Z_1, \ldots, Z_N)(d\theta) \in \mathcal{M}_+^1(\mathcal{Z})$ is a *posterior* parameter distribution, that is a probability distribution on the parameter set depending on the observed sample.

We will investigate three kinds of oracle inequalities, under different sets of hypotheses. To simplify notations, let us put

$$\hat{R}(Z_1, \ldots, Z_N) = R\big(\hat{P}(Z_1, \ldots, Z_N)\big) \qquad [\text{resp. } R\big(\hat{f}(Z_1, \ldots, Z_N)\big)],$$

and $R_\theta = R(P_\theta)$ [resp. $R(f_\theta)$].

- **Upper bounds on the cumulated risk of individual sequences of observations**. In the pattern recognition case, these bounds are of the type :

$$\frac{1}{N+1} \sum_{k=0}^{N} \mathbb{1}\left[Y_{k+1} \neq \widehat{f}(Z_1, \ldots, Z_k)(X_{k+1})\right]$$

$$\leq \inf_{\theta\in\Theta}\Big\{C\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{1}\left[Y_{k+1}\neq f_\theta(X_{k+1})\right]+\gamma(\theta,N)\Big\}.$$

Similar bounds can also be obtained in the case of least square regression and of density estimation. Integrating with respect to a product probability measure $P^{\otimes(N+1)}$ leads to

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}_{P^{\otimes k}}\left[\hat{R}(Z_1,\ldots,Z_k)\right]\leq\inf_{\theta\in\Theta}\left\{CR_\theta+\gamma(\theta,N)\right\}.$$

Here, $\gamma(\theta,N)$ is an upper bound for the estimation error, due to the fact that the best approximation of P within $\{P_\theta:\theta\in\Theta\}$ is not known to the statistician. From a technical point of view, the size of $\gamma(\theta,N)$ depends on the *complexity* of the model $\{P_\theta:\theta\in\Theta\}$ in which an estimator is sought. In the extreme case when Θ is a one point set, it is of course possible to take $\gamma(\theta,N)=0$. The constant C will be equal to one or greater, depending on the type of risk function to be used and on the type of the estimation bound $\gamma(\theta,N)$. These inequalities for the cumulated risk will be deduced from lossless data compression theory, which will occupy the first chapter of these notes.

- **Upper bounds for the mean non cumulated risk**, of the type

$$\mathbb{E}\left[\hat{R}(Z_1,\ldots,Z_N)\right]\leq\inf_{\theta\in\Theta}\left\{CR_\theta+\gamma(\theta,N)\right\}.$$

Obtaining such inequalities will not come directly from compression theory and will require to build specific estimators. Proofs will use tools akin to statistical mechanics and bearing some resemblance to deviation (or concentration) inequalities for product measures.

- **Deviation inequalities**, of the type

$$P^{\otimes N}\left\{\hat{R}(Z_1,\ldots,Z_N)\geq\inf_{\theta\in\Theta}\left[CR_\theta+\gamma(\theta,N,\epsilon)\right]\right\}\leq\epsilon.$$

These inequalities, obtained for a large class of *randomized* estimators, provide an *empirical* measure $\gamma(\theta,N,\epsilon)$ of the *local* complexity of the model around some value θ of the parameter. Through them, it is possible to make a link between randomized estimators and the method of penalized likelihood maximization, or more generaly penalized empirical risk minimization.

In chapter 7, we will study the behaviour of Markov chains with "rare" transitions. This is a clue to estimate the convergence rate of stochastic simulation and optimization methods, such as the Metropolis algorithm and simulated annealing. These methods are part of the statistical learning program sketched above, since the posterior distributions on the parameter space

$\hat{\rho}_{(Z_1,\ldots,Z_N)}$ we talked about have to be estimated in practice and cannot, except in some special important cases, be computed exactly. Therefore we have to resort to approximate simulation techniques, which as a rule consist in simulating some Markov chain whose invariant probability distribution is the one to be simulated. Those posterior distributions used in statistical inference are hopefully sharply concentrated around the optimal values of the parameter when the observed sample size is large enough. Consequently, the Markov chains under which they are invariant have uneven transition rates, some of them being a function of the sample size converging to zero at exponential speed. This is why they fall into the category of (suitably generalized) Metropolis algorithms. Simulated annealing is a variant of the Metropolis algorithm where the rare transitions are progressively decreased to zero as time flows, resulting in a nonhomogeneous Markov chain which may serve as a stochastic (approximate) maximization algorithm and is useful to compute in some cases the mode of the posterior distributions we already alluded to.

1

Universal lossless data compression

1.1 A link between coding and estimation

1.1.1 Coding and Shannon entropy

We consider in this chapter a finite set E, called in this context the *alphabet*, and a E valued random process $(X_n)_{n\in\mathbb{N}}$.

The problem of lossless compression is to find, for each input length N, a "code" c that is a one to one map

$$c : E^N \longrightarrow \{0,1\}^*$$

where $\{0,1\}^* = \bigcup_{n=1}^{+\infty}\{0,1\}^n$ stands for the set of finite sequences of zeros and ones with arbitrary length. Given any $s \in \{0,1\}^*$, its length will be noted $\ell(s)$. It is defined by the relation $s \in \{0,1\}^{\ell(s)}$.

We will look for codes with the lowest possible *mean length*

$$\mathbb{E}\Big(\ell\big(c(X_1,\ \ldots,\ X_N)\big)\Big). \tag{1.1.1}$$

If no other requirements are imposed on c, the optimal solution to this problem is obviously to sort the blocks of length N, $(x_1,\ \ldots,\ x_N)$ in decreasing order according to their probability to be equal to $(X_1,\ldots,X_N)$. Let $(b_i)_{i=1}^{|E|^N}$ be such an ordering of E^N, which satisfies

$$\mathbb{P}\Big((X_1,\ \ldots,\ X_N) = b_i\Big) \leq \mathbb{P}\Big((X_1,\ \ldots,\ X_N) = b_{i-1}\Big), \qquad i = 2,\ \ldots,\ |E|^N.$$

Let us introduce

$$\mathcal{B}(i) \stackrel{\text{def}}{=} \left(\left\lfloor \frac{i+1}{2^j} \right\rfloor \bmod 2\right)_{j=0}^{\lfloor \log_2(i+1)\rfloor - 1}, \qquad i = 1,\ \ldots,\ |E|^N,$$

the binary representation of $i+1$ from which the leftmost bit (always equal to 1) has been removed. The code

$$c(b_i) \stackrel{\text{def}}{=} \mathcal{B}(i) \tag{1.1.2}$$

obviously minimizes (1.1.1). Indeed,

$$\begin{aligned} \mathbb{N}^* &\longrightarrow \{0,1\}^* \\ i &\longmapsto \mathcal{B}(i) \end{aligned}$$

is a bijection and $i \mapsto \ell\big(\mathcal{B}(i)\big)$ is non decreasing. Starting with any given code, we can modify it to take the same values as c by exchanging binary words with shorter ones taken from the values of (1.1.2). We can then exchange pairs of binary words without increasing the code mean length to make it eventually equal to (1.1.2), which is thus proved to be optimal.

The *mean length* of the optimal code, defined by equation (1.1.1), is linked with Shannon's entropy, defined below.

Definition 1.1.1. The Shannon entropy $H(p)$ of a probability distribution p, defined on a finite set $\mathcal{X}$ is the quantity

$$H(p) = -\sum_{x\in\mathcal{X}} p(x)\log_2\big(p(x)\big).$$

The notation $\log_2$ stands for the logarithm function with base 2. Entropy is thus measured in *bits*. It is a concave function of p, linked with the Kullback Leibler divergence function with respect to the uniform distribution μ on $\mathcal{X}$ by the identity

$$H(p) = \log_2(|\mathcal{X}|) - \frac{1}{\log(2)}\mathcal{K}(p,\mu).$$

It cancels on Dirac masses and is equal to $\log_2(|\mathcal{X}|)$ for μ.

Let us recall a basic fact of ergodic theory :

Proposition 1.1.1. *For any stationary source $(X_n)_{n\in\mathbb{N}}$, the map $N \mapsto H\big(\mathbb{P}(dX_1^N)\big)$ is* sub-additive, *proving the existence of the limit*

$$\lim_{N\to+\infty} \frac{H\big(\mathbb{P}(dX_1^N)\big)}{N} = \inf_{N\in\mathbb{N}^*} \frac{H\big(\mathbb{P}(dX_1^N)\big)}{N} \stackrel{\text{def}}{=} \bar{H}\big(\mathbb{P}(dX_1^{+\infty})\big).$$

which is called the (Shannon) entropy *of the source $(X_n)_{n\in\mathbb{N}}$.*

Next proposition shows that Shannon's entropy measures in first approximation the optimal compression rate.

Proposition 1.1.2. *For any finite source X_1^N distributed according to $\mathbb{P}$, the mean length of the optimal code is such that*

$$\begin{aligned} &H\big(\mathbb{P}(dX_1^N)\big)(1-1/N) - 1 - \log_2(N) \\ &\leq \sup_{\alpha>1} \frac{1}{\alpha} H\big(\mathbb{P}(dX_1^N)\big) - 1 + \frac{\log_2(\alpha-1)}{\alpha} \leq \mathbb{E}\Big(\ell\big(c(X_1^N)\big)\Big) \leq H\big(\mathbb{P}(dX_1^N)\big) + 1. \end{aligned}$$

Thus, for any infinite stationary source $(X_n)_{n\in\mathbb{N}}$ *with distribution* $\mathbb{P}$,

$$\lim_{N\to+\infty}\frac{1}{N}\mathbb{E}\Big(\ell\big(c(X_1^N)\big)\Big) = \bar{H}(\mathbb{P}).$$

Proof. Let us adopt the short notation

$$\mathbb{P}\Big((X_1,\ \ldots,\ X_N) = b_i\Big) \stackrel{\text{def}}{=} p(b_i).$$

The chain of inequalities

$$\begin{aligned} p(b_i) &\leq i^{-1} \\ \mathbb{E}\Big(\ell\big(c(X_1^N)\big)\Big) &= \sum_{i=1}^{|E|^N} p(b_i)\,(\lfloor\log_2(i+1)\rfloor) \\ &\leq \sum_{i=1}^{|E|^N} -p(b_i)\log_2\big(p(b_i)\big) + 1 \\ &= H(p) + 1 \end{aligned}$$

shows that the mean length of the optimal code is upper-bounded by the Shannon entropy of the distribution of blocks of length N (up to one bit at most).

On the other hand, for any $\alpha > 1$,

$$\begin{aligned} \mathbb{E}\Big\{\ell\big[c(X_1^N)\big]\Big\} &\geq \sum_{i=1}^{|E|^N} p(b_i)\Big(\log_2(i+1) - 1\Big) \\ &\geq -\frac{1}{\alpha}\sum_{i=1}^{|E|^N} p(b_i)\log_2\left(\frac{\alpha-1}{(i+1)^\alpha}\right) - 1 + \frac{\log_2(\alpha-1)}{\alpha}. \end{aligned}$$

We can then notice that $b_i \mapsto \frac{\alpha-1}{(i+1)^\alpha}$ is a sub-probability distribution. This shows, along with the fact that the Kullback divergence is non-negative, that

$$-\sum_{i=1}^{|E|^N} p(b_i)\log_2\left(\frac{\alpha-1}{(i+1)^\alpha}\right) \geq -\sum_{i=1}^{|E|^N} p(b_i)\log_2\big(p(b_i)\big),$$

and consequently that

$$\mathbb{E}\Big(\ell\big(c(X_1^N)\big)\Big) \geq -\frac{1}{\alpha}\sum_{i=1}^{|E|^N} p(b_i)\log_2\big(p(b_i)\big) - 1 + \frac{\log_2(\alpha-1)}{\alpha}.$$

Then we can e.g. choose $\alpha = (1-1/N)^{-1}$, to obtain

$$\mathbb{E}\Big\{\ell\big[c(X_1^N)\big]\Big\} \geq H\big(\mathbb{P}(dX_1^N)\big)(1-1/N) - 1 - \log_2(N).$$

□

It is to be remembered from this first discussion about lossless data compression that it requires, to be done efficiently, a fairly precise knowledge of the distribution of the source, and that for any stationary source of infinite length, the optimal compression rate per symbol tends towards the entropy $\bar{H}$.

1.1.2 Instantaneous codes

In the previous sections, we have considered arbitrary binary codes, our purpose being to show that the compression rate could not be significantly lower than Shannon's entropy, whatever choice of code is made. We are now going to focus on the restricted family of *prefix* codes, which share the property to be *instantaneously* separable into words, when the codes for many blocks are concatenated together before being sent through a transmission channel.

The optimal code (1.1.2), described previously, indeed suffers from a major drawback : it cannot be used to code more than a single block. Indeed, if the codes for two successive blocks of length N (or more) are concatenated to be sent through a transmission channel, the receiver of this message will have no mean to find out how it should be decomposed into block codes. Such a code is called non separable, or non uniquely decodable or decipherable. Moreover, even if only one block is to be transmitted, the receiver has no mean to know whether the received message is completed, or whether more data should be waited for (such a code is said to be non *instantaneous*).

Instantaneous codes have a very simple characterization : a code is instantaneous if and only if no codeword is the beginning of another codeword. For this reason, such a code is also said to be a *prefix* code.

Definition 1.1.2. We will say that a finite subset $\mathcal{D} \subset \{0,\ 1\}^*$ of finite binary words is a *prefix dictionary* if any two distinct words a_1^r and b_1^s of $\mathcal{D}$ of respective lengths $r \leq s$ are always such that $a_1^r \neq b_1^r$. In other words we require that no word of the dictionary should be the prefix of another one. We will say that the map

$$c : E^N \longrightarrow \{0,\ 1\}^*$$

is a prefix code if it is one to one with values in a prefix dictionary.

Proposition 1.1.3 (Kraft inequality). *For any prefix dictionary* $\mathcal{D}$

$$\sum_{m \in \mathcal{D}} 2^{-\ell(m)} \leq 1.$$

Proof. This inequality can be proved using a construction which will lead also to *arithmetic coding*, to be described further below.

The set of finite binary sequences $\{0,1\}^* \cup \{\varnothing\}$ (where $\varnothing$ is the "void sequence" of null length) can be put into one to one correspondence with the set $\mathfrak{D}$ of *dyadic intervals*, defined by

$$\mathfrak{D} = \big\{[k\,2^{-n},\ (k+1)\,2^{-n}[:\ n \in \mathbb{N},\ k \in \mathbb{N} \cap [0, 2^n[\big\},$$

putting for any sequence

$$s = (s_i)_{i=1}^{\ell(s)} \in \{0,1\}^*,$$

$$\mathfrak{I}(s) \stackrel{\text{def}}{=} [k\,2^{-n},\ (k+1)\,2^{-n}[,$$

where $n = \ell(s)$, $k = \sum_{i=1}^{\ell(s)} s_i 2^{n-i}$ and putting moreover $\mathfrak{I}(\varnothing) = [0,1[$.

It is then easy to see that no one of two codewords s and s' is the prefix of the other if and only if the corresponding dyadic intervals $\mathfrak{I}(s)$ and $\mathfrak{I}(s')$ are disjoint. It immediately follows that the sum of the lengths of the intervals attached to the words of a prefix dictionary cannot exceed one (because all these dyadic intervals are subsets of the unit interval): this is precisely the Kraft inequality. This proof also shows that the Kraft inequality remains true for infinite dictionaries. □

Proposition 1.1.4 (Inverse Kraft inequality). *For any sequence* $(r_i)_{i=1}^T \in (\mathbb{N}^*)^T$ *of positive integers such that*

$$\sum_{i=1}^{T} 2^{-r_i} \leq 1,$$

there exists a prefix code

$$c : \{1,\ \ldots,\ T\} \longrightarrow \{0,1\}^*$$

such that

$$\ell\big(c(i)\big) = r_i, \qquad i = 1, \ldots, T.$$

Proof. Without loss of generality, the sequence $(r_i)_1^T$ may be assumed to be non decreasing. It is then easy to check that the intervals $[\alpha_i,\ \beta_i[$ defined by

$$\alpha_i = \sum_{j=1}^{i-1} 2^{-r_j},$$

$$\beta_i = \alpha_i + 2^{-r_i},$$

are dyadic in the sense defined in the proof of the Kraft inequality. Moreover, they are obviously non overlaping. The code $c(i) \stackrel{\text{def}}{=} \mathfrak{I}^{-1}([\alpha_i,\ \beta_i[)$ is therefore a prefix code. Let us remark here again that this proof holds also for infinite sequences. □

Definition 1.1.3. A prefix dictionary $\mathcal{D}$ is said to be complete (as well as any code using such a dictionary) if it is maximal for the inclusion relation, that is if any prefix dictionary $\mathcal{D}'$ such that $\mathcal{D} \subset \mathcal{D}'$ is in fact equal to it.

To a prefix dictionary $\mathcal{D}$ corresponds a binary tree, whose vertex set and edge set are defined as

$$\mathcal{N} = \bigcup_{s\in\mathcal{D}} \{I \in \mathfrak{D} \,:\, \mathfrak{I}(s) \subset I\},$$
$$\mathcal{A} = \big\{(I, I') \in \mathcal{N}^2 \,:\, I' \subset I\big\}.$$

It is the same thing to say that the dictionary $\mathcal{D}$ is complete or that the corresponding binary tree $(\mathcal{N}, \mathcal{A})$ is complete in the usual sense that its vertices have either zero or two sons (in other words, any interior vertex has exactly two sons).

Proposition 1.1.5 (Kraft equality). *A prefix dictionary $\mathcal{D}$ is complete if and only if*

$$\sum_{m\in\mathcal{D}} 2^{-\ell(m)} = 1.$$

Proof. The tree $(\mathcal{N}, \mathcal{A})$ is complete if and only if the set $\mathcal{F} = \big\{\mathfrak{I}(s) \,:\, s \in \mathcal{D}\big\}$ of its leaves is a partition of the unit interval $[0, 1[$.

Indeed, the two sons of $I \in \mathcal{N}$ are the two half length dyadic intervals into which it decomposes.

An interval I belongs to $\mathcal{N}$ if and only if it contains a leaf $I' \in \mathcal{F}$. In the case when $\mathcal{F}$ is a partition, it is either a leaf itself, or its two sons belong to the tree.

On the other hand, let us assume that some point $x \in [0, 1[$ does not belong to any leaf. Let I be the largest dyadic interval containing x and not overlaping with $\mathcal{F}$. The "father" of this interval thus meets $\mathcal{F}$, therefore its other son belongs to $\mathcal{N}$, proving that the tree $(\mathcal{N}, \mathcal{A})$ is not complete. □

Proposition 1.1.6. *Any prefix code satisfies*

$$\mathbb{E}\Big\{\ell\big[c(X_1^N)\big]\Big\} \geq H\big[\mathbb{P}(dX_1^N)\big].$$

Proof. It is enough to notice that, from the Kraft inequality,

$$E^N \longrightarrow [0, 1]$$
$$x_1^N \longmapsto 2^{-\ell\left(c(x_1^N)\right)}$$

defines a subprobability, and to follow the same reasoning as in the proof of proposition 1.1.2. □

Theorem 1.1.1. *There exists a complete prefix code c such that*

$$H\Big[\mathbb{P}\big(dX_1^N\big)\Big] \leq \mathbb{E}\Big\{\ell\big[c(X_1^N)\big]\Big\} < H\Big[\mathbb{P}\big(dX_1^N\big)\Big] + 1.$$

Proof. The sequence of positive integers

$$\big\{r(b) = \lceil -\log_2\big(p(b)\big)\rceil \ : b \in E^N, p(b) > 0\big\}$$

satisfies the Kraft inequality. Thus, there exists a prefix code c' defined on the support of p such that

$$\ell\big[c'(b)\big] = \lceil -\log_2\big[p(b)\big]\rceil, \qquad b \in E^N,\ p(b) > 0.$$

From any prefix code c', a complete prefix code c can be built in such a way that $\ell\big[c(b)\big] \le \ell\big[c'(b)\big]$, for any block $b \in E^N$ such that $p(b) > 0$, by "erasing the non coding bits" (the ith bit of $(s_1^r) \in \mathcal{D}'$ is "non coding" if $\mathfrak{I}\big[(s_1^{i-1}, 1 - s_i)\big] \notin \mathcal{N}'$). This code c is therefore such that

$$\mathbb{E}\Big\{\ell\big[c(X_1^N)\big]\Big\} < \sum_{b\in E^N} p(b)\Big[-\log_2\big(p(b)\big) + 1\Big] = H\Big[\mathbb{P}\big(dX_{1^*}^N\big)\Big] + 1.$$

□

Remark 1.1.1. One sees that looking for an optimal prefix code is, up to some rounding errors, strictly equivalent to estimating the distribution of the blocks of length N.

Proposition 1.1.7 (Huffman code). *Let p be a probability distribution on the set $\{1,\dots,M\}$ giving a positive weight to each point. Let i and j be any two indices satisfying*

$$\max\{p(i), p(j)\} \le \min\big\{p(k) \ : k \notin \{i,j\}\big\}.$$

Let c be an optimal prefix code for the probability vector $\big[p(k)_{k\notin\{i,j\}}, p(i) + p(j)\big]$. The prefix code $c' : \{1,\dots,M\} \to \{0,1\}^*$ defined by

$$\begin{cases} c'(k) &= c(k), \qquad k \notin \{i,j\}, \\ c'(i) &= \Big[c\big((i,j)\big), 0\Big], \\ c'(j) &= \Big[c\big((i,j)\big), 1\Big] \end{cases} \tag{1.1.3}$$

is optimal.

Proof. Let c be a code on $\{1,\dots,M\} \setminus \{i,j\} \cup \{(i,j)\}$. Let c' be the code on $\{1,\dots,M\}$ defined by (1.1.3). Obviously

$$\mathbb{E}\big[\ell(c')\big] = \mathbb{E}\big[\ell(c)\big] + p(i) + p(j).$$

The code c is thus optimal if and only if the code c' is optimal within the set $\mathcal{C}'$ of prefix codes such that the codewords for i and j differ only in their last bit. In the case when $p(i) + p(j)$ is minimal, $\mathcal{C}'$ contains a subset of the set of optimal prefix codes. Indeed, let c'' be an optimal prefix code for $(p_k)_{k=1,\dots,M}$