ERRATA

The Method of Maximum Entropy

H. Gzyl

Even though to err is human, some "misprints" are plain dum, if not worse. Here are some of the worst that I caught a bit too late.

1. I. Csiszar's name is uniformly misspelt. Sorry for that.
2. P. 41, the first term on the R.H.S. of (5.2) should be $-\ln Z_{Q,\Phi}(t)$.
3. P 42, the L.H.S. of (5.3) should be

$$\int \mathrm{L}(f_1/f_2)\mathrm{d}Q$$

4. P. 82, line 6 from below, I should have written

$$q_1 = q_2 = q_3 = q_4 = 1/4.$$

THE METHOD OF MAXIMUM ENTROPY

This page is intentionally left blank

Series on Advances in Mathematics for Applied Sciences – Vol. 29

THE METHOD OF MAXIMUM ENTROPY

Henryk Gzyl

Facultad de Ciencias
Universidad Central de Venezuela

World Scientific
Singapore • New Jersey • London • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.

P O Box 128, Farrer Road, Singapore 9128

USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Gzyl, Henryk, 1946–
The method of maximum entropy / Henryk Gzyl.
p. cm. -- (Series on advances in mathematics for applied sciences; vol. 29)
ISBN 9810218125
1. Maximum entropy method. I. Title. II. Series.
Q370.G97 1995
511'.42--dc20 94-23122
CIP

Printed in Singapore by Uto-Print

Preface

This book is an outgrowth of a set of lecture notes on the maximum entropy method delivered in 1988 at the 1st Venezuelan School of Mathematics. This yearly event aims at acquainting graduate students, and university teachers with trends, techniques and open problems of current interest. It takes place during the month of September at the Universidad de los Andes, in the city of Merida.

At the same time I was being invited to give lectures, Didier Dacunha-Castelle passed by and reported on his work on the subject. This happened not long after some astronomers friends of mine from the CEDA (also in Merida) had asked me to go with them over some methods for reconstructing images based on a maximum entropy procedure. So what else was left for me to do but collect material for that course?

The more I looked around, the more applications of the method I found. My original goal was to organize the material in such a way that the underlying philosophy of the method became transparent and to try to understand myself why it works. I hope to convey some of that to you, even though some of the whys are still a mystery (at least to me).

This page is intentionally left blank

Table of Contents

Chapter 0

INTRODUCTION

The Method of Maximum Entropy is an offspring of the Maximum Entropy Principle introduced in 1957 in statistical physics by E. Jaynes. That principle has the esthetic appeal of all variational principles in physics and its basic role is to characterize equilibrium states. It works as follows: a functional which is a Lyapunov function for the dynamics of the system is defined on the set of states. It is postulated that the equilibrium states are those yielding a maximum value of the functional, compatible with a set of given values of some extensive variables.

In chapter 2 we shall explain these things further, here we direct the reader to [0.1] where the original papers by Jaynes are reprinted together with many other interesting ones.

Actually, the possibility of characterizing probability densities by variational methods was already noticed by statisticians and information theorists well before 1957. Take a look at [0.2], especially chapter 3. By now, the list of probability distributions derived via the maximum entropy method is pretty long. We can go back even further. The germ of the idea is already presented in Boltzmann's writings. See [0.24] in the commemorative volume dedicated to his life and work. The germ of the idea is also present in Gibb's work. See the paper on the Carnot's principle by Jaynes in [0.7].

See the volume [0.3] by Kapur which devotes several chapters to characterization of standard probability densities by maximum entropy methods. Besides, a large variety of applications in which the notion of entropy enters is presented. And, speaking of applications, we bring the collection [0.4]-[0.11] to the reader's attention, where not only many applications of the M.E.M. are collected, but a lot of space is dedicated to foundational matters, and to explain the word "Bayesian" in the title of many of the volumes.

To explain the general philosophy of the M.E.M., and the underlying common approach of the long list of successful applications, let us begin by saying that many inverse and/or direct problems, all of which we shall call reconstruction problems, lead to searching for solutions to equations like

$$Ax = y \tag{0.1}$$

where A: $V_1 \to V_2$ is a linear transformation between two appropriate vector spaces V_1 and V_2, and we may be looking for solutions in some cone (or convex set) $C_1 \subset V_1$ while the data lies in

some other cone (or convex set) $C_2 \subset V_2$.

It may happen that the number of variables in (0.1) is much larger than the number of equations. For example, you may know a Laplace of a Fourier transform at a small number of values of the transform parameter only. The question is how to find $\mathbf{x}$ in (0.1).

The M.E.M. enters at two different levels. At level 1, one way to tackle (0.1) is to define a concave functional on an appropriate convex set $C_1 \subset V_1$

$$S: C_1 \to \Re \tag{0.2}$$

which will be the entropy functional and instead of solving (0.1) one sets up the following maximization problem

$$\max \{S(\mathbf{x}): \mathbf{x} \in C_1, A\mathbf{x}=\mathbf{y}\} \tag{0.3}$$

thus, we see that if $\mathbf{x}^*$ is such that $S(\mathbf{x}^*)$ reaches a maximum value, and the constraint is satisfied, one automatically has a solution to (0.1). The beauty about (0.3) is that many times solving (0.3) is equivalent to finding a minimum of a convex functional

$$H(\boldsymbol{\lambda})=\ln Z(\boldsymbol{\lambda})+(\boldsymbol{\lambda},\mathbf{y}) \tag{0.4}$$

where $Z(\boldsymbol{\lambda})$ will make its appearance below. In general $H(\boldsymbol{\lambda})$ is defined on a convex $D \subset V_2^*$, and when we are lucky $D= V_2^*$ $H(\boldsymbol{\lambda})$ is some sort of dual to $S(\mathbf{x})$ although not quite. Physicists have nice interpretations for it. In (0.4) the $\boldsymbol{\lambda}$ are the Lagrange multipliers for (0.3).

We shall write $(\mathbf{x},\mathbf{y})$ for the scalar product of vectors $\mathbf{x}$, and $\mathbf{y}$, and when $\mathbf{x} \in V$ and $\boldsymbol{\lambda} \in V^*$, $(\boldsymbol{\lambda},\mathbf{x}) \equiv \boldsymbol{\lambda}(x)$ as usual.

The value $\boldsymbol{\lambda}^*$ of the $\boldsymbol{\lambda}$ that makes $x(\boldsymbol{\lambda})$ a solution x^* to (0.3) is obtained by minimizing the convex functional $H(\boldsymbol{\lambda})$, which depends on as many variables as there are equations in (0.1).

We have thus transformed solving linear problem with more unknowns than equations into solving a smaller minimization problem, hopefully without constraints.

Many of the initial applications consisted in looking for positive probability densities yielding prescribed mean values for a finite collection of functions, taking S as the Gibbs-Boltzmann entropy associated to a density was natural.

In 1967 Burg proposed another entropy functional which has proven very useful for reconstructing densities when information about time series is given by a few correlations. We shall come back to this below.

At what we call a level 2 reconstruction problem, the M.E.M. enters the following way: On some appropriate measurable space $(\Omega,\Im)$, see chapter 1, we consider a class $\mathbf{P}$ of probability

measures, possibly absolutely continuous with respect to some fixed, preassigned a priori measure, and a family of random variables $\mathbf{X}$: $\Omega \to V_1$. Thus if $P \in \mathbf{P}$, the expected value of $\mathbf{X}$ with respect to $\mathbf{P}$, $E_P(X)$, is an element $\mathbf{x} \in V_1$. Instead of considering equation (0.1) we shall think of random variables $A\mathbf{X}$ with expected value

$$E_P AX = AE_P X = \mathbf{y}$$

Instead of solving (0.1) we will search for measures satisfying $E_P A\mathbf{X} = \mathbf{y}$. Now this becomes a level 1 problem on a different space namely, we want to find

(0.5) $$\sup\{S(P): P \in \mathbf{P}, E_P A\mathbf{X} = y\}$$

The rest of the comments made when describing the level 1 version of the M.E.M. apply here as well.

I have been able to trace this approach to reconstruction problems at least to the work of Rietsch [0.12], where it was applied to reconstruct the earth's density given its mass and moment of inertia.

After recalling some basic notations and definitions from probability theory, in chapter 1, we devote chapter 2 to a watered down presentation of equilibrium statistical mechanics. Chapter 3 consists of some heuristic arguments backing up the M.E.M..

In chapter 4 we introduce the most used entropy functionals and examine some of their properties.

Finally, it is in chapter 5 where the M.E.M. is explained. There we borrow from the important work by Cszizar, Dacunha and Gamboa, and make a few comments about the work by Bowrein and Lewis.

Surely the appeal of the M.E.M. has to do with its success in a large variety of applications, in some of which an entropy like concept is natural. But in many cases it is just something you pull out of your hat and it solves a problem for you. It may be this fact what prompts much of the work of explaining what the M.E.M. is about. Besides that, there is the appeal of the concept of entropy that comes through the second law of thermodynamics in understanding irreversibility; see [0.13]-[0.14]. To see how entropies help to understand issues related to self-organization and/or chaotic behavior as explained in [0.14]-[0.16]. Some uses of the notion of entropy in biology and economics have generated strong and sarcastic criticism. See [0.17]-[0.18] and the reviews in [0.19]-[0.20].

An interesting collection in which the thermodynamic notion of entropy plays a role is compiled in [0.21] and, of course, we should not fail to list at least one reference on the use of the

concept of entropy in information theory, [0.22] and in the theory of dynamical systems [0.23], connections between entropy, complexity and several quantum issues are reviewed in [0.25].

It will be up to the reader to decide whether there is or there is not a common thread in this list of references, many not directly related to the M.E.M., which explains its appeal beyond the mere: it just works.

To conclude, it must be clear that we are citing references by square brackets, numbered almost always by order of appearance, listed by chapter. Also, formulae, definitions and results will be cited sequentially in each chapter within round brackets.

I would like to thank my colleague Aldo Tagliani for writing sections (6.19)-(6.21).

And last, but not least, my thanks go to Ms. Leda Calderón, who typed the manuscript and went along nicely with my changing of mind now and then about a paragraph here and there. The editorial staff at WSP did a fabulous job weeding out uncountable mispirints

To finish I want to acknowledge the support of the Facultad de Ciencias, U.C.V., and of CONICIT for financial support during the preparation of the book.

Two references which I obtained during a brief visit to the CWI in Amsterdam, and added at the last minute are [0.25]-[0.26].

REFERENCES

[0.1] E. T. Jaynes: Papers on Probability, *"Statistics and Statistical Physics"*. Ed. Rosenkrantz, E.D. Kluwer Acad. Publi., Dordrecht, 1983.

[0.2] Kullback, S. "*Information Theory and Statistics*". Dover Publi., New York, 1968.

[0.3] Kapur, J. N. "*Maximum Entropy Models in a Science and Engineering*" John Wiley, New York, 1989.

[0.4] Justice, J. H. (Eds) "*Maximum Entropy and Bayesian Methods in Applied Statistics*" Cambridge Univ. Press, 1986.

[0.5] Ray Smith C. and Grandy, W. T., Jr. (Eds) "*Maximum Entropy and Bayesian Methods in Inverse Problems*". D. Reidel Publi. Co., Dordrecht, 1987.

[0.6] Ray Smith C. and Erickson, G. J. (Eds) "*Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*". D. Reidel Publi. Co., Dordrecht, 1987.

[0.7] Erickson, G. and Ray Smith C. (Eds) "*Maximum Entropy Methods and Bayesian Methods in Science and Engineering*": Vol I-Foundations. Kluwer Acad. Publi., Dordrecht, 1988.

[0.8] Erickson, G. and Ray Smith C. (Eds) "*Maximum Entropy Methods and Bayesian Methods in Science and Engineering*": Vol II-Applications. Kluwer Acad. Publi., Dordrecht, 1988.

[0.9] Skilling, J. (Eds) *"Maximum Entropy and Bayesian Methods"* Kluwer Acad. Publi. Co.,Dordrecht, 1989.

[0.10] Fougere, P. F. (Eds) *"Maximum Entropy and Bayesiam Methods"* Kluwer Acad. Publi. Co., Dordrecht, 1990.

[0.11] Grandy, W. T., Jr. and Schick, L. H. (Eds) "*Maximum Entropy and Bayesian Methods*" Kluwer Acad. Publi., Dordrecht, 1991.

[0.12] Rietsch, E. "*A Maximum Entropy approach to inverse problems*". Journ. of Geophysics. 42, pp. 489-506, 1977.

[0.13] Atkins, P. W. "*The Second Law*" W. H. Freeman. New York, 1984.

[0.14] Prigogine, I. and Stengers, I. "*Order out of Chaos*". Bantham Books, New York, 1984.

[0.15] Klimontovich, Yu. L. "*Turbulent Motion and The Structure of Chaos*" Kluwer Acad. Publi., Dordrecht, 1991.

[0.16] Mackey, M. C. *"The Origin of Thermodynamic Behaviour"* Springer Verlag, Berlin, 1992.

[0.17] Rifkin, J. "*Entropy: Into the Greenhouse World*". Bantham Books, New York, 1989.

[0.18] Brooks, D. R. and Wiley, E. O. "*Evolution as Entropy*".Univ. of Chicago Press, Chicago, 1988.

[0.19] Morowitz, H. "*Entropy Anyone*"., in "*Mayonnaise and the Origin of Life*" Berkeley Books, New York, 1985.

[0.20] Rothman, T. "*Science à la Modè: Physical Fashions and Fictions*". Princeton Univ. Press, Princeton, 1989.

[0.21] Leff, H. S. and Rex, A. F. "*Maxwell's Demon. Entropy, Information, Computing*" Princeton Univ. Press, Princeton, 1990.

[0.22] McEliece, R. J. " *The Theory of Information and Coding"* Vol 3, Encyclop. Math., Addison - Wesley, Reading, 1981.

[0.23] Martin, N. F. G. *"Mathematical Theory of Entropy"*. Vol 12, Encyclop. Math., Addison-Wesley, Reading, 1981.

[0.24] Klein, M. J. "*The development of Boltzmann's statistical ideas*". Acta Phys. Aust. Suppl. X, pp 53-106, 1973.

[0.25] Gelfand, I. M. and Yaglom, A. M. "Calculation of the amount of information about a random function contained in another such function". Amer. Math. Soc. Transl. Series 2 , A.M.S., Providence, 1959, pp. 199-246.

[0.26] "Maximum Entropy and Bayesian Methods" 3 volumes edited by Grandy, W. T. and Schick, L. H.; Ray-Smith, C. et al; Mohamed-Djafary, A. and Demoment, G. Published by Kluwer Acad. Publishers respectively in 1991, 1992 and 1993 in Dordrecht, Holland.

Chapter 1

BASIC CONCEPTS FROM PROBABILITY THEORY

We will recall some basic concepts from measure theory and from probability theory. The purpose of this chapter is to provide applied and other scientists with some standard vocabulary. Most of the concepts and results are intuitive and obvious at times, even though the proper names are not widely known.

A measurable space $(E,\mathcal{E})$ consist of a set E and a σ-algebra $\mathcal{E}$ of subsets of E. In $\mathcal{E}$ are the sets to which we will assign a measure (or, later on, a probability). It is a collection of subsets of E closed with respect to:

i) taking complements: if $A\in\mathcal{E}$ then $E-A=A^c\in\mathcal{E}$

ii) forming denumerable unions: if $\{A_n\in\mathcal{E}, n\geq 1\}$ then $\cup A_n\in\mathcal{E}$.

These sets operations when viewed abstractly, correspond with the logical operations not and or This is what makes σ-algebras a convenient realizations of events to which we want to assign probabilities.

A **measure** m on a measurable space $(E,\mathcal{E})$ **is a function** $m: \mathcal{E}\to[0,\infty)$ satisfying

$$m\left(\overset{\infty}{\underset{1}{\cup}} A_n\right)=\sum_{i=1}^{\infty} m(A_n) \tag{1.1}$$

where $\{A_n: n\geq 1\}$ is any countable collection in $\mathcal{E}$ such that $A_i\cap A_j=\varnothing$. We add here that instead of $[0,\infty)$ the range of values of m can be taken to be any space X on which an additive operation and a notion of convergence are defined such that the right hand side of (1.1) exists. In such cases one says that m is an X-valued measure.

Consider two measurable spaces $(E,\mathcal{E})$ and $(F,\mathcal{F})$. We shall say that a function $X: E\to F$ is $\mathcal{E}/\mathcal{F}$measurable if

$X^{-1}(A)=\{x: X(x)\in A\}=\{X\in A\}\in\mathcal{E}$ for any $A\in\mathcal{F}$.

The σ-algebra $B(\Re)=B$ generated by the open (or the closed, or the compact) subsets of $\Re=(-\infty,\infty)$ is called the Borel σ-algebra, since real valued functions appear all the time, one usually writes $X\in\mathcal{E}$ instead of $X\in\mathcal{E}/B$.

We shall say that f=g a.e(m) (almost everywhere with respect to m) whenever $\{f>g$ or $f<g\}=\{x: f(x)>g(x)$ or $f(x)<g(x)\}$ is such that $m(\{f>g$ or $f<g\})=0$.

It is left for the reader to verify that the basic arithmetic properties, performed on Borel measurable functions yield Borel measurable functions, that is linear combinations, products, quotients (whenever defined), infima, suprema of measurable functions are measurable.

Also, pointwise limits and other limiting operations performed on sequences on measurable functions yield measurable functions. The reader is directed to either [1.1] or [1.2] for this and much more about the basics on real analysis and probability .

The process of integration is introduced stagewise. First of all the integral of X is defined for $X=\Sigma a_n I_{A_n}$ where $\{A_n: n\geq 1\}$ is a countable (disjoint) partition of E and the a_n are real. Such X's are called simple functions. For such an X we set

$$\int X(x)\,m(dx) = \Sigma a_n\, m(A_n). \tag{1.2}$$

For this to make sense it is required that the right hand side converges. This is easily achieved when all except finitely many of the A_n are empty.

The second step is to realize that any positive $X\in\mathcal{F}$ can be approximated by an increasing sequence of simple functions, that is $X=\lim X_n$ where X_n is simple,

$$X_n = \Sigma a_{n,k} I_{A_{n,k}},\ X_{n+1} \geq X_n .$$

Now define

$$\int X(x)\,m(dx) = \lim_n \int X_n(x)\,m(dx). \tag{1.3}$$

The third step is to take an arbitrary function X, decompose it as $X=X_+-X_-$, with $X_\pm=(X\pm|X|)/2$ and compute $\int X_\pm$ dm as in (1.3). When both are finite we define

$$dm = \int X_+ dm - \int X_- dm. \tag{1.4}$$

We shall also say that X is integrable if

$$\int |X|\, dm = \int X_+ dm + \int X_- dm$$

is finite and write $X\in L_1$ $(E,\mathcal{F},m)$ (or $X\in L_1$ whenever $E,\mathcal{F}$ and m are understood from the context). Also we say that $X\in L_p$ whenever $\int (X)^p$ dm is finite.

Let m and n be two measures on $(E,\mathcal{F})$. We shall say that m is absolutely continuous with respect to n, and write $m<<n$, if whenever $n(A)=0$ for $A\in\mathcal{F}$, then $m(A)=0$. In this case there exists $p\in\mathcal{F}$ such that

$$m(A) = \int_A p(x)\,n(dx) = \int I_A(x)\,p(x)\,n(dx). \tag{1.5}$$

The convention is to write p(x)=dm(x)/dn(x) or dm(x)=p(x)dn(x) and call p(x) the Radon-Nikodym derivative of m with respect to n. It is reasonable straightforward to verify that if m<<n and n<<q then m<<q and dm/dq=(dm/dn)(dn/dq) (almost everywhere q).

Consider now X: E→F and X∈ $\mathscr{E}/\mathscr{F}$. Let m be a measure on (E,$\mathscr{E}$) and define n on (F,$\mathscr{F}$) by

$$n(B) = m(X^{-1}(B)) = m(X \in B) \tag{1.6}$$

and note that since X^{-1}: $\mathscr{F} \to \mathscr{E}$ preserves set operations, n is a well defined measure on (F,$\mathscr{E}$). Also, using (1.6) one can prove (going stagewise from simple functions onwards) that for any positive measurable Y: F→ $\mathfrak{R}$

$$\int Y(y)\, n(dy) = \int Y(X(x))\, m(dx).$$

Let us proceed to rewrite some of the former in terms of probabilistic language. We shall say that (Ω,$\mathscr{F}$,P) is a probability space if (Ω,$\mathscr{F}$) is a measurable space and P is a measure with range [0,1], such that P(Ω)=1.

The points ω in Ω are called elementary events, they may or may not be such that {ω} is in $\mathscr{F}$. The usual interpretation of ω is that it represents an experiment (sequence of measurements in continuous or discrete times) performed on some system. The elements of $\mathscr{F}$ are called events and, questions about experiments are described by the set operations of union, intersection and complementation.

A (real valued) random variable X is a (Borel) measurable function defined on (Ω,$\mathscr{F}$).

From now on we shall consider a given (Ω,$\mathscr{F}$,P). Let X be a (real valued) random variable. The distribution function of X is defined to be the function F_X: $\mathfrak{R} \to [0,1]$

$$F_X(x) = P(X \le x)$$

and when X : $\Omega \to \mathfrak{R}^n$ is such that each component function X_i is measurable, we define

$$F_X(X) = P(X_1 \le x_1, \ldots, X_n \le x_n).$$

From now on we shall follow the standard notational convention and write

$$\bigcap_{\alpha \in I} \{X_\alpha \in A_\alpha\} = \{X_\alpha \in A_\alpha ; \alpha \in I\}$$

where X_α : (Ω,$\mathscr{F}$) → (E_α,$\mathscr{E}_\alpha$) are (E_α-valued) random variables, I is some set of indices and A_α ∈E_α for all I. (Warning: the set described above may not be in $\mathscr{F}$ when I is not countable.)

Assume that X is an integrable random variable. We introduce the symbols

$$EY \equiv \langle Y \rangle \equiv \int Y(\omega) dP(\omega)$$

and we will call it the mathematical expectation of Y (respect to P) or average value of Y.

When X is an $\mathfrak{R}^n$ random variable represented by a column vector with components X_i which are L_2, and X^t denotes the transposed (row) vector, the covariance matrix of X is defined to be the matrix $C=E(X-\langle X\rangle)(X-\langle X\rangle)^t$ which is positive definite, i.e.

$$\sum_{i,j=1}^{n} \xi_i \xi_j C_{ij} = E\left| \sum_{i=1}^{n} (X_i - \langle X_i \rangle)\xi_i \right|^2 \geq 0.$$

When m is a measure on $\mathfrak{R}^n$, we shall say that then $\mathfrak{R}^n$-valued **X** has a density **p** with respect to X_n if the measure induced on $\mathfrak{R}^n$ by F_X is absolutely continuous respect to m and $dF_X/dm=p$. Then, for any bounded measurable $G: \mathfrak{R}^n \to \mathfrak{R}$

$$EG(\mathbf{X}) = \int G(\mathbf{x}) p(\mathbf{x})\, dm.$$

As an exercise for the reader, we leave the proof of

Lemma 1.7. Let X be a random variable with strictly positive distribution, absolutely continuous with respect to the Lebesgue measure on $\mathfrak{R}$. Then

$$EG(X) = \int_{-\infty}^{\infty} G(x) p(x)\, dx = \int_0^1 G(F^{-1}(u))\, du$$

where G is a bounded measurable function and F^{-1} is the compositional inverse of F.

Probability theory begins to be different from real analysis (measure theory) when we come down to define independence and conditional probability.

Definition 1.8. Let $\mathscr{F}_1$, $\mathscr{F}_2$ be sub-σ-algebras of $\mathscr{F}$. We shall say that they are (P) independent if

(1.9) $$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

for any $A_1 \in \mathscr{F}_1$, $A_2 \in \mathscr{F}_2$.

This notion extends trivially to any countable family of σ-algebras. When $X_1 \in \mathscr{F}_1/\mathscr{E}_1$, $X_2 \in \mathscr{F}_2/\mathscr{E}_2$ we say that they are independent if

$$Ef(X_1)g(X_2) = Ef(X_1)Eg(X_2)$$

for any bounded measurable functions defined on E_1, E_2. We leave for the reader to verify that if X_1 and X_2 are uncorrelated, i.e.

$$EX_1X_2 = EX_1EX_2$$

they may not be necessarily independent. Also, for independence, it suffices to verify that

$$p\{X_1 \in A_1, X_2 \in A_2\} = p\{X_1 \in A_1\}p\{X_2 \in A_2\}$$

for classes of sets A_1, A_2 generating the sub-σ-algebras

$$\sigma(X_i) = X_i^{-1}\mathbf{B} = \left\{X_i^{-1}(B) : B \in \mathbf{B}(\mathfrak{R})\right\}.$$

A related notion, that of conditional expectation is to be introduced in (almost) full generality by

Definition 1.10. Let $(\Omega,\mathcal{F},P)$ be a probability space and , let $\boldsymbol{G}$ be a sub-σ-algebra of $\mathcal{F}$ and X be either a positive or integrable random variable. Then, there exists a $\boldsymbol{G}$-measurable random variable, defined up to a e. equivalence, denoted by $E[X|\boldsymbol{G}]$, such that for any bounded, $\boldsymbol{G}$-measurable function Y, the following holds

$$E[XY] = E[YE[X|\mathbf{G}]]. \tag{1.11}$$

The proof of the following lemma, asserting some basic properties of conditional expectations is left for the reader.

Lemma 1.12. The notations are as above.

a) If $X \geq 0$, $E[X|\boldsymbol{G}] \geq 0$.

b) If X_1 is bounded, $\boldsymbol{G}$-measurable, $E[X_1X|\boldsymbol{G}]=X_1E[X|\boldsymbol{G}]$.

c) Let g be a bounded function on $\mathfrak{R}^2$ and $X_1 \in \boldsymbol{G}$, then

$$E[g(X,X_1)|\mathbf{G}] = E[G(X,z)|\mathbf{G}]_{z=X_1}$$

where z is a constant.

d) $E[E[X|\boldsymbol{G}]]=EX$.

e) For a_1, a_2 in $\mathfrak{R}$, $E[a_1X_1+a_2X_2|\boldsymbol{G}] = a_1E[X_1|\boldsymbol{G}]+a_2E[X_2|\boldsymbol{G}]$.

f) If $\{X_n\}$ is either a monotonic sequence of positive functions or a uniformly bounded, pointwise convergence sequence, $E[\lim X_n|\boldsymbol{G}]=\lim E[X_n|\boldsymbol{G}]$.

g) If the σ-algebras σ(X) and $\boldsymbol{G}$ are independent, then

$$E[X|\mathbf{G}] = E[X].$$

h) Let $\boldsymbol{G}_1 \subset \boldsymbol{G}_2$ be two sub-σ-algebras of $\mathcal{F}$. Then

$$E[E[X|\mathbf{G}_2]|\mathbf{G}_1].$$

i) Let Φ be a convex function, infinite on the range of X, then

$$\Phi(E[X|\mathbf{G}]) \leq E[\Phi(X)|\mathbf{G}].$$

Comment. Properties a, e, f, i, are shared by all integration like functionals. And properties b, e, h, show that conditional expectations are projection operators. As a matter of fact, restricted to $L_2(\Omega,\mathcal{F},P)$, the operator

$$X \to E[X|\mathbf{G}]$$

is an orthogonal projection onto $L_2(\Omega,\mathbf{G},P)$. This result is important when dealing with Gaussian processes and computing predictors optimal in quadratic distances.

Let us now recall some important factorization results.

Lemma 1.13. Let Y: $(E,\mathcal{E})\to(E',\mathcal{E}')$ be a $\mathcal{E}/\mathcal{E}'$-measurable and then X: $E\to\Re$ is measurable $\sigma(Y)/B$ if and only if there exists g: $E'\to\Re$, $\mathcal{E}'/B$-measurable such that $X=g(Y)$.

Therefore, when $\mathbf{G}$ in definition (1.10) is $\sigma(Y)$, then there is h_X: $E'\to\Re$ such that

$$E[X|\sigma(Y)] = h_X(Y).$$

According to properties b, e, h of Lemma (1.12) the correspondence $X\to h_X$ is linear, and behaves like some sort of integration with respect to X. To round things up we need the following

Definition 1.14. Let $(E,\mathcal{E})$ and $(E',\mathcal{E}')$ be two measure spaces. The function $N(A,Y)$: $\mathcal{E}\times E\to[0,\infty)$ is called a positive kernel (or probability kernel when N takes values in [0,1]) if:

a) For every $y\in E'$, $A\to N(A,y)$ is a measure on $(E,\mathcal{E})$.

b) For every $A\in\mathcal{E}$, $y\to N(A,y)$ is $\mathcal{E}'$-measurable.

When the random variables considered take values in complete, separable, metric spaces. The conditional distributions can be computed from a regular conditional distribution, that is:

Lemma 1.15. Let $X:(\Omega,\mathcal{F},P)\to(E,\mathcal{E})$ be measurable and let E be a complete, separable, metric space and $\mathcal{E}$ denote its Borel sets. Let Y be a random variable and f: $E\to\Re$ be any bounded function. Then there exists a kernel $N(A,y)$ on $\mathcal{E}\times\Re$ such that

$$E[f(X)|Y] = \int f(x)\,N(dx,Y).$$

Comment. Usually life is good with us and there is a measure m(dx) on E with respect to $N(.,y)$ which is absolutely continuous with a jointly measurable density n(x,y) then

$$E[f(X)|Y] = \int f(x)\,n(x|y)\,m(dx)$$

and the notation

$$E[f(X)|Y=y] = \int f(x)\,n(x|y)\,m(dx)$$

is usually employed. See [1.3] for nuances about constructing kernels and [1.1] or [1.4] for the necessary measure theoretic results.

We shall need in chapter 4 a slight extension of the conditional expectation operator to the case when $(\Omega,\mathcal{F},\mu)$ is such that μ is a σ-finite measure with $\sigma(\Omega)=+\infty$.

Let $G \subset \mathcal{F}$ be a sub-σ-algebra. Let $f\in L_1(\mu)$, and **assume that** (Ω,G,μ) is σ-finite. Let us state

Definition 1.16. We shall denote by $E_\mu[f|G]$ the unique (up to appropriate sets of measure zero) element of $L_1(\mu)$ such that

$$\int gf\,d\mu = \int gE_\mu[f|G]\,d$$

for every bounded function $g\in G$.

The properties of the conditional expectation operator $f\to E[f|G]$ introduced above when μ was a probability hold true in this case as well.

Comment. When (Ω,G,μ) is not σ-finite these things do not make much sense. Consider an atomic measure with infinite atoms.

Comment. It is important to realize that as a consequence of the next proposition we have $E_\mu[1|G]=1$.

Proposition 1.17. Let $h_1, h_2\in G$, be strictly positive and such that $dP_1=h_1d\mu$ and $dP_2=h_2d\mu$ are probability measures. Then, for $f\in\mathcal{F}$, $g\in G$

i) $\int gE_{P_1}[f|G]\,d\mu = \int gf\,d\mu = \int gE_{P_2}[f|G]\,d\mu.$

ii) $E_{P_1}[f|G] = E_{P_2}[f|G]$ *a.e.*μ.

Proof: (i) $\int fg\,d\mu = \int (g/h_1)\,fdP_1 = \int (g/h_1)E_{P_1}[f|G]dP_1 = \int gE_{P_1}[f|G]d\mu$. Part (ii) follows from (i) by taking $g=(E_{P_1}[f|G]-E_{P_2}[f|G])$ and using the fact that $\int g^2d\mu=0$ implies $g=0$ a. e. μ.

Let us consider some simple examples. To begin with note that if $G=\{\Phi,\Omega\}$ is the trivial σ-algebra, then

$$E[X|G] = E[X]$$

for any integrable or positive X. Assume now that G is the σ-algebra generated by a partition $\{A_k: k\geq 1\}$ of Ω. That is, its elements are countable unions of sets of $\{A_k: k\geq 1\}$ and any G-measurable function is of the type $\Sigma a_k I_{A_k}$ for appropriate constants a_k. In this case E[X|G] must be something like

$$E[X|G] = \Sigma a_k I_{A_k}$$

and we have to determine the a_k. For that, multiply both sides of the identity by I_{A_j} for some j and use (1.11) to obtain

$$a_k = \frac{E[X;A_j]}{P(A_j)} = \frac{1}{P(A_j)} \int_{A_j} X(w)dP(w).$$

Here we convene to define $E[X;A_j]/P(A_j)=0$ whenever $P(A_j)=0$.

When $G = \sigma(Y)$ for $Y: \Omega \to E$, E being a countable set, instead of the notation above we have

$$E[X|Y = e_i] = (P(Y = e_i))^{-1} E[X;\ Y = e_i]$$

and correspondingly

$$E[X|Y] = \sum_i (P(Y = e_i))^{-1} E[X;\ Y = e_i] I_{(Y=e_i)}$$

that is, the function on the right-hand side is constant on the sets $\{Y=e_i\}$ taking the value $E[X|Y=e_i]$ as specified above.

The other very common case is the following. Let X and Y be respectively $\mathfrak{R}^n$ and $\mathfrak{R}^m$ valued random variables such that the distribution of (X,Y) has density p(x,y) with respect to the product Lebesgue measure dx dy on $\mathfrak{R}^{n+m}$. From the factorization lemma above, we know that the bounded random variables, measurable with respect to $\sigma(Y)$ are of the form g(Y) for $g:\mathfrak{R}^n \to \mathfrak{R}$. Therefore computing both sides of (1.11) we see that for bounded $f: \mathfrak{R}^n \to \mathfrak{R}^n$

$$E[f(X)g(Y)] = \int f(x)\, g(y)\, \rho(x,y)\, dx\, dy$$

where we denoted by $h_X(y)$ the function introduced right after Lemma (1.13). Since both sides are equal for any g(y) we conclude that

$$\rho(X|Y) = \left(\int (x, Y)dx\right)^{-1} \rho(x, Y)\,. \tag{1.18}$$

For the sake of future reference, we shall now present two variations on the theme of Bayes formula, which are at the basics of both applications and interpretations of the maximum entropy method.

In what follows, we shall denote the intersection $C_1 \cap C_2$ of two sets C_1 and C_2 by C_1C_2. Let $\{A_i\}$ be a countable partition of Ω, i.e., a denumerable exhaustive collection of mutually exclusive events. Then for any events B, C

$$P(B|C) = \sum_i P(A_i|BC)\, P(B|C)$$

and also

$$P(B|C) = \sum_i P(A_i B|C) = \sum_i P(B|A_i C) P(A_i|C)$$

since

$$P(A_i B|C) = \frac{P(A_i BC)}{P(C)} = \frac{P(BA_i C)\,P(A_i C)}{P(A_i C)P(C)} = P(B|A_i C)P(A_i|C).$$

Exchanging the roles of A_i and B we obtain for any i

$$P(A_i|BC) = P(B|A_i C)\frac{P(A_i|C)}{P(B|C)}$$

which is known as Bayes identity. Substituting P(B|C) by the second summation displayed above we have

$$P(A_i|BC) = \frac{P(B|A_i C)P(A_i|C)}{\sum_j P(B|A_j C)P(A_j|C)}.$$

In this identity $P(A_i|C)$ is interpreted as the **"a priori"** probability of A_i given C, taken to describe the knowledge we have about the event A_i given the preliminary information in event C. The left-hand side tells us how much does our knowledge about A change when we collect the information contained in event B. The right-hand side is the recipe for computing the change. The left-hand side is called the **" a posteriori"** probability of A_i.

REFERENCES

[1.1] Bauer, H. "*Probability Theory and Elements of Measure Theory*". Holt, Rinehart and Wilson, Inc., New York, 1972.

[1.2] Gihman, I. I. and Skorohod A. V. "*The Theory of Stochastic Processes I*" Springer-Verlag, New York, 1974.

[1.3] Getoor, R. K. "*On the Construction of Kernels*" Lect. Notes in Math. N°465, pp. 443-463, Springer-Verlag, Berlin, 1977

[1.4] Rudin, W. "*Real and Complex Analysis*". McGraw-Hill, New York, 1966.

Chapter 2

EQUILIBRIUM DISTRIBUTIONS IN STATISTICAL MECHANICS

Statistical physics owes its birth to the inconvenience and impossibility of describing systems having very large numbers of particles by specifying the behavior of each individual. Loosely speaking, the aim of statistical physics is to describe the "collective" or "macroscopic" properties of a system of particles in terms of appropriate averages of its "microscopic" motions.

The words in quotation marks are the key ones. Macroscopic refers to the "properties of the system as a whole", the properties "visible by the naked eye". Microscopic means to the description based on the exact evolutions laws (classical or quantum) describing the motions of the particles.

Even though today's supercomputers can follow the individual motions of large numbers of independent particles, there is yet nothing able to handle 10^{23} individuals.

To begin with, we will only consider systems to which equilibrium thermodynamics applies, i.e., systems whose external, macroscopic changes are very slow compared to the microscopic, internal motions and can be considered at every instant to be in equilibrium.

One of the cornerstones in physics, important for the outlook at the world that it provides, is the second law of thermodynamics. According to this law, whenever an isolated system evolves its entropy can only increase and, when an equilibrium state is reached its entropy attains the highest possible value compatible with the values of the macroscopic extensive parameters of the system.

I want to emphasize that this is not the standard formulation, see [2.1] or [2.2], for I have made the characterization of the equilibrium state part of the statement of the second law. Actually, when the entropy functional happens to be a Lyapunov functional for the evolution law of the system, the characterization of equilibrium states as maxima of the entropy functional is obvious. See section (6.5) for some more on these issues.

The connection between the macroscopic and microscopic descriptions can be traced down to ideas of Maxwell, Boltzmann and Gibbs. The ingredients involved depend on the kind of system under study: how do we describe its microscopics states and the changes of state, i.e., its evolution on one hand and, on the other, on the method we choose to average over the microscopical states. To be more specific, we may consider either classical or quantum description of the particles making up a system and in the latter case, we may consider the

particles to be distinguishable or not. But regardless of all these subtleties, the basic philosophy is always the same.

The presentation that follows is essentially due to Jaynes [2.3] see also [2.4].

Consider, to simplify as much as possible, a system whose states can be described by a countable set E, which are unaccessible to observation. The microscopic observables describing the properties of the system are described by real valued functions

$$F: E \to \Re.$$

The basic assumption is that the macroscopic values of these observables are obtained as average values

$$\langle F \rangle = \sum_{i \in E} F(i) P_i,$$

where the P_i are the probabilities of finding the system in state $i \in E$, or the fraction of time the system spends at state i when it is in equilibrium. With all these ingredients we state the

Principles of Maximum Entropy

The state of equilibrium is characterized by the assignment of probabilities $\{P_i : i \in E\}$ such that

(2.1) $$s(P_i) = -k \Sigma P_i \ln P_i$$

attains its maximum possible value among all distributions $\{P_i : i \in E\}$ such that

(2.2) $$\Sigma A_k(i) P_i = \langle A_i \rangle = a_i . \qquad i = 1, 2, 3, \ldots, M.$$

If we proceed as most physicists do and apply elementary variational analysis using Lagrange multiplies we would obtain

(2.3) $$P_i^* = \frac{1}{z_A(\lambda)} \exp -\frac{1}{k} \sum_{k=1}^{M} \lambda_k A_k(i),$$

where the partition function $Z_A(\boldsymbol{\lambda})$ is defined by

$$Z_A(\boldsymbol{\lambda}) = \sum_{i \in E} \exp -\frac{1}{k} \sum_{k=1}^{M} \lambda_k A_k(i).$$

Observe that $S(\{P_i\})$ defined by (2.1) is a concave function defined on the convex set $P=\{\{P_i : i \in E, \Sigma P_i = 1\}\}$ of all probability measures on E. Also, $Z_A(\boldsymbol{\lambda})$ is only defined on a subset of $\Re^M$ specified by

$$D_A = \{\boldsymbol{\lambda} \in \Re^M : Z_A(\boldsymbol{\lambda}) < \infty\}.$$

When E is finite, $D_A = \Re^M$ but when E is not finite D_A is only a convex subset of $\Re^M$ One should always know how big the set D_A is.

Anyway, when $\{P_i^*\}$ given by (2.3) is substituted in (2.1) we obtain

$$H(\boldsymbol{\lambda}) = k \ln Z_A(\boldsymbol{\lambda}) + (\boldsymbol{\lambda}, a) \tag{2.4}$$

which is convex on $D_A(\boldsymbol{\lambda})$. We shall verify these assertions below, in a more general setting. We have set

$$(\boldsymbol{\lambda}, \mathbf{a}) \equiv \sum_{i=1}^{M} \lambda_i a_i$$

and ln(s) denotes the natural logarithm of s>0.

What is the beauty of this set up from the point of view of optimization theory? Well $H(\boldsymbol{\lambda})$ is S written in terms of the extensive variables $\boldsymbol{\lambda}$ and is a kind of dual to (S{P}) and minimization of $H(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ yields a value $\boldsymbol{\lambda}^*$ such that $\partial H(\boldsymbol{\lambda}^*)/\partial \boldsymbol{\lambda}_j = 0$ which is equivalent to

$$\sum_{i \in E} A_j(i) (Z_A(\lambda))^{-1} \exp-\Big(\sum_N \lambda_n A_n(i)/k\Big) = a_j, \qquad j = 1, \ldots, M.$$

That is, the value $\boldsymbol{\lambda}^*$ at which $H(\boldsymbol{\lambda})$ reaches its minimum is that of the Lagrange multiplier which insures that $\{P_i^* | i \in E\}$ given by (2.3) satisfy the constraints (2.2). Note that the number M may be much smaller than the cardinality of E, and $H(\boldsymbol{\lambda})$ is convex. These are the two facts that lie behind the appeal of the M.E.M..

We shall remark as well that the condition for $\boldsymbol{\lambda}^*$ to be a minimum, namely that $\partial^2 H(\boldsymbol{\lambda}^*)/\partial\lambda_i \partial\lambda_j$ be a positive definite form has strong consequences in thermodynamics. See [2.2] and [2.4]. In terms of covariances of the observables A_k it looks like this: the matrix

$$\frac{\partial^2 H(\lambda^*)}{\partial\lambda_i \partial\lambda_j} = \langle A_i A_j \rangle - \langle A_i \rangle \langle A_j \rangle = \langle (A_j - \langle A_j \rangle)(A_i - \langle A_i \rangle) \rangle$$

is (obviously) positive definite. Actually, in some cases the function $H(\boldsymbol{\lambda})$ may be very flat near $\boldsymbol{\lambda}^*$ making the numerical search for $\boldsymbol{\lambda}^*$ hard (especially for large M).

A similar procedure can be followed to obtain the probability distribution in the (simplest) quantum case. For that we assume we are studying a system of noninteracting particles, each of which may be found in any of the states of a denumerable set. To distinguish between the two classes of identical particles we introduce two possible sets of states for the system.

We shall assume that the bosons have as (denumerable) set of states the set

$$H_b = \{\Psi : S \to \mathbf{N},\ \Psi(i) \neq 0 \quad \textit{finitely many} \quad i \in S\}$$

whereas the state of the fermions will be described by

$$H_f = \{\Psi : S \to \{0, 1\},\ \Psi(i) = 1 \quad \textit{finitely many} \quad i \in S$$

In each case we shall assume there are two observables whose average values are accessible to us, namely

$$E : H \to \mathfrak{R}, \quad E(\psi) = \sum_{i \in S} E(i)\psi(i)$$

(2.5)

$$N : H \to \mathfrak{R}, \quad N(\psi) = \sum_{i \in S} \psi(i)$$

where $\mathcal{E}$: S$\to \mathfrak{R}$ is to be interpreted as a microscopic energy of the individual quantum states.

Now (2.1) looks like

(2.6) $$S(\{P(\Psi)\}) = -k \sum_{\Psi \in H} P(\Psi) \ln P(\Psi),$$

where H stands for either H_b or H_f.

Again, a small computation leads to the partition functions

$$Z(\boldsymbol{\lambda}) = \sum_{\Psi \in H} \exp\left(-\frac{\lambda_1}{k}\xi(\psi) - \frac{\lambda_2}{k}N(\psi)\right)$$

which have to be computed separately for bosons and fermions.

For the first case notice that

$$Z(\lambda) = \sum_{\psi \in H_b} \exp -\frac{1}{k}\left[\sum_i \lambda_1 \xi(i)\psi(i) + \sum_i \lambda_2 \psi(i)\right]$$

$$= \sum_{\psi \in H_b} \prod_{i \in H} \exp -\frac{1}{k}(\lambda_1 \xi(i) + \lambda_2)\psi(i)$$

$$= \prod_{i \in H} \sum_{n \geq 0} \exp\ -\frac{n}{k}(\lambda_1 \xi(i) + \lambda_2)$$

$$= \prod_{i \in H} (1 - \varsigma \exp\ -(\xi(i)/kT))^{-1}$$

where in the last line we set ς =exp-(λ_2/k) and λ_1 = 1/T to conform with standard notation in the physical literature.

To compute the partition function for the fermionic case, we proceed in the same way, except that now instead of summing over all n as in step 3, we sum only over n=0 and n=1 obtaining

$$Z(\lambda) = \prod_{i \in H} (1 + \varsigma \exp-(\xi(i)/kT)).$$

The point of these exercises was to show how the specification of the set of states and the choice of the observables determine the final result.

To complete this mock approach to equilibrium statistical mechanics we shall verify that the functional (2.1) is a Lyapunov functional. We shall assume that the probability distributions $\{P_i(t): i \in E\}$ evolve in time according to

$$\frac{d}{dt}P(i) = \sum_j P_j P_{ji} - P_i P_{ij} \tag{2.7}$$

starting from some initial distribution $P_i(0)$. The P_{ij} are given in advance and we assume them to satisfy the microscopic reversibility condition $P_{ij}=P_{ji}$. If we compute the entropy of $\{P_i(t)\}$ according to (2.1) then its rate of change satisfies

$$\begin{aligned}
\frac{dS}{dt} &= -k \sum \frac{d}{dt} P_i \ln P_i - k \sum_i \frac{d}{dt} P_i \\
&= -k \sum_{ij} P_{ij}(P_j - P_i) \ln P_i \\
&= -\frac{k}{2} \sum_{ij} P_{ij}(P_j - P_i) \ln P_i + \frac{k}{2} \sum P_{ij}(P_i - P_j) \ln P_j \\
&= -\frac{k}{2} \sum_{ij} P_{ij} (P_j - P_i) \ln (P_i / P_j)
\end{aligned}$$

where we used the symmetry of P_{ij} and the fact that $\Sigma P_i(t)=1$ to go from the first line to the second, a simple symmetrization to go from the second to the third line. It is clear that the last line is always positive. (Verify that (1-s)lns is always negative for $s > 0$.)

The interesting fact here is that in equilibrium, a distribution P_e such that the left-hand side of (2.7) vanishes, provides a local maximum for $S(\{P_i\})$. In his book [2.5] Gibbs somehow postulates a distribution like (2.3) as an equilibrium distribution for the evolution provided by the Hamilton (Newton) equations of motion in phase space.

It was Boltzmann who used (2.1) as a Lyapunov function for his equation describing the time evolution of particle a density function. See the reprint [2.6]. And, as we said in the

introduction it was Jaynes who characterized equilibrium distribution in terms of a variational principle, even though the ideas were already present in Boltzmann's work.

To finish, we direct the readers who are interested in statistical physics to [2.7] and [2.8] for more on entropy and to [2.9] for an application of conditional probabilities to obtain all equilibrium distributions from the grand-canonical distribution. Also check with [2.10] for more references on the different entropies and for the role of entropy production in the characterization of stationary states. For recent work on these issues, refer to [2.11] and references therein.

REFERENCES

[2.1] Atkins, P. W. "*The Second Law*" W. H. Freeman and Co., New York, 1984.

[2.2] Callen, H. "*Thermodynamics and Introduction to Thermostatistics*" John Wiley, New York, 1985.

[2.3] Jaynes, E. T. "*Information theory and statistical mechanics*" Phys. Rev. 106, pp. 620 - 630, 1957. See [0.1] for more along related lines.

[2.4] Tribus, M. "*Micro and macro-thermodynamics*" Am. Scientist 54, No.2, pp. 201-211, 1966.

[2.5] Gibbs J. W. "*Elementary Principles in Statistical Mechanics*". Dover Books, New York, 1960. Reprint of the 1902 U. of Yale Edition.

[2.6] Boltzmann, L. "*Lectures on Gas Theory*". California Univ. Press, Berkeley, 1964.

[2.7] Wehrl, A. "*General properties of entropy*" Rev. Mod. Phys. 50, No.2, pp. 221-260, 1978.

[2.8] Lindblad, G. "*Non-Equilibrium Entropy and Irreversibility*" D. Reidel Publishing Co., Dordrecht, 1983.

[2.9] Gzyl, H. "*A unified presentation of equilibrium distributions in classical and quantum mechanics*". Ann Inst. Henri Poincare. 32, 1980.

[2.10] Jaynes, E. T. "*The minimum entropy production principle*" Ann. Rev. Phys. Chem. 31, pp. 579-601, 1980 (Reprinted in [0.1]).

[2.11] Garcia-Colin, L. S. "*Entropy and irreversibility macroscopics issues*". Rev. Mex. Física. Supl. 1, pp. 198-201, 1992.

Chapter 3

SOME HEURISTICS

Here we follow Jaynes [3.1] and Papoulis [3.2] in developing some heuristics that sheds light on the concept of entropy and on the method of maximum entropy.

Let X be a discrete random variable taking n values $x_1, x_2, \ldots, x_n$ and let p_i be the probabilities of observing the events $A_i = \{X = x_i\}$. We shall write

$$H(x) = H(p_1, \ldots, p_n) = -\sum_1^n p_i \ln p_i \tag{3.1}$$

and call it the entropy of X or the entropy of $\{p_1, \ldots, p_n\}$.

We may think of p_i as relative frequencies

$$p_i = \lim_{N\to\infty} \frac{1}{N} N_i = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} I_{\{x_i\}}(X_j). \tag{3.2}$$

N_i being the number of times the value x_i appears in a run of N independent observations of X.

A different way of looking at (3.2) is the following. Each of the possible results of observing X N consecutive times is an element of $E^N = E \times \ldots \times E$, $E = \{x_1, \ldots, x_n\}$. If $\mathbf{t} \in E^N$ then

$$P(\mathbf{t}) = P(X_1 = t_1, \ldots, X_n = t_n) = p_1^{N_1} \ldots p_n^{N_n}$$

where $X_1, \ldots, X_N$ are independent copies of X and N_i denotes the number of times the value x_i is repeated in the list $t_1, \ldots, t_N$.

Now, if N is large enough so that $p_i \sim N_i/N$ then

$$P(\mathbf{t}) = p_1^{Np_1} \ldots p_n^{Np_n} = \exp N \sum_{i=1}^{N} p_i \ln p_i = \exp -NH(X).$$

If we say that **the configuration t is typical** whenever $N_i \sim NP_i$, it follows that the number of typical configurations W(typ) given by

$$W(typ) \approx \frac{1}{P(\mathbf{t})} = \exp NH(X) \tag{3.3}$$

from which we obtain

$$H(X) \approx \tfrac{1}{N} \ln W(typ).$$

Below, and in the next chapter, we shall discuss these relations in some detail. For the time being let us compare the number of typical configurations for two distributions corresponding to a random variable having six possible outcomes, that is a die.

The first distribution $\{p_1,...,p_n\}$ is the distribution that maximizes

$$H(p_1,...,p_6) = -\sum_1^6 p_i \ln p_i$$

subject to the constraint

$$\Sigma\, i P_i = 4.5$$

instead of the "classical" on corresponding to the equiprobable situation

$$\tfrac{1}{6} \sum_{i=1}^{6} i = 3.5 .$$

Simple optimization shows that

$$p_i = e^{-\lambda i} \sum_{j=1}^{6} e^{-\lambda j}$$

and with the aid of a computer one obtains the λ such that $\Sigma ip_i = 4.5$. It is $\lambda = -0.37105$ from which it follows that

(3.4) $$(p_1,...,p_6) = (0.0543, 0.0788, 0.1142, 0.1654, 0.2398, 0.3476)$$

to which corresponds an entropy

(3.5) $$H_{\max} = 1.61358 .$$

We leave for the reader to verify that the distribution

(3.6) $$\bar{p}_i = \binom{6}{i-1} p^{i-1}(1-p)^{6-i} \quad 1 \le i \le 6$$

also satisfies $\Sigma ip_i = 4.5$ for $p = 0.7$. But this distribution has entropy

(3.7) $$\bar{H} = 1.4136.$$

The quotient of the two numbers of typical configuration corresponding to (3.5) and (3.7) is given by

$$\frac{\overline{W}_m}{W} = \frac{e^{NH_m}}{e^{N\bar{H}}} = e^{N(\Delta H)} \approx e^{200} > 10^{84}$$

for a sequence of N=1000 throws of the die.

Had we taken N=50 and computed $N_i=p_iN$, the numbers would be (approximately) {3,4,6,8,12,17} and {0,1,7,16,18,8} and since the number of ways a configuration $N_1,...,N_6$ comes up is $N!/N_1!,N_2!,...,N_6!$ in this case the corresponding quotient is

$$\frac{\overline{W}_m}{W} = 38220.$$

That is, the number of microscopical configurations (i.e., chains of 50 throws) for which the frequencies N_i/N correspond to the maximum entropy distribution (3.5) is 38220 times more frequent than the microscopic configurations corresponding to the distribution (3.6).

This is a statement about how different assignments of probabilities are reflected in the outcome of an experiment.

In a statistical physics, in systems with 10^{20} particles the use of asymptotic methods is quite justified. This is the reason why, in almost the first line of any book in statistical physics, the following typical phrase appears: the number of microscopical states accessible to the system, compatible with its macroscopic constraints is [1]

$$W = \exp S/k$$

which is equivalent to assert how disordered a system is, or equivalently, the higher the entropy, the higher the more states it can occupy. By the way, according to the third law of thermodynamics, which asserts that at 0°K the entropy vanishes i.e. S=0. Therefore at 0°K there is only 1 configuration available to the system.

To read about Boltzmann's ideas about these subjects check with [0.24] and with Jayne's essay in [0.7].

Later on we shall see how does the entropy concept appear with the theory of large deviations. The issue is to find the bridge between the two aspects of probability assignments.

To finish, I cannot but help directing the reader to chase back from [3.3] where an application of the maximum entropy method to find "dishonest" dice is explained.

---

[1] This is written in Ludwig Boltzmann's epitaph.

REFERENCES

[3.1] Jaynes, E. T. "*On the rationale of maximum entropy methods*". Proc. IEEE 79, No.2, pp. 939-952, 1982.

[3.2] Papoulis, A. "*Maximum entropy and spectral estimation*". IEEE Trans. Acoust. Speech and Signal Processes. ASSP- 29, No.6, pp. 1176-1186, 1981.

[3.3] Fougere, P. F. "*Maximum entropy calculations on a discrete probability space*".See [0.9].

Chapter 4

ENTROPY FUNCTIONALS

1. Basics.

We shall take a look at some properties of several functionals defined on the set of all positive, σ-finite measures and on the set of all probability measures on a measurable space $(\Omega,\mathcal{F})$.

The definitions and results are variations on the themes developed in [4.1]-[4.2]. We direct the reader to [0.2] where besides the results, many references to original work and applications to statistics are developed. We shall also describe briefly some of the results from the review on entropy inequalities compiled by Dembo, Cover and Thomas, [4.3].

For any measurable space $(\Omega,\mathcal{F})$, $\mathbf{M}(\Omega)$ and $\mathbf{P}(\Omega)$ will denote the sets

$\mathbf{M}(\Omega)$={positive σ-finite measures on (Ω,F)}

$\mathbf{P}(\Omega)$={probability measures on $(\Omega,\mathcal{F})$}.

Definition 4.1. Let $\gamma \in \mathbf{M}(\Omega)$ and $P \in \mathbf{P}(\Omega)$. We define

$$S_\gamma(P) = -\int \frac{dP}{d\gamma} \ln \frac{dP}{d\gamma} d\gamma$$

$P \ll \gamma$ and $\ln dP/d\gamma \in L_1(P)$. We set $S_\gamma(P) = -\infty$ otherwise.

Comment. We shall from now on convene on setting

$$0 \ln 0 = 0, \quad \ln \frac{X}{0} = +\infty, \quad 0(+\infty) = 0$$

unless explicitly computed to be something else.

Definition 4.2. Let X: $(\Omega,\mathcal{F},P) \to (M,\mathbf{M},\mu)$ be an M-valued random variable such that the P-distribution of X is absolutely continuous with respect to the σ-finite measure μ, having density $p(y) = P(X \in dy)/\mu(dy)$. We set

$$S_\mu(X) = -\int p(y) \ln p(y) d\mu(y)$$

if $\ln p(X) \in L_1(P)$ and $-\infty$ otherwise.

The following two cases are the most frequent. When M is a countable set, $\mu(m_i)=1$ (i.e. μ is the counting measure) and $P(X = m_i) = p_i$. Then

$$S_\mu(X) = -\Sigma p(i) \ln p(i).$$

The other case being $M = \Re^n$ and $\mu(dy) = dy$ being the Lebesgue measure. In this case we shall just write

$$S(X) = -\int p(y) \ln p(y)\, dy.$$

Definition 4.3. Let $\gamma,\mu,\nu \in \mathbf{M}(\Omega)$ be such that $\mu<<\gamma$, $\nu<<\gamma$ and μ, ν are finite. Define

$$K_\sigma(\mu, \nu) = \int \frac{d\mu}{d\sigma} \ln\left(\frac{d\mu/d\sigma}{d\nu/d\sigma}\right) d\sigma + \nu(\Omega) - \mu(\Omega)$$

whenever $\ln((d\mu/d\sigma)/(d\nu/d\sigma))$ is in $L_1(d\mu)$ and $+\infty$ otherwise.

Some times the particular case corresponding to $\mu = P$ and $\nu = Q$ being probability measures has a different symbol associated:

Definition 4.4. For $P,Q \in \mathbf{P}(\Omega)$ and $s \in \mathbf{M}(\Omega)$ we set

$$I_\sigma(P, Q) = \int \frac{dP}{d\sigma} \ln\left(\frac{dP/d\sigma}{dQ/d\sigma}\right) d\sigma$$

again when $\ln((dP/d\sigma)/(dQ/d\sigma))$ is in $L_1(dP)$ and $+\infty$ otherwise.

The proof of the following obvious lemma is left for the reader.

Lemma. When $\mu << \nu$ (and $P << Q$) $K(\mu,\nu)$ (and $I(P,Q)$) are independent of σ and denoted by $K(\mu,\nu)$ (and $I(P,Q)$ resp.).

The functionals K and/or I have many names: Kulback Leibler information number, information for discrimination, information distance, information gain or entropy gain of μ (or P) with respect to ν (or Q). The functionals $S_\nu(P)$ or $S_\nu(X)$ are called μ-entropy of P (or X).

Lemma 4.5. With the notation introduced above we have

i) $S_\nu(P)$ is concave in P.

ii) $K(\mu,\nu)$ is convex in μ. When $\mu(\Omega) = \nu(\Omega)$, $K(\mu,\nu)\geq 0$, the identity holds true when

$$d\mu/d\sigma = d\nu/d\sigma .$$

Proof:

i) The function $-x \ln x$ being concave on $[0,\infty)$ yields $S_\nu(aP_1+bP_2)\geq bS_\nu(P_1)+bS_\nu(P_2)$.

ii) The convexity of $K(\mu,\nu)$ can be obtained similarly. When $\mu(\Omega)=\nu(\Omega)$ and setting $c=\{dP/d\sigma>0\}$ we have

$$-K_\sigma(\mu,\nu) = \int_C \frac{dP}{d\sigma} \ln \frac{dQ/d\sigma}{dP/d\sigma} d\sigma \le \ln \int_C \frac{dP}{d\sigma}\left(\frac{dQ/d\sigma}{dP/d\sigma}\right) d\sigma \le \ln \int_C dQ \le 0.$$

When $K(\mu,\nu)=0$

$$0 = \ln \int_D \frac{dP}{d\sigma}\left(\frac{dQ/d\sigma}{dP/d\sigma}\right) d\sigma \le \ln\left(\int_{C\cap\left\{\frac{dQ}{d\sigma}>\frac{dP}{d\sigma}\right\}} \frac{dP}{d\sigma} d\sigma \right) \le 0$$

and the result follows from the strict concavity of lnx and the following lemma. (Nevertheless, see the simpler proof in Theorem 3.1 of [0.1].)

Lemma 4.6. Let g be a positive function defined on $(\Omega,\mathcal{F},P)$. Then $\ln\int g dP \ge \int \ln g\, dP$ with the identity holds and only if g is constant a.s-P.

Proof: The inequality is the obvious concavity of lnx. Recall that lnx is strictly concave, i.e., $\ln(\Sigma a_i x_i)=\Sigma a_i \ln x_i$ when $\Sigma a_i=1$, if and only if $x_1=x_2=\ldots=x_n$.

For any a such that $P\{g\le a\}>0$ and $P(g>a)>0$ we have, when the identity $\ln\int g\, dP=\int \ln g\, dP$ holds, then

$$\ln\left\{ P(g \ge a) \int_{\{g\ge a\}} g\, dP/P\{g \ge a\} + P\{g < a\} \int_{\{g<a\}} g\, dP/P\{g < a\} \right.$$

$$\ge P\{g \ge a\} \ln \frac{1}{P(g\ge a)} \int_{\{g\ge a\}} g\, dP + P\{g < a\} \ln \frac{1}{P\{g<a\}} \int_{\{g<a\}} g\, dP$$

$$= \int_{\{g\ge a\}} \ln g\, dP + \int_{\{g<a\}} \ln g\, dP = \int g\, dP.$$

The assumption implies that the middle term equals the first term and therefore, the strict concavity of the logarithm function implies that

$$\int_{\{g>a\}} g\, dP/P\{g > a\} = \int_{\{g\le a\}} g\, dP/P(g \le a)$$

$$P(g \le a) \int_{\{g>a\}} g\, dP = P(g > a) \int_{\{g\le a\}} g = (1 - P(g \le a)) \int_{\{g\le a\}} g\, dP.$$

From which we obtain

$$P(g \le a)\int g\, dP = \int_{\{g\le a\}} g\, dP \quad or \quad \int g\, dP \le a.$$

Similarly we would obtain

$$P(g > a)\int g\,dP = \int_{\{g \le a\}} g\,dP \quad or \quad \int g\,dP > a.$$

Since this is impossible, we conclude that for some a, $P(g = a) = 1$.

Lemma 4.7. Let $\mu,\nu \in \mathbf{M}(\Omega)$ be such that $P \ll \nu$, $P \ll \mu$ and $\ln (dP/d\mu)$, $\ln(dP/d\nu)$ are in $L_1(dP)$. Then

$$S_\nu(P) - S_\mu(P) = E_P\left[\ln \frac{dP/d\mu}{dP/d\nu}\right] = E_P\left[\ln \frac{d\nu/d\lambda}{d\mu/d\lambda}\right],$$

where $\lambda = a\mu + b\nu$ and $0 < a,b$, $a+b=1$. When μ,ν are finite

$$S_\nu(P) + \nu(\Omega) - S_\mu(P) - \mu(\Omega) = K_\lambda(\nu,\mu).$$

Comment. Instead of $\lambda = a\mu + b\nu$ we could use any σ such that $\lambda \ll \sigma$.

Lemma 4.8. If $P \ll Q_1$, $P \ll Q_2$, $Q_2 \ll Q_1$ then

$$K(P,Q_1) - K(P,Q_2) = E_P[\ln dQ_2 \mid dQ_1]$$

We leave these as exercises for the reader. We only add that when $\nu = Q$ is a probability measure

$$S_\nu(P) = -K_\nu(P,\nu).$$

The next lemma contains the basic behavior relative to changes of variables.

Lemma 4.9. a) Let Φ: $(\Omega,\mathscr{F}) \to (\Omega',\mathscr{F}')$ be a measurable mapping. Let σ, μ, $\nu \in \mathbf{M}(\Omega)$, P, $Q \in \boldsymbol{P}(\Omega)$ and σ', μ', $\nu' \in \mathbf{M}(\Omega')$ and $P', Q' \in \boldsymbol{P}(\Omega)$ be related by $\sigma' = \sigma(\Phi^{-1})$, etc... Then

$$S_{\mu'}(P') \ge S_\mu(P);\ K_{\sigma'}(\mu',\nu') \le K_\sigma(\mu,\nu), I_{\sigma'}(P',Q') \le I_\sigma(P,Q).$$

b) Let X: $(\Omega,\mathscr{F}) \to (M,\mathbf{M})$ be as in definition 4.4 and f: $(M,\mathbf{M}) \to (M',\mathbf{M}')$ be a measurable transformation. Set $X_1 = f(X_1)$, $Q = P \circ X^{-1}$, $Q_1 = P \circ X_1^{-1}$, $\mu_1 = \mu(f^1)$. Then $Q \ll \mu$ implies $Q_1 \ll \mu_1$ and $S_{\mu_1}(X_1) = S_\mu(X)$.

c) As in (b) but now $Q_1 \ll \nu \in \mathbf{M}(M')$ with density q. Then $S_\nu(f(X)) \ge S_\mu(X)$.

Comment. These results are a variation on the theme of section 4, chapter 2 of [0.2].

Proof : It is easy to verify that when P is restricted to $\Phi^{-1}(\mathcal{F})$ (see lemma (1.13)) $dP/d\mu=(dP'/d\mu')\circ\Phi$. Thus

$$S_{\mu'}(P') - S_{\mu}(P) = -\int \frac{dP'}{d\mu'} \ln \frac{dP'}{d\mu'} d\mu' + \int \frac{dP}{d\mu} \ln \frac{dP}{d\mu} d\mu =$$

$$-\int \frac{dP}{d\mu} \ln\left(\frac{dP'}{d\mu'}\right) \circ \Phi \, d\mu + \int \frac{dP}{d\mu} \ln \frac{dP}{d\mu} d\mu = \int \frac{dP}{d\mu} \ln \frac{dP/d\mu}{(dP'/d\mu')\circ\Phi} \geq 0$$

Similarly

$$K_{\sigma'}(\mu', \nu') - K_{\sigma}(\mu, \nu) = -\int \left(\frac{d\mu/d\sigma}{d\nu/d\sigma}\right) \ln \frac{(d\mu/d\sigma)/(d\nu/d\sigma)}{[(d\mu'/d\sigma')/(d\nu'/d\sigma')]\circ\Phi} \frac{d\nu}{d\sigma} d\sigma$$

which ≤ 0 due to (4.5-ii). Instead of proving (b) we prove (c).

$$S_{\nu}(f(X)) - S_{\mu}(X) = -\int \ln q(f(X)) dP + \int \ln P(X) dP = \int P(x) \ln \frac{P(x)}{q(f(x))} \mu(dx)$$

which is positive.

Next we shall introduce some variations on the theme of conditional entropies.

Definition 4.10. In the same setting as in (4.1) assume further that ν restricted to the sub σ-algebra G is σ-finite. Then we shall define

$$S_{\nu}(P/G) = -\int E_{\nu}\left[\left(\frac{dP}{d\nu}|G\right)\right] \ln E_{\nu}\left[\left(\frac{dP}{d\nu}\right)|G\right] d\nu = -E_P\left[\ln E_{\nu}\left[\left(\frac{dP}{d\nu}\right)|G\right]\right]$$

similarly.

Definition 4.11. Let μ,ν,σ be in $\mathbf{M}(\Omega)$ with μ,ν finite and absolutely continuous with respect to σ. Let G be a sub-σ-algebra of $\mathcal{F}$ such that s is σ-finite on G. Set

$$K_{\sigma}^{G}(\mu|\nu) = \int \frac{d\mu}{d\sigma} \ln \frac{E_{\sigma}[d\mu/d\sigma|G]}{E_{\sigma}[d\nu/d\sigma|G]} + \nu(\Omega) - \mu(\Omega).$$

When μ=P, ν=Q are probability measures we obtain

$$I_{\sigma}^{G}(P|Q) = \int dP \ln \frac{E_{\sigma}[dP/d\sigma|G]}{E_{\sigma}[dQ/d\sigma|G]}$$

and we set (compare with chapter II of [4.1])

$$I_{\sigma}^{\mathcal{F}/G}(P|Q) = \int dP \ln \frac{(dP/d\sigma)/E_{\sigma}[(dP/d\sigma)|G]}{(dQ/d\sigma)/E_{\sigma}[(dP/d\sigma)|G]}.$$

Notice that $E_s[dP/d\sigma|G]$ is a density for P restricted to G with respect to σ restricted to G. Also as measures on (Ω,G), $(dP/d\sigma)/E_s[dP/d\sigma)|G]$ is the density of $E_s[dP/d\sigma|G]\, d\sigma$ with respect to $(dP/d\sigma)d\sigma$.

The analogue of theorem (2.3) in chapter II of [4.1] is

Theorem 4.12. For $G \subset \mathcal{F}$ and P,Q, σ as above

$$I_{\sigma}^{\mathcal{F}}(P|Q) = I_{\sigma}^{G}(P|Q) + I_{\sigma}^{\mathcal{F}/G}(P|Q)$$

and $I_s^{\mathcal{F}|G}(P|Q) \geq 0$, the identity being satisfied when

$$\frac{dP}{d\sigma} \Big/ \frac{dQ}{d\sigma} = \frac{E_{\sigma}\left[\frac{dP}{d\sigma}|G\right]}{E_{\sigma}\left[\frac{dQ}{d\sigma}|G\right]} \qquad a.e.P.$$

Proof: The first assertion is obvious. For the second use a variant of lemma (4.5) and the comment preceding the statement of the theorem.

Comment. When $Q=\sigma$ and $P<<Q$, we restate (4.11) as

$$I^{\mathcal{F}}(P|Q) = I^{G}(P|Q) + I^{\mathcal{F}|G}(P|Q).$$

Also $I^{\mathcal{F}|G}(P|Q) \geq 0$, the identity sign occurs whenever $dP/dQ=E_Q[dP/dQ|G]$.

Comment. Similarly, when $G_1 \subset G_2$ are such that s restricted to G is finite, then

$$I_{\sigma}^{G_2}(P|Q) = I_{\sigma}^{G_1}(P|Q) + I_{\sigma}^{G_2|G_1}(P|Q) \tag{4.13}$$

and similar assertions to those in (4.11) also apply here.

The previous notions of conditional are of interest to statisticians. Associated with definition (4.2) there are two notions of conditional entropy of interest in information and coding theory and in the theory of dynamical systems. Let us rewrite things as follows.

Let $(M_i,\mathbf{M}_i,n_i)$ be two σ-finite spaces and let $M=M_1 \times M_2$, $\mathbf{M}=\mathbf{M}_1 \otimes \mathbf{M}_2$, $\nu=\nu_1 \otimes \nu_2$ be their product space.

Definition 4.14. Let $(\Omega,\mathcal{F},P)$ be a probability space (X,Y): $\Omega \to M$ be a random variable such that

$$P((X,Y) \in dx\,dy) = P(x,y)\nu_1(dx)\nu_2(dy).$$

Let $P(x|y)=P(x,y)/\int P(x,y)\nu_1(dx)$ be the conditional density as in (1.18). Then set

$$S_\nu[X|Y](y) = -\int P(x|y)\ln P(x|y)\nu_1(dx)$$

$$S_\nu(X|Y) = \int S_\nu[X|Y](y)P_2(y)\nu_2(dy) = -\int P(x,y)\ln P(x|y)\nu_1(dx)\nu_2(dy),$$

where $P_2(y)=\int P(x,y)\nu_1(dx)=P(Y\in dy)/\nu_2(dy)$. Notice the difference of notations. $S_\nu[X|Y]$ is a function of y where $S_\nu(X|Y)$ is a number. Some arithmetic is enough to verify

Lemma 4.15. With the notations introduced above

$$S_\nu(X,Y) = S_\nu(X|Y) + S_\nu(Y).$$

Lemma 4.16. Also $S_\nu(X|Y) \le S_\nu(X)$. The identity holds when X and Y are independent.

Proof:

$$S_\nu(X) - S_\nu(X|Y) = \int P(x,y)\ln\frac{P(x|y)}{P(x)}\nu_1(dx)\nu_2(dy)$$

$$= \int P(x,y)\ln\frac{P(x,y)}{P_1(x)P_2(y)}\nu_1(dx)\nu_2(dy) \ge 0$$

where the last step follows from lemma (4.5). Here $P_1(x)=\int P(x,y)\nu_2(dy)$. Again, applying lemma (4.5) to the last term we conclude that the identity $S_\nu(X/Y)=S(X)$ holds whenever $P(x,y)=P_1(x)P_2(y)$, i.e., when X and Y are independent.

A repeated application of this lemma yields.

Lemma 4.17. Let $\{X_i | 1\le i\le n\}$ be random variables with values in $(\Omega_i,\mathbf{M}_i,\nu_i)$ respectively, etc. Then

$$S_\nu(X_1,\ldots,X_n) = \sum_{i=1}^{n} S_\nu(X_i|X_{i-1},\ldots,X_1) \le \sum_{i=1}^{n} S_{\nu_i}(X_i)$$

the identity holds when the X_i are independent.

Let us finish this section with some definitions.

For $\mu\in\mathbf{M}(\Omega)$ and Φ: $\Omega\to\Re^k$ a measurable function, the (μ,Φ)-partition function is defined by

(4.18) $$Z_{\mu,\Phi}(t) = \int_\Omega \exp -(t,\Phi)\,d\mu$$

for $t \in \mathfrak{R}^k$ when the integral exists. Whenever (μ,Φ) remain fixed, we will not mention them explicitly. We shall introduce the notation

$$D(\mu, \Phi) = \{\mathbf{t} \in \mathfrak{R}^k | Z_{\mu\Phi}(\mathbf{t}) < \infty\}. \tag{4.19}$$

The Φ - **Hellinger arc of** m is the family of measures, absolutely continuous with respect to $d\mu$, defined by

$$d\mu_t^{\phi} = (\exp -(\mathbf{t}, \phi)/Z_{\mu\phi}(\mathbf{t}))d\mu \tag{4.20}$$

and for fixed $\mathbf{c} \in \mathfrak{R}^k$, we set

$$P(\mathbf{c}, \Phi) = P \in \mathbf{P}(\Omega) | E_{\mathbf{P}}(\Phi) = \mathbf{c}. \tag{4.21}$$

Observe that for $\mathbf{t} \in D(\mu, \Phi)$ and $P \in \mathbf{P}(\mathbf{c}, \Phi), P << \mu$.

$$K_\mu(P, \mu_t^{\Phi}) = H_{\mu\Phi}(\mathbf{t}) - S_\mu(\mathbf{P}) \tag{4.22}$$

where

$$H_{\mu\Phi}(\mathbf{t}) = \ln Z_{\mu\Phi}(\mathbf{t}) + \langle \mathbf{t}, \mathbf{c} \rangle \tag{4.23}$$

is related to the physicists free energy.

The important comment to make here is that, for given and fixed $\mathbf{t}$, $\mathbf{c}$ in $\mathfrak{R}^k$ to minimize $K_\mu(P, \mu_t^{\Phi})$ over $\mathbf{P}(\mathbf{c},\Phi)$ is the same as to maximize $S_\mu(P)$ over the same set.

Whenever $D(\mu,\Phi)$ is convex and nonempty, and $\mathbf{t}^*$ is the point at which $H_{\mu\Phi}(\mathbf{t})$ reaches a minimum $H^*(c,\Phi)$, then, $K_\mu(P,\mu_t^{\Phi})$ reaches its minimum at μ_t^{Φ} and $S_\mu(P)$ reaches its maximum at μ_t^{Φ} as well and both sides of (4.23) vanish. We shall have more about this below when we study the maximum entropy methods.

Lemma 4.24. The set $D(\mu,\Phi)$ is convex when nonempty and $\ln Z_{\mu\Phi}(t)$ is convex.

Proof: Let $0 \le a,b$ be such that $a + b = 1$ and set $p = 1/a$, $q = 1/b$. For $\mathbf{t}_1, \mathbf{t}_2 \in D$ we have $\exp a(\mathbf{t},\Phi)$ is in $L_q(\mu)$. Then

$$\ln Z(at_1 + bt_2) = \ln \int \exp +(at_1 + bt_2, \Phi)\, d\mu \le \ln Z(t_1)^a Z(t_2)^b = a \ln Z(t_1) + b \ln Z(t_2)$$

where the inequality is obtained by applying Hoelder's inequality.

Comment. This lemma implies that $H(\mu,\Phi,t)$ is a convex function of t.

Consider the Cauchy distribution on $-\infty < x < \infty$ with density $c/(c^2+x^2)\pi$ with respect to Lebesgue measure dx. For $\Phi(x)=x$ we have $D(dx,x) = \phi$ the empty set and when we restrict x to $[0,\infty)$ then $D(dx,x)=(-\infty,0)$. Notice that in this case

$$\int x\, d\mu_t < \infty \quad for \quad t > 0 \quad but \quad \int x\, d\mu_0 = \infty$$

Lemma 4.25. When $D(\mu,\Phi)$ has a nonempty interior $D^*(\mu,\Phi)$, the mapping $D^*(\mu,\Phi) \to \Re^k$ given by $t \to \int \Phi\, d\mu_t$ is defined and $\int \Phi_1\, d\mu_t = \partial \ln Z(t)/\partial t_1$. When the covariance matrix exists and is of full rank, the mapping is 1:1.

Proof: Let h be any fixed vector in $\Re^k$ and $c > 0$ such that $t + sh \in D^*(\mu,\Phi)$ for $t \in D^*(\mu,\Phi)$ for $0 < |s| < c$. Since Z(t+sh) is finite

$$s\int (h,\Phi)\, d\mu_t = \int \ln e^{s(h,\Phi)}\, d\mu_t \le \ln \int e^{s(h,\Phi)}\, d\mu_t < \infty$$

because of the concavity of lnx. Since the sign of s is arbitrary we conclude that $\int (h,\Phi)\, d\mu_t < \infty$ for any h.

It is not hard to see that Z(t+sh) is twice differentiable at s=0 and

$$\frac{d^2 \ln}{ds^2} Z(t+sh) \Big|_{s=0} = \int (h,\Phi)^2\, d\mu_t - \left(\int (h,\Phi)\, d\mu_t\right)^2 = (h, Ch)\,,$$

where C is the covariance matrix of Φ, which happens to be the Jacobian of the map $t \to \int \Phi\, d\mu_t$.

Even though $D^*(\mu,\Phi)$ is convex, the range of the map described above is not necessarily convex. For an example take a look at the third page of [4.4] where some properties of exponential families are investigated. For some other metric properties associated to the Kullback I-divergence see [4.5] and [4.6]. Below we shall be quoting extensively from the last one.

2. *Entropy inequalities.*

Let us now describe, somewhat scantily, a few results from [4.3]. We commented at the end of chapter 3 that setting

(4.26) $$N_\mu(\mathbf{P}) = b \exp a S_\mu(P)$$

(which will be called the entropy power of P relative to μ) gives us an intuitive way of understanding how big is the support of P. For example, when Ω is a finite set $\Omega=\{1,2,\ldots,n\}$ and

μ is the counting measure and P{i}=1/n then S=lgn and, when a=b=1 then $N_m(P)=n$, the number of states that can be occupied.

When $(\Omega,\mathcal{F})=(\Re^n,\boldsymbol{B})$, $\mu(dx)$ is the Lebesgue measure and P has Gaussian density with covariance $K_{ij}=E_p X_i X_j$ and zero mean, setting a=2/n, b=1/2πe, we obtain $N_\mu(P)=|K|^{1/n}$

When Ω is a product space $\Omega_1 \otimes \Omega_2$ and P *is a probability on* Ω, *absolutely continuous* with respect to a product $\nu = \nu_1 \otimes \nu_2$, then Lemma (4.17) asserts that

$$S_\nu(\mathbf{P}) \le \Sigma S_{\nu i}(P_i)$$

therefore, from (4.26) we have

(4.27) $$N_\nu(P) \le N_{\nu 1}(P_1) N_{\nu 2}(P_2)$$

with the identity holds whenever P is actually a product of its projections P_1 and P_2.

Notice that (4.27) does not depend on the nature of the measures involved. In the papers by Shanon and by Stam quoted in [4.3] the following is proved: let X,Y be two independent $\Re^n$ -valued variables having density with respect to Lebesgue measure, such that S(X) and S(Y) exist. Then, the Shanon's entropy power inequality asserts that

(4.28) $$N(X+Y) \ge N(X) + N(Y).$$

But, notice that when X,Y take finitely many values the opposite inequality seems to hold. For example, let X,Y be independent such that P(X = ±1) = P(Y = ±1) = 1/2. In this case S(X) = S(Y) = ln2 and S(X+Y) = 3/2 ln2. Using (4.26) with a = b = 1 we obtain

(4.29) $$N(X+Y) = 2^{3/2} \le N(X) + N(Y) = 4.$$

One way to understand this is to consider finite sets Ω_1 and Ω_2 on which probabilities are defined and a map Φ: $\Omega \to \Omega'$ such that F is into and $|\Omega'| = |\Omega_1| + |\Omega_2| - 1$ = number of diagonals (or antidiagonals). Thus the conjecture here is that if we set $P = (P_1 \otimes P_2)\circ\Phi'$ then $N(P) \le N(P_1) + N(P_2)$ as suggested by (4.29).

When the density of the distribution of the $\Re^n$ valued random variable is continuously differentiable and together with its first partial derivatives decays sufficiently rapidly at infinity, then between **the Fisher information** of X defined by

(4.30) $$J(X) = \int \left(\frac{(\nabla p)^2}{p(x)} \right) dx$$

and the entropy N(X) the following relationship exists.

Theorem 4.31. (De Bruijn's Identity) Let X be defined as above and Z be a N(0,1), $\Re^n$ valued random variable. Then

$$\frac{d}{d\varepsilon}S(X+\sqrt{\varepsilon}Z)=\tfrac{1}{2}J(X) \tag{4.32}$$

and furthermore, the **isopermetric inequality for entropies** states that

$$\tfrac{1}{n}J(X)N(X)\geq 1 \tag{4.33}$$

The **Fisher information** matrix **J**(X) is defined by

$$\mathbf{J}(X)_{ij}=\int\left(\frac{\partial p}{\partial x_i}\right)\left(\frac{\partial p(x)}{\partial x_i}\right)\frac{dx}{p(x)}. \tag{4.34}$$

Let $\psi(x)$, $\Phi(y)$ be conjugate elements in $L_2(\Re^n)$, i.e. $\Phi(y)$ is the Fourier transform of $\psi(x)$. Define X, Y to be random variables with densities

$$\rho_\psi(x)=\frac{|\psi(x)|^2}{\|\psi\|_2^2},\quad \rho_\Phi(y)=\frac{|\Phi(y)|^2}{\|\Phi\|_2^2}$$

Let K_X and K_Y be the covariance matrices of X and Y respectively. Then **Stam's uncertainty principle** asserts that

$$\begin{aligned}16\pi^2K_Y-\mathbf{J}(X)&\geq 0\\ 16\pi^2K_X-\mathbf{J}(Y)&\geq 0\end{aligned} \tag{4.35}$$

(where 0 for matrices means positive definite). The **Cramer-Rao** inequality asserts that

$$\begin{aligned}\mathbf{J}(X)-K_X^{-1}&\geq 0\\ \mathbf{J}(Y)-K_Y&\geq 0\end{aligned} \tag{4.36}$$

and the combination of these two yields the analogue of the **Heisenberg - Weyl uncertainty relations** in quantum mechanics in four possible equivalent statements

$$16\pi^2K_Y-K_X^{-1}\geq 0$$

$$16\pi^2K_X-K_Y^{-1}\geq 0$$

(4.37)

$$16\pi^2 K_X^{1/2} K_Y K_X^{1/2} - I \geq 0$$

$$16\pi^2 K_Y^{1/2} K_X K_Y^{1/2} - I \geq 0.$$

To conclude we shall present an important lower bound to Kullback's information again. As in definition (4.4), let $P,Q \in \mathbf{P}(\Omega)$ and $\mu \in \mathbf{M}(\Omega)$ be such that $K_\mu(P,Q)$ is finite, then

$$\tfrac{1}{4}\left(\int \left|\frac{dP}{d\mu} - \frac{dQ}{d\mu}\right| d\mu\right)^2 \leq K_\mu(P,Q)$$

There are several proofs of this inequality. See [5.5] for references. Here we present Kullback's version. It appears as a set of exercises at the end of chapter 3 of [0.2] in which references to the original papers can be found. Let us set $f_1 = dP/d\mu$ and $f_2 = dQ/d\mu$. They are positive, integrable functions (with respect to $d\mu$). An easy application of Cauchy-Schwartz inequality yields

$$\int (f_1 f_2)^{1/2}\, d\mu \leq 1$$

the identity holds only when $f_1 = f_2$. Rewriting the integrand above as $f_1 (f_2/f_1)^{1/2}$ and using the concavity of the logarithm functions we obtain

$$-2 \ln \int (f_1 f_2)^{1/2}\, d\mu \leq K\mu(P,Q)$$

and since for $x \leq 1$ $\log x \leq x-1$ we obtain, using the normalization $\int f_1\, d\mu = \int f_2\, d\mu = 1$, that

$$K_\mu(P,Q) \geq 2\left(1 - \int (f_1 f_2)^{1/2}\, d\mu\right) = \int \left(f_1^{1/2} f_2^{1/2}\right)^2 d\mu.$$

The last step consists of verifying the inequality

$$\int \left(f_1^{1/2} - f_2^{1/2}\right)^2 d\mu \geq \tfrac{1}{4}\left(\int |f_1 - f_2|\right)$$

for which you rewrite $|f_1 - f_2| = |f_1^{1/2} - f_2^{1/2}|\, |f_1^{1/2} + f_2^{1/2}|$ and use Cauchy-Schwartz inequality again. Though lengthy, this is straightforward. Actually, a slick, but hard to hit upon, trick is presented in [4.13]. Observe that for $v > 0$, $u > 0$

$$(u-v)^2 \leq ((2u/3) + (4v/3))(u \ln (u/v) + v - u)$$

substitute $f_1 = u$, $f_2 = v$, take square roots and use Cauchy-Schwartz inequality to obtain the inequality.

This was just a sample of an interesting class of inequalities. I hope to have wetted your appetite enough.

3. *Axiomatic characterization of entropies.*

Certainly the entropy functionals we introduced in section 1 do not seem to be the obvious convex (or concave) functionals, having metric-like properties, to be defined on $\mathbf{P}(\Omega)$ or $\mathbf{M}(\Omega)$. This has prompted many people, to **postulate "natural" assumptions** on the functionals to be studied, which would lead to functional equations to be satisfied by the desired functionals. Then they would prove that either the entropy functional or the Kullback-like directed divergence were the unique functionals having these properties.

But then again, one is always left wondering why the chosen postulates are natural.

Anyway, a line of research traceable to Shannon's work of 1948 was summarized in the book by Aczel and Daróczy [4.7]. The results there concern entropy functionals on probability spaces having finitely many atoms.

To obtain $S_\mu(P)$ or $I_\mu(P,Q)$ from axioms on functionals defined on $\{P\in \mathbf{P}(\Re^n)\colon P<<\mu\}$ or $\{P\in \mathbf{P}(\Re^n)\colon P<<\mu\}^2$ see the work of Forte and Sastri [4.8]-[4.9] and that of Johnson and/or Shore.

Assume that a concave functional

$$F\colon \{P \in \mathbf{P}(\Re^n)\colon P << \mu\} \to \Re$$

i) Is in subadditive with respect to projections, i.e. when P_1 and P_2 are restrictions to $\Re^{n_1}$ and to $\Re^{n_2}$ (identifying $\Re^n = \Re^{n_1} \otimes \Re^{n_2}$) then

$$F_\mu(P) \le F_{\mu_1}(p_1) + F_{\mu_2}(p_2),$$

where the identity holds whenever $\mu = \mu_1 \otimes \mu_2$ and $P = P_1 \otimes P_2$.
ii) If $T\colon \Re^n \to \Re^n$ is μ-preserving and f(T) is the density of $P \circ T^{-1}$ (see lemma 4.9) then

$$F(P(T^{-1})) = F(P)$$

iii) If P_n is such that $dP_n/d\mu \uparrow dP(d\mu)$, then $F(P_n) \to F(P)$.
Then $F(P) = -aS_\mu(P)+bn+c\ \ln\mu(dP/d\mu>0)$.

If finiteness is insisted upon, then c = 0 whenever $\mu(\Re^n) = \infty$. The constant b can be different from zero in applications where n can vary. For example, in statistical physics the entropy is an extensive quantity depending on the number of particles.

In [4.10] the postulates of uniqueness, invariance, system independence and subset independence are manipulated to obtain not only the functional forms of $S_\mu(P)$ or $I_\mu(P,Q)$ but Jaynes principle of maximum entropy and Kullback's principle of minimum cross-entropy, as the **uniquely correct methods** for inductive inference when information is given in the form of expected values.

REFERENCES

[4.1] Kullback, S., Keegel, J. and Kullback, J. "*Topics in Statistical Information Theory*". Lect. Notes in Stat. No.42. Springer - Verlag, Berlin, 1987.

[4.2] Dacunha, D. and Gamboa, F. "*Maximum d"entropie et probleme des moments*" Ann. Inst. Poincaré. 26, No.9, pp. 576-596, 1990.

[4.3] Dembo, A., Cover, T. and Thomas, J. "*Information theoretic inequalities*". IEEE Transac-Info. Theory. 37, No.6, pp. 1501-1518, 1991.

[4.4] Efrom, B. "*The geometry of exponential families*" Ann. of Statistics. 6, No.2, pp. 362-376, 1978.

[4.5] Rodriguez, C. C. "*The metrics induced by the Kullback number*". In [0.9].

[4.6] Cizár, I. "*I-Divergence geometry of probability distributions and minimization problems*". Ann. of Probability. 3, No.1, pp. 146-158, 1975.

[4.7] Aczel, J. and Daróczy, Z. "*On measures of information and their characterization*". Acad. Press, New York,1975.

[4.8] Forte, B. and Sastri L. "*Is there something missing in the Boltzmann entropy*". Jour - Math. Phys. 16, No.7, pp. 1453-1456, 1975.

[4.9] Forte, B. and Sastri L. "*Representation of the entropy functional for a grand canonical ensemble in classical statistical mechanics*" J. Math. Phys. 18, No.7, pp. 1299-1302, 1975.

[4.10] Johnson, R. W. "*Axiomatic characterization of the directed divergences and their linear combinations*". IEEE Trans. Info. Theory. IT-25, No.6, pp. 709-716, 1979.

[4.11] Shore, J. E. and Johnson, R. W. "*Axiomatic derivation of the principle of maximum entropy...*". IEEE Trans. Info. Theory. IT-26, No.1, pp. 26-37, 1980.

[4.12] Borwein, J. M. and Lewis, A. S. "*Convergence of Best entropy estimates*" SIAM Jour. Optimizat. 1, No.2, pp. 191-205, 1991.

Chapter 5

THE METHOD OF MAXIMUM ENTROPY

1. *Kullback's and Jaynes' reconstruction methods.*

In this chapter we carry out the program described in the introduction for the reconstruction problems at levels 1 and 2 of the M.E.M.. For the sake of a presentation that more or less follows the chronological development of the results,we will be somewhat repetitive.

To begin at the beginning, one of our basic reconstruction problems was to find the measure $P << \mu$ realizing

$$(P1) \qquad \inf\{K_\mu(P, Q) \mid P << \mu;\ EP[\Phi] = \mathbf{c}\},$$

where Φ: $\Omega \to \Re^k$ is a given measurable, finite valued, function and $c \in \Re^k$, $Q << \mu$ is a fixed measure. Our first result dates back to the fifties. The following theorems or variations on the themes of theorems 2.1 and 2.2 of chapter 4 of [0.2].

Theorem 5.1. With the notations, and assumptions introduced above, assume that $\mathbf{P}_\mu(c,\Phi)=\{P \in \mathbf{P}(\Omega) | P << \mu,\ E_P(\Phi)=\mathbf{c}\}$ is not empty, that $\mathbf{c} \in \text{int}(D(c,\Phi))=\text{int}\{\mathbf{t} \in \Re^k | Z_{Q,\Phi}(t) < \infty\}$. Then, i)

$$(5.2) \qquad K_\mu(P, Q) \geq \ln Z_{Q,\Phi}(t) - (c, t),\ c = \nabla_t \ln Z_{Q,\Phi}(t)$$

and (ii), the solution to (P1) is given by

$$\frac{dP^k}{d\mu} = \exp -(\mathbf{t}, \Phi)\frac{dQ}{d\mu} / Z_{Q,\Phi}(t^*)$$

where t^* is the element in $intD(Q, \Phi)$ such that

$$\mathbf{c} = \nabla_t \ln Z_{Q,\Phi}(t)|_{t^*}$$

Proof: Fix $\mathbf{t} \in D(Q,\Phi)$, $t_0 \in \Re$ and define L on $(0,\infty) \times \Omega$

$$L(s) = s \ln s + s(\mathbf{t}, \Phi) + st_0$$

where we dropped the variable ω and we shall be dropping as many super and subscripts as possible. L(s) has, for each given ω, a maximum at $s_0()$=exp-{(**t**,Φ)+t_0+1} at which $L(s_0)$=-s_0.

Since L"(s)=1/s, there is $s_1(\omega)$, lying between $s_0(\omega)$ and s such that

$$L(s) = L(s_0) + (s - s_0)L'(s_0) + (s - s_0)^2/2s_1$$

$$= -s_0 + (s - s_0)^2/2s_1 .$$

Actually, this identity defines s_1. Set dP/dμ=f_1, dQ/dμ=f_2 and substitute f_1/f_2 for s in the identity above. After integrating with respect to dQ=f_2dμ , we obtain

$$\ln(f_1/f_2)\,dQ = -\int \exp-((\mathbf{t},\phi) - t_0 - 1)dQ + \tfrac{1}{2}\int\left(\frac{f_1}{f_2} - \exp-((\mathbf{t},\phi) - t_0 - 1)\right)^2 \frac{dQ}{s_1} \tag{5.3}$$

from which we obtain

$$K_\mu(P, Q) \geq -\ln Z(\mathbf{t}) - (\mathbf{t}, \mathbf{c})$$

where we used the fact that ∫Φf_1dμ=**c**. In the lemmas below we prove that

$$H(\mathbf{t}) = (\mathbf{t}, \mathbf{c}) + \ln Z(\mathbf{t})$$

is negative, analytic and, whenever the hypothesis of the theorem are met, (ii) is fulfilled.

Before doing the lemmas, we extend Theorem 5.1 as

Theorem 5.4. With the same notations as above, but now we consider P<<μ and Q<<μ with respect to which Φ is integrable and

$$E_P(\Phi, \mathbf{t}) \leq (\mathbf{t}, \Theta)$$

for **t**∈ int D(Q,Φ), then

$$K_\mu(P, Q) \geq -\ln Z(\mathbf{t}) - (\mathbf{t}, \Theta)$$

where $\Theta = \nabla_t \ln Z(\mathbf{t})$.

Proof: Denote by Q^* the measure with density

$$(\exp-(\mathbf{t}, \phi)/Z(\mathbf{t}))(dQ/d\mu)$$

with respect to μ. Then

$$K_\mu(P, Q) = K_\mu(P, Q^k) - \int (\mathbf{t}, \Phi) f_1 d\mu - \ln Z(\mathbf{t}).$$

Lemma 5.5. For $\mathbf{t} \in int\ D(Q,\Phi)$, the complex valued extension of $Z(\mathbf{t})$ obtained by replacing $\mathbf{t}$ by $\mathbf{t}+i\mu=\mathbf{Z}$ is analytic in $\mathbf{Z}$ and the following hold

(*a*) $$\nabla_t Z(\mathbf{t}) = -\int \Phi e^{-(\mathbf{t},\Phi)} dQ$$

(*b*) $$Hess(Z)(t) = \int \Phi\Phi^t e^{-(\mathbf{t},\Phi)} dQ \geq 0$$

where the identity holds only when $Q(\{\Phi=0\})=1$.

Proof: From Jensen's inequality it follows that $Z(\mathbf{t})$ is well defined, continuous in $\mathbf{Z}$ when $Re\mathbf{Z}_i \in intD(Q,\Phi)$ and from Morera's theorem applied to each component we get the analyticity.

Let $\mathbf{t} \in intD(Q,\Phi)$, $\mathbf{n} \in \Re^k$ and $\zeta>0$ be such that $\mathbf{t}+s\mathbf{n} \in intD(Q,\Phi)$ for all $|s|<\zeta$. Since

$$-s(\mathbf{u},\Phi)\exp(\mathbf{t},\Phi) \leq \exp(\mathbf{t}+s\mathbf{u},\Phi) - \exp-(\mathbf{t},\Phi)$$

the first identity drops out. To obtain (b) start from (a) and invoke differentiability through analyticity.

Lemma 5.6. When the inequality in (5.5)(b) holds, the function $\theta(t)= -\nabla_t \ln Z(\mathbf{t})$ defined on $intD(Q,\Phi)$ is one-to-one.

Comment. The range of the mapping thus defined may not be convex. See example at the end of section 2 of [4.4]. It is nevertheless an open set.

Proof: For the reader. It is based on (5.5b).

We shall denote by $\mathbf{t}(\theta)$ the value of $\mathbf{t} \in intD(Q,\Phi)$ for which $\theta= -\nabla \ln Z(\mathbf{t})$. It is also easy to see that.

Lemma 5.7. $E_{Q^t}(T-\theta)^2 = Hess(\ln Z(t)) = J(\theta)$. Also $J(\theta)J(\mathbf{t})=I$, where $J(\theta)$ and $J(\mathbf{t})$ denote the Jacobian matrices of the mappings $t \to \theta(\mathbf{t})$ and $\theta \to \mathbf{t}(\theta)$.

Lemma 5.8. Assuming that the inequality in (5.5-b) holds, let $\mathbf{t}(\theta)$ be the inverse function to that defined in Lemma 5.6. Then $H(\theta) = H(\mathbf{t}(\theta))$ is negative and has a strict maximum at $\theta(0)= \int \Phi dQ$.

Proof: Note that

$$K_\mu(Q_t^\Phi, Q) = -H(\mathbf{t}) \geq 0 .$$

Then so is $-H(\theta)$. Using $H(\mathbf{t})= -(\theta,t)-\ln Z(\mathbf{t})$ with $\theta= -\nabla_t \ln Z(\mathbf{t})$. Solving for $\mathbf{t}(\theta)$, differentiating, using (5.6) and the chain rule one obtains $\nabla\theta H(\theta)=\mathbf{t}(\theta)$. Using Lemma 5.7 and the assumptions, it follows that the quadratic form associated to Hess $H(\theta)$ is strictly positive. Therefore $H(\theta(0))=0$ is the maximum value of $H(\theta)$.

To completely finish the proof of theorem 5.1, notice that when the datum $\mathbf{c}$ is in the image of $int(D(Q,\Phi))$ by $-\nabla_t \ln Z(t)$, then

$$\inf\{K_\mu(P,Q)/P << \mu, E(P) = c\} = -\sup\{\mathbf{t}/(t,c) + \ln Z(\mathbf{t})\}$$

and the supremum is reached at $\mathbf{t}^*$ such that $\mathbf{c} = -\nabla_t \ln Z(\mathbf{t})$.

The reconstruction problem corresponding to the method of maximum entropy at level 1 consists of finding

(P2) $$\sup\{S_\mu(P)/P << \mu; EP(\Phi) = c\},$$

where the meaning of the symbols is as above. Also, now $Z(\mathbf{t})$ stands for $Z_{\mu,\Phi}(\mathbf{t})$. The solution to (P2) is contained in

Theorem 5.9. For a given $\mathbf{c}$ in the image of $\mathrm{intD}(\mu,\Phi)$ by $-\nabla_t \ln Z(\mathbf{t})$, let $\mathbf{t}^*$ in $\mathrm{intD}(\mu,\Phi)$ be the point at which $H(\mathbf{t}) = (\mathbf{c},\mathbf{t}) + \ln Z(\mathbf{t})$ reaches its maximum value. Then (P2) is solved by $d\mu_{t^*}^{\phi} = \exp -(\mathbf{t}^*, \phi) d\mu / Z(\mathbf{t})$.

Proof: Flip a few pages back, and check (4.23). It asserts that if $\mathbf{t} \in \mathrm{intD}(\mu,\Phi)$ and $P << \mu$, $E_P(\Phi) = c$, then

$$K_\mu(P, \mu_{t^*}^{\Phi}) = H(t^*) - S_\mu(P),$$

where $H(\mathbf{t}^*) = (\mathbf{c},\mathbf{t}^*) + \ln Z(\mathbf{t}^*)$. Since $K_\mu(P,\mu_{t^*}^{\Phi}) \geq 0$ we have $H(\mathbf{t}^*) \geq S_\mu(P)$. Also, since at $\mathbf{t}^*$. $H(\mathbf{t}^*)$ reaches its maximum value, we see that $K_\mu(P,\mu_{t^*}^{\Phi}) = 0$ only when $P = \mu_{t^*}^{\Phi}$ or $S_\mu(P)$ reaches its maximum there.

It is not hard to realize that things have been set up so that (P1) and (P2) have as solutions specific elements of the Φ-Hellinger arc of Q or μ. The interesting results by Czizar in [5.1] assert that i) if the solution to (P1) exists it is an element of the Φ-Hellinger arc of Q and (ii) whenever $D(Q,\Phi)$ has a nonempty interior, then (P1) has a solution.

In an appendix at the end of this Chapter we present, without proof, a few results taken from Chapter 9 of [5.8] about the relationship between the range of the map $\nabla_t \ln Z(t)$ and the support in $\Re^k$ of the induced measure $\mu \circ \Phi^{-1}$

2. *Czizar's results.*

Here we present a summary of [5.1], with some obvious changes. In section II of chapter 4 we obtained the lower bound

(5.10) $$\tfrac{1}{2}\int \left|\frac{dP}{d\mu} - \frac{dQ}{d\mu}\right| d\mu \leq (K_\mu(P,Q))^{\frac{1}{2}}$$

from which the proof to the following result rapidly follows.

But first some notation. We will set $S_\mu(Q,\rho)=\{P\in \mathbf{P}(\Omega)|K_\mu(P,Q)<\rho\}$ and will call it the K_μ-sphere of radius ρ and center Q. Also, if $\mathcal{E}$ is a convex set intersecting $S_\mu(Q,\infty)$, a $P\in\mathcal{E}$ such that

$$K_\mu(P/Q)=\inf\{K_\mu(P',Q) / P'\in E\}$$

will be called the K_μ* projection of Q on $\mathcal{E}$. Now we are ready to state.

Theorem 5.11. If the convex set $\mathcal{E}$ is variation closed (i.e., closed in the distance specified by the left hand side of (5.10)), then each Q such that $S_\mu(Q,\infty)\cap\mathcal{E}\neq\varnothing$ has a K_μ-projection on $\mathcal{E}$.

The proof of this result is similar to the proof of projections on closed subspaces of Hilbert spaces. It depends on an application of (5.10), the triangle inequality and Fatou's lemma.

For the statement and proof of the following lemma we need the following variation on the theme of (4.8)

$$K_\mu(P,R)-K_\mu(P,Q)=E_P\ln\frac{dQ/d\mu}{dR/d\mu} \tag{5.12}$$

which, when Q<<R reduces to $E_P\ln(dQ/dR)$.

Lemma 5.13. Assume that $K_\mu(P,Q)$ and $K_\mu(Q,R)$ are finite. The segment joining P and Q (in the class of measures absolutely continuous with respect to μ) does not intersect $S_\mu(R,\rho)$ with $\rho=K_\mu(Q,R)$. That is, $K_\mu(P,R)\geq\rho$ for each $P_\alpha=\alpha P+(1-\alpha)Q$ if and only if $E_P(\ln(dQ/d\mu)/(dR/d\mu))\geq\rho$.

Also, if $Q=\alpha P+(1-\alpha)P'$, $0<\alpha<1$, then $K_\mu(Q,R)<\infty$ implies $K_\mu(P,R)<\infty$, and the segment joining P and P' does not intersect $S_\mu(R,\rho)$ (with $\rho=K_\mu(Q,R)$) if and only if

$$E_P\left(\ln\left(\frac{dQ/d\mu}{dR/d\mu}\right)\right)=K_\mu(Q,R).$$

Proof: Let $\rho_\alpha=\alpha(dP/d\mu)+(1-\alpha)(dQ/d\mu)$ denote the density of P with respect to μ. From the convexity of $\rho\ln\rho$ it follows that

$$f_\alpha=\frac{1}{\alpha}\left(\rho_\alpha\ln\rho_\alpha-\frac{dQ}{d\mu}\ln\frac{dQ}{d\mu}\right)$$

decreases to

$$\lim_{\alpha\downarrow 0}f_\alpha=\left(\frac{dP}{d\mu}-\frac{dQ}{d\mu}\right)\left(\ln\frac{dQ}{d\mu}+1\right).$$

Note that, f_1 is μ-integrable by assumption, and by the monotone convergence theorem

$$\frac{d}{d\alpha}K_\mu(P_\alpha,R)=E_P\left(\ln\left(\frac{dQ}{d\mu}/\frac{dR}{d\mu}\right)\right)-K_\mu(Q,R)$$

from which it follows that if $E_P(\ln(dQ/d\mu)/(dR/d\mu))<K_\mu(Q,R)$ then there is $0<\alpha<1$ such that

$$K_\mu(P_\alpha,R)<K_\mu(P_0,R)=K_\mu(Q,R).$$

The converse is easier. The hypothesis implies that $K_\mu(P,R)\geq K_\mu(Q,R)$ by (5.12), and therefore $K_\mu(P_\alpha,R)\geq K_\mu(Q,R)$ because of the convexity of $K_\mu(P,R)$ in P.

The second half is left for the reader. It is also left for the reader to verify that the lemma and (5.12) imply the next proposition.

Proposition 5.14. A probability P is the K_μ-projection of Q on the convex set of probabilities $\mathcal{E}$ if and only if every $P'\in\mathcal{E}\cap S_\mu(Q,\infty)$ satisfies

$$K_\mu(P',Q)\geq K_\mu(P',P)+K_\mu(P,Q). \tag{5.15}$$

If the K_μ -projection P is an algebraic inner point of $\mathcal{E}$ (i.e. if for every $P'\in\mathcal{E}$, there exists $P''\in\mathcal{E}$ with $P=\alpha P'+(1-\alpha)P''$, $0<\alpha<1$) then $\mathcal{E}\subset S_\mu(Q,\infty)$ and $E_P(\ln(dP/d\mu)/(dQ/d\mu))=K_\mu(P,Q)$ and (5.15) holds with the equal sign.

Before we consider the first of the results we want, observe that if $\mathcal{E}$ is any set of measures, and if there exists $P\in\mathcal{E}$ with μ-density $c\exp g(x)(dQ/d\mu)$ where $\int g dP_1=\int g dP_2$ for any P_1, P_2 in $\mathcal{E}$, then $K_\mu(P|Q)=\inf\{K_\mu(P',Q)|P'\in\mathcal{E}\}$. More exactly, in this case

$$K_\mu(P',Q)=K_\mu(P',P)+K_\mu(P,Q). \tag{5.16}$$

The particular case we are interested in can be rephrased as: Let $\mathcal{E}=\{P\in\mathbf{P}(\Omega)|\ P<<\mu, E_P(\Phi)=\mathbf{c}\}$, where Φ: $\Omega\to\mathfrak{R}^k$ is given measurable mapping and $\mathbf{c}\in\mathfrak{R}^k$, then if a $P\in\mathcal{E}$ exist and is of the form $dP/d\mu=c\exp-(\mathbf{t},\Phi)(dQ/d\mu)$, then it is the K_μ-projection of Q and (5.16) holds. We are now ready for

Theorem 5.17. Let $\{f_a|a\in A\}$ be an arbitrary set of real valued, measurable functions defined on Ω and $\{c_a|a\in A\}$ a set of real constants. Let $\mathcal{E}=\{P\in\mathbf{P}(\Omega)|E_P(f_a)=c_a,\ a\in A\}$. Then, if a probability $Q<<\mu$ has K_μ-projection P on $\mathcal{E}$, its μ-density is of the form

$$\frac{dp}{d\mu}=\begin{cases} c\exp g(\alpha)\frac{dq}{d\mu} & on\ N \\ 0 & off\ N \end{cases} \tag{5.18}$$

where N has P'(N)=0 for all $P' \in \mathcal{E} \cap S_\mu(Q,\infty)$ and g(·) belongs to the closed subspace of $L_1(Q)$ spanned by the f_a's. Conversely, if a $P \in \mathcal{E}$ has Q-density of the form (5.18) with g belonging to the linear space spanned by the f_a's, then P is the K_μ-projection of Q on $\mathcal{E}$ and (5.16) holds.

Proof: It follows from Proposition 5.14 that P is the K-projection of Q on $\mathcal{E}$, then for $N=\{dP/d\mu=0\}$ it is necessary to have P'(N) = 0 for any $P' \in \mathcal{E} \cap S_\mu(Q,\infty)$.

Let $\mathcal{E}' \subset \mathcal{E}$, the class of $P' \in \mathcal{E}$ with $dP'/dP \leq 2$. If $P' \in \mathcal{E}'$, there is $P'' \in \mathcal{E}'$ with $dP''/d\mu = 2\text{-}dP'/d\mu$ and $P=(P'+P'')/2$. (Given $P' \in \mathcal{E}$, define P"=2P-P' and verify it is in $\mathcal{E}'$.) Thus P is an algebraic inner point of $\mathcal{E}'$.Applying Proposition 5.14 to $\mathcal{E}'$ instead of $\mathcal{E}$ we obtain $E_P[\ln(dP/d\mu)/(dQ/d\mu)] = K_\mu(P,Q)$ or

$$\int \ln\left(\frac{dP/d\mu}{dQ/d\mu}\right)\left(\frac{dP'}{dP}-1\right)dP = 0 \qquad \forall P' \in \mathcal{E}'. \tag{5.19}$$

Now for any measurable $h: \Omega \to \Re$ such that $|h| \leq 1$ and

$$\int h dP = 0, \qquad \int f_a h dP = 0 \qquad \forall a \in A \tag{5.20}$$

there exists $P' \in \mathcal{E}'$ with dP'/dP=1+h. Thus, (5.19) yields

$$\int \lg\left(\frac{dP/d\mu}{dQ/d\mu}\right) h dP = 0$$

for all such h, and therefore for all $h \in L(dP)$ satisfying (5.20).

Therefore, $\ln((dP/d\mu)/(dQ/d\mu))$ belongs to the (closed) subspace of $L_1(P)$ spanned by 1 and the f_a's. For, were this not the case, the Hann-Banach theorem ([1.4]) would imply the existence of a bounded linear functional on $L_1(P)$ vanishing on the said subspace but not at $\ln((dP/d\mu)/(dQ/d\mu))$. Since the dual of $L_1(P)$ is $L_\infty(P)$, this is a contradiction.

To prove the second part, suppose that $(dP/d\mu)$ is of said form. Since g is a finite linear combination of f_a's, $\int g\, dP$ is constant on $\mathcal{E}$ and

$$\int \ln\left(\frac{dP/d\mu}{dP'/d\mu}\right) dP' = \lg c + \int g dP' = \text{constant} = K_\mu(P',P) \quad \textit{for} \quad P' \in \mathcal{E}.$$

But for $P' \in \mathcal{E}$ both $K_\mu(P,Q)<\infty$ (by hypothesis) and $K_\mu(P',P)<\infty$. Therefore $K_\mu(P',Q)=K_\mu(P',P)+K_\mu(P,Q)$ as desired.

So far we know that if a solution to (P1) exists it has the desired form. Let us see how Czizar settles the existence question.

Theorem 5.21. Let $\mathbf{P}_\mu(\mathbf{c},\Phi)$ be as in statement of Theorem 5.1 and $A=\{\mathbf{c}\in \Re^k | P\in \mathbf{P}_\mu(\mathbf{c},\Phi), K_\mu(P,Q)<\infty\}$. Assume that $D(Q,\Phi)$ is open. Then, the K_μ-projection of Q on $P_\mu(\mathbf{c},\Phi)$ exists for every **c** in the interior of A.

Comment. The obvious question is: what is the relationship between the interior of A and the image of $D(Q,\Phi)$ by $-\nabla_t \ln Z_{Q,\Phi}(\mathbf{t})$?

During the proof we shall need the following lemma, the proof of which is carried out in [5.1].

Lemma 5.22. For any measurable function Φ such that $e^{(t,\Phi)}$ is Q-integrable for small $|\mathbf{t}|$, $K_\mu(P_n,P)\to 0$ implies $\int \Phi dP_n \to \int \Phi\ dP$.

Proof of Theorem 5.21. Since $K_\mu(P)$ is convex in P, the set A is convex, and

$$F(a) = \inf\{K_\mu(P',P) | P' \in P_\mu(a,\Phi)\}$$

is a finite, convex function on A. Hence, if **a** is an inner point of A, there exists **t** such that

$$F(\mathbf{b}) \geq F(\mathbf{a}) - (\mathbf{t}, \mathbf{b}-\mathbf{a}) \qquad \forall \mathbf{b} \in A \tag{5.23}$$

Let us verify that $\mathbf{t}\in D(\mathbf{c},\Phi)$. Let $P_n \in \mathbf{P}_\mu(\mathbf{c},\Phi)$ be such that $K_\mu(P'_n,Q)\to F(\mathbf{a})$. Then by (5.10), P_n converges in variation to some P. Set $\Phi_i^{(n)} = \Phi_i$ if $-t_i \Phi_i \leq K_n$ and $\Phi_i^{(n)} = 0$ elsewhere. Here $K_n \uparrow \infty$. Let P_n be probability distribution with

$$\frac{dP_n}{d\mu} = N_n \exp -(\mathbf{t}, \Phi^{(n)}) \frac{dQ}{d\mu} \tag{5.24}$$

From (5.24) and the definitions

$$K_\mu(P_n, Q) = E_{P'_n}\left(\ln\left(\frac{dP_n/d\mu}{dQ/d\mu}\right)\right) - (t, E_{P'_n}(\Phi^{(n)})) + (t, E_{P'_n}\Phi^{(n)}). \tag{5.25}$$

Since, $0\in D(Q,\Phi)$, the components of Φ are Q-integrable, (see Lemma (5.5a)) and P_n is integrable as well. Therefore, for n large enough, $\int \Phi^{(n)}\, dP_n$ is arbitrary close to $\int \Phi dP_n \equiv \mathbf{b}^n$ say.

Choosing the K_n property, get the $\int \Phi^{(n)} dP'_n$ close to $\int \Phi dP'_n = \mathbf{a}$. Compare (5.25), (5.12) to (5.23) with $\mathbf{b}^n$ in the role of **b** and obtain $K_\mu(P'_n,Q_n)\to 0$. Thus, on account of (5.10), the P_n with densities (5.24) converge to P with density $N \exp -(\mathbf{t},\Phi)(dQ/d\mu)$ with respect to μ. And also, $\mathbf{t}\in D(Q,\Phi)$. Setting $\mathbf{c}=E_P\Phi$, similarly to (5.25) we have

$$K_\mu(P,Q) = E_{P'_n}\left(\ln \frac{dP/d\mu}{dQ/d\mu}\right) - (\mathbf{t}, \mathbf{c}-\mathbf{a})$$

from which, using (5.12) and (5.23) we obtain $K_\mu(P_n,P)\to 0$. Since we are assuming that $D(\mathbf{c},\Phi)$ is open, lemma stated above implies that $\int \Phi\, dP=\lim \int dP'_n=\mathbf{c}$. This completes the proof.

3. *Borwein and Lewis's extensions.*

The problems we have been dealing with are actually particular cases of linear programming problems which can roughly be described as follows:

$$minimize\{\int \Psi(\rho(s))\mu(ds) | \rho \in K,\ \int \Phi(s)d\mu(s) = \mathbf{c}\}.$$

where Ψ is an appropriate convex function defined on $\Re$, K is a convex set of functions ρ defined on a measure space $(\Omega,\mathcal{J},\mu)$, Φ: $\Omega \to \Re^k$ is a measurable mapping and $\mathbf{c}\in \Re^k$.

Even though I am hardly describing the results of this interesting line of work, the reader should at least take a look at [5.2] and at some of the references there, in particular to the pioneering work by Rockafellar in [5.3]-[5.4].

As an appetizer, I will only mention a few of the examples described by them.

Let Ω, be [0,1] and $\mathcal{J}$ its Borel sets. Let $\Psi(u)=u^p$ $(p>1)$ or $\Psi(u)=1/u$ for $u>0$ or $0\le u\le 1$. Or, let $\Psi(u)=u\ lnu$ or $\Psi(u)=-lnu$ for $u>0$. Supply with appropriate linear constraints to obtain a problem as above.

Anyway, their treatment relies heavily on convex analysis, particularly on duality theory. The general idea is always to go from the original problem (on an infinitely dimensional space) to a dual (finitely dimensional) problem, and to verify that there are enough conditions under which both lead to the same solution. Heavy stuff, but quite general, and useful!

4. *Dacunha-Castelle and Gamboa's approach to level 2 M.E.M.*

Even though the original idea behind the present approach seems to date back to Rietsch's paper in 1977, see [0.12], the development of the method in its full french generality is due to Dacunha-Castelle and Gamboa, see [4.2] and, for further extensions and generalizations by Gamboa and Gassiat see references [5.5]-[5.7]. Before presenting the main results in [4.2], and to motivate further, consider the following problem.

Suppose you want to solve

$$\sum A_{ij}x_j = y_i \qquad 1\le i\le M, \qquad 1\le j\le N \tag{5.26-a}$$

in which perhaps N>M and to make it worse, it is known that

(5.26 – *b*) $$x_i \in \{0,1\}.$$

The level 2 M.E.M. way of solving this problem is the following. On $\Omega=\{0,1\}^N$ with the obvious σ-algebra . For $\omega \in \Omega$, $P(\omega)$ is the probability of configuration ω . We define X_i: $\Omega \to \{0,1\}$ by the obvious thing $X_i(\omega)=\omega(i)$. On $\{0,1\}$ we define a probability $m(0)=m_0$,$m(1)=m_1$ and on Ω we define $\mu_N= m\otimes...$ m in the obvious way. This μ_N is some "a priori" measure on Ω.

We shall define Φ_i on Ω by $\Phi_i=\Sigma A_{ij}X_j$ and we shall look for $P'\in \mathbf{P}(\Omega)$ such that

(5.26 – *c*) $$E_P[\Phi_i] = y_i$$

and such that the entropy

(5.26 – *d*) $$S_\mu(P) = -\int \rho \ln \rho d\mu(\omega) = -\sum_\omega \rho(\omega)\ln \rho(\omega)\mu(\omega)$$

where $P(\omega)=\rho(\omega)\mu(\omega)$ for every $\omega \in \Omega$. Having established the notation, there is no problem in verifying that

$$Z(t) = \sum_\omega e^{-(t,\Phi)(\omega)}\mu(\omega) = \prod_{j=1}^{N}\left(m_0 + m_1 e^{-\sum t_i a_{ij}}\right).$$

Now see that if we find $t^*\in \Re^M$ such that

$$H(\mathbf{t}) = (\mathbf{t},\mathbf{y}) + \ln Z(\mathbf{t})$$

assumes its minimum value there, then $\rho^*(\omega)=\exp-(t^*,\Phi)/Z(t^*)$ and the x_i we want are given by

$$x_j = E_{P^*}X_j = \sum_\omega X_j(\omega)\rho^*(\omega)\mu(\omega)$$

which, taking into account the product nature of ρ^*,μ,Z, yields

$$x_j = m_1 \exp-\sum_{i=1}^{M} t_i^* A_{ij}\left(m_0 + m_1 \exp-\sum_{i=1}^{M} t_i^* A_{ij}\right)^{-1}$$

for $1\le j\le M$. If all goes well, the numbers t^*_i, $1\le i\le M$ will be such that x_j is very near zero or very near one, (in practice it is somewhat hard to go beyond such statements).

In its most general form, the reconstruction problem via the M.E.M. goes like this: Let B be a locally compact, topological vector space and B^* its dual. Let μ be a reference measure on B and X a B-valued random variable and P its distribution.

Let $\Phi \in (B^*)^k$ and $\mathbf{c} \in \Re^k$ The M.E.M. reconstruction consists of finding P that maximizes $S_\mu(P)$ subject to $E_P(\langle\Phi,X\rangle)=c$, $P \in C$.

But we shall not aim at such generality here. For us B will be $\mathbf{C}([0,1])$ the class of all continuous, real valued functions defined on [0.1]. To be specific we shall consider the cases

$$C_1 = \{g \in C([0,1]) - \infty \le a < g(x) < b \le +\infty\}$$

$$C_2 = \{g \in C([0,1]) | \int g^2 dx < 1\}$$

Also, let Φ: $[0,1] \to \Re^k$ be continuous, and set $\langle \Phi, g\rangle = \int_0^1 \Phi(x)g(x)dx$.

Since dealing with measures on infinitely dimensional spaces is hard, the thing to do is to discretize and verify that a solution to our problem exists in the limit . This explains the reason behind our regularity assumptions.

We want to solve the problems: For i=1 or 2

(5.27 – *a*) $\qquad$ *find* $\quad \{g \in C_i | \langle \Phi, g\rangle = \mathbf{c}\}$.

To discretize, we shall consider the discretization $I_n=\{i/n|\ i=0,...,n-1\}$ of size n of [0,1] and for $h \in \mathbf{C}([0,1])$ the trace h_n of h on I_n is defined to be $\{h(i/n)|\ i=0,...,n-1\}$.

Let $C(i)_n=(a,b)^n$ and for any measure m on (a,b) we set $\mu_n=m\otimes...\ m$. And, when dealing with C_2, $C(2)_n$ will be the unit ball B_n in $\Re^n$ and we shall take μ_n to be the uniform measure on B_n.

In any case, $X_u=(X^{(n)}_1,...,X^{(n)}_n)$ will denote the obvious coordinate vector and we will search for measures P_n on $C(i)_n$ such that $P_n \ll \mu_n$ and

(5.27 – *b*) $$\frac{1}{n}E_{P_n}\left[\sum \Phi(i/n)X_i^{(n)}\right] = \mathbf{c}$$

The P_n will be provided for us by the M.E.M. and it is natural to state

Definition 5.28. Let $(C(i)_n, \mu_n, \Phi, c)_{n\in N}$ the datum for a sequence of maximum entropy problems. We shall say that it is convergent if

a) For any arbitrary large n, the ME-problem of size n admits a unique solution P_n^*.

b) For all x, and i(n) such that $(i(n)/n) \to x$ as $n \to \infty$,

$$E_{P_n^*}\left(X_i^{(n)}(n)\right) \to g^*(x)$$

c) $$\int_0^1 \Phi(x)g^k(x)dx = \mathbf{c}\,.$$

The following lemma asserts that discretization yields feasible solution at every stage. The proof is in [4.2].

Lemma 5.29. Let C be an open convex of $\mathbf{C}([0,1])$. Assume that the constraint is realizable, i.e., there exists g such that $\langle\Phi,g\rangle = \mathbf{c}$. Furthermore, assume Φ is of rank k, i.e., for any $\alpha \in \Re^k$ such that $(\alpha,\Phi(x)) = 0$ for all x in $[0,1]$, we must have $\alpha = 0$. Then, the constraint is realizable in $C_n(i)$ that is, there exists $\sigma_n \in C_n(i)$ such that $\frac{1}{n}\Sigma\,\Phi(i/n)\sigma_n^i = \mathbf{c}$.

The following lemma asserts the existence of solutions to the M.E.M. problems of size n.

Lemma 5.30. If the hypotheses of Lemma 5.29 hold, and if the convex envelope of the support of μ_n contains C_n, and if

$$D(\mu_n) = \left\{u \in \Re^k \,|\, \int \exp\left[-\tfrac{1}{n}\Sigma\left(u, \Phi\left(\tfrac{i}{n}\right)\right)x_i^n\right]d\mu_n(x_n) < \infty\right\}$$

is a non empty open set, the ME-problem of size n admits a unique solution defined by

$$dP_n^*(x_n) = \tfrac{1}{Z_n(u)}\exp\left[-\tfrac{1}{n}\sum_{i=1}^{n}\left(u, \Phi\left(\tfrac{i}{n}\right)\right)x_n^i\right]d\mu_n(x_n)\,.$$

Proof: According to Lemma 5.30, there exists σ_n such that $(1/n)\Sigma\,\Phi(i/n)\sigma_v^{\,i} = \mathbf{c}$. Thus we only have to quote Theorem 5.21 above to obtain a measure $\mathbf{P}^*_n$ such that

$$E_{P_n^*}\left(X_i^{(n)}\right) = \sigma_n^i$$

Let us now concentrate on problem (5.27-a) for C_1. Under the following assumptions, we shall prove that the sequence of ME-problems yields a solution to our (5.27-a).

Assumptions on the measure m on (a,b)

A1) (a,b) is contained in the convex envelope of the support of m.

A2) The set $D(m) = \{\, t \in \Re \,|\, \int \text{exp-ty } m(dy) < \infty \,\}$ is a non empty, open set on which we define $\zeta(t) = \int \text{exp-ty } m(dy)$ and $\Psi(t) = \ln\zeta(t)$.

A3) The set $V = \{\mathbf{u} \in \Re^k \,|\, (u, \Phi(x)) \in D(m),\ \forall x \in [0,1]\}$ is non empty and coincides with $V' = \{\mathbf{u} \in \Re^k \,|\, \Psi((u,\Phi(x)) \in L_1([0,1]),dx)\}$.

Denoting the $\mathbf{u}$ in Lemma 5.30 by $\mathbf{u}_n$ and setting $\Lambda_n = \mathbf{u}_n/n$ we can restate Lemma 5.30 as

where N has P'(N)=0 for all $P' \in \mathcal{E} \cap S_\mu(Q,\infty)$ and g() belongs to the closed subspace of $L_1(Q)$ spanned by the f_a's. Conversely, if a $P \in \mathcal{E}$ has Q-density of the form (5.18) with g belonging to the linear space spanned by the f_a's, then P is the K_μ-projection of Q on $\mathcal{E}$ and (5.16) holds.

Proof: It follows from Proposition 5.14 that P is the K-projection of Q on $\mathcal{E}$, then for N= {dP/dμ=0} it is necessary to have P'(N) = 0 for any $P' \in \mathcal{E} \cap S_\mu(Q,\infty)$.

Let $\mathcal{E}' \subset \mathcal{E}$, the class of $P' \in \mathcal{E}$ with dP'/dP≤2. If $P' \in \mathcal{E}'$, there is $P'' \in \mathcal{E}'$ with dP''/dμ=2-dP'/dμ and P=(P'+P'')/2. (Given $P' \in \mathcal{E}$, define P''=2P-P' and verify it is in $\mathcal{E}'$.) Thus P is an algebraic inner point of $\mathcal{E}'$.Applying Proposition 5.14 to $\mathcal{E}'$ instead of $\mathcal{E}$ we obtain $E_P[\ln(dP/d\mu)/(dQ/d\mu)] = K_\mu(P,Q)$ or

$$\int \ln\left(\frac{dP/d\mu}{dQ/d\mu}\right)\left(\frac{dP'}{dP}-1\right)dP = 0 \qquad \forall P' \in \mathcal{E}'. \tag{5.19}$$

Now for any measurable $h: \Omega \to \Re$ such that $|h| \le 1$ and

$$\int h dP = 0, \qquad \int f_a h dP = 0 \qquad \forall a \in A \tag{5.20}$$

there exists $P' \in \mathcal{E}'$ with dP'/dP=1+h. Thus, (5.19) yields

$$\int \lg\left(\frac{dP/d\mu}{dQ/d\mu}\right) h dP = 0$$

for all such h, and therefore for all $h \in L(dP)$ satisfying (5.20).

Therefore, ln((dP/dμ)/(dQ/dμ)) belongs to the (closed) subspace of $L_1(P)$ spanned by 1 and the f_a's. For, were this not the case, the Hann-Banach theorem ([1.4]) would imply the existence of a bounded linear functional on $L_1(P)$ vanishing on the said subspace but not at ln((dP/dμ)/(dQ/dμ)). Since the dual of $L_1(P)$ is $L_\infty(P)$, this is a contradiction.

To prove the second part, suppose that (dP/dμ) is of said form. Since g is a finite linear combination of f_a's, ∫g dP is constant on $\mathcal{E}$ and

$$\int \ln\left(\frac{dP/d\mu}{dP'/d\mu}\right) dP' = \lg c + \int g dP' = \text{constant} = K_\mu(P',P) \quad \textit{for} \quad P' \in \mathcal{E}.$$

But for $P' \in \mathcal{E}$ both $K_\mu(P,Q)<\infty$ (by hypothesis) and $K_\mu(P',P)<\infty$. Therefore $K_\mu(P',Q)=K_\mu(P',P)+K_\mu(P,Q)$ as desired.

So far we know that if a solution to (P1) exists it has the desired form. Let us see how Czizar settles the existence question.

Proof: Let $\mathbf{x} \in W$. Since W is open, there is $\mathbf{x}^{**}$ and $0 < \lambda < 1$ such that $\mathbf{x} = \lambda\mathbf{x}^{*} + (1-\lambda)\mathbf{x}^{**}$ From the concavity of K we obtain

$$K(\Lambda, \mathbf{x}) \geq \lambda K(\Lambda, \mathbf{x}^{*}) + (1-\lambda)K(\Lambda, \mathbf{x}^{**}) \geq \lambda K(\Lambda, \mathbf{x}^{*}) + (1-\lambda)L(\mathbf{x}^{**}) .$$

Now use assumption (iii), and the fact that $\mathbf{x}^{*}$ is fixed to obtain the desired result.

Proof of Lemma 5.32: Set

$$V_n = \left\{ \Lambda \in \Re^k \middle| \left(\Lambda, \Phi\left(\frac{i}{n}\right)\right) \in D(m), \qquad i = 1, 2, \ldots n \right\}$$

$$\Phi(C_1) = \left\{ C \in \Re^k \middle| C = \int_0^1 \Phi(x) g(x) dx, \quad some \quad g \in C_1 \right\}$$

which is an open, convex set by the open mapping theorem.

Define H_n: $V_n x \Phi(C_1) \rightarrow \Re$ and H: $V x \Phi(C_1) \rightarrow \Re$ by

$$H_n(\Lambda, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \Psi\left(\left(\Lambda, \Phi\left(\frac{i}{n}\right)\right)\right) + (\Lambda, \mathbf{c}) \quad \Lambda \in V_n$$

$$H(\Lambda, \mathbf{c}) = \int_{-\infty}^{+\infty} \Psi((\Lambda, \Phi(x)))dx + (\Lambda, \mathbf{c}) \qquad \Lambda \in V$$

Certainly $H_n(\Lambda, \mathbf{c}) \rightarrow H(\Lambda, \mathbf{c})$ uniformly on compact sets. Thus if we show that $H(\Lambda, \mathbf{c})$ has a minimum at $\Lambda_\infty \in V$, we will be through. For that, let us begin by verifying that $H(\Lambda, \mathbf{c})$ satisfies the assumptions of Lemma 5.33.

$$For \quad \Lambda^{*} \in V \quad set \quad \mathbf{c}^{*} = \int_0^1 \Phi(x)\Psi'(\Lambda^{*}, \Phi(x))dx .$$

Then Λ^{*} is the minimum of $H(\Lambda, \mathbf{c}^{*})$ and (i) of the lemma holds. Let $g \in C_1$ be such that $\mathbf{c} = \langle \Phi, g \rangle$. Set

$$\frac{1}{n} \sum_{i=1}^{n} \Phi\left(\frac{i}{n}\right) g\left(\frac{i}{n}\right) = \mathbf{c}_n$$

As we have seen above, in Lemma 5.31,

$$\mathbf{c}_n = \frac{1}{n} \sum_{i=1}^{n} \Phi\left(\frac{i}{n}\right) \Psi'\left(\left(\Lambda_n, \Phi\left(\frac{i}{n}\right)\right)\right)$$

and also

$$\frac{1}{n} S_{\mu_n}(P_n^{*}) = H_n(\Lambda_n, \mathbf{C}_n).$$

Consider the law $Q_n << \mu_n$ with a density defined to be

$$\prod_{i=1}^{n} \exp\left(-x_n^i(\Psi^1)^{-1}\left(g\left(\frac{i}{n}\right)\right) - \Psi \circ (\Psi')^{-1}\left(g\left(\frac{i}{n}\right)\right)\right)$$

and introduce, for y in (a,b)

$$\gamma_m(y) = y(\Psi^1)^{-1}(y) - \Psi \circ (\Psi^1)^{-1}(y)$$

the Cramer transform of the measure m.

Therefore

$$E_{Q_n}\left[X_i^{(n)}\right] = \Psi' \circ (\Psi^1)^{-1}\left(g\left(\frac{i}{n}\right)\right) = g\left(\frac{i}{n}\right)$$

and also

$$S_{\mu_n}(Q_n) = -\sum_{i=1}^{n} \gamma_m\left(g\left(\frac{i}{n}\right)\right)$$

The law Q_n verifies the constraints $\mathbf{c}_n$, which implies that

$$-\frac{1}{n}\sum_{i=1}^{n} \gamma_m\left(g\left(\frac{i}{n}\right)\right) \le \frac{1}{n} S_{\mu_n}(P_n^*) = H(\mathbf{\Lambda}_n, \mathbf{c}^{(n)}) \le H_n(\mathbf{\Lambda}, \mathbf{c}^{(n)})$$

for every $\Lambda \in \Re^k$ Taking limits as $n \to \infty$ we obtain

$$\Gamma_m(g) = -\int_0^1 \gamma_m(g(x))dx \le H(\mathbf{\Lambda}, \mathbf{c})$$

for any $\mathbf{\Lambda} \in \Re^k$. Since $\Gamma_m(g)$ is finite on $g \in C_1$, due to the continuity of γ_m on (a,b). Therefore, (ii) holds as well. Fatou's lemma yields (iii), and therefore Lemma 5.32 can be invoked to conclude that the minimum of $H(\Lambda,\mathbf{c})$ cannot be reached at U^*

All these scattered lemmas amount to proving

Theorem 5.34. Let $\Phi:[0,1] \to \Re^k$ be a continuous mapping. Let $\mathbf{c} \in \Re^k$ and consider problem (5.27-a) for C_1. The following are equivalent

1) (5.27-a) has a solution.

2) These exists $\Psi(t)$, such that $\exp\Psi$ is the Laplace transform of a measure m on (a,b) satisfying A1, A2 and A3 such that (5.27-a) has $\Psi'(\mathbf{\Lambda}_\infty, \Phi(x))$ as solution, where $\mathbf{\Lambda}_\infty$ verifies

$$\mathbf{c} = \int_0^1 g(x)\Psi'(\mathbf{\Lambda}_\infty, g(x))dx.$$

3) For any Ψ such that $\exp\Psi$ is the Laplace transform of a positive measure satisfying A1, A2 and A3, (5.27-a) has a solution $\Psi'(\Lambda_\infty, g(x))$ with Λ_∞ satisfying

$$\int_0^1 g(x)\Psi'(\Lambda_\infty, g(x))dx = \mathbf{c}$$

And to finish we have

Theorem 5.35. Define Γ: $C_1 \to \Re$ by

$$\Gamma(h) = -\int_0^1 \gamma[h(y)]dy$$

where $\gamma(y)=y(\Psi^1)^{-1}(y)-[\Psi\circ(\Psi^1)^{-1}](y)$. Then $g^*(x)=\Psi'(<\Lambda_\infty,\Phi>)$ is the unique element at which $\sup\{\Gamma(h)|h\in C^1,<\Phi,h>=c\}$ is achieved.

Proof: We saw in the course of the proof of Lemma 5.32 that $\Gamma(h) \le H(\Lambda_\infty,\mathbf{c})$ of every h in C_1 such that $<\Phi,h> = \mathbf{c}$. But a simple computation shows that $H(\Lambda_\infty,\mathbf{c}) = \Gamma(g^*)$, which asserts that g^* is a candidate to solve the maximization problem stated above. Since $\Gamma(h)$ is strictly convex the solution has got to be unique.

This rounds up things for problem (5.27-a) for C_1. We shall now consider the equivalent problem, but for C_2. Recall that now the reference measure μ_n is the uniform measure on the unit ball B_n in $\Re_n$, of volume v_n.

In order to compute

$$Z_n(\mathbf{u}) = \frac{1}{u_n}\int_{B_n} \exp\left[-\frac{1}{n}\langle(\mathbf{u},\Phi_n),\mathbf{x}^n\rangle\right]d\mathbf{x}^n$$

$$\left(beware,\quad \langle(\mathbf{U},\Phi_n),\mathbf{x}^n\rangle = \sum_{i=1}^n \left(\mathbf{U},\Phi\left(\frac{i}{n}\right)x_i^n\right)\right)$$

it suffices to note that for $\mathbf{t}\in\Re^n$

$$\int \exp(\mathbf{t},\mathbf{x}^n)dx^n = \int_{-1}^1 \exp[\|\mathbf{t}\|y](1-y^2)^{\frac{n-1}{2}}v_{n-1}dy$$

where $\|\mathbf{t}\|=(\mathbf{t},\mathbf{t})^{1/2}$. Denote this last integral by $I_n(\mathbf{t})$. Therefore

$$Z_n(\mathbf{u}) = \frac{1}{v_n}I_n\left(\frac{1}{n}\|(\mathbf{u},\Phi)\|\right)$$

from which we obtain that

$$E_{P_n^*}[X_i^n] = \left(\|(\mathbf{U}_n,\phi_n)\|\right)^{-1}\left\langle \mathbf{u}_n, \Phi\left(\frac{i}{n}\right)\right\rangle G_n(\mathbf{u}_n)$$

where $\mathbf{u}_n$ has to satisfy

$$\tfrac{1}{n}(\|\langle \mathbf{u}_n, \Phi_n\rangle\|)^{-1} G_n(\mathbf{u}_n) \sum_{i=1}^{n} \left(\mathbf{u}_n, \Phi\left(\tfrac{i}{n}\right)\right)\Phi\left(\tfrac{i}{n}\right) = \mathbf{c}$$

and $G_n(u)$ is defined by

$$I_n\left(\tfrac{1}{n}(\mathbf{u}, \Phi_n)\right)^{-1} \int_{-1}^{1} y \exp[y\|(\mathbf{u}, \Phi_n)\|n](1-y^2)^{\frac{(n-1)}{2}}\, dy$$

Set now $\Lambda_n = \mathbf{u}_n G_n(\mathbf{u}_n)/\|(\mathbf{u}_n, \Phi_n)\|^{-1}$ and the n-th maxentropic approximant to the desired g^*_k is given by

$$g_n^*(x) = \Sigma\left(\Lambda_n, \Phi\left(\tfrac{i}{n}\right)\right)\chi\left(\left[\tfrac{i-1}{n}, \tfrac{i}{n}\right]\right)(x)$$

where we set $\chi(A)(x)$ equal 1 or 0 depending on whether x is in A or not (i.e. the indicator function of A in the parlance of measure theorists but not in that of convex theorists).

Anyway, above, Λ_n is the unique minimum of the convex functional

$$H_n(\Lambda, \mathbf{c}) = \tfrac{1}{2n} \sum_{i=1}^{n} \left(\Lambda, \Phi\left(\tfrac{i}{n}\right)\right)^2 + (\Lambda, \mathbf{c})$$

which converges, uniformly on compacts, $n \to \infty$, to

$$H(\Lambda, \mathbf{c}) = \tfrac{1}{2} \int_0^1 (\Lambda, \Phi(x))^2 dx + (\Lambda, \mathbf{c}).$$

This achieves its unique minimum at $-M_\Phi^{-1}\mathbf{c}$, where M_Φ is the matrix

$$\int_0^1 \Phi(x)\Phi^+(x)dx$$

which we assume invertible. Thus the $g^*_n(x)$ tends as $n \to \infty$ to

$$g^*(x) = -\left(M_\Phi^{-1}\mathbf{c}, \Phi(x)\right)$$

which is the least square solution to the problem under study.

APPENDIX.

Before starting to quote results from [5.8], it is convenient to translate the scheme of section 1 to $\mathfrak{R}^k$. By means of $\Phi : \Omega \to \mathfrak{R}^k$ we can associate with each measure P, U, μ on Ω a corresponding measure $P \circ \Phi^{-1} \equiv \Phi(P)$, etc. on $\mathfrak{R}^k$ We shall assume that the range S of Φ is a Borel set in $\mathfrak{R}^k$ and we shall denote by C the closure of the convex set generate by S. And we shall write $T(\mathbf{x}) = \mathbf{x}$ for the identity mapping on $\mathfrak{R}^K$

Instead of considering the translates of $\wp(\mu) = [P \in \wp(\Omega) | P << \mu]$ by Φ we shall consider only the translates of the Hellinger arc

$$d\mu_\alpha^\Phi = \frac{1}{Z(\lambda)} \exp -(\boldsymbol{\lambda}, \Phi) d\mu$$

and if, for short, we denote by μ the measure $\Phi(\mu)$ on $\mathfrak{R}^k$ we have the exponential family $H(\mu) = \{\mu_\lambda : \boldsymbol{\lambda} \in D\}$where

$$d\mu_\lambda = \frac{1}{Z(\lambda)} \exp -(\boldsymbol{\lambda}, \mathbf{T}) d\mu$$

where as usual $Z(\lambda) = \int \exp -(\lambda, T) d\mu$ and $D = \{\boldsymbol{\lambda} \in \mathfrak{R}^k | Z(\boldsymbol{\lambda}) < \infty\}$.

We shall set $K(\lambda)$=ln$Z(\boldsymbol{\lambda})$. Throughout we shall assume that D° =int(D) is nonempty and that $H(\mu)$ is full, i. e., that the Hellinger arc generated by any element of $H(\mu)$ is $H(\mu)$.

Let us write $h(\mathbf{c}) = \inf\{(\boldsymbol{\lambda}, \mathbf{c}) + k(\boldsymbol{\lambda}) | \boldsymbol{\lambda} \in \mathfrak{R}^k\}$ for any $\mathbf{c} \in \mathfrak{R}^k$ and set $s(\mathbf{c}) = -h(-\mathbf{c}) = \sup\{(\boldsymbol{\lambda}, \mathbf{c}) - k(\boldsymbol{\lambda}) | \boldsymbol{\lambda} \in \mathfrak{R}^k\}$ for the convex conjugate of $k(\boldsymbol{\lambda})$ now we can restate the results we need from section 9.1 of [5.8] as

Theorem 5.36. With notation introduced above we have

i) $s^*(\boldsymbol{\lambda}) = k(\boldsymbol{\lambda})$.

ii) k is closed, strictly convex on dom k=int D.

iii) $s(\mathbf{c})$ is a closed, essentially smooth, convex function and

$$\inf C < dom\, s < C.$$

Theorem 5.37. $k(\boldsymbol{\lambda})$ is steep if and only if $\mathcal{J}(D^\circ) = \inf C$, where $\mathcal{J}$: $D^\circ \to \mathfrak{R}^k$ is given by $\mathcal{J}(\boldsymbol{\lambda})$=$\nabla_\lambda k(\boldsymbol{\lambda}) = \nabla_\lambda \ln Z(\boldsymbol{\lambda})$.

Theorem 5.38. Let $\mathbf{t}$ be a boundary point of C. If there is a hyperplane H supporting C at $\mathbf{t}$ and satisfying μ(H)=0, then $\mathbf{t} \notin dom\ s$ in particular, whenever μ is absolutely continuous with respect to the Lebesgue measure on $\mathfrak{R}^k$, we have that *dom s= inf C*.

To explain a bit some of the words, we mention that a closed convex function is just a convex lower semicontinuous function. That $s(\mathbf{c})$ is essentially smooth means that $dom\, s = \{\mathbf{c} \in \mathfrak{R}^k | s(\mathbf{c}) < \infty\}$ is open and s is differentiable on *int*(*dom s*). In this case, $s(\mathbf{c})$ comes out being steep, that is $\frac{ds}{d\lambda}(\mathbf{c}' + \lambda(\mathbf{c} - \mathbf{c}'))$ tends to infinity as $\lambda \downarrow 0$, where $\mathbf{c}'$ is in the boundary of *dom s* and $\mathbf{c} \in$ *int*(*dom s*).

For more about these facts, the reader is directed to Chapter 5 of [5.8] where appropriate references to the treatise on convexity by Rockafellar are given.

In our situation $K(0)$ is the natural logarithm of a Laplace transform, and therefore it will be differentiable on the interior of D whenever it is not empty. According to Theorem 5.37 this will happen whenever *int* C is not empty. Then, the first thing to examine is the support of μ in $\Re^k$. If it has a nonempty interior we proceed to find D. If S is a finite or countable set, then we proceed according to

Theorem 5.39. Let S be a finite or countable set. Then *conv* $S \subset$ *dom* S. In particular *dom* $S=C$ if S is finite.

REFERENCES

[5.1] Czizar,I."*I-divergence geometry of probability distributions and minimization problems*" The Ann. of Prob. 3, No 1, pp. 146-158, 1975.

[5.2] Borwein, J. M. and Lewis, A. S. "*Partially-finite programming in L_1: entropy maximization*". Res. Rep. C00R 91-05, Faculty of Mathematics, Univ. of Waterloo, 1991.

[5.3] Rockafellar, R. I. "*Integrals which are convex functionals*". Pacific Journal of Math. 24, pp. 525-539, 1968.

[5.4] Rockafellar, R. T. "*Integrals which are convex functionals*". Pacific Journal of Math. 39, pp. 439-469, 1971.

[5.5] Gamboa, F. and Gassiat, E. "*Maximum d'entropie et problem des moments: cas multidimensionel*". Probab. and Math. Statistics. pp. 67-83, 1991.

[5.6] ibid. "*Large deviations and generalized moment problems*" To appear in Probability Theory and Related Fields.

[5.7] ibid. "*M.E.M. for solving moment problems*" Technical Report 91-23. Univ. Paris-Sud. Orsay.

Chapter 6

APPLICATIONS AND EXTENSIONS

This chapter is made up of bits and pieces. It is a collection of sections, not related in any logical order, the contents of which can be considered as either comments on the material of the preceding chapters, extensions or variations on the theme of some of the topics, or applications mostly taken from the literature, and presented in no particular order at all, hopefully to break the monotony.

Since this chapter is very long, results and formulae will be numbered by section.

6.1 Entropy maximization under quadratic constraints, or constraint relaxation.

Many reconstruction problems have inexact data, and instead of wanting to solve for $\mathbf{x}$ in $\mathbf{Ax}=\mathbf{y}$ one decides to look for $\mathbf{x}$'s such that

(6.1.1) $$\mathbf{Ax} \in V_M(\mathbf{y}, \gamma)$$

where

(6.1.2) $$V_M(\mathbf{y}, \gamma) = \{\mathbf{z} : (z-y, M(z-y)) \leq \gamma\}$$

where M is a symmetric, positive definite matrix.

Besides the issue of undefined constraints, the technique of relaxed constraints is useful in solving infinite dimensional reconstruction problems by maximum entropy methods. The basic references for this section are [6.1]-[6.2]. To make (6.1.1) more precise, let A: $\mathfrak{R}^n \to \mathfrak{R}^k$ be identified with its matrix in the canonical basis, say.

Assume, to keep things simple, that we do not have constraints on the set of possible values of $\mathbf{x}$ and let $\mathbf{X}$: $\Omega \to \mathfrak{R}^n$ be an $\mathfrak{R}^n$-valued random variable.

Let μ be a σ-finite measure on $\mathfrak{R}^n$ and consider measures P on $(\Omega, \mathscr{F})$ with $P \circ X^{-1}$ having a density $\rho(\mathbf{x})$ with respect to μ, and set as before

$$Z_\mu(\boldsymbol{\lambda}) = E_P[\exp-(\boldsymbol{\lambda}, A\mathbf{X})] = \int \rho(\mathbf{x})\, e^{-(\boldsymbol{\lambda}, A\mathbf{X})} d\mu(\mathbf{x})$$

for $\boldsymbol{\lambda} \in \Re^k$ and, of course, $Z_\mu(\boldsymbol{\lambda})$ is finite for $\boldsymbol{\lambda}$ in

$$D(\mathbf{M}) = \{\boldsymbol{\lambda} \in \Re^k : Z_\mu(\boldsymbol{\lambda}) < \infty\}.$$

Let $K(\mu) = \{ P \in \boldsymbol{P}(\Omega): S_\mu(P) < \infty \}$ and

$$B_M(\mathbf{y}, \gamma) = \{P \in \mathbf{P}(\Omega) : AE_P X \in V_M(\mathbf{y}, \gamma)\}$$

and to finish the list of symbols

$$\Sigma_M(\mu, \gamma) = \{\mathbf{y} \in \Re^k | B_M(\mathbf{y}, \gamma) \cap K(\mu) \neq \varnothing\}$$

With all these notations Gamboa proved the

Theorem 6.1.3. If $D(\mu)=\Re^k$ and $\Sigma_M(\mu, \gamma)$ is open with $\mathbf{y} \in \Sigma_M(\mu, \gamma)$, the problem of finding

$$\sup \{S_\mu(P)/P \in B_M(y, \gamma)\}$$

has the unique solution

i) $(\int d\mu)^{-1} P \circ \mathbf{X}^{-1}$ *if* $\int_{\Re^n} d\mu < \infty$ and $(\int d\mu)^{-1} P \circ \mathbf{X}^{-1} \in B_M(\mathbf{y}, \gamma)$

ii) Otherwise $dP_{\lambda^*} = \frac{\exp-(\lambda^*, AX) dP \circ \mathbf{X}^{-1}}{Z_{\lambda^*}(\mu)}$.

Here $\boldsymbol{\gamma}^*$ minimizes $H_\gamma(\boldsymbol{\lambda}, \mathbf{y}) = \ln Z_\lambda(\mu) + (\boldsymbol{\lambda}, \mathbf{y}) + \gamma(\mathbf{M}^{-1}\boldsymbol{\lambda})^{1/2}$

Proof: It is clear that when $(\int d\mu)^{-1} P \circ X^{-1} \in B_M(\mathbf{y}, \gamma)$ it maximizes $S_\mu(P)$. Assume that (i) does not hold and let $\boldsymbol{\eta} \in V_M(0, \gamma)$ such that $\mathbf{y} + \boldsymbol{\eta} \in V_M(y, \gamma)$ We know from chapter 5 that $\boldsymbol{\lambda}_1(\mathbf{y} + \boldsymbol{\eta})$ minimizing

$$H_0(\boldsymbol{\lambda}, \mathbf{y} + \boldsymbol{\eta}) = \ln Z_1(\mu) + (\boldsymbol{\lambda}, \mathbf{y} + \boldsymbol{\eta})$$

exists and

$$H_0(\boldsymbol{\lambda}_1(\mathbf{y} + \boldsymbol{\eta}), \mathbf{y} + \boldsymbol{\eta}) = S_\mu(P_{\lambda_1(y+\eta)}).$$

Then to find the $sup\{S_\mu(P): P \in B_M(\mathbf{y}, \gamma)\}$ it suffices to find $sup\{S_\mu(P_{\lambda_1(y,\eta)}): \boldsymbol{\eta} \in V_M(0, \gamma)\}$. The final step consist in applying the min-max exchange theorem in [6.3].

Comments. Actually, instead of $V_M(\mathbf{y},\gamma)$ we could have considered any convex set K. The issue would then be to find the analogue of $H_\gamma(\boldsymbol{\lambda},\mathbf{y})$.

There is one very important sense in which relaxation is of real help. Notice that when $\mathbf{y} \notin R(A)$ the range of A, then there will be no hope of finding a minimum of $\ln Z_\lambda(\mu)+(\boldsymbol{\lambda},\mathbf{y})$. We have to consider finding $\mathbf{x}$ such that $A\mathbf{x} \in V_M(\mathbf{y},\gamma)$ and $V_M(\mathbf{y},\gamma) \cap R(A)$ is not empty. There will be a critical value of γ below which no solution will exist.

6.2 *Failure of maximum entropy methods for reconstruction in infinite systems.*

Here we present some examples, borrowed from [6.2], in which the maximum entropy solution to a linear reconstruction problem does not satisfy the associated dual problem. We also present, without proof, the way around this difficulty proposed by Bowrein and Lewis.

Consider a measure space $(\Omega,\mathscr{F},\mu)$ and a vector subspace $\mathbf{X}$ of $L_p(\Omega,\mu)$ $1 \le p \le \infty$, on which a functional S_φ is defined by

$$S_\varphi(\mathbf{x}) = \int \varphi(\mathbf{x}(s))\mu(ds) \tag{6.2.1}$$

where $\varphi : \Re \to [-\infty, \infty]$ is a closed concave function. The maximum entropy problem consists of finding

$$\sup \{S_\varphi(x) : A\mathbf{x} = \mathbf{y}\} \tag{6.2.2}$$

where A: $\mathbf{X} \to \mathbf{Y}$ is some continuous linear operator. Some examples of φ are

a) Burg entropy
$$\varphi(x) = -\ln x$$

b) Boltzmann entropy
$$\varphi(x) = -x \ln x$$

c) Fermi-Dirac entropy
$$\varphi(x) = -x \ln x - (1-x) \ln (1-x)$$

d) L_p norm
$$\varphi(x) = -x^P/P$$

e) L_p entropy

$$\varphi(x) = \begin{cases} x^P/P & x \geq 0 \\ -\infty & x < \infty \end{cases}.$$

The Fenchel conjugate of S_φ is defined on $\mathbf{X}^*$ (the dual of) and is given by

$$S_{\varphi^*}(\xi) = \int \varphi^*(\xi(s))\mu(ds)$$

where $\varphi^*(\xi)=inf\{\varphi(x)-(\xi,x)\}$ is the Fenchel conjugate of φ. Here we use (ξ,x) to mean $\xi(x)$ for $\xi\in\mathbf{X}^*,\mathbf{x}\in\mathbf{X}$.

The conjugates of the functions listed above are respectively

a) $\varphi^*(\xi) = 1 + \ln(-\xi)$

b) $\varphi^*(x) = e^{-\xi-1}$

c) $\varphi^*(\xi) = \ln(1 + e^{-\xi})$

d) $\varphi^*(\xi) = \xi^q/q \qquad \frac{1}{p} + \frac{1}{q} = 1$

e) $\varphi^{*(x)} = \max\{0, \xi\}^q/q$.

When life is nice, we have

Theorem 6.2.3. Assume that φ^* is everywhere finite and differentiable, $\mathbf{X}=L_1(\Omega,\mu)$ and $\mathbf{Y}$ is *finite dimensional*. Then,if the sup in (6.2.2) is finite, it is attained whenever the following qualification constraint holds:

(C.Q) There is a measurable function $\hat{\mathbf{x}}$ solving

$$\Lambda\mathbf{x} = \mathbf{y} \quad with \ \inf dom\,\varphi < \inf\hat{\mathbf{x}} \leq \sup\hat{\mathbf{x}} < \sup dom\,\varphi.$$

Then (6.2.2) equals

(6.2.2)′ $$\inf\{S_{\rho^*}(A^*\lambda) + (\lambda, \mathbf{y}) : \lambda \in Y^*\}.$$

Moreover, the unique optimal solution to (6.2.1) equals

(6.2.4) $$x^* = -\frac{d\varphi^*}{dt}(A^*\lambda^*)$$

where $\boldsymbol{\lambda}^*$ realizes (6.2.2)'.

This result is recalled just to make explicit where it fails when $\mathbf{Y}$ is infinite dimensional. It is the identity of (6.2.2) and (6.2.2)' that breaks down.

To begin with we shall consider $L_2(\Omega,\mu)$ and assume it has a base. On $L_2(\Omega,\mu)$ we define S_φ as above with $\varphi(t)=-t^2/2$ and $(A,\mathbf{x})_n=x_n/4^n$, where the components are taken with respect to the given base. Also, define $\mathbf{y}$ as the vector with components $y_n=1/8^n$. The solution to (6.2.1) is given by

$$\mathbf{x}^* = -A\boldsymbol{\lambda}^*$$

where $\boldsymbol{\lambda}^*$ has to satisfy (6.2.4), i.e.

$$A(A^*\boldsymbol{\lambda}^*) = -y.$$

Let us verify that the range of AA^* can be larger than the range of A. Notice that $x^*_n=1/2^n$ satisfies our problem, but no $\boldsymbol{\lambda}^*$ **such that A** $(A^*\boldsymbol{\lambda}^*)=\mathbf{y}$ can be found. Since $(A\,A^*\boldsymbol{\lambda})_\nu=\lambda_\nu/4^{2\nu}$ and $y_n=1/8^n$ we would have $\lambda_n=2^n$ or $\boldsymbol{\lambda}$ not in $L_2(\Omega,\mu)$. In other words the maximum entropy problem cannot be solved using duality theory, a real handicap.

One may consider solving the finite dimensional problems $(A\mathbf{x})_n=\mathbf{y}_n$ for $0\le n\le N$ and then attempt taking limits. But observe that in this case $\mathbf{x}_N=(1,1/2,\ldots,1/2^N,0,\ldots,0\ldots)$ and $\boldsymbol{\lambda}^*_N=(1,2,\ldots,2^N,0,\ldots,0)$. Even though the solution to the full primal (6.2.2) is the limit of the $\mathbf{x}^*_N$, the $\boldsymbol{\lambda}_N$ cannot converge. This is related to the fact that

$$H(\boldsymbol{\lambda},\mathbf{y}) = S_\varphi\cdot(A^*\boldsymbol{\lambda}) + (\lambda,y) = \Sigma\,\lambda_n^2/2^{4n+1} + \lambda_n/2^{3n}$$

is strictly convex, has a unique minimum at zero, but notice that if $\mathbf{e}_n$ is the n-th basis vector, $H(n\mathbf{e}_n,\mathbf{y})\to 0$ and $\|n\mathbf{e}_n\|\to\infty$, that is $H(\boldsymbol{\lambda},\mathbf{y})$ is not coercive.

A different, but similar example, is the following: let Ω_1 and Ω_2 be compact metric spaces endowed with Borel measures μ and ν respectively. We shall assume that Ω_2 is separable as well. Let $\mathbf{X}=C(\Omega_1)$, $\mathbf{Y}=C(\Omega_2)$ denote the continuous functions on Ω_1 and Ω_2 respectively considered as vector subspaces of $L_1(\Omega_1,\mu)$ and $L_1(\Omega_2,\nu)$.

Define $A:\mathbf{X}\to\mathbf{Y}$ by

$$(Ax)(\omega_2) = \int_{\Omega_1} a(\omega_2,\omega_1)\mu(d\omega_1).$$

If $\boldsymbol{\lambda}$ denotes a measure on Ω_2, $(A^*\boldsymbol{\lambda})(d\omega_1)$ is given by

$$(A^*\boldsymbol{\lambda})(d\omega_1) = \left(\int_{\Omega_2} a(\omega_2,\omega_1)\lambda(d\omega_2)\right)\mu(d\omega_1)$$

that is $A^* \lambda$ is absolutely continuous with respect to μ with a density in $L_1(\Omega_1, \mu)$.

If the infimum in (6.2.4) were attained, an $\mathbf{x}^*$ given by

$$\mathbf{x}^* = -\frac{d\varphi^*}{dt}(A^*\lambda^*)$$

would have to satisfy $A\mathbf{x}^*=\mathbf{y}$, but again, $A((d\varphi^*)/(dt)(R(A^*)))$ may be smaller than $R(A)$ and duality will fail. To use Theorem 6.2.3 and verify that the situation is not circumvented by going first to the finite dimensional case, take $\mathbf{X}=L_1(\Omega,\mu)$ and $\mathbf{Y}=L_1(\Omega_2,\nu)$ but the rest as above. Let $\{\omega^n_2: n\geq 1\}$ be a dense subset in Ω_2 and consider the problems of size N, i.e. find

$$\sup\left\{S_\varphi(x): (\mathbf{A}x)\left(\omega_2^k\right) = \mathbf{y}\left(\omega_2^k\right);\ k = 1, 2, \ldots, N\right\}.$$

By invoking (6.2.3) we obtain

$$\mathbf{x}_N = -\frac{d}{dt}\varphi^*(A^*\lambda_N)$$

(6.2.5)
$$\mathbf{x}_N(\omega_1) = -\frac{d}{dt}\varphi^*\left(\sum_{k=1}^{n} a\left(\omega_2^k, \omega_1\right)\lambda_N^k\right).$$

If we set

$$\lambda_N = \sum_{k=1}^{N} \lambda_N^k \delta_{\omega_2^k}$$

where $\delta_{\omega_2^k}$ denotes point mass at ω^k_2. If

$$\sup \sum_{k=1}^{N} \left|\lambda_N^k\right| < \infty$$

then there is a subsequence of λ_N converging to a λ^* in the weak-star topology on $\mathbf{M}(\Omega_2)$. The same subsequence would ensure the pointwise convergence of $A^*\lambda_N$ to $A^k\lambda^*$ and therefore, if $(d\varphi^*)/(dt)$ were continuous

$$\mathbf{x}_N = -\frac{d\varphi^*}{dt}(A^k\lambda^N) \to \frac{d\varphi^k}{dt}(A^*\lambda^*) = \mathbf{x}^*.$$

That is, if the $\mathbf{x}_N$'s are a bounded sequence in $L_1(\Omega,\mu)$, $\mathbf{x}^*$ would be as in Theorem 6.2.3. But when $A(d/dt\ \varphi^*(R(A^*)))$ is smaller than $R(A)$, that could not happen and λ_N cannot be bounded.

Borwein proposed in [6.2] two ways of going around these difficulties. We shall cite one of them and urge the reader to see [6.2] for details and for the method of penalization.

Theorem 6.2.6. (Relaxation). Assume φ^* is everywhere finite and differentiable. Take $\mathbf{X}=L_1(\omega_1,\mu)$ on a complete measure space and (Y, $\|\ \|$) to be some normed space. Then the supremum in (6.2.2) is attained (when finite). In this case consider, for $\varepsilon > 0$, the relaxed problem

$$(ME)_\varepsilon \qquad \max\{S_\varphi(x) : \|Ax - y\| \le \varepsilon\}.$$

The value of $(ME)_\varepsilon$ equals the value of the dual problem

$$(DE)_\varepsilon \qquad \min\{S_{\varphi^*}(A^*\boldsymbol{\lambda}) + (\boldsymbol{\lambda}, \mathbf{y}) : \boldsymbol{\lambda} \in \mathbf{y}^*\}$$

and the unique, optimal solution to $(ME)_\varepsilon$ equals

$$\mathbf{x}_\varepsilon^* = -\frac{d\varphi^*}{dt}(A^*\boldsymbol{\lambda}_\varepsilon)$$

where $\boldsymbol{\lambda}_\varepsilon$ is any solution to $(DE)_\varepsilon$. Moreover as $\varepsilon \to 0$, $\mathbf{x}_\varepsilon^{\mathbf{k}}$ converges in mean to the unique solution $\mathbf{x}^*$ of $(ME)_\varepsilon$ and

$$S_\varphi(\mathbf{x}_\varepsilon^*) \to S_\varphi(\mathbf{x}^*).$$

6.3 *Some finite dimensional, linear reconstruction problems.*

Not too long ago I attempted to show algebraists how to use standard maximum entropy methods to solve linear equations. Much to my surprise, since many years ago a journal like Linear Algebra and Applications has been publishing papers on the subject. Besides passing down some more references, missed in [0.3], given to me via R. Brualdi, it will be the gist of this section to compare the level 1 and level 2 approaches.

Consider for example the problem of finding $\{P_i: i=1,\ldots,n\}$ such that

(6.3.1) $$\Sigma P_i C_i = c \qquad \textit{and} \qquad \Sigma P_i = 1, \qquad P_i \ge 0.$$

We shall generalize shortly. For some applications see chapter 13 of [0.3] for example. You could think of (6.3.1) as a problem of resource allocation, the P_i being the fraction of the total resource allocated to mode i, or think of (6.3.1) as the problem of determining how loaded a

die may be from the knowledge of the mean earnings of a player that has bet (enough) on each possibility.

Anyway, the standard approach consists of finding the $\{P_i\}$ maximizing

$$\{-\Sigma P_i \ln P_i : \ \Sigma P_i = 1, \ \Sigma C_i P_i = c, \quad P_i \geq 0\}$$

As we have seen many times so far, a candidate for P_i is $exp\text{-}\lambda C_i / Z(\lambda)$ with $Z(\lambda)=\Sigma exp\text{-}\lambda C_i$. The Lagrange multiplier λ is to be determined by minimizing the Hamiltonian (the name physicists employ for the dual of the Lagrangian) $H(\lambda)=\ln Z(\lambda)+\lambda c$.

Again, in chapter 13 of [0.3] the analysis on conditions for c to be in the range of $-\delta Z(\lambda)/\delta\lambda$ is carried out. Certainly, since $\Sigma C_i P_i = C$ is a convex combination, we have to have $\min C_i < C < \max C_i$. If c is generated by experimental data, as in the second motivational situation above, that will certainly be the case.

But when the consistency condition is not satisfied, you either drop your towel, or relax your constraints and look for $\{P_i\}$ such that

$$\Sigma P_i = 1, \quad |\Sigma P_i - c| < \varepsilon$$

which leads to an extended Hamiltonian

$$H(\lambda, \varepsilon) = \ln Z(\lambda) + \lambda C + \varepsilon|\lambda|.$$

Here, unless $|c-\min\{C_i\}|<\varepsilon$ or $|c-\max\{C_i\}|<\varepsilon$ we will not have a solution.

Instead of proceeding as above we could assume that P_i is the mean value of a random variable X_i with respect to a probability density $\rho(x_1,\ldots,x_n)$ on $D=[0,\infty)x\ldots x[0,\infty)$. Instead of (6.3.1) we have

$$\int_D (\Sigma C_i X_i)\,\rho(x)d\mu(x) = c, \quad \int_D (\Sigma x_i)\,\rho(x)d\mu(x) = 1 \tag{6.3.2}$$

where $d\mu(x)=\exp\text{-}(\Sigma x_i)dx$. We now look for densities $\rho(x)$ maximizing

$$S_\mu(\rho) = -\int_D \rho(x)\ln\rho(x)d\mu(x)$$

under the constraint (6.32).

Here we set $D=[0,\infty)^n$ instead of $[0,1]^n$ only because it facilitates an explicit computation that comes below. The partition function is now

$$Z(\lambda_1,\lambda_2)=\int_D \exp(-\lambda_1(\Sigma C_i x_i)+\lambda_2(\Sigma x_i))\ \exp(-\Sigma x_i)dx=\prod_{i=1}^{n}(\lambda_1 C_i+\mu)^{-1}$$

where we set $\mu=\lambda_2+1$. Once the values λ^*_1 and λ^*_2 that minimize

$$H(\lambda_1,\lambda_2)=\ln Z(\lambda_1,\lambda_2)+\lambda_1 C+\lambda_2$$

are found, the P_i we want are

$$P_i=1/\lambda_1^* C_i+\lambda_2^*+1$$

and the maximum entropy density yielding these values is

$$\rho(x)=\exp-\Big(\sum_i(\lambda_1^* C_i+\lambda_2^*+1)x_i\Big)/Z(\lambda_1^*,\lambda_2^*).$$

When n=2, a simple but lengthy computation shows that for $c_1<c<2$, the direct solution or, any of the two maxentropic methods yields

$$P_i=(c_2-c)/(c_2-c_1),\quad P_2=(c-c_1)/(c_2-c_1).$$

Actually the same is true for any square, invertible, reconstruction problem.

It may happen that minimizing $H(\lambda_1,\lambda_2)$ becomes too difficult for it may be too flat near the minimum. In this case it may be convenient to use a genetic algorithm to minimize something like

$$\left|\sum_i \frac{1}{\lambda_i c_i+\mu}-1\right|W+\left|\sum_i \frac{c_i}{\lambda_i c_i+\mu}-c\right|W'$$

where the W, W' are weights chosen "a piacere"

To make critics and detractors happy, notice that if instead of taking $D=[0,\infty)^n$, had we taken $D=\Re^n$ and $d\mu(x)=\exp-(\Sigma x^2/2)\,dx$, we would have obtained

$$Z(\lambda_1,\lambda_2)=\exp\tfrac{1}{2}\sum_{i=1}^{n}(\lambda_1 c_1+\lambda_2)^2$$

which yields

$$H(\lambda_1,\lambda_2)=\sum_{i=1}^{n}(\lambda_1 c_1+\lambda_2)^2+\lambda_1 c+\lambda_2$$

which has a minimum at

$$\lambda_1^*=\left(\sum_{ij}c_i c_j\right)^{-1}c-\Sigma c_i,\quad \lambda_1^*=\left(\sum_{ij}c_i c_j\right)^{-1}\Sigma c_i^2-c\Sigma c_i$$

which yield for the P_i the values

$$P_i = \left(\frac{\sum c_j - c}{\sum_{j,k} c_k c_j}\right) c_i + c\left(\frac{\sum c_j - \sum c_j^2}{\sum c_j c_k}\right) = \lambda_1^* c_i + \lambda_2^*$$

which is what you'd get solving (6.3.1) by proposing the solution $P_i=\lambda_1 c_i+\lambda_2$ and finding the right λ_1, λ_2.

Before returning to the mainstream of this section, let us recap what we have done in the following.

Comments.

i) Level 1 and level 2 approaches to reconstruction problems may yield the same answer to a reconstruction problem.

ii) When using level 1 approach the choice of $S_\mu(P)$ is arbitrary whereas, when using a level 2 approach, one agrees up on the Boltzmann-Gibbs-Shannon functional and plays with the a priori knowledge one has, or can assume, about the range and distribution of the x_i. But, of course, the choice of the entropy functional and the reference measures is totally arbitrary.

The following problem: find $\{x_i: 0 \le x_i \le 1, i=1,...,n\}$ such that

(6.3.3) $$A\mathbf{x} = \mathbf{b}$$

where the nxm matrix A and the vector $b \in \Re^m$ are given. For a review about work in this problem see [6.4], and for conditions for the existence of a minimum of the dual problem, i.e., a minimum of $H(\lambda)$ associated with the standard maximum entropy problem see [6.5].

The way the second level maximum entropy method applies to (6.3.3) consists of assuming an a priori measure, say $d\mu(x)=dx$ on $\Omega=[0,1]^n$. By $\varphi_i : \Omega \to \Re$ we denote the coordinate map $\varphi_i(x)=x_i$. We look for measures P on Ω (equipped with the obvious Borel σ-algebra $\mathcal{B}$) having density $\rho(x)$, that maximize $S_\mu(P)=-\int \rho(x)\ln\rho(x)dx$ subject to the constraints $(\sum A_{ji}\varphi_i(x))\rho(x)dx=b_j$, $j=1,...,m$.

The usual arguments provide us with

(6.3.4) $$\rho(x) = \frac{1}{Z(\lambda)} \exp -(\mathbf{C}, \varphi)(x)$$

where $$C_i = \sum_{j=1}^{m} \lambda_j A_{ji} \qquad i = 1, 2, ..., n, \qquad \textit{of course}$$

$$Z(\lambda) = \int_\Omega e^{-\{\mathbf{C},\varphi\}(x)} = \prod_{i=1}^{n} \left(\frac{1-e^{-C_i}}{C_i}\right)$$

and the Hamiltonian to be minimized to find the λ is

(6.3.5) $$H(\lambda) = \sum_{i=1}^{n} \ln\left(\frac{1-e^{-C_i}}{C_i}\right) + (\lambda, b).$$

Whenever life is nice to us and the minimum in (6.3.5) is reached, we know from chapter 5 that

$$\rho_{ME}(x) = \frac{1}{Z(\lambda^*)} \exp-(C^*, \varphi)(x)$$

is the distribution on the set of all images that maximizes $S_\mu(P)$. The maxentropic reconstruction is

(6.3.6) $$x_i = \int \varphi_i(x)\rho_{ME}(x)dx = -\frac{\partial}{\partial C_i} \ln Z(\lambda) = 1/C_i^* - 1/\left(e^{C_i^*} - 1\right).$$

To finish this section with a calculation that we shall make use of in the next one, assume that we know or we have to impose the condition that the solution to (6.3.3) belongs to the set $\{0,1\}^n$. In this case the measure $d\mu(x)$ on $\Omega=\Re^n$ is

$$d\mu(x) = \prod_{i=1}^{n} \left\{\frac{1}{2}(\delta_0(dx) + \delta_1(dx))\right\}$$

where $\delta_a(dx)$ is the Dirac measured concentrated at $a \in \Re$. Here, what we take Ω to be is irrelevant as long as it contains the convex closure of the support $\{0,1\}^n$ of $d\mu(x)$. In this case the partition function is

$$Z(\lambda) = \int e^{-(C,\varphi)(x)} d\mu(x) = \prod_{i=1}^{n} \frac{1}{2}(1 + e^{-C_i})$$

and once the λ that minimizes

$$H(\lambda) = \sum_{i=1}^{n} \ln(1 + e^{-C_i}) + (\lambda, \mathbf{b})$$

is found, provided it exists, the analogue of (6.3.6) is

(6.3.7) $$x_i = \left(e^{C_i^*} + 1\right)^{-1}$$

which suggests one should look for values of λ that make the absolute value of $C_i=\Sigma\lambda_j A_{ji}$ very large when searching for λ's that minimize $H(\lambda)$.

6.4 *Maxentropic approach to linear programming.*

When we put out [6.6] and [6.7] we neglected to look through the published literature for related work. To patch this up a bit, here we mention some maxentropic approaches at the linear programming problem. Consider references [6.8]-[6.11] for example. Our approach is nevertheless quite different. We consider the problem of finding

(6.4.1) $$\sup\{(\mathbf{A}_0, x)/\mathbf{x} \in D_1,\ Ax = c\}$$

where $D_1=\{x\in \mathfrak{R}^n\ \ 0\le x_i\le 1,\ \ i=1,\ldots,n\}$, $\mathbf{A}_0$ is a fixed vector in $\mathfrak{R}^n$, $\mathbf{A}$ is an nxk-matrix and $\mathbf{c}$ is a fixed vector in $\mathfrak{R}$.

We shall assume that the nx(k+1)-matrix

$$\begin{pmatrix} A_0 \\ A \end{pmatrix}$$

obtained by adding $\mathbf{A}_0$ to $\mathbf{A}$ as first row is of rank k+1. We shall also assume that $\mathbf{A}$ has at least one row, say the first one, with all entries positive.

We leave for the reader to verify:

i) By an obvious transformation the domain

$$D_2 = \{x \in \mathfrak{R}^n : a_i \le x_i \le b_i;\quad 1 \le i \le n\}$$

can be transformed into D_1.

ii) Using the assumptions we see that for each $1\le i\le n$,

$$0 \le x_j \le C_1/A_{1j}$$

or, in other words, our choice of D_1 is natural.

We shall set

$$\tilde{A} = \begin{pmatrix} \mathbf{A}_0 \\ \mathbf{A} \end{pmatrix},\quad \tilde{C} = \begin{pmatrix} c_0 \\ \mathbf{c} \end{pmatrix}$$

The way we go about solving (6.4.1) is to find a c_0 in $\mathfrak{R}$ for which the maxentropic solution to $\tilde{\mathbf{A}}\mathbf{x}=c$ fails to exist. The first such c_0 will be the one we need.

So, pick a c_0 and consider the reconstruction problem

(6.4.2) $$\tilde{\mathbf{A}}x = \tilde{c}, \quad x_i \in \{0,1\}$$

and on $\{0,1\}$ put the measure $m=1/2\{\delta_0+\delta_1\}$, which induces an obvious Q on $\Omega=\{0,1\}^n$. The elements ω of $\{0,1\}^n$ are configurations with probabilities $P(\omega)$. As always, we denote by $X_i(\omega)$ the i-th element of ω and think of (6.4.2) as

(6.4.3) $$E_P \Sigma A_{ji} X_i = c_j \qquad j = 0, 1, ..., k.$$

Now, maximization of $-\Sigma P(\omega)\ln((P(\omega))/(Q(\omega)))$ subject to (6.4.3) leads to a partition function

$$Z(\Lambda) = \prod_{i=1}^{n} \left[\left(1 + \exp - \left(\Lambda, \tilde{A}_i \right) \right) /2 \right]$$

where Λ^t is the vector $(\lambda_0, \lambda_1, ..., \lambda_{k+1})$ and $\tilde{\mathbf{A}}_i$ is the vector corresponding to the i-th column of $\tilde{\mathbf{A}}$.

Again, Λ is to be found by minimizing

$$H(\Lambda, \tilde{c}) = \lg Z(\Lambda) + (\Lambda, \tilde{c}).$$

The main result in [6.6], modulo misprints, is

Theorem 6.4.4. For a given c_0, there will be an $\mathbf{x} \in D_1$ such that $\tilde{\mathbf{A}}\mathbf{x}=\mathbf{c}$ if and only if $H^*(c_0,\mathbf{c})=\min\{H(\Lambda,\mathbf{c}): \Lambda \in \Re^{k+1}\}$, and the problem will have no solution whenever $H^*(c_0,\mathbf{c})=-\infty$. More precisely, let K be the compact convex set

$$K = \left\{ \tilde{c} \in \Re^{k+1};\ \hat{\mathbf{A}}x = \tilde{c} \quad \textit{for some} \quad x \in D_1 \right\}$$

then

1) If $c \in K$, there exist infinitely many solutions to $\tilde{\mathbf{A}}\mathbf{x}=c$, one of which is of the form

(6.4.5) $$x_i^* = \left(1 + \exp - \sum_{j=0}^{k} \lambda_j A_{ji} \right)^{-1} \qquad i = 1, 2, ..., n.$$

2) If $c \in \overset{0}{K} - K$, then there exists a solution x^j to $\tilde{\mathbf{A}}\mathbf{x}=\mathbf{c}$ of the form

$$\begin{cases} x_i^k = 1 & i \in P \\ x_i^k = 0 & i \in N \\ x_i = \lambda_i & 0 \le \lambda_i \le 1,\ i \notin P \cup N \end{cases}$$

where P and N are disjoint subsets of $\{1,2,...,n\}$.

Proof: Part (1) is easy and left for the reader. To see (2) let c_0 be such that $c \in \overset{0}{K}-\overset{0}{K}$. Then there is a sequence c_n in $\overset{0}{K}$ with $c_n \to c$ as $n \to \infty$. For each for them there are a Λ_n **and a** x^k_n such that $h(x^k_n)=H^*(\Lambda_n)\geq 0$. Here $h(\mathbf{x})=-\Sigma x_i \ln x_i+(1-x_i)\ln(1-x_i)$. By passing to a subsequence if necessary, let $x^* \in D_1$ be such that $x^k_n \to x^*$. Now $h(x^*)= H^*(c)=\lim H(\Lambda_n, \tilde{c}_n) \geq 0$

From (6.4.5) we see that

$$P = \{i : x_i^* = 1\} = \left\{i : \sum_{j=0}^{n} \Lambda_j^n A_{ji} \to \infty\right\}$$

and depending on the vectors $\mathbf{A}_i$, we may have $(\Lambda_n, \mathbf{A}_i)$ converging to a finite limit despite $\|\Lambda^n\| \to \infty$. These comprise the third case.

$$N = \left\{i : x_i^k = 0\right\} = \left\{i : \sum_{i=0}^{n} \Lambda_j^n A_{ji} \to \infty \quad as \quad n \to \infty\right\}.$$

If c_0 is such that $\overset{0}{c} \in K$, then $H^*(\mathbf{c})= -\infty$. This is due to the fact that $H(\Lambda,\mathbf{c})$ is convex on $\Re^{k+1}$, and if

$$\sum_{i=1}^{n}\left[(\exp-(\Lambda, A_i)A_{ji})\left(1+\exp-\left(\Lambda, \tilde{A}_i\right)\right)^{-1}\right] = \tilde{\mathbf{c}}_j \qquad j = 0, ..., k+1$$

then the minimum of $H(\Lambda,\mathbf{c})$ cannot be achieved and $H^*(\mathbf{c})= -\infty$.

The practical way to use this result would be to start from a variable $\mathbf{x}_0$, i.e. such that $\mathbf{A}\mathbf{x}_0=\mathbf{c}$ holds. Then compute $c^0_0=(\mathbf{A}_0,\mathbf{x}_0)$ and start solving maximum entropy problems for c_0 larger than c^0_0 until the max-ent procedure breaks down.

Instead of (6.4.1) we could have considered the problem of maximizing $(\mathbf{A}_0,\mathbf{x})$ when the constraints are known with uncertainty, namely we want to find

$$\max\{(\mathbf{A}_0, \mathbf{x}) | \mathbf{x} \in D_1 \cap \{\xi | \mathbf{A}\xi - \mathbf{c} \in B_M(\delta)\}\}$$

where $\mathbf{c} \in \Re^k$, D_1 is the unit cube in $\Re^n$ and $B_M(\delta)=\{y \in \Re^k | (y,My) \leq \delta\}$ for a positive definite, symmetric kxk matrix M.

Here instead of minimizing $H_0(\Lambda,\mathbf{c})=\ln Z(\Lambda)+(\Lambda,\mathbf{c})$ to find the vector Λ of Lagrange multipliers, standard procedure leads us to minimize

$$H_\delta(\tilde{\Lambda}, \tilde{\mathbf{c}}) = H_0(\tilde{\Lambda}, \tilde{\mathbf{c}}) + \delta(\Lambda, M^{-1}\Lambda)^{\frac{1}{2}}.$$

Recall that $\tilde{\Lambda} = \begin{pmatrix} \Lambda_0 \\ \Lambda \end{pmatrix}$ with $\Lambda \in \Re^k$

6.5 *Entropy as Lyapunov functional. Further comments.*

This section provides more substance to the way the second law of thermodynamics was phrased in chapter 3. To make things simple we shall assume that the states of our system are discrete and that microscopic dynamics are described by an infinitesimal transition matrix (or rate matrix) W_{ij}. We shall denote the set of microscopic states by S.

Thus, if $P_i(t)$ denotes the probability of finding the system in state i at time t, when at time t=0 the distribution was known to be $P_i(0)$, then

(6.5.1) $$\frac{d}{dt}P_i = \sum_j W_{ij}P_j(t) = \sum P_j(t)W_{ji}.$$

When there is no loss of probability,

$$\sum_{j \neq i} W_{ji} = -W_{ji}$$

and setting $\Lambda_i = W_{ij}$ (this is the mean holding time at state i) we have $Q_{ij} = W_{ij}/\Lambda_i$ for jump distribution (see [6.12] for more on this and other stuff).

Although we do not need to assume symmetry $W_{ij} = W_{ji}$, we do need to assume the existence of a measure μ_i with respect to which our dynamics satisfies the detailed balance condition

(6.5.2) $$\mu_i W_{ij} = \mu_j W_{ji}, \qquad all \quad i,j$$

which, summing over i implies the invariance condition for μ

(6.5.3) $$\sum_i \mu_i W_{ij} = 0.$$

For references to detailed balance and applications see [6.13]-[6.14].

If we denote by $\mathbf{P}_{ij}(t)$ the solution to (6.5.1) with $P_i(0) = \delta_{ij}$, then setting

$$\mu_i(t) = \sum_j \mu_j P_{ji}(t)$$

we see from (6.5.3) and (6.5.1) that $d\mu_i(t)/dt = 0$.

We shall say that a function f(i) defined on the state space is invariant (or harmonic) if

$$\sum_j \mathbf{P}_{ij}(t)f(j) = f(i) \tag{6.5.4 – a}$$

and making use of (6.5.1)

$$\Sigma W_{ij} f(j) = 0 \tag{6.5.4 – b}$$

and hence the name harmonic. We assume we have at least a few invariant functions. If we set for any probability distribution $\{P_i\}$ and a given invariant distribution $\{\mu_i\}$

$$S_\mu(P) = -\sum_i P_i \ln P_i/\mu_i \tag{6.5.5}$$

we see that if $P_i(t)$ satisfies (6.5.1)

$$\frac{d}{dt}S_\mu(P) = -\sum_i \frac{dP_i}{dt} \ln P_i/\mu_i = -\sum\sum_{i\,j\neq i} (P_j W_{ji} - P_i W_{ij}) \ln P_i/\mu_i$$

$$= -\sum\sum_{i\,j\neq i} \left\{ \frac{P_j}{\mu_j}\mu_j W_{jx} - \frac{P_i}{\mu_i}\mu_i W_{ij} \right\} \ln P_i/\mu_i$$

$$= -\frac{1}{2} \sum\sum_{i\,j\neq i} \mu_j W_{ji}((P_j/\mu_j) - (P_i/\mu_i))(\ln P_i/\mu_i - \ln P_j/\mu_j) \geq 0.$$

Thus, for any initial value $p_i(0)$, $S_\mu(p)$ increases until $p_i=\mu_i$ for all i. From the point of view of physical applications, we need a supply of invariant measures μ.

Let $f_1,\ldots,f_N$ be N invariant functions, let m be any invariant measure on the set of states and, as above, let

$$Z(\boldsymbol{\lambda}) = \sum_i m_i \exp -(\boldsymbol{\lambda}, \mathbf{f}_i)$$

$$\mathbf{P(F)} = \{P \in \mathbf{P(S)} | P << m, \quad \langle P, \mathbf{f} \rangle = \mathbf{F}\}$$

where **F** is an element in $\mathfrak{R}^N$ with components F_j, Λ and **f**(i) are in $\mathfrak{R}^N$ with components Λ_j, $f_j(i)$, j=1,...,N. Also, we set <**P**,**f** > for the vector with components $\Sigma \mathbf{P}_i f_j(i)$for j=1,2,...,N. If we think of **P** as a row vector, then **P**(t)= **P**$\mathbf{P}$(t) is also a row vector.

Notice that for $P \in \mathbf{P}(\mathbf{F})$, $P(t) \in \mathbf{P}(\mathbf{F})$ for $\langle P(t), \mathbf{f}\rangle = \langle P, \mathbf{P}(t)\mathbf{f}\rangle = \langle P, \mathbf{f}\rangle$ since the f_j are invariant.

Consider $S_m(P)$ restricted to $\mathbf{P}(\mathbf{F})$. From what we know from before, there is a unique P^* in $\mathbf{P}(\mathbf{F})$ such that $S_m(P^*) = \sup\{S_m(P) | P \in \mathbf{P}(\mathbf{F})\}$. Also

$$S_m(P^*) = H(\Lambda^*) = \ln Z(\Lambda^*) + (\Lambda^*, \langle \mathbf{f} \rangle)$$

where Λ^* minimizes

$$H(\Lambda) = \ln Z(\Lambda) + (\Lambda, \langle \mathbf{F} \rangle).$$

Assume, which is reasonable for physical applications, that for $P \in \mathbf{P}(\mathbf{f})$ $\lim P(t) = P_{eq}$ exists, and denote by $P^*(t) = P^*\mathbf{P}(t)$ the time evolved of P^* Since $S_m(P^*(t))$ is increasing on $\mathbf{P}(\mathbf{F})$ and its smallest value $S_m(P^*)$ is already the largest value of $S_m(\circ)$ on $\mathbf{P}(\mathbf{F})$ it follows; from the uniqueness of P^*, that

$$P^*(t) = P^*\mathbf{P}(t) = P^*$$

Theorem 6.5.6. The measure P^* yielding a maximum for the entropy $S_m(P)$ over $\mathbf{P}(\mathbf{F})$ is an equilibrium measure for the microscopic dynamics given by $\mathbf{P}(t)$.

It is not hard to conceive all sort of extensions of these results.

Let us say a few things about the use of the entropy as a Lyapunov functional.

Assume that $\{\mu_i\}$ is an invariant distribution and $\{P_i\}$ is any distribution. We saw in Lemma 4.5 that $-S_\mu(P) \geq 0$ (here we let counting measure on S to play the role of what we denote by μ there). Above we saw that $dS_\mu/dt \geq 0$, when we let $P(t) = \mathbf{P}P(t)$. Notice that when P happens to be invariant, then $S_\mu(P)$ is constant in time. We shall set

$$A(\mu) = \{P \in \mathbf{P}(S) | dS_\mu(P(t))/dt > 0 \quad some \quad t \geq 0, \quad P \neq \mu\} \tag{6.5.7}$$

and we shall call it the attractor of μ. Notice that we exclude μ from it.

Theorem 6.5.8. If $P \in A(\mu)$ then $P(t)$ tends to μ as t tends to infinity whenever $S_\mu(P(t)) \uparrow 0$.

Proof: Consider first $t = inf\{t > 0 | dS/dt = 0\}$. Note from the computation of dS/dt given above that the right hand side vanishes if and only if $P_i(t_0) = \mu_i$. Therefore if $t_0 < \infty$, $P_i(t) = \mu_i$ for all $t \geq 0$.

Consider now the case $dS_\mu/dt > 0$ for all $t > 0$. From (4.38) we obtain that

$$\tfrac{1}{4}\Big(\sum_i |P_i(t) - \mu_i|\Big)^2 \leq -S_\mu(P(t)).$$

Since the right hand side goes to zero, by passing to a subsequence if required, we obtain $P_i(t) \to \mu_i$ for all i.

Comment. Note that assuming that $P(t)\to P_{eq}$ is not enough, it may happen that $S_\mu(P_{eq})\neq 0$, and $(\Sigma|P_i(t)-\mu_i|)^2$ may only oscillate in the interval $(0,-S_\mu(P_{eq}))$.

6.6 *Solving matrix equations.*

Let us state a few problems leading to search for solution of the matrix equation

(6.6.1) $$AX = C$$

where A, X, C are respectively $n\times m$, $m\times k$ and $n\times k$ matrices. Here A and C are given and we shall require the unknown matrix X to have its components in a preassigned convex set. We direct the reader to [6.4] and [6.16] for more on related issues, namely, different problems leading to matrix equations like (6.6.1) and their solution via the level 1 maximum entropy method.

Example 1. Let A_{ij} denote the intensity of spectral band i, $1\le i\le M$ of a substance j, $1\le j\le N$. Assume that the intensity c_i in the i-th band for mixture is known and we want to know the concentration x_j of substance j in the mixture. Certainly the normalization $0\le x_j\le 1$, for $1\le j\le N$ is natural in this case.

Example 2. Consider the problem of finding the generalized inverse X of a matrix A. The whole thing here is that A may not be a square matrix. The matrix equation defining X is

(6.6.2) $$AXA = A$$

For a very fast review and analysis of best solutions in norms other than l_2 see [6.16].

Example 3. Consider the extension of (6.6.2) to either of

(6.6.3) $$AXB = C \quad or \quad \sum_{i=1}^{L} A_i X B_i = C.$$

Example 4. Relating stimuli to responses by means of linear maps. Suppose you encode stimuli by vectors in certain $\mathfrak{R}^n$ and have m of them, described by $\{S_{ij}\ 1\le i\le n,\ 1\le j\le m\}$. Assume that the system under scrutiny responds linearly to the stimuli to produce k different responses encoded by vectors in $\mathfrak{R}^m$ You want to know the mechanism, or transfer matrix such that

(6.6.4) $$SX = R.$$

You may need different k and m because, say, the independent or different stimuli may yield common or related responses. Besides that you may know in advance, or need to assume that

the x_{ij} are to take values in some preassigned set; {-1,1} say.

For the fun of it, we shall look at the problem of finding a 1xn-matrix **X** , the inverse of the nx1-matrix **A**, such that (6.6.2) holds and $-\|\mathbf{A}\|<X_i<\|\mathbf{A}\|$. On $\Omega=[-1,1]^n$ equipped with the Borel σ-algebra we shall define the measure m(dx) with density 2^{-n} with respect to $dx=dx_1,\ldots,dx_n$.

Denote by ξ_i the coordinate maps $\xi_i(\omega)=\omega_i$ and by $\Phi_i(\omega)$ the map

$$a_i \sum_j a_j \xi_j(\omega)$$

where a_j denotes the j-th component of **A**. We shall look for measures P on Ω such that

$$E_P \Phi_j = a_j$$

for j=1,2,...,n. The standard level 2 procedure yields

$$dP = (Z(\lambda))^{-1} \exp-(\lambda, \Phi) m(dx)$$

with $Z(\boldsymbol{\lambda})$ such that

$$\ln Z(\boldsymbol{\lambda}) = \sum_i \left[(a_i(\lambda, \mathbf{A})\|\mathbf{A}\|)^{-1} \sinh a_i(\lambda, A)\|\mathbf{A}\| \right].$$

Then the minimum of $H(\boldsymbol{\lambda})=\ln Z(\boldsymbol{\lambda}) + (\boldsymbol{\lambda},\mathbf{A})$ is reached at a $\boldsymbol{\lambda}^*$ such that

$$\|\lambda\| \Sigma a_i \coth (a_i(\boldsymbol{\lambda}^*, \mathbf{A})\|\mathbf{A}\|) - n/(\boldsymbol{\lambda}^*, \mathbf{A}) = 1$$

and when $X_i=E_P \xi_i$ is computed, it comes out as

$$X_i = \|\mathbf{A}\| \coth (a_i(\boldsymbol{\lambda}^*, \mathbf{A})\|\mathbf{A}\|) - 1/a_i(\boldsymbol{\lambda}^*, \mathbf{A})$$

which satisfies $(X_i, a_i)= 1$ so that AXA=A (So, we can go to sleep with a feeling of being consistent. Some may think that this is a dum way of writing $x_i=A_i/\|\mathbf{A}\|^2$)!. This would be true if $\|X\|=1/\|\mathbf{A}\|$ which is not apparent from the result found above. Also, try to find X using singular values decomposition.

6.7 *Estimation of transition probabilities.*

The following is a variation on the theme of a nice paper by Bard, [6-18], in which state probabilities are estimated, using the M.E.M., from the knowledge of probabilities of a collection of sets.

Suppose S is a finite set, with atoms $J_1,...,J_n$ and the σ-algebra S consists of the collection of subsets of S. Any probability P on S is then determined by the $P(\{S_i\})$.

Also, if (X_n) denotes a time homogeneous Markov chain, having S as state space, then the transition matrix

$$P_{ij} = P(X_1 = J_j | X_0 = J_i) \tag{6.7.1}$$

determines the time evolution of (X_n).

It may happen that for every starting point s_i, only the aggregate transition matrix

$$\tilde{P}_{ij} = P(X_1 \in C_{ij} | X_0 = J_i) \tag{6.7.2}$$

where the C_{ij} are a collection of not necessarily exclusive nor exhaustive events. In terms of the P_{ij}, the P_{ij} can be rewritten as

$$\tilde{P}_{ij} = \sum_{k \in C_{ij}} P_{ik} I_{C_{ij}}(S_k) \tag{6.7.3}$$

where, for any set A, I_A is the indicator function of A.

Observe as well that the C_{ij} may be different for each starting point i, but that is not an essential assumption. Also, for each fixed i,

$$\sum_j P_{ij} = 1 \tag{6.7.4}$$

is a condition satisfied if the chain is conservative (When it does not hold, we throw in a cemetery state to enforce it).

So our problem becomes that of determining for each i, a collection P_{ij} satisfying (6.7.4) when all that is known is (6.7.3). Dropping any reference to the index i, we are in the situation discussed by Bard. So, we will follow him.

Given sets C_j, j=1,2,...,K; we denote by D_j; $1 \le j \le M$ the partition of S induced by the C_j, that is, D_j is a mutually exclusive, exhaustive collection such that any set in the σ-algebra generated by $\{C_j\}$ is a union of sets from $\{D_j\}$. In particular

$$C_j = \bigcup_{K=1}^{M} (C_j \cap D_K), \quad \textit{obviously} \quad M \ge K.$$

Observe that if we know, for each j=1,...,K

$$\tilde{P}_j = \sum_{S_k \in C_j} P(S_K) \qquad j = 1, 2, ..., K \tag{6.7.5}$$

the most we can hope for is to be able to find

$$Q_l = \sum_{S_k \in D_l} P(S_K) \qquad l = 1, 2, ..., M.$$

The constraints being, of course

$$\sum_{i=1}^{M} Q_l = 1. \tag{6.7.6-a}$$

From the solution to this problem, the original problem drops out, for the procedure can be carried out for each starting point i.

Observe as well that once the Q_l are known, we can setup the problem of finding P_j, j=1,...,N such that

$$\sum_{j=1}^{N} P_j = 1 \tag{6.7.7}$$

$$\sum_{D_l \cap C_j \neq \phi} Q_l = \tilde{P}_j \qquad j = 1, ..., k. \tag{6.7.6-b}$$

So Bard's technique, applied twice in succession, provides us with the complete collection of (transition) probabilities. Let us apply the M.E.M. to solve the set (6.7.6). Again, denoting by λ the Lagrange multiplier corresponding to (6.7.6-b) we would obtain, after an application of the level 1 routine

$$Q_l = \exp -\sum_j \lambda_j \chi(D_l \cap C_j)/Z(\lambda) \tag{6.7.8}$$

where we set $\chi(A)=1$ or 0 depending on A being empty or not. Finding the $\boldsymbol{\lambda}^*$ such that (6.7.6-b) are met provides us with the Q^*_1 that maximize the entropy and satisfy (6.7.6-a).

If you compare (6.7.8) with Bard's results you will notice some differences, stemming from the fact that he does not have condition (6.7.6-a). As a simple minded application, that can be worked out by hand consider the problem of figuring out the probabilities of the different outcomes of a die throw when you only know

$$\tilde{P}_1 = P(2,3,4) = \tfrac{1}{2} \qquad \tilde{P}_2 = (3,4,5) = \tfrac{1}{2}.$$

The sets $C_1=\{2,3,4\}$ and $C_2=\{3,4,5\}$ determine the partition $D_1=\{1,6\}$, $D_2=\{2\}$, $D_3=\{3,4\}$ and $D_4=\{5\}$ of the sample space.

According to (6.7.8) the outcomes of the throw fall in these sets with frequencies

$$Q_1 = 1/Z, \quad Q_2 = e^{-\lambda_1}/Z, \quad Q_3 = e^{-(\lambda_1+\lambda_2)}/Z, \quad Q_4 = e^{-\lambda_2}/Z$$

where $Z = 1 + e^{-\lambda_1} + e^{-(\lambda_1+\lambda_2)} + e^{-\lambda}$

An easy computation shows that

$$e^{-\lambda_1} = \frac{\tilde{P}_1}{1-\tilde{P}_1} \qquad e^{-\lambda_2} = \tilde{P}_2/\left(1-\tilde{P}_2\right)$$

$$Z = 1/\left(1-\tilde{P}_1\right)\left(1-\tilde{P}_2\right)$$

which, when $\tilde{P}_1 = \tilde{P}_2 = 1/$, yield $\boldsymbol{\lambda}_1=\boldsymbol{\lambda}_2=0$ and therefore

$$q_1 = \tfrac{1}{8}, \quad q_2 = \tfrac{1}{4}, \quad q_3 = \tfrac{1}{8}, \quad q_4 = \tfrac{1}{4}$$

and repeating this procedure yields the frequencies

$$P_1 = \tfrac{1}{8}, \quad P_2 = \tfrac{1}{4}, \quad P_3 = \tfrac{1}{8}, \quad P_4 = \tfrac{1}{8}, \quad P_5 = \tfrac{1}{4}, \quad P_6 = \tfrac{1}{8}$$

for the different throws of the die.

But had we begun with $P(\{1,2,3,4\})=P(\{3,4,5,6\})=2/3$, we would have obtained $P_1=P_2=P_3=P_4=P_5=P_6$. So knowledge counts after all. Curious huh?!

A larger scale application of this technique could be the following random search algorithm. Let 1=1,...,N label the points of a grid and let P_1 denote the probability of finding the particle, individual, oil, or water at i.

Assume you have a way of assigning areas to detection procedures and you determine

$$\tilde{P}_i = P(C_i) = \sum_{k \in C_i} P_k$$

by some experimental procedure. For example, P_1 is the fraction of successful detections in region C_i. The C_i, i=1,2,...,k are some not necessarily disjoint nor necessarily covering of the whole domain. The procedure outlined above would yield the P_i.

6.8 *Maxentropic reconstruction of velocity profiles.*

Even though we shall be following [6.19] we urge the reader to take a look at [6.20] on which it is based. Especially the section devoted to the choice of the a priori profile. Instead of directly applying the results in section 4 of chapter 5 I shall repeat myself a bit and, restate the results of [6.19]

At given instants $t_1,...,t_n$ during an interval [0,T] the following mean square averages are some how determined

(6.8.1) $$d_j = t_j V^2(t_j) = \int_0^{t_j} V^2(s)ds = \int_0^T \rho_j(s)V^2(s)ds$$

where $\rho_j(s) = 1$ for $s \le t_j$ and 0 for $t_j < s \le T$. These averages are somehow computed from 1D seismograms. Our problem is to find a positive V(s) such that (6.8.1) holds. Actually we could refine things a bit and have $V_l < V(s) < V_u$ for two physically reasonable upper and lower bounds.

Again, since we do not want to fool around with measures on function spaces, we will discretize [0,T] by introducing a partition of size N and replace (6.8.1) by

(6.8.2) $$t_j V^2(t_j) = \sum \tfrac{T}{N} \rho_{jk} x_k = d_j$$

where $\rho_{jk} = \rho_j\left(\tfrac{T}{N}(k-1)\right)$, $x_k = V^2\left(\tfrac{T}{N}(k-1)\right)$ $k = 1, \ldots, N$.

Also, to avoid ridiculous complications we assume that the t_j happen to coincide with points of the partition. And we will want to think of the x_k as $E_p X_k$ and, as usual of (6.8.2) as

(6.8.3) $$E_P\Big[\Sigma \tfrac{T}{N}\rho_{jk}X_k\Big] = d_j.$$

Even though somewhat unrealistic from the physical point of view, we shall assume that the random variables X_k take values in the interval $[L,\infty)$ and on $\Omega_N=[L,\infty)^N$ we shall define an a priori reference measure $dQ(\mathbf{x})$ with density

(6.8.4 – *a*) $$q_N(\mathbf{x}) = \prod_{i=1}^{N} \tfrac{1}{x_0(i)-L} \exp-(x_i - L)/(x_0(i) - L)$$

with respect to the density $d\mathbf{x}=dx_1 \ldots dx_N$ on $[L,\infty)^N$. The $x_0(i)$ are chosen so that $x_0(i)>L$ and therefore

(6.8.4 – *b*) $$\int_{\Omega_N} \mathbf{x}_i q_N(\mathbf{x})d\mathbf{x} = x_0(i), \quad i = 1, \ldots, N.$$

The Q-entropy of P is defined by

$$S_Q^N(P) = -\tfrac{T}{N} \int_{\Omega_N} \rho(\mathbf{x}) \ln p(\mathbf{x})/q_N(\mathbf{x})d\mathbf{x}$$

and the partition function is now

$$Z_N(\lambda) = \int_{\Omega_N} \exp-\Big\{\sum_{i=1}^{N} (\lambda, \varphi(i))x^i\Big\} q(x)dx$$

$$= \prod_{i=1}^{N} (x_0(i) - L)^{-1}\Big((x_0(i) - L)^{-1} + (\lambda_1, \varphi(i))\Big)^{-1} \exp-(\lambda, \varphi(i))$$

where $\varphi(i)$ is the n-vector with components ρ_{ji} and $\boldsymbol{\lambda}$ is in $\Re^N$ The maxentropic P will have density $P_N(\mathbf{x})$ given by

$$P_N(\mathbf{x}) = \prod_{i=1}^{N} a_i e^{-a_i(x_i - L_i)}$$

where $a_i = \tfrac{1}{x_0(i)-L} + (\lambda, \varphi(i))$.

The maxentropic estimate or reconstruction is

$$\tilde{x}_i = \int_{\Omega_N} x_i P_N(\mathbf{x}) d\mathbf{x} = L + 1/a(i). \tag{6.8.5}$$

Except for the fact that we have to find the $\boldsymbol{\lambda}$ minimizing

$$H_N(\boldsymbol{\lambda}) = \tfrac{T}{N} \ln Z(\boldsymbol{\lambda}) + (\lambda, \mathbf{d}) \tag{6.8.6}$$

where $\mathbf{d}$ is the n-vector with components d_j, j=1,2,...,n. We are through. Setting $x_0(i)=v^2_0(i)$, $L=v_1^2$ and letting N tend to infinity we would have

$$V^2(t) = V_1^2 + \left(\left(V_0^2(t) - V_e^2 \right)^{-1} + (\lambda, \varphi(t)) \right)^{-1} \tag{6.8.7}$$

instead of (6.8.5) and $\boldsymbol{\lambda}$ is to be determined minimizing

$$H_\infty(\boldsymbol{\lambda}) = \int_0^T \ln \left\{ \frac{\exp -(\lambda, \varphi(t))}{1 + \left(V_0^2(t) - V_e^2 \right) + (\lambda, \varphi(t))} \right\} dt + (\boldsymbol{\lambda}, \mathbf{d})$$

where of course $\varphi(t)$ is the n-vector with components $\rho_j(t)$. To simplify, as in [6.19], we set $V_1^2=0$, $V_0^2(t)=V_u^2$ constant, which is to be added to $\boldsymbol{\lambda}_n$.

It is easy to verify that

$$d_j = \int_0^{t_j} \frac{ds}{V_0^2 + \sum_1^n \lambda_k \varphi_k(s)} = d_{j-1} + \left(t_j - t_{j-1} / \left(V_0^2 + \sum_j^n \boldsymbol{\lambda}_k \right) \right).$$

Inserting this into (6.8.6), we note that for $t_{j-1} < u < t_j$

$$V(u) = \frac{d_j - d_{j-1}}{t_j - t_{j-1}} = \frac{1}{V_0^2 + \sum_j^n \lambda_k} = t_j V_R^2 t_{j-1} V_R^2 (t_{j-1})(t_j - t_{j-1})^{-1}$$

which is a standard result in models of layered earth .

6.9 *Fragmentation in a nuclear reaction.*

This is a part of a project, once started with L. Dohnert, based on [6.21]. The problem studied there is to understand the fragmentation of a heavy nucleus by a fast light nucleus.

The everyday language description of the process consists in supposing that the large nucleus gets "hot" when it absorbs the kinetic energy of the smaller nucleus. Upon cooling down, it condenses in globules that fly away. The problem is to find the distribution of the fragments.

To be precise, we specify the outcome of a reaction by giving {n(i,j): $0 \le i < j$, i,j integers}, where n(i,j) is the number of fragments of mass j (measured by the number of nucleons) and charge i (measured by the number of protons).

The macroscopic constraints on n(i,j) are

$$F(\{n(i,j)\}) = \sum_{j>i} n(i,j)$$

(6.9.1)

$$Q(\{n(i,j)\}) = \sum_{j>i} i n(i,j)$$

$$M(\{n(i,j)\}) = \sum_{j>i} j n(i,j)$$

The meaning being: F({n}), Q({n}) and M({n}) stand for the number of fragments, the charge, the mass of the distribution {n} respectively. We want to find the measure P({n}) defined on the set of all possible configurations, and such that

$$\sum_{\{n\}} P(\{n\}) = 1$$

$$\sum_{\{n\}} P(\{n\}) F(\{n\}) = N_0$$

(6.9.2)

$$\sum_{\{n\}} P(\{n\}) Q(\{n\}) = Z_0$$

$$\sum_{\{n\}} P(\{n\}) M(\{n\}) = A_0$$

where the numbers on the right hand side denote the average number of fragments, charge and mass respectively. The maxentropic procedure would yield a probability

(6.9.3)

$$P(\{n\}) = \exp -[\boldsymbol{\lambda}_1 F(\{n\}) + \boldsymbol{\lambda}_2 Q(\{n\}) + \boldsymbol{\lambda}_3 M(\{n\})]/Z(\boldsymbol{\lambda})$$

where, as usual

$$
\begin{aligned}
Z(\boldsymbol{\lambda}) &= \sum_{\{n\}} \exp-\{\lambda_1 F(\{n\}) + \lambda_2 Q(\{n\}) + \lambda_3 M(\{n\})\} \\
&= \sum_{\{n\}} \prod_{j>i} \exp-(\lambda_1 + \lambda_2 i + \lambda_3 j) n(i,j) \\
&= \prod_{j>i} \left(\sum_{m=0}^{\infty} \exp-(\lambda_1 + \lambda_2 i + \lambda_3 j) n \right) \\
&= \prod_{j>i} (1 - \exp-(\lambda_1 + \lambda_2 i + \lambda_3 j))^{-1}
\end{aligned}
\tag{6.9.4}
$$

from which, differentiating with respect to the appropriate $\boldsymbol{\lambda}$, we obtain

$$
\begin{aligned}
N_0 &= -\frac{\partial}{\partial \lambda_1} \ln Z(\boldsymbol{\lambda}) = \sum_{j>i} (\exp(\lambda_1 + \lambda_2 i + \lambda_3 j) - 1)^{-1} \\
Z_0 &= -\frac{\partial}{\partial \lambda_2} \ln Z(\boldsymbol{\lambda}) = \sum_{j>i} i(\exp(\lambda_1 + \lambda_2 i + \lambda_3 j) - 1)^{-1} \\
A_0 &= -\frac{\partial}{\partial \lambda_3} \ln Z(\boldsymbol{\lambda}) = \sum_{j>i} j(\exp(\lambda_1 + \lambda_2 i + \lambda_3 j) - 1)^{-1}
\end{aligned}
\tag{6.9.5}
$$

Obviously, we can interpret $(\exp(\lambda_1+\lambda_2 i+\lambda_3 j)-1)^{-1}$ as the probability of finding a fragment of charge i and mass j.

What is left to do is to rewrite $\ln Z(\boldsymbol{\lambda})$ as

$$
\ln Z(\boldsymbol{\lambda}) = \int_0^{\infty} du \int_0^{\infty} dv \ln [1 - \exp-(\lambda_1 + \lambda_2 u + \lambda_3 v)]
$$

which is what any decent physicist would do. By differentiating we obtain the integral analogues of (6.9.5) and the $\boldsymbol{\lambda}$ can be found by minimizing

$$
H(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \lambda_1 N_0 + \lambda_2 Z_0 + \lambda_3 A_0
$$

over the set $D=\{\boldsymbol{\lambda} | Z(\boldsymbol{\lambda}) < \infty\}$, which has to be precisely determined. This set seems to be the positive orthant in $\Re^3$ Can you enlarge it?

6.10 *Maxentropic inversion of Laplace transforms.*

Suppose you want to recover a continuous function f(t) defined on $[0,\infty)$ such that either f(t) tends to zero as t goes to infinity or, that its growth rate is such that for some $\alpha_0 > 0$ $f(t)\exp-\alpha_0 t$ tends to 0 as t goes to infinity. Suppose that you know

$$f(\alpha_i) = \int_0^\infty f(t)\exp-(\alpha_i t)dt \qquad i = 1, \ldots, M$$

where $0 \le \alpha_0 < \alpha_1, \ldots, \alpha_M$, and you want to recover f(t). Note that the change of variables t=-ln s transforms that problem into recovering x(s)=f(-ln s) from

$$\tilde{x}(\alpha_i) = \int_0^1 x(s)\mathbf{s}^{\alpha_i - 1}ds \qquad i = 1, \ldots, M. \tag{6.10.1}$$

The analysis below is modeled on the computations in example 2 of [4.2]. Let W be Wiener's measure on the class Ω of all continuous functions $\omega(s):[0,1] \to \Re$ and, let $\mathcal{F}$ denote the σ-algebra generated by all cylindrical sets.

We denote by $E_P[H]$ the expected value of the functional $H:(\Omega,\mathcal{F}) \to (\Re, \boldsymbol{B})$ with respect to measure dP.

Let $x_0(s)$ be a continuous function on [0,1], such that $x_0(0)=0$ and we define $T:\Omega \to \Omega$ by $T(\omega)(t)=\omega(t)+\int x_0(s)ds$. The famous Cameron-Martin theorem asserts that the measure W on Ω changes according to

$$E_{P_0}[H] = E_W[HM_0]$$

where

$$M = \exp \int_0^1 x_0(s)d\mathbf{B}(s) - \tfrac{1}{2}\int_0^1 x_0(s)^2 ds,$$

and $\boldsymbol{B}(s)$: $\Omega \to \Re$, $\boldsymbol{B}(s)(\omega)=\omega(s)$ denotes the standard brownian motion on [0,1]. Here,

$$\int_0^1 x_0(s)d\mathbf{B}(s)$$

denotes the standard $\hat{I}t_0$ integral of $x_0(s)$ with respect to $\boldsymbol{B}(s)$. All these probabilistic constructions are described in [6.22].

Again, standard reasoning yields

(6.10.2) $$E_{P_0}[\mathbf{B}(t)] = \int_0^t x_0(s)ds$$

and it follows that under P_0, and if we denote by $\Phi(s)$ the $\mathfrak{R}^N$-valued function on [0,1] with components $\Phi_i(s) = s^{\alpha_i}$, note that

(6.10.3) $$E_{P_0} \int_0^1 \frac{\Phi(s)}{s} d\mathbf{B}(s) = \int_0^1 x_0(s)\Phi(s)\frac{ds}{s} = \tilde{\mathbf{x}}_0$$

where $\tilde{x}_0$ is the vector whose components are the Laplace transform of (the initial guess) $\mathbf{x}_0(s)$.

Ours maxentropic problem now is to find a law P on $(\Omega,\mathscr{F})$ such that $S_{P0}(P)$ achieves the maximum value over the class of measures Q on $(\Omega,\mathscr{F})$ such that $Q << P_0$ and

(6.10.4) $$E_Q \int_0^1 \tfrac{1}{s}\Phi(s)d\mathbf{B}(s) = \tilde{\mathbf{x}}$$

where $\tilde{\mathbf{x}}$ is the M-vector with components $\tilde{x}_i$ as in (6.10.1).

Standard argument says that dP/dP_0 is

$$\frac{dP}{dP_0} = \exp-\left(\int_0^1 \tfrac{1}{s}\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}\rangle d\mathbf{B}(s)\right)/Z(\boldsymbol{\lambda})$$

where $Z(\boldsymbol{\lambda})$ can be explicitly computed (fly me in and I'll tell you how, but it is really simple)

$$Z(\boldsymbol{\lambda}) = E_{P_0}\left[\exp-\left(\int_0^1 \langle \boldsymbol{\lambda}, \Phi(s)\rangle d\mathbf{B}(s)\right)\right] = \exp \tfrac{1}{2}\int_0^1 \left\{\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}(s)\rangle^2/s^2 - 2\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}\rangle x_0(s)\right\}ds$$

from which it follows that to obtain $\boldsymbol{\lambda}^*$ that makes P the measure that satisfies (6.10.4) we have to minimize

$$H(\boldsymbol{\lambda}) = \tfrac{1}{2}\int_0^1 \left\{\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}(s)\rangle^2/s^2 - 2\langle \boldsymbol{\lambda}, \boldsymbol{\Phi}(s)\rangle x_0(s)\right\}ds + \langle \boldsymbol{\lambda}, \tilde{\mathbf{x}}\rangle$$

$$= \tfrac{1}{2}\left\{\langle \boldsymbol{\lambda}, C\boldsymbol{\lambda}\rangle^2 - 2\langle \boldsymbol{\lambda}, \tilde{\mathbf{x}}_0\rangle\right\} + \langle \boldsymbol{\lambda}, \tilde{\mathbf{x}}\rangle$$

where C is the matrix with elements

$$C_{ij} = \int_0^1 \tfrac{1}{s^2}\Phi_i(s)\Phi_j(s)ds = (\alpha_i + \alpha_j - 1)^{-1}$$

The minimum of $H(\lambda)$ is achieved at

(6.10.5) $$\boldsymbol{\lambda}^* = C^{-1}(\tilde{\mathbf{x}}_0 - \tilde{\mathbf{x}}).$$

Note that when the Laplace transform of $x_0(e^{-t})$ coincides with that of x, then $\boldsymbol{\lambda}^*$ is 0 and P_0 is already the maxentropic measure on $(\Omega,\mathscr{F})$. The maxentropic reconstruction of x(s) is

$$\hat{x}(s) = \frac{d}{ds}E_P[X(s)] = x_0(s) - \langle \boldsymbol{\lambda}^*, \boldsymbol{\Phi}(s)\rangle$$

and the reader is invited to do the arithmetics involved in verifying that

$$\int_0 \hat{x}(s)\boldsymbol{\Phi}(s) = \tilde{\mathbf{x}}.!$$

In the original variables on $[0,\infty)$ the maxentropic reconstruction of f(t) is

$$f(t) = x_0(e^{-t}) - \sum_{i=1}^{M} \lambda_i^* e^{-\alpha_i t}.$$

6.11 *Maxentropic inversion of Fourier transforms.*

In this section we shall present two approaches to the problem of reconstructing a function x(s) on [0,1] from the values of its first 2M+1 Fourier coefficients

(6.11.1) $$b_k = \int_0^1 e_k(s)x(s)ds, \qquad b_{-k} = \bar{b}_k, \qquad k = 1,\ldots,M$$

where $e_k(s) = \exp 2\pi iks/\sqrt{2\pi}$.

The first approach is a variation on the theme developed in the previous section, the second evolves according to a discretization procedure as in section (5.4). We direct the reader to [6.23] for a level-1 like approach.

The (inessential) difference with the setup of (6.10) is that we allow for the initial point $\boldsymbol{B}(0)$ of the brownian motion on [0,1] to be started with a distribution such that the Wiener measure, W^μ on $(\Omega,\mathscr{F})$ satisfies $W^\mu(\boldsymbol{B}(0)\in A)=\mu(A)$ and $\int\xi\mu(d\xi)=0$.

The measure P_0^μ is similarly defined, i.e., for any measurable functional H

$$E_P^\mu[H] = E_W^\mu[HM_0]$$

with $M_0 = \exp\left\{\int_0^1 x_0(t)d\mathbf{B}(t) - \frac{1}{2}\int_0^1 x_0(t)dt\right\}$.

Again, $x_0(s)$ is the a priori knowledge we have of x(s). Instead of (6.10.2) we now have

$$E_P^{\mu}[\mathbf{B}(t)] = \int \xi\mu(d\xi) + \int_0^t x_0(s)ds = \int_0^t x_0(s)ds \tag{6.11.2}$$

by assumption, and analogously to (6.10.3) we have

$$E_P^{\mu}\left[\int_0^1 \mathbf{\Phi}(s)d\mathbf{B}(s)\right] = \int_0^1 x_0(s)\mathbf{\Phi}(s)ds = \mathbf{b}^0 \tag{6.11.3}$$

where $\mathbf{\Phi}(s)$ is the (2M+1)-vector valued function with components $\mathbf{\Phi}_k(s)=e_k(s)$, $-M\le k\le M$, and $\mathbf{b}^0$ is the (2M+1) Fourier components of x(s).

As always, we search for a measure P^{μ} on and analogously to (6.10.3) we have $(\Omega,\mathcal{F})$ such that $S_P(P^{\mu})$ achieves its maximum possible value over those P^{μ} satisfying

$$E_P^{\mu}\left(\int_0^1 \mathbf{\Phi}(s)d\mathbf{B}(s)\right) = \mathbf{b} \tag{6.11.4}$$

where $\mathbf{b}$ is the vector in (6.11.1). Again, mutatis mutandum, everything is as before, with

$$\frac{\partial P^{\mu}}{\partial P_0^{\mu}} = \exp -\int_0^1 \langle \boldsymbol{\lambda}^*, \boldsymbol{\phi}(s)\rangle dB(s)/Z(\boldsymbol{\lambda})$$

where $\boldsymbol{\lambda}^*$ is the point at which

$$H(\boldsymbol{\lambda}) = \tfrac{1}{2}(\boldsymbol{\lambda}, \mathbf{C}\boldsymbol{\lambda}) - (\boldsymbol{\lambda}, \mathbf{b} - \mathbf{b}^0)$$

achieves its minimum. Now things are simpler due to the orthogonality properties of the $\{e_k(s):-M\le k\le M\}$. Actually $C=C^{-1}$ is a rather simple matrix, to wit.

$$\mathbf{C} = \begin{pmatrix} 0 & & I \\ & 1 & \\ I & & 0 \end{pmatrix}$$

Again, as in the previous section, the maxentropic reconstruction $x^*(s)$ of the function x(s) is

(6.11.5) $$x^* = x_0 + (\mathbf{C}(\mathbf{b} - \mathbf{b}^0), \mathbf{\Phi})$$

which, when written in terms of coordinates becomes

$$x^*(s) = x_0(s) + \sum_{-M}^{M} \left(b_k - b_k^0\right) e_{-k}(s).$$

The (nice) interpretation of this version is left for the reader.

Quite a different approach is obtained when a discretized version of the constraint is used and some a priori bounds for x(s) are brought in. Let us rewrite (6.11.1) as

(6.11.6) $$\frac{1}{N} \sum_{j=0}^{N-1} e_k(j) x_j = b_k, \quad b_{-k} = \bar{b}_k, \qquad k = 0_1, \ldots, M_1$$

where $x_j = x(j/N)$ and $e_k(i) = e_k(j/N)$. We shall assume that each x_j is the mean value of the j-th element of a collection $(X_0, \ldots, X_{N-1})$ defined on $\Omega = [-1,1]^N$ as $X_j(\mathbf{x}) = x_j$, and the a priori reference measure $Q(d\mathbf{x}) = d\mathbf{x}/2^N$ is defined on the Borel sets of Ω.

We want to find a probability $P(d\mathbf{x})$ having density $\rho(\mathbf{x})$ with respect to $Q(d\mathbf{x})$, yielding a maximum value for

$$S_n^N = -\frac{1}{N} \int \rho(\mathbf{x}) \ln \rho(\mathbf{x}) d\mathbf{x}$$

subject to the constraints

(6.11.7) $$E_P\left[\frac{1}{N} \sum_{j=0}^{n-1} e_k(j) X_j\right] = \frac{1}{N} \sum_{j=0}^{N-1} e_k(j) \int \rho(\mathbf{x}) X_j d\mathbf{x} = b_k.$$

The corresponding partition function $Z(\boldsymbol{\lambda})$ is

$$Z(\boldsymbol{\lambda}) = E_Q\left[\exp -\left(\sum_0^{N-1} (\boldsymbol{\lambda}, e(j)) X_j\right)\right] = \prod_{j=0}^{N-1} \left[\sinh (\boldsymbol{\lambda}, \mathbf{e}(j))(\boldsymbol{\lambda}, \mathbf{e}(j))^{-1}\right]$$

where $\boldsymbol{\lambda}$ is the 2M+1 vector with components $\lambda_{-M}, \ldots, \lambda_0, \ldots, \lambda_M$ and $\mathbf{e}(j)$ is, for each j=0,...,N-1, the 2M+1 vector with components $e_{-M}(j), \ldots, e_0(j), \ldots, e_M(j)$.

The value of $\boldsymbol{\lambda}$ that makes the usual maxentropic $\rho(\mathbf{x})$ satisfy (6.11.7) is obtained by minimizing

$$H_n(\boldsymbol{\lambda}) = \frac{1}{N} \ln Z(\boldsymbol{\lambda}) + (\lambda, \mathbf{b}_k)$$

which when N tends to infinity tends to

$$H_\infty(\boldsymbol{\lambda}) = \int^1 \ln\left[\frac{\sinh(\lambda, \mathbf{e}(s))}{(\lambda, \mathbf{e}(s))}\right] ds + (\boldsymbol{\lambda}, \mathbf{b}).$$

The maxentropic reconstruction of X_j is

$$X_j = \left[\langle \boldsymbol{\lambda}, \mathbf{e}(j) \rangle^{-1} - \tanh \langle \boldsymbol{\lambda}, \mathbf{e}(j) \rangle \right]$$

which, in the limit $N \to 0$, $j/n \to s$ for an appropriate sequence, tends to

$$x(s) = \left\{ \frac{1}{\langle \boldsymbol{\lambda}, \mathbf{e}(s) \rangle} - \tanh \langle \boldsymbol{\lambda}, \mathbf{e}(s) \rangle \right\}$$

where, in this case, $\boldsymbol{\lambda}$ is to minimize the $H_\infty(\boldsymbol{\lambda})$.

6.12 *Maxentropic spectral estimation.*

Here we present a few basic results on the problem of reconstructing a time series, or to be more precise of reconstructing a second order stationary process, from the observation of this values at a finite set of times. In almost any of [0.3]-[0.11] there is at least one paper on this issue, but what follows is lifted mainly from [4.1] and [6.25]-[6.27]. For even more references and applications see [6.28].

To establish notation, let us recall a few basic facts, the proof of which appears in chapter 9 of [6.29] and chapter 9 of [1.2].

Let $\{X_n : n \in Z\}$ be a sequence of random variables. We shall say that it is a weakly stationary, centered process if for any n

$$EX_n = 0; \quad EX_n X_{n+k} = C(k).$$

We shall say that $\{X_n\}$ is Gaussian (or $\{X_n\}$ is a Gaussian process) whenever for any finite collection $\{n_1, n_2, \ldots, n_m\}$ of integers and any finite collection $\{a_1, \ldots, a_m\}$ of real numbers, the random variable $a_1X(n_1)+a_2X(n_2)+\ldots+a_mX(n_m)$ has a Gaussian distribution. Some authors phrase it like: {Xn} is a Gaussian process if and only if the vector $(X(n_1),\ldots,X(n_m))$ is a Gaussian $\Re^m$ valued random variable.

Anyway, in terms of the Fourier transform of the distribution of $(X(n_1),\ldots,X(n_m))$, the Gaussian property is stated as

$$E[\exp i \Sigma a_k X(n_k)] = \exp -\tfrac{1}{2}(\mathbf{a}^*, C\mathbf{a})$$

where C is the symmetric, positive definite matrix with elements

(6.12.1) $$C_{ij} = EX(n_i)X(n_j) = R(n_i - n_j).$$

The joint distribution of the $(X(n_1),\ldots,X(n_k))$ has density

(6.12.2) $$\rho(\mathbf{x}) = \left[(2\pi)^N \det C\right]^{-1/2} \exp -\tfrac{1}{2}(\mathbf{x}, C^{-1}\mathbf{x}).$$

The following two basic results show why the correlation function is important for the reconstruction of the process $\{X_n\}$.

Theorem. (Bochner-Herglotz). There exists a positive bounded measure μ on $I=[0,2\pi)$ such that

(6.12.3 – *a*) $$R(n) = \int_I e^{in\alpha} d\mu(\alpha).$$

To state the next result, we need the concept of random measure or random kernel. Let $(\Omega,\mathscr{F},P)$ be a probability space, let $B(I)$ denote the Borel σ-algebra of subsets of $I=[0,2\pi)$. Then

Definition. Z: $B(I)\times\Omega\to[0,\infty)$ is the random kernel associate with the measure μ on I if and only if

a) $A\to Z(A)$ is a finitely additive function and $\Sigma Z(A_n)$ converges to $Z(A)$ in $L_2(\Omega,dP)$.

b) $EZ(A)=0$, $E[Z(A)Z(B)]=EZ(A\cap B)^2=\mu(A\cap B)$.

Bochner-Herglotz theorem is to be complemented with the following.

Theorem. Let $\mu(d\lambda)$ be the spectral measure associated to R(n). Then

(6.12.3 – *b*) $$X_n = \int_I e^{in\alpha} Z(d\alpha).$$

Comment. In terms of an auxiliary dimensional brownian motion $\{B(t): t>0\}$

$$Z([a,b)) = \mathbf{B}(F(b)) - \mathbf{B}(F(a)), \qquad F(a) = \mu([0,a))$$

for any $[a,b)\in I$.

Let us cite two of the standard examples.

i) Let ε_n be an uncorrelated (independent due to Gaussianness) collection, having $E\varepsilon_n=0$, $E\varepsilon_n^2 = \sigma^2$. Thus $R(n) = \sigma^2\,\delta_{v,0}$ and in this case $\mu(d\alpha)=d\alpha/2\pi$.

ii) Let (X_n) be the AR(p) (autoregressive of order p) process satisfying

(6.12.4) $$\sum_{k=0}^{p} a_k X_{n-k} = \varepsilon_n$$

where ε_n is as in (I). We leave for the reader to verify that $\mu(d\alpha)=g(\alpha)d\alpha$ with

(6.12.5) $$g(\alpha) = \frac{\sigma^2}{2\pi}\left|q(e^{-i\alpha})\right|^{-2}$$

where

$$q(x) = \sum_{k=0}^{p} a_k x^k$$

the relationship between $g(\lambda)$ and R(n) is contained in the identity

(6.12.6) $$g(\alpha) = \frac{1}{2\pi}\Sigma R(n)e^{i\alpha_n}$$

Now we can state the reconstruction problem as:

Given a sequence $X_0,\ldots,X_N$ of observation of the process (X_n), we want to find, estimate or compute R(n). Of course from this, using (6.12.3-a) we compute $\mu(d\alpha)$ and using (6.12.3-b) and the comment we reconstruct (or construct a realization of) X_n.

How to proceed, being the good maxentropists that we are? From the observation $X_0,\ldots,X_N$ we have to reconstruct the density $\rho_N(x_0,\ldots,x_N)$. Assume that the measurements allow us to compute the mean values

(6.12.7) $$\tilde{A}_k = \int \rho_N(x_0,\ldots,x_N)A_k(x_0,\ldots,x_N)dx_0,\ldots,dx_N, \qquad 1 \le k \le M.$$

Then the maxentropic distribution, compatible with the knowledge provided by (6.12.7) is given by

(6.12.8) $$P_N(x_0,\ldots,x_N) = \exp-(\boldsymbol{\lambda}, \mathbf{A}(\mathbf{x}))/Z(\boldsymbol{\lambda})$$

where of course

$$Z(\boldsymbol{\lambda}) = \int \exp-(\boldsymbol{\lambda}, \mathbf{A}(\mathbf{x}))d\mathbf{x}, \qquad (\boldsymbol{\lambda}, \mathbf{A}) = \sum_{0}^{N} \lambda_k \mathbf{A}_k.$$

The $\boldsymbol{\lambda}$ which makes $P_N(\mathbf{x})$ satisfy (6.12.7) can be found by minimizing

$$H_N(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \left(\boldsymbol{\lambda}, \tilde{\mathbf{A}}\right)$$

or solving (6.12.7) for $\boldsymbol{\lambda}$ as usual.

Since we want our process to be stationary, Gaussian, it is reasonable to assume that

$$A_k = \frac{1}{N+1}\sum_{0}^{N-k} X_j X_{j+k} \qquad 0 \le k \le M \tag{6.12.9-a}$$

and the left-hand side of (6.12.7) can be completed from the data as

$$\tilde{A}_k = \frac{1}{N+1}\sum_{0}^{N-k} \tilde{X}_j \tilde{X}_{j+k}. \tag{6.12.9-b}$$

To symmetrize things a bit, we shall set

$$\tilde{A}_{-k} = \tilde{A}_k, \qquad A_{-k} = A_k$$

and have a (2M+1)-dimensional Lagrange multiplier $\boldsymbol{\lambda}=(\lambda_{-M},\ldots,\lambda_{-1},\lambda_2,\ldots,\lambda_M)$ instead, in terms of which

$$P_N(\mathbf{x}) = \exp \sum_{-M}^{N} \boldsymbol{\lambda}_j A_j / Z(\boldsymbol{\lambda})$$

which after an obvious rearrangement becomes

$$P_N(\mathbf{x}) = \exp -\tfrac{1}{2}(\mathbf{x}, \Lambda\mathbf{x})/(2\pi)^{\frac{N+1}{2}} (\det \Lambda)^{1/2} \tag{6.12.10}$$

where Λ is the (N+1)x(N+1)-matrix with elements

$$\Lambda_{ij} = \begin{cases} \lambda_{i-j} & |i-j| \le M \\ 0 & |i-j| > M. \end{cases}$$

If we denote the eigenvalues of Λ by g_j and if we set

$$g(Z) = \sum_{-M}^{M} \lambda_k Z^k$$

then for N>>M, the eigenvalue g_j tends to $g(Z_j)$ with Z_j=exp(2πj/N+1)i and therefore

$$\frac{2}{N+1} \ln Z(\boldsymbol{\lambda}) = -\frac{1}{N+1}\sum_{0}^{M} \ln g(Z_j) - \ln 2\pi \to -\frac{1}{2\pi}\int_{0}^{2\pi} \ln 2\pi g(e^{i\theta}) d\theta.$$

The first identity follows from computing the entropy of $P_N(x)$ given by (6.12.10), and the limit is an exercise in Riemann integration theory. The important fact is that $g_j \to g(Z_j)$, which can be found in [6.30].

The nonlinear relationship between the R(k) and the λ_k is contained in

$$R(k) = -\frac{2}{N+1}\frac{\partial}{\partial \lambda_k} \ln Z(\boldsymbol{\lambda}) = \int_0^{2\pi} \frac{e^{ik\theta}}{g\left(e^{i\theta}\right)} d\theta \qquad -M \le k \le M \tag{6.12.11}$$

where, recall $g(Z) = \sum_{-M}^{M} \lambda_k Z^k$.

The left-hand side of (6.12.11) can be computed from the data, and once the λ_k are known for $|k| \le M$, the right-hand side of (6.12.11) determines R(k) for $|k| > M$. Therefore, the procedure outlined above can be applied.

Let us now consider two more, equivalent, inductive ways of getting the R(n). The first approach consists of computing the entropy of the joint distribution of the first N+2 variables $X_0, \ldots, X_{N+1}$ of a Gaussian process as

$$S(N+1) = \lg (2\pi e)^{\frac{N+2}{2}} [\det \Lambda(N+1)]^{\frac{1}{2}} \tag{6.12.12}$$

where $\Lambda(N+1)$ is the correlation matrix

$$\Lambda(N+1) = \begin{pmatrix} R(0) & R(1) & & R(N) & R(N+1) \\ R(1) & R(0) & \cdots & & R(N) \\ & \cdot & \cdots & & \cdot \\ & \cdot & & & \cdot \\ & & \cdot & & \cdot \\ R(N+1) & R(N) & \cdots & R(1) & R(0) \end{pmatrix}.$$

Remember that our problem is to find the R(k) for $|k| > M$. Take $N \ge M$ to begin with and assume that R(0),...,R(M) are known. It is clear that the value of R(M+1) that maximizes det Λ (M+1) is the same that maximizes S(M+1). We denote also that

$$d^2[\det \Lambda(M+1)] / \left(dR(M+1)^2\right) = -2 \det \Lambda(M-1) \le 0.$$

Since det Λ (M+1) is a quadratic function of R(M+1) with a negative derivative, then it has a unique maximum. The allowed values of R(M+1) fall between the values of γ(N+1) that make the detΛ(M+1) zero.

Choosing the R(M+1) that maximizes det(Λ(M+1)) we maximize the entropy S(M+1). It is clear then that this procedure yields the R(k) for |k|>M.

The other procedure that produces the same result is to assume that our process is an AR(M), autoregressive process of order M, satisfying

$$X(n)+b_1X(n-1)+\ldots+b_MX(n-M)=\varepsilon(n) \tag{6.12.13}$$

where ε_n is as in example (i) above.

Since $E\{\varepsilon_n X(n-k)\}=0$ for k>0, it follows that

$$\tilde{R}(k)+b_1\tilde{R}(k-1)+\ldots+b_M\tilde{R}(k-M)=0 \qquad if \quad k>M$$

where we have set $E[X(n)X(n-k)]=\tilde{R}(k)$. In other words

$$\begin{array}{ccccccc}
\tilde{R}(1) & + & b_1\tilde{R}(0) & +\ldots+ & b_M\tilde{R}(M-1) & = & 0 \\
\tilde{R}(2) & + & b_1\tilde{R}(1) & +\ldots+ & b_M\tilde{R}(M-2) & = & 0 \\
\vdots & & \vdots & & \vdots & & \vdots \\
\tilde{R}(M+1) & + & b_1\tilde{R}(M) & +\ldots+ & b_M\tilde{R}(1) & = & 0
\end{array}$$

which can be possible only if

$$\det\begin{pmatrix}
\tilde{R}(1) & \tilde{R}(0) & \ldots & \tilde{R}(M-1) \\
\tilde{R}(2) & \tilde{R}(1) & & \tilde{R}(M-2) \\
\vdots & \vdots & \ldots & \vdots \\
\tilde{R}(M+1) & \tilde{R}(M) & \ldots & \tilde{R}(1)
\end{pmatrix}=0. \tag{6.12.14}$$

Now, if we know R(0),...,R(M) we could use (6.12.14) to solve for R(M+1). But a simple computation shows that if $\tilde{R}(0)$=R(0), $\tilde{R}(1)$=R(1),...,$\tilde{R}$(M)=R(M) are known, then

$$\det\begin{pmatrix} R(1) & . & . & R(M-1) \\ & & & \\ R(M+1) & . & . & R(1) \end{pmatrix} = \frac{1}{2}\frac{d\det\Lambda(M+1)}{dR(N+1)}\Big|_{\tilde{R}(M+1)} = 0$$

which, as we saw above, is the condition determining the R(M+1) that maximizes S(N+1). Observe that the values $b_1,...,b_M$ can be obtained from the first M equations above. Thus if covariances are our only information, and AR(M) process is the candidate from process having the given covariances.

We could arrive at the same conclusion by yet another way. To wit, consider the (differential) entropy rate defined by

$$S = \lim_{N\to\infty} \frac{\ln S(N+1)}{N+1} = \frac{1}{2}\ln(2\pi e) + \frac{1}{4\pi}\int_0^{2\pi} \ln(2\pi g(\alpha))d\alpha$$

where the density $g(\alpha)$ is related to the R(n) by

$$g(\alpha) = \frac{1}{2\pi}\sum_{-\infty}^{\infty} R(n)\exp(-in\alpha)$$

and we have

Theorem. The random process $\{X_n\}$ which maximizes the differential entropy rate S, subject to the constraints

$$EX_nX_{n+k} = R(k) \qquad |k| \le M$$

is the Gauss-Markov AR(M) process satisfying the same constraints.

Proof: Let $\{Y_n\}$, $\{Z_n\}$, $\{X_n\}$ denote arbitrary, Gaussian, and Gauss-Markov processes satisfying the constraints. Then for any N≥M

$$S_N(Y_1,...,Y_N) = -\int \rho(y_1,...,y_N)\ln\rho(y_1,...,y_N)dy_1,...,dy_N$$

$$\le S_N(Z_1,...,Z_N) = S(Z_1,...,Z_M) + \sum_{M+1}^{N} S(Z_k|Z_{k-1},...,Z_1)$$

$$\le S(Z_1,...,Z_N) + \sum_{M+1}^{M} S(Z_k|Z_{k-1},...,Z_{k-M})$$

$$= S(X_1,...,X_M) + \sum_{M+1}^{M} S(X_k|X_{k-1},...,X_{k-M})$$

$$= S(X_1, \ldots, X_M) + \sum_{M+1}^{M} S(X_k | X_{k-1}, \ldots, X_1)$$

$$= S_N(X_1, \ldots, X_N).$$

Proof: We know that for any positive f,g on $\mathfrak{R}^N$, $\int f \ln(f/g)d\mathbf{x} \geq 0$. To verify the first inequality let $\rho_N(\mathbf{x})$ stand for $f(\mathbf{x})$ and put $g(\mathbf{x})=[(2\pi)^N \det C]^{-1/2}\exp\text{-}\frac{1}{2}(\mathbf{x},C\mathbf{x})$ where C is the correlation matrix of $Y_1,\ldots,Y_N$ computed with their joint density $\rho_N(\mathbf{x})$. The next one is an application of Lemma 4.15 and the one right after follows from Lemma 4.16 (actually, a simple variation on the theme thereof).

The following identity is obtained when we exchange the Gaussian families. The next to the last identity follows from the Markov property and the last step is justified by the same reasoning that implied the second step. Therefore

$$\lim \tfrac{1}{N} \ln S_N(Z) \leq \lim \tfrac{1}{N} S_N(X)$$

which almost completes the proof. It only remains to show that $\{X_n\}$ satisfying (6.12.13) is a Gauss-Markov process having the correct correlations and its spectral density is obtained inverting (6.12.3-a) for the appropriate R(n)'s.

6.13 Maxentropic solution of integral equations.

We shall consider a problem examined in [6.31]. It should be interesting to explore the mathematics of it more thoroughly.

We want to solve the following equation for x(t),

(6.13.1) $$x(t) = f(t) - \int_a^b K(t, \mathbf{s})x(s)ds$$

where, to make things easy we assume $a \leq s$, $t \leq b$ and the regularity assumptions needed on f and K will become clear below. This set up can be extended in several obvious ways.

Let $\{M_n(t): n \geq 1\}$ be a collection of linearly independent functions. Multiply both sides of (6.13.1) by $M_n(t)$, integrate over (a,b) with respect to dx (or any appropriate m(dt)) and obtain

(6.13.2) $$a_n = \int_a^b x(s)G_n(s)ds$$

where

$$a_n = \int_a^b M_n(t) f(t) dt \qquad (6.13.3-a)$$

are known coefficients and the constraints $G_n(y)$ are

$$G_n(s) = M_n(s) + \int_a^b M_n(t) K(t,s) dt. \qquad (6.13.3-b)$$

Here, we see what the minimal assumptions on f and K are. The functions f, K, M_n have to be such that the integrals above exist, that all exchanges of integrals make sense. What else?.

Under the assumption of positivity on x(t), Mead replaced the problem of solving (6.13.1) by the problem of finding a maxentropic solution $x_N(t)$ maximizing

$$-\int_a^b p(t) \ln p(t) dt$$

subject to (6.13.3-a) for n=1,...,M, and obtains the estimate

$$p_M(t) = \exp - \sum_{n=0}^{N} \lambda_i G_i(t) \qquad (6.13.4)$$

where the λ_i, i=1,...,M are such that (6.13.3-a) holds. A few examples are provided in [6.31] as well.

Here we mix approaches a bit to further illustrate maxentropic reconstruction techniques. To begin with consider the discretized version of (6.13.2)

$$a = \Delta \sum_{i=1}^{N} \Phi(i) x(i-1) \qquad (6.13.5)$$

where Δ = (b-a)/N, $\Phi_n(i) = G_n(i-1)$, n = 1,...,M. We shall consider on $\Omega \to \Re^N$ the Gaussian density $p_0(\xi) = \exp-(\xi^2/2)/[2\pi]^{N/2}$ which makes the coordinate maps $X_i : \Omega \to \Re$, $X_i(\xi) = \xi_i$ independent, centered Gaussian random variables with covariance $E_0[X_i X_j] = \Delta\delta_{ij}$.

Given an initial guess $x_0(i)$, i = 0,...,N-1, of x(i) we introduce a new auxiliary measure $P_a(d\xi)$ on Ω such that

$$dP_a / dP_0 = \exp \tfrac{1}{2} \sum_1^N x_0(i-1) X_i - \Delta \sum_1^N x_0^2(i-1)$$

with respect to P_a we have $E_a(X_j) = x_0(j-1)\Delta$. We now ask for a measure $dP(\xi)$, having density $\rho(\xi)$ with respect to $dP_a(\xi)$, yielding a maximum for $S_{P_a}(P)$ over the set of P' such that

$$E_P\left(\sum_{i=1}^{N} \Phi(i)X_i\right) = \mathbf{a}$$

Again, the candidate has a density

$$P(\xi) = \exp -\sum_i \langle \boldsymbol{\lambda}, \Phi(i)\rangle X_i(\xi)/Z(\boldsymbol{\lambda})$$

where $Z(\boldsymbol{\lambda})$ happens to be such that

$$\ln Z(\boldsymbol{\lambda}) = \tfrac{1}{2}\langle \boldsymbol{\lambda}, \mathbf{C}\boldsymbol{\lambda}\rangle - \Delta \sum_i \langle \boldsymbol{\lambda}, \Phi(i)\rangle x_0(i-1)$$

with the matrix $\mathbf{C}$ having elements

$$C_{nm} = \Delta \sum_{i=1}^{M} G_n(i-1)G_m(i-1) \qquad 1 \le m, n \le M$$

and $\boldsymbol{\lambda}$ given by

$$\boldsymbol{\lambda} = C^{-1}\left(\Delta \sum_{i=1} \Phi(i)x_0(i-1) - \mathbf{a}\right) = C^{-1}(\mathbf{a}_0 - \mathbf{a}) \tag{6.13.6}$$

where $\mathbf{a}_0$ is the $\mathbf{G}$-transform of x_0.

Again, the maxentropic reconstruction of x is given by

$$x(i-1) = \Delta E_P(X_i) = x_0(i-1) - \langle \boldsymbol{\lambda}, \Phi(i)\rangle.$$

And when N tends to infinity and i/N tends to t via an appropriate sequence, we obtain

$$x(t) = x_0(t) - \langle \boldsymbol{\lambda}, \Phi(t)\rangle$$

where $\boldsymbol{\lambda}$ is given by (6.13.6) with $\mathbf{C}$ given by

$$C_{nm} = \int_a^b G_n(s)G_m(s)ds.$$

6.14 *Maxentropic image reconstruction.*

Maxentropic Image Reconstruction methods have made it to movies and may have, perhaps, contributed a lot to popularize maximum entropy. The references [6.32]-[6.37] are to serve as starting or guide to literature. Below we present a variation on the theme, in which the set up is taken from the literature, but we apply to it a level 2 reconstruction technique.

The standard formulation of the problem consists of assuming the (compact) domain containing the picture to be divided into N cells and imagining the intensity C_n in the n-th cell to be superposition of the impinging unknown intensities x(j) according to a blurring function b_{n-j}. To make things worse, there is noise contaminating the background in an additive way. Thus C_n is actually

$$C_n = \sum_{j=1}^{N} b_{n-j}x_j + v_n \qquad n = 1, 2, ..., N \tag{6.14.1}$$

where the v_n describe the noise measured in the n-th cell. The stochastic nature of the v_n is part of the data, or of the assumed a priori knowledge. Here we shall assume the v_n to be centered, Gaussian random variables with variance σ_k.

We will assume the x_i to be the mean values of random variables X_i with respect to a distribution $dP(\xi)$ on $\mathfrak{R}^N$, and we shall consider an a priori distribution $dP_0(\xi)$ on $\mathfrak{R}^N$ with respect to which the X_i are independent and gamma distributed as

$$P_0(X_i) \in d\xi = \frac{\alpha_i}{\Gamma(\beta_i)}\xi^{\beta_i^{-1}} e^{-\alpha_i\xi} d\xi \qquad i = 1, ..., N. \tag{6.14.2}$$

To deal with the random nature of the constraint, notice that (6.14.1) implies that

$$\sum_{i=1}^{N} \frac{1}{\alpha_k}\left(C_n - \sum_{j=1}^{N} b_{n-j}E_P\chi_j\right) = \chi^2 \tag{6.14.3}$$

is a χ-distribution with N-1 degrees of freedom. Analogously to [6.32], we shall consider a reconstruction to be admissible whenever $\chi^2 \le \chi^2_{0.95}$ or some other level.

We can state our reconstruction problem thus: we seek to find the measured $P(\xi)$ realizing

$$\sup\left\{S_{P0}(P) : \chi^2(P) \le \chi^2_{0.95}\right\}.$$

Now, this is the set up dealt with in section (6.1) there we proved that the maxentropic $dP^*(\xi)$ was such that

$$dP(\xi) = (Z(\lambda))^{-1}\exp-\langle \lambda, B\xi\rangle dP_0(\xi)$$

where $B_{ij}=b_{i-j}$ denotes the blurring matrix, and $Z(\lambda)$ is given by

$$Z(\lambda) = E_{P_0}\left[e^{\langle \lambda, B\xi\rangle}\right] = E_{P_0}\left[e^{-\langle B^*\lambda, \xi\rangle}\right] = E_{P_0}\left[e^{-\langle \mu, \xi\rangle}\right] = \prod_{i=1}^{N}\left[\frac{\alpha_i}{\mu_i+\alpha_i}\right]\beta_i$$

where of course $\mu_i = \Sigma B_{ji}\lambda_j = \Sigma b_{j-i}\lambda_j$. Notice that we let the parameters α and β depend on i to allow for different "illuminations" of the picture. (By the way, perhaps more physically reasonable candidates for the a priori distribution could be used. Different situations may merit doing so.)

The value of λ that makes $dP(\xi)$ satisfy the constraints can be found by minimizing

(6.14.4) $$H(\lambda) = \ln Z(\lambda) + \langle \lambda, C\rangle + \chi_{0.95}\left(\sum_i^N \lambda_j^2\alpha_j\right)^{\frac{1}{2}}$$

as was proved in Theorem 6.1.3.

Once the right $\lambda=(\lambda_1,\ldots,\lambda_N)$ is found, the maxentropic image is provided by

$$x_i = E_P[X_i] = \left[\frac{\alpha_i}{\alpha_i+\mu_i}\right]^{\beta_i+1}.$$

Just for the fun of it, had we assumed that each X_i is uniformly distributed on [0,M], M for maximum, then, in this case

$$Z(\lambda) = \prod_{j=0}^{N}\left(\frac{1-e^{\mu_i M}}{\mu_i M}\right)$$

and, once the corresponding version of (6.14.4) had been minimized for λ, the corresponding maxentropic image is

$$x_i = E_P[X_i] = (1/\mu_i) - M/e^{\mu_i M} - 1$$

where $\mu_i=\Sigma b_{j-i}\lambda_j$ as before.

I wish I could take a peep at images reconstructed by any of these recipes.

6.15 *An application in systems analysis.*

When analyzing queuing or network systems with exponentially distributed arrival or service rates it is convenient to assume that the system is in stationary regime. This simplifies many computations. When the system is Markovian and has discrete states, the holding times for each state (the times it takes to make a transition) have exponential distribution as consequence of the Markov property.

Below we shall follow [6.40] in constructing the dynamics of a special Markov chain and that a special maxentropic distribution in its equilibrium distribution.

Let S be a discrete set of states and assume that we have defined γ_k: S→**N** be N functions such that $\gamma_k(x)$ is interpreted as the number of elements of type k in state x. The number N is the number of different types. Then

$$\sum_1^N \gamma_k(x)$$

is the number of elements in state x.

Let A(x) and B(x) be the states adjacent to x, i.e., accessible from x by adding an element to those already present in state x or removing an element from those present at x. The states in A(x) are said to be above x and those in B(x) are said to be below x. We shall also put A(k,x) for the subset of A(x) consisting of elements of type k and B(k,x) for the subset of B(x) consisting of elements of type k. Then

$$A(x) = \bigcup_1^N A(k,x), \quad B(x) = \bigcup_1^N B(k,x) .$$

For any set c we denote by |c| the cardinality of c. Let X(t) denote the state of the system at time t, we shall define the transition matrix by

$$Q_{xy} = \lim_{\Delta t \to 0} P(X(t+\Delta t) = y | X(t) = y)(\Delta t)^{-1}$$

(6.15.1)

$$Q_{xy} = \begin{cases} -\sum_{k=1}^N \{U_k |B(k,x)| + V_k |A(k,x)|\} & y = x \\ V_k & y \neq x, \quad y \in B(k,x) \\ U_k & y \neq x, \quad y \in A(k,x) \\ 0 & \textit{otherwise} \end{cases}$$

where the rate function V_k is assumed positive and, it is interpreted as the mean service or discharge time and U_k is the mean arrival time.

A probability distribution q on S is an equilibrium distribution if $\Sigma q_y Q_{xy}=0$ for all x in S. This condition may be rewritten as

$$q_x Q_{xx} = \sum_{y \neq x} q_y Q_{yx} \text{ , or}$$

$$(6.15.2) \qquad q_x \sum_{k=1}^{N} \{U_k |B(k,x)| + V_k |A(k,x)|\} = \sum_{k=1}^{N} \left(\sum_{y \in B(k,x)} V_k q_y + \sum_{y \in A(k,x)} q_y U_k \right)$$

for all $x \in S$.

To guess a candidate for q(x) we invoke the following

Lemma 6.15.3. The probability distribution on S that maximizes

$$S(q) = -\sum_{x \in S} q(x) \ln q(x)$$

subject to the constraint $\Sigma \gamma_k(x) q(x) = m_k$ (the mean number of individuals of type k) is given by

$$q(x) = \prod_{k=1}^{N} y_k^{\gamma_k(x)} / Z(\mathbf{y}).$$

Proof: Do as usual but denote $\exp-\lambda_j$ by y_j where λ_j is the usual Lagrange multiplier.

If we substitute the q(x) given by the lemma in (6.15.2) we see that the candidate for y_k is (V_k/U_k) thus we arrive at

Lemma 6.15.4. The equilibrium distribution on S satisfying (6.15.2) is

$$q(x) = \prod_{k=1}^{N} \left(\frac{V_k}{U_k} \right)^{\gamma_k(x)} / Z(\mathbf{U}, \mathbf{V}),$$

where

$$Z(\mathbf{U}, \mathbf{V}) = \sum_{x \in S} \prod_{k=1} \left(\frac{V_k}{U_k} \right)^{\gamma_k(x)}$$

Comment. If we do not want to be inconsistent, the m_k in Lemma 3 cannot be arbitrary. They should be

$$(6.15.5) \qquad m_k = \sum_{S} \gamma_k(x) \prod_{k=1}^{N} \left(\frac{V_k}{U_k} \right)^{\gamma_n(k)} / Z(\mathbf{U}, \mathbf{V})$$

which will be mean occupancy numbers of type k in equilibrium.

Comment. If we insisted in preassigning the m_k's, then the V_k are to be determined from (6.15.5) since the arrival rate U_k is out of control a priori.

For a bunch of nice applications of these ideas the reader is directed to [6.40].

6.16 Distributions with preassigned marginals and related problems.

Apparently it was E. Schrödinger in 1931 who first set up the problem of finding the p(i,j) = P(X = i, Y = j) as "similar or close" to a given $P_0(i,j)$ such that the marginals

$$f(j) = \sum_j p(i,j) \quad \text{and} \quad g(i) = \sum_j p(i,j)$$

are known before hand. Take a look at [6.41]-[6.42] for some history on this problem and for its analysis without using max-ent procedures.

Here the closeness between p(i,j) and $p_0(i,j)$ will be measured in terms of

$$K(P, P_0) = -S_{P_0}(P) = \sum_{ij} P_{ij} \ln P_{ij}/P_0(i,j) \tag{6.16.1}$$

This problem is also dealt with in [5.1] where it appears as corollary to the theorem we presented there as (5.17), and we direct the reader to [6.43] for recent work on related subjects.

Before we look into that problem, let us look at two other problems that have usually been reduced to the one we started with.

Suppose we know the distribution of income of a population, that is we split the income spectrum into M groups and we know that I_i people are in group $i \le M$. Suppose we also know that the total consumption of j-th good is G_j, for $j \le N$. If we denote by P_{ij} the number of persons of income i buying good j, then certainly

$$\sum_j P_{ij} = I_i; \quad \sum_i P_{ij} = G_j. \tag{6.16.2}$$

Actually, instead of I_i we should put $a_i I_i$, where a_i is the (known) fraction of income of group i spend on goods.

Certainly $(1-a_i)I_i$ is saved or invested and we assume it does not determine the consumption pattern.

The following two examples were reviewed in [0.3]. The first one consists of assuming P_{ij} to be the number of trips between origin i and destination j, $i \le M$, $j \le N$. The number of trips originating from i is known to be O_i and the number of trips coming into destination j is known to

be D_j. The trip pattern is useful for urban planners when deciding where to build roads, gas stations or whatever.

The second similar situation we describe consists of the problem of determining an international trade pattern P_{ij} measuring the amount of commerce between country i and country j, when the only assumed informations are the total exports of country i and the total imports of country j.

We direct the reader to [0.3] for original references and for a description of how to convert the reconstruction problem into the problem of finding a density given its marginals. Below we will do it as an application of the level 2 procedure.

If we introduce constraints $F_i(n,m)=\delta_m$; $G_j(n,m)=\delta_{jm}$, then $E_PF_i=f(i)$, $E_PG_j=g(j)$ and the candidate for maximizing $S_{P_0}(P)$ or minimizing $K(P,P_0)$ given by (6.17.1)

$$P(n,m) = e^{-\lambda_n} e^{-\mu_m} P_0(n,m)/Z(\lambda,\mu)$$

where λ_i , i=1,...,M are the Lagrange multipliers corresponding to constraint $E_P(F_i)=f_i$ and μ_j are similarly defined. Also,

$$Z(\lambda,\mu) = \sum_{n,m} e^{-\lambda_n} e^{-\mu_m} P_0(n,m)$$

If we set $\Phi_n=\exp-\lambda_n$ and $\psi_m=\exp-\mu_m$ then the Φ's and ψ's are to be determined solving

(6.16.3) $$g_i = \Phi_i \sum_j \psi_j P_0(i,j)/Z; \quad f_j = \psi_i \sum_i \Phi_i P_0(i,j)/Z$$

or equivalently, minimizing

$$H(\Phi,\psi) = \ln \sum_{nm} \Phi_n \psi_m P_0(n,m) - \sum_n f_n \ln \Phi_n - \sum_m g_m \ln \psi_m .$$

When $P_0(i,j)=P_1(i)P_2(j)$ this problem has the obvious solution $\Phi_i=f_i/P_1(i)$, $\psi_j=g_j/P_2(j)$ which yields the obvious $P(i,j)=f_ig_j$ as an answer. The set (6.16.3) can be "simplified" a bit by setting up a max-ent problem to determine $P(i|j)=P(i,j)/g(j)$ or $P(j|i)=P(i,j)/f(i)$.

To begin with note that P(i|j) satisfies the constraints

$$\sum_i p(i|j) = 1 \qquad j = 1,\ldots,N; \qquad \sum_j P(i|j)g(j) = \sum_{n,m} \delta_{in} g(m) p(n,m) = f_i$$

the partition function is

$$Z_\lambda = \sum_{nm} P_0(n,m) e^{\lambda_n g(m)}/g(m)$$

and the maxentropic conditional distribution

$$\hat{P}(n|m) = Z_1^{-1} P_0(n,m)\exp-\lambda_n g(n)/g(m).$$

Similarly, we would have obtained

$$\hat{P}(m|n) = Z_2^{-1} P_0(n,m)\exp-\mu_m f(m)/f(n)$$

and once the λ_n's or μ_m's are found we would have

$$\hat{P}(n,m) = \hat{P}(n|m)g(m) \quad or \quad \hat{P}(n,m) = \hat{P}(m|n)f(n).$$

To treat the problem as a reconstruction assume that the X_{ij} in (6.17.2) are positive and let $B_{ij}=\min(f_i,g_j)$. Thus $0\le P_{ij} \le B_{ij}$. We shall consider a collection of random variables, each taking values in $[0,B_{ij}]$, uniformly distributed there, and mutually independent relative to a law P_0.

$$P_0(X_{ij} \in A_{ij};\ i \le M, j \le N) = \prod_{ij} \frac{|A_{ij}|}{B_{ij}}$$

where $|A_{ij}|$ denotes the length of the subinterval $|A_{ij}|$ of $[0,B_{ij}]$.

We want to find a density $\rho(x_{11},\ldots,x_{MN})$ of a probability law $P<<P_0$ such that

$$S_{P_0}(P) = -\int \rho(\mathbf{x})\ln \rho(\mathbf{x})d\mathbf{x}$$

is maximum at the same time that

$$\sum_j E_P X_{ij} = f_i, \quad \sum_i E_P X_{ij} = g_j.$$

Now, the partition function is

$$Z(\boldsymbol{\lambda}, \boldsymbol{\mu}) = E_{P_0}\exp-\sum_{ij}(\lambda_i + \mu_j)X_{ij}$$

$$= \prod_{ij}\left(1 - e^{-(\lambda_i+\mu_j)B_{ij}}\right)/(\lambda_i + \mu_j)B_j.$$

The maximum entropy density is

$$p(\mathbf{x}) = \exp -\sum_{ij} (\lambda_i + \mu_j) x_{ij} / Z(\boldsymbol{\lambda}, \boldsymbol{\mu})$$

where $(\boldsymbol{\lambda},\boldsymbol{\mu})$ is obtained, as usual, by minimizing

$$H(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{ij} \ln \frac{\left(1-\exp-\left(\lambda_j+\mu_i B_{ij}\right)\right)}{\left(\lambda_i+\mu_j\right) B_{ij}} + \sum_i \lambda_i g_i + \sum_j \mu_j f_j .$$

Again, once the $(\boldsymbol{\lambda},\boldsymbol{\mu})$ is known, P_{ij} the maxentropic reconstruction is

$$\hat{P}_{ij} = E_P X_{IJ} = 1/(\lambda_i + \mu_j) - B_{ij} / \left(e^{\left(\lambda_i+\mu_j\right) B_{IJ}} - 1 \right).$$

6.17 *Maxentropic approach to the moment problem.*

The problem of reconstructing a measure defined on an interval (a,b) with $-\infty \leq a \leq b \leq \infty$ has a long history and a lot of mathematics has been denoted to it. See the nice review in [6.45]. Here we will lift some results from [6.47] and [6.48] on the convergence of maxentropic estimates, and we direct the readers' attention to [6.49]-[6.50] for related results using more functional analytic techniques and to [5.5]-[5.7] for a more probabilistic approach.

We shall denote by **P** the set of probability densities on [0,1]. It is known that

Theorem 6.17.1. Given a sequence $\{M_n\}$ of positive numbers such that $\mu_0=1$, there exists a bounded $f \in \mathbf{P}$ such that

$$\int_0^1 f(x) x^n dx = \mu_n \qquad n \geq 0$$

if and only if μ_n is completely strictly monotonic and there is a constant M such that

$$\binom{n}{m}^{n-m} \mu_m < \frac{M}{n+1} \qquad m, n = 0, 1, \ldots$$

Comments. $\{\mu_n\}$ is completely strictly monotonic if $\mu_0=1$,

$$\Delta^k \mu_n = \Sigma (-1)^m \binom{k}{m} \mu_{m+n} > 0$$

When >0 is replaced by ≥0 we obtain a completely monotonic sequence and we know that a measure (dx) exists on B[0,1] such that

$$\int_0^1 x^n d\mu(x) = \mu_n \qquad n \geq 0$$

For all about this see Widder's [6.51].

For a given f∈ **P** having moments $\{\mu_n\}$ we set

$$\mathbf{P}_n(f) = \{\textit{densities } g \textit{ on } [0,1] : \int g(x)x^k dx = \mu_k;\ k \leq n\}$$

and if we put S(f)=-∫f(x) lnf(x) dx for f∈ **P**, then a sequence $f_n \in \mathbf{P}_n(f)$ maximizing S(f) over $\mathbf{P}_n(f)$ is called a sequence of maximum entropy estimators for the moment problem.

Each $f_n(x)$ is of the form

$$\mathbf{f}(x) = \exp -\sum_0^n \lambda_k x^k,$$

where $-\lambda_0 = \ln Z(\lambda)$,

$$Z_n(\boldsymbol{\lambda}) = \int_0^1 \exp -\left(\sum_1^n \lambda_k x^k\right) dx$$

and as usual $\boldsymbol{\lambda}$ is found by minimizing

$$H_n(\boldsymbol{\lambda}) = \ln Z_n(\boldsymbol{\lambda}) + \sum_1^n \lambda_k \mu_k.$$

Comment. To be consistent we should have put $\boldsymbol{\lambda}^n = (\lambda_1^n, \ldots, \lambda_n^n)$ but that the heck.

The following first Lemma is proved in [6.47]

Lemma 6.17.2. A necessary and sufficient condition for $H_n(\boldsymbol{\lambda})$ to have an absolute minimum is that $\{\mu_n\}$ is completely strictly monotonic.

The idea of the proof is to write $\boldsymbol{\lambda}$ as $\lambda\mathbf{u}$ with $\|\mathbf{u}\|=1$ and rewrite $H(\boldsymbol{\lambda})$ as

$$H(\boldsymbol{\lambda}) = \ln \int_0^1 dx \exp \lambda F_n(x)$$

with

$$F_n(x) = \sum_1^n U_k(\mu_k = x^k)$$

and then to prove that $H(\boldsymbol{\lambda})$ grows linearly with $\|\boldsymbol{\lambda}\|$.

Comment. This proof has a drawback: it needs the existence and complete strict monotonicity of the whole sequence of moments. The reader should go to the references of [6.19] to see how to proceed when only $\mu_0,\ldots\mu_n$ are known. The next nice result in [6.47] is

Theorem 6.17.3. Let f(x) be a non-negative integrable function on [0,1] having moments $\mu_0,\ldots,\mu_1$. (Dividing f(x) by μ_0 we obtain a density.) Now let $f_n(x)$ be the maxentropic densities described above. Then for any bounded function F(x) on [0,1]

$$\lim_n \int_0^1 f_n(x)F(x)dx = \int_0^1 f(x)F(x)dx.$$

Proof: Construct the sequence

$$\psi_n(x) = \int_0^x (f(t) - f_n(t))dt$$

of total bounded variation

$$V(\psi_n(x)) = \int_0^x (f(t) + f_n(t))dt \le 2\mu_0.$$

Since the absolutely continuous (with respect to dx) measures of finite total variation are the dual of the bounded functions on [0,1] and the unit sphere is weakly-*-compact, given $\psi_n(k)$ there exists a subsequence, denote it by $\psi_n(x)$ again, and a function of finite total variation $\psi(x)$ such that $\psi_n(x)\to\psi(x)$. Since for all k

$$0 = \int_0^1 x^k d\psi_n(x) \to \int_0^1 x^k d\psi(x)$$

and since the right-hand side is zero for all k, the uniqueness of the moment problem asserts that $\psi(x) = 0$ on [0,1] which amounts to what we want.

Note that given the special form of $f_n(x)$ and that $f_n(x)$ and f(x) have the same moments

$$S(f_n) - S(f) = \int f(x)\ln(f(x)/f_n(x)) \ge 0.$$

Actually, when $n_2 > n_1$ we also have $S(f_{n1}) \ge S(f_{n2})$ (which can also be guessed from the fact that $\mathbf{P}_{n2}(f) < \mathbf{P}_{n1}(f)$). The gist of [6.48] is to prove

Theorem 6.17.4. Let f be a bounded density. Then the maxentropic sequence f_n introduced above satisfies

$$\lim_n \sup S(f_n) \le S(f).$$

From this it is clear that $S(f_n) \to S(f)$ since]

$$\lim_n \inf S(f_n) \ge S(f) \ge \lim_n \sup S(f_n).$$

In view of the results of section 6.2 and of Lemma 6.17.2, if we attempted to reconstruct f(x) by a max-ent procedure we would have to prove that (for example)

$$H(\boldsymbol{\lambda}) = \ln \int_0^1 dx \exp -\left(\sum_1^\infty \lambda_n x^n\right) + \sum_1^\infty \lambda_n \mu_n$$

achieves a minimum over some appropriate (candidates?) set of infinite dimensional $\boldsymbol{\lambda}$'s.

6.18 Maxentropic taxation policies.

The following is a variation on a subject developed by Theil in [6.52] and reviewed in [0.3]. Assume we are part of a population in which there are n_i individuals in stratum i making an income I_i. Assume that the state has to raise an amount T by taking a fraction f_i off individuals in stratum i, that is

(6.18.1) $$T = \sum_{i=1}^N f_i \mathbf{n}_i I_i$$

where N is the number of strata.

After tax, individuals making I_i are left with $R_i = I_i(1-f_i)$ and parliament for congress decides "to be fair" and preassigns limits $\alpha_1, \ldots, \alpha_N$; $0 < \alpha_i < 1$ such that

$$\alpha_1 I_1 \le \alpha_2 I_2 \le \ldots \le \alpha_n I_n \quad \textit{and} \quad \alpha_i I_i \le I_i.$$

In this fashion a maximum tax $1-\alpha_i$ is preassigned, which if applied does not make the richer poorer.

The question is how to rise T within these constraints. So how to find f_i such that (6.18.1) holds and

(6.18.2) $$0 \le f_i \le 1 - \alpha_i, \qquad i = 1, 2, \ldots, N$$

is satisfied. This is a level 2 reconstruction problem as described in section 6.3, but let us redo it from scratch here.

Define on $\Omega=[0,1-\alpha_1]x...x[0,1-\alpha_n]$ the reference measure

$$Q(dx) = \prod_{i=1}^{N} (dx_i/1-\alpha_i) = dx/A, \quad \textit{where } A = \prod_{1}^{N} (1-\alpha_i).$$

The coordinate maps $X_i(\mathbf{x})=x_i$ are independent and each is uniformly distributed in the corresponding interval. The partition function corresponding to the constraints

$$E_P \Sigma n_i X_i I_i = T$$

is given by

$$Z(\boldsymbol{\lambda}) = \prod_{i=1}^{N} \left(\frac{1-\exp-(\lambda M_i)}{\lambda M_i}\right),$$

where we set $M_i=n_iI_i(1-\alpha_i)$ for brevity. Correspondingly, since $-\partial \ln Z(\lambda)/\partial\lambda=T$ we conclude that

$$f_i^* = E_P[X_i] = (1-\alpha_i)\left((\lambda M_i)^{-1} - (\exp(\lambda M_{i-1}))^{-1}\right)$$

where as always $\boldsymbol{\lambda}$ has to be such that

$$\Sigma n_i I_i f_i^* = T.$$

For such $\boldsymbol{\lambda}$ to exist, your representatives will have to find α_i's such that

$$T \leq \sum_{1}^{N} n_i I_i (1-\alpha_i)$$

Note that for every x, $0 < 1/x-1/e^x-1 < 1$ thus $0< f_i^* /1-\alpha_i < 1$ and the constraints are satisfied.

Comment. The next three sections were written by Aldo Tagliani, and contain a summary of a line of research bearing on an important practical issue: are there a priori restrictions to be satisfied by the moment of a distribution when only a finite number of them are used in the reconstruction problem?

6.19 *The Stieltjes moment problem.*

The moment problem on the semi-finite interval $[0,+\infty)$ consists of producing a positive density p(x) such that

(6.19.1) $$\int_0^\infty x^n p(x)dx = \mu_n \qquad n \geq 0.$$

It is well known, see [6.51], that

Theorem. Given a sequence $\{\mu_n, n\geq 0\}$ of positive numbers such that $\mu_0=1$, there exists a positive density p(x) such that (6.20.1) holds if and only if the Hankel determinants

(6.19.2) $$\mu_0,\ \mu_1,\ \begin{vmatrix}\mu_0 & \mu_1\\ \mu_1 & \mu_2\end{vmatrix},\ \begin{vmatrix}\mu_1 & \mu_2\\ \mu_2 & \mu_3\end{vmatrix},\ \begin{vmatrix}\mu_0 & \mu_1 & \mu_2\\ \mu_1 & \mu_2 & \mu_3\\ \mu_2 & \mu_3 & \mu_4\end{vmatrix},\ \begin{vmatrix}\mu_1 & \mu_2 & \mu_3\\ \mu_2 & \mu_3 & \mu_4\\ \mu_3 & \mu_4 & \mu_5\end{vmatrix}$$

are positive.

If we consider the problem of finding a positive p(x) on $[0,+\infty)$ such that (6.19.1) holds **only** for n=0,1,...,N, then the standard maxentropic reconstruction procedure suggests that we look for $\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_N$ such that

(6.19.3) $$P_N(x) = \exp-\left(\sum_0^N \hat{\lambda}_j x^j\right)$$

satisfies

(6.19.4) $$\int_0^\infty x^l P_N(x)dx = \mu_n, \qquad n = 0, 1, \ldots, N.$$

We shall consider the change of variables

$$e^{-\hat{\lambda}_0} = e^{-\lambda_0}/\mu_1$$

$$\hat{\lambda}_i/\mu_1^i = \lambda_i$$

and the corresponding change of variable in x, we rewrite (6.20.2) and (6.20.3) as

(6.19.3)′ $$P_N(x) = \exp-\Sigma\, \lambda_i x^i$$

(6.19.4)′ $$\int_0^\infty x^l P_N(x)dx = \mu_i/\mu_1^i = \mu_i$$

the μ_i being normalized moment. Introducing the standard a dimensional statistical parameters (variation, skewness, kurtosis,...) $\gamma,\nu,\kappa,\ldots$ we can rewrite the μ_i as

$$M_0 = 1,\ M_1 = 1,\ M_2 = 1+\gamma^2,\ M_3 = 1+3\gamma^2+\nu\gamma^3,\ M_4 = 1+6\gamma^2+4\nu\gamma^3+\kappa\gamma^4,\ \ldots\ .$$

Integrating (6.19.4)' by parts, the following identities are obtained

(6.19.5) $$mM_{m-1} = \sum_{j=1}^{N} j\lambda_j M_{m+j-1} \qquad m \geq 1$$

which can be used to obtain the moments M_j for j>N as functions of the first N moments and the Lagrange multipliers λ_j .

We leave for the reader to work out the details for the case N=1. The case N=2 was dealt with by Dowson and Wragg in [6.53]-[6.59] based on previous work by Barrow-Cohen [6.55]. They introduced the Mill's function B(x) defined by

(6.19.6) $$1/B(x) = \exp(x^2/2)\int_x^{\infty} \exp-(t^2/2)dt$$

such that $B'=(B\text{-}x)$ and $B''=B'(2B\text{-}x)\text{-}1$. They prove that there is an x such that

$$\gamma^2 = 1 - BB''/(B')^2,\ \lambda_2 = \tfrac{1}{2}(B'/B)^2,\ \lambda_1 = x\sqrt{2\lambda_2},\ \lambda_0 = -\ln\left(B\sqrt{2\lambda_2}\right)$$

Barrow and Cohen proved that

$$M_2 = \frac{1-x(1/B-x)}{(1/B-x)^2}$$

has to satisfy

(6.19.7) $$1 < M_2 \leq 2 \quad or \quad 0 < \gamma \leq 1.$$

In other words, when N=2 the positivity of the Hankel determinants (6.19.1) is only a necessary condition for the existence of $P_N(x)$. The condition (6.19.7) must be imposed to obtain the existence of $P_N(x)$ satisfying (6.19.3)' and (6.19.4)'.

The cases N=3 and N=4 were discussed by Tagliani in [6.56] extending previous work in [6.38]-[6.39]. The computations and arguments are intrincate. The case N=3 is briefly summarized and the results for N=4 have just been presented.

The basic philosophy consist of transforming (6.19.4)' or (6.19.5) into a system of differential equations, by varying continuously one of the moments, say μ_j, and keeping the others constant. The dependence of $\lambda_0,\lambda_1,\ldots,\lambda_N$ in terms of μ_j is studied and of particular interest is the range of μ_j making λ_N positive.

From now, on we shall write κ_i, $i \geq 1$, for the standard statistical coefficients γ,ν,κ, etc. and we shall denote by $D(\kappa_j,N)$ the domain of acceptable values of the coefficient κ_j when N moments are preassigned. It was proved in [6.56] that

$$D(\kappa_j,N) \subseteq D(\kappa_j N+1). \tag{6.19.8}$$

Also the following inequalities hold

$$\kappa_{2j+1}\kappa_2^{2j+1} = \int_0^\infty (x-1)^{2j+1} P_N(x)dx < \kappa_{2j+3}\kappa_2^{2j+3} + \int_1^2 (x-1)^{2j+1}(2x-x^2)P_N(x)dx$$

$$\kappa_{2j}\kappa_2^{2j} = \int_0^\infty (x-1)^{2j} P_N(x)dx < \kappa_{2j+2}\kappa_2^{2j+2} + \int_0^2 (x-1)^{2j}(2x-x^2)P_N(x)dx .$$

From these, we see that if we let $\kappa_2,\kappa_{2j},\kappa_{2j-1}$ become arbitrarily large, then

$$\kappa_{2j+1} \to \infty \quad \textit{and} \quad \kappa_{2j+2} \to \infty \qquad j \geq 2 \tag{6.19.9}$$

which are obviously interpreted as saying that if for a particular value of N, say N^*, none of the coefficients $\kappa_2,\ldots,\kappa_{N^*}$, admits an upper bound, then for any $N>N^*$, the coefficients $\kappa_2,\ldots,\kappa_N$ are bounded as well.

In other words, if for given $\mu_0,\ldots,\mu_{N^*}$, a $P_{N^*}(x)$ satisfying (6.19.3)' and (6.19.4)' exists, then $P_N(x)$ exists for $N>N^*$ and (6.19.8) represents a necessary and sufficient condition for the existence of a maximum entropy reconstruction of a density with the first N moments preassigned.

Let us now look at the details for N=3. In this case γ and ν are preassigned and we want to determine $D(\gamma,3)$ and $D(\nu,3)$.

It can be seen that $D(\gamma,3)=(0,\infty)$ but $(\gamma,3)$ depends on the value for γ. From Schwartz's inequality written as

$$\mu_{j-1}^2 \leq \mu_j \mu_{j-2} \qquad j \geq 2$$

it follows that whenever $\gamma \leq 1$

$$\gamma - \tfrac{1}{\gamma} < \nu \leq \nu_{max}$$

where ν_{max} is a constant depending on ν and λ_2, and whenever $\gamma > 1$ we obtain

$$\gamma - \tfrac{1}{\gamma} < \gamma < +\infty$$

All for which can be summarized as

$$D(\gamma, 3) = (0, +\infty)$$

(6.19.10)

$$D(\nu, 3) = (\gamma - 1/\gamma, \nu_{max}) \quad \textit{for} \quad \gamma \leq 1$$

$$D(\nu, 3) = (\gamma - 1/\gamma, +\infty) \quad \textit{for} \quad \gamma > 1$$

in which the final comment that the positivity of the Hankel determinants (6.19.2) represent a necessary and sufficient condition for the existence of $P_3(x)$ if $\gamma > 1$ but only a necessary condition for $\gamma \leq 1$.

Let us take a brief look at the case N=4 where γ, ν, κ are preassigned. From (6.19.8) we obtain that

$$D(\gamma, 4) = (0, \infty)$$

(6.19.11)

$$D(\nu, 4) = \left(\gamma - \tfrac{1}{\gamma}, +\infty\right) \quad \textit{when} \quad \gamma > 1$$

but $D(\nu,4)$ for $\gamma \leq 1$ and $D(\kappa,4)$ are yet to be determined. After a quite cumbersome analysis one arrives at

$$D(\nu, 4) = (\gamma - 1/\gamma, +\infty) \quad \gamma \leq 1$$

(6.19.12)

$$D(\kappa, 4) = (1 + \nu^2, +\infty)$$

where the quantity $1+\nu^2$ is related to the positivity of the Hankel determinant

$$\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix}$$

These results solve completely the case N=4. No upper bound exists for the coefficients γ,ν,κ, or the positivity of the Hankel determinants (6.19.2) represents a necessary and sufficient condition for the existence of $P_4(x)$.

By taking into account (6.19.8) and (6.19.9) the same is valid for $N\geq 4$. We summarize the discourse of this section in

Theorem. Given a sequence $\mu_0,\mu_1,\ldots,\mu_N$, $N\geq 4$, of positive numbers, a necessary and sufficient condition for the existence of $P_N(x)$ is the positivity of the Hankel determinants (6.19.2). For N=2 or N=3, the positivity of the determinants is only a necessary condition and auxiliary constraints have to be introduced for the existence of $P_N(x)$.

6.20 *The Hamburger moment problem.*

Here we present a brief description of the results in [6.57], and we will be concerned with finding a density $P_N(x)$ on $(-\infty,\infty)$ whose first moments $\mu_0,\ldots,\mu_N$ are assigned. That is we want $P_N(x)$ such that

$$P_N(x) = \exp-\sum_0^N \lambda_j x^j$$

(6.20.1)

$$\int_{-\infty}^{\infty} x^n P_N(x)dx = \mu_n \qquad n = 0, 1, \ldots, N.$$

To begin with let us recall the result in [6.51] asserting the existence of solution to the moments problem.

Theorem. Given a sequence $\{\mu_n: n\geq 0\}$ of numbers such that $\mu_0=1$, there will exist a positive measurable density p(x) on $(-\infty,\infty)$ such that

$$\int x^n p(x)dx = \mu_n \qquad n \geq 0$$

if and only if the Hankel determinants

(6.20.2)

$$\mu_0,\ \begin{vmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{vmatrix},\ \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix},\ \ldots$$

are strictly positive.

To begin with our problem let us consider first the symmetric case, i.e. , let us assume we know $P_N(x)=P_N(-x)$. This will be possible if and only if N=2M and $\mu_0=\ldots=\mu_{2M-1}=0$.

As in the Stieltjes case, the even moments can be expressed in terms of adimensional coefficients $\gamma,\nu,\kappa,...$(also labeled as κ_j below) such that

$$\mu_0 = 1,\ \mu_2 = 1,\ \mu_4 = 1+\gamma^2,\ \mu_6 = 1+3\gamma^2+\nu\gamma^3,\ \mu_8 = 1+6\gamma^2+4\nu\gamma^3+\kappa\gamma^4$$

And as above we want to determine what restrictions, besides the positivity of (6.20.2) does the finite size of the moment problem impose.

Begin with N=2. This case was analyzed by Powles and Carranza in [6.58]. They obtain after transforming (6.20.2) into a differential equation solved with the aid of weber functions that

$$1 < \frac{\mu_0\mu_4}{\mu_2^2} \le 3$$

or equivalently

(6.20.3) $$D < \gamma \le \sqrt{2}$$

and from a relation similar to (6.19.5) the coefficient

(6.20.4) $$\nu_{max} = \gamma - 1/\gamma + 2 - \gamma^2/4\lambda_4\gamma^3$$

to be used below is deduced.

Also, as above one can prove with some effort that

(6.20.5) $$D(\kappa_J, N) \subseteq D(\kappa_J, N+1)$$

as in the Stieltjes problem, together with

$$\kappa_{2j+1}\kappa_{2j}^{2j+1} = \int_0^\infty (x^2-1)^{2j+1} p_N(x)dx < \kappa_{2j+3}\kappa_2^{2j+3} + \int_1^{\sqrt{2}} (x^2-1)^{2j+1}(2x^2-x^4)p_N(x)dx$$

$$\kappa_{2j}\kappa_2^{2j} = \int_0^\infty (x^2-1)^{2j} p_N(x)dx < \kappa_{2j+2}\kappa_2^{2j+2} + \int_0^{\sqrt{2}} (x^2-1)^{2j}(2x^2-x^4)p_N(x)dx$$

and as before , when either $\kappa_2,...,\kappa_{2j-1},\kappa_{2j}$ are unbounded, then so are κ_{2j+1}, κ_{2j+2} and so on.

The results for N=3 and N=4 for the symmetric case are similar to the corresponding results for the Stieltjes case and are obtained by Tagliani in [6.39].

Let us now look at the general, non-symmetric, case. As usual begin with the case N=2. Here we have

(6.20.6) $$D(\gamma, 2) = (0, +\infty)$$

and hence

(6.20.7) $$D(\gamma, N) = (0, +\infty) \qquad n > 2.$$

Now consider N=4 when γ,ν,κ are preassigned. Again (6.21.1) or its equivalent (6.20.5) is transformed into a system of differential equations. Examination of this solution yields

(6.20.8)
$$D(\nu, 4) = (-\infty, +\infty)$$
$$D(\kappa, 4) = (1 + \nu^2, +\infty)$$

and no upper bound exists for γ,ν,κ. As above we sum up with

Theorem. The conditions for the existence of a maxentropic solution on the full Hamburger problem, plus the additional constraints that have to be imposed when symmetry is required.

6.21 *Applications to data analysis.*

In many cases when applying statistical modeling in applied sciences, the analytical representation of probability distributions is essentially empirical.

Experience suggests that whenever a particular mathematical distribution gives a good fit to experimental data under limited information, it is also reasonable to base the estimation of probabilities on the maximum entropy method. See for example the work by Siddall and Diab [6.59]. From their work one would conclude that almost all well known analytical distributions can be accurately reconstructed from the convergence of their first four or five moments.

In other words, the probabilistic nature of the random variable can be reasonably well captured by these moments. Or, the question whether the first few moments are a good representation of information supplied by the sample data, appears to have a positive answer.

In general, however, the result of the testing program is a set of N measured values $\{x_1,...,x_N\}$ rather than a set of population moments. From these one could compute N independent sample moments by

$$\hat{\mu}_k = \frac{1}{N}\sum_{j=1}^{N}(x_j)^k \qquad k = 1, 2, ..., N_S. \tag{6.21.1}$$

And in applications it is frequently assumed that the unknown population moments μ_k can be replaced by the known sample moments $\hat{\mu}_k$.

By replacing μ_k by $\hat{\mu}_k$ it would appear that the entropy, which is presumable a measure of information, does not depend on the number of tests used to compute the sample moments. Besides that, it is not clear how many moments should be included as constraints in the maximum entropy formalism when the available information is a sample of N measured values.

Such questions have been raised by Baker in [6.60] in a vivid paper and his approach is applied to the case of a random variable takes values in $D \subset \Re$ (typically D=[0,1], as in the Hausdorff moments case).

Making use of Kullback's relative information, we would obtain $p^*(x)$ as

$$\inf\left\{K(p,p_0) \middle| \int_D p(x)x^k dx = \hat{\mu}_k;\ k = 0, .., N\right\}$$

that is $p^*(x)=p_0(x)\exp-\left(\sum_0^N \lambda_k x^k\right)$, with the λ_k's determined as usual.

On the other hand, Akaike's estimation procedure determines both the number of parameters N and their values in such a way that the resting probability reflects property the information contained in a given sample.

Baker introduces a "differential entropy" $\Delta H(\lambda_0, ..., \lambda_m, M)$ depending on the number M of moments $\hat{\mu}_k$ to be considered and on the Lagrange multipliers $\lambda_0,...,\lambda_\mu$ by

$$\Delta H(\lambda_0, ..., \lambda_\mu, M) = \frac{M}{N} - \sum_0^M \lambda_k \hat{\mu}_k. \tag{6.21.2}$$

With this

a) One solves for the $\{\lambda_0,...,\lambda_\mu\}$ for different M as usual.

b) The "best" number of moments corresponds to that value of M making (6.21.2) smallest.

REFERENCES

[6.1] Gamboa, F. "Minimization de L'information de Kullback et Maximization de L'entropie sous une contraint quadratique". C.R.A.S. Paris t 306, Serie 1, pp. 425-427, 1988.

[6.2] Borwein, J.M. "On the failure of the maximum entropy reconstruction for Fredholm equations and other infinite systems". To appear.

[6.3] Rockafellar, R.T. "Convex Analysis". Princeton Univ. Press, Princeton, 1970.

[6.4] Schneider, M.H. "Matrix scaling, Entropy minimization and conjugate duality". Lin. Alg. Appl. 114/115, pp. 785-813, 1989.

[6.5] Charnes, A. and Cooper, W.W. "Constrained Kullback-Leibler estimation". Accad. Naz. Dei Lincei, Serie VIII, LVIII, Fasc 4. pp. 568-576, 1975.

[6.6] Gamboa, F. and Gzyl, H. "Linear Programming with the Maximum Entropy" Mathl. Comp Modelling. 13, pp. 49-52, 1990.

[6.7] Gzyl, H. "The max-ent approach to linear programming with quadratic errors". ibid. 15, pp. 43-45, 1991.

[6.8] Elfving, T. "On some methods for entropy maximization and matrix scaling". Lin. Alg. Appl. 34, pp. 321-339, 1980.

[6.9] Erickson, J. "A note on the solution of large sparse maximum entropy problems with linear equality constraints". Math. Prog. 18, pp. 146-154, 1980.

[6.10] Erlander, S. "Entropy in linear programs". Math. Prog. 21, pp. 137-151, 1981.

[6.11] Censor, Y. "On linearly constrained entropy maximization". Lin. Alg. Appl. 80, pp. 191-195, 1986.

[6.12] Cinlar, E. "Introduction to stochastic processes" Prentice Hall. Englewood Cliffs, New Jersey, 1976.

[6.13] Kelly, F. "Reversibility and Stochastic Networks". John Wiley, New York, 1979.

[6.14] Doyle, P.G. and Snell, J.L. "Random Walks and Electric Networks" Carus Monograph N^0 22, Math. Assoc. Am, 1984.

[6.15] Koopman, B.O. "Entropy increase and symmetry" In "The Maximum Entropy Formalism". Levine, R and Tribus, M. Eds, M.I.T. Press, Cambridge, 1979.

[6.16] Flores de Chela, D. "Generalized inverses in normed linear spaces" Lin. Alg. Appl. 26, pp. 243-263, 1977.

[6.17] Levine, R.D. "An information theoretical approach to linear inversion problems". J. Phys. A. Math. Gen. 13, pp. 91-108, 1980.

[6.18] Bard, Y. "Estimation of state probabilities using the maximum entropy principle". IBM J Res. Develop. 24, pp. 563-569, 1980.

[6.19] Ulrich, T., Bassrei, A. and Lane, M. "Minimum relative entropy inversion of 1D data with applications". Geoph. Prosp. 38, pp. 465-487, 1990.

[6.20] Rietsch, E. "The maximum entropy approach to the inversion of 1d seismograms". Geoph. Prosp. 36, pp. 365-382, 1988.

[6.21] Aichelin, J. and Huefner, J. "Fragmentation reactions on nuclei: condensation of vapor, shattering of glass". Phys. Lett. 136 B, pp. 15-17, 1984.

[6.22] Varadham, S.R.S. "Diffusion problems and partial differential equations" Tata Lect. Notes. N^0 64, Springler-Verlag, Berlin, 1980.

[6.23] Gassiat, E. "Probleme sommatoire par maximum d'entropie" C.R.A.S. Paris. t 303. Serie I, pp. 675-680, 1986.

[6.24] Landau, H. J. "Maximum entropy and the moment problem". Bull. Am. Math. Soc. l. 16, pp. 47-77, 1987.

[6.25] Choi, B.S and Cover, T. M. "An information theoretic proof of Burg's maximum entropy spectrum". Proc. IEEE. 72, pp. 1094-1096, 1984.

[6.26] Grandel, J., Hamrud, H. and Toll, P. "A remark on the correspondence between the max-entropy method and the autoregressive model" IEEE. Trans. Inf. Th. IT-26, pp. 750-751, 1980.

[6.27] Van den Bos, A. "Alternative interpretation of maximum entropy spectral analysis". IEEE. Trans. Inf. Th. IT-17, pp. 493-494, 1971.

[6.28] Lin, Dh. and Wong, E.K. "A survey on the maximum entropy method and parameter spectral estimation". Phys. Reports. North Holland, 193, pp. 41-135, 1990.

[6.29] Karlin, Sand Taylor, H.M. "A first course in stochastic processes". 2nd. Ed. Acad. Press, New York, 1975.

[6.30] Grenander, V and Szegö, G. "Toeplitz forms and their applications". Univ. Calif. Press, Berkeley, 1958.

[6.31] Mead, L. R. "Approximate solution of Fredholm integral equations by the maximum entropy method? Jour. Math. Phys. 27, pp. 2903-2907, 1986.

[6.32] Bryan, L. K. and Skilling, J. "Deconvolution by maximum entropy, as illustrated by application to the jet of M87". Mon. Not. R. Art. Soc. 191, pp. 69-79, 1980.

[6.33] Birch, S. F., Gull, S. F. and Skilling, J. "Image restoration by a powerful maximum entropy method" Comp. Vis. Graph and Im. Proc. 23, pp. 113-128, 1983.

[6.34] Wenecke, S. J. and D'Addario, L. R. "Maximum entropy image reconstruction". IEEE. C. 26, pp. 351-369, 1977.

[6.35] Geman, D. and Geman, S. "Bayesian image analysis". Nato ASI Series, F20, Disord. Syst.and Biol. Organize, Springer-Verlag, Berlin, 1986.

[6.36] Zuang, X., Ostelvold, E. and Haralick, R. M. "A differential equation approach to maximum entropy image reconstruction". IEEE AS.5.P 35, pp. 208-218, 1987.

[6.37] Elfving, T. "An algorithm for maximum entropy image reconstruction from noisy data". Math. Comp. Modeling, 12, pp. 729-745, 1989.

[6.38] Rosenblueth, E. Karmesh and Hong, H. P. "Maximum entropy and discretization of probability distributions". Probab. Engin. Mech. 2, pp. 58-63, 1987.

[6.39] Tagliani, A. "On the existence of maximum entropy distributions with four or more assigned moments". Probab. Engin. Mech.

[6.40] Ferdinand, A. E. "A statistical mechanical approach to systems analysis". I.B.M. Jour. Res. Dev. Sept., pp. 539-547, 1970.

[6.41] Jamison, B. "A Martin boundary interpretation of the maximum entropy method". Zeit. f. Warsch. 30, pp. 265-272, 1974.

[6.42] Jamison, B. "Reciprocal processes". Zeit f. Warsch. 30, pp. 65-86, 1974.

[6.43] Aebi, R. and Nagasawa, M. "Large deviations and the propagation of chaos for Schroedinger processes". Zeit. f. Warsch. 94, pp. 53-68, 1992.

[6.44] Arnold, G. S. and Kinsey, J. L. "Information theory for marginal distributions applications to energy disposal in an exothermic reaction". Jour. Chem. Phys. 67, pp. 3530-3532, 1977.

[6.45] Rebick, C., Levine, R. D. and Bernstein, R. B. "Energy requirements and energy disposal" Jour. Chem. Phys. 60, pp. 4977-4989, 1974.

[6.46] Landau, H. J. "Maximum entropy and the moment problem". Bull. Am. Math. Soc. 16, pp. 47-71, 1987.

[6.47] Mead, L. R. and Papanicolau, N. "Maximum entropy in the problem of moments". Jour. Math. Phys. 25, pp. 2404-2417, 1984.

[6.48] Forte, B., Hughes, N. and Pales, Z. "Maximum entropy and the problem of moments" Rendiconti di Matematica, Serie VII, 9, pp. 689-699, 1989.

[6.49] Borwein, J. M. and Lewis, A. S. "Convergence of best entropy estimates" SIAM Jour. Optim. 1, pp. 191-205, 1991.

[6.50] Lewis, A. S. "The convergence of entropic estimates for moment problems". Workshop on Functional Analysis/Optimization. Fitzpatrick S. and Giles, J. Eds, Centre for Mathem. Analysis, Australian Nat. Univ. Canberra, pp. 100-115, 1988.

[6.51] Widder, D. V. "The Laplace transform". Princeton Univ. Press, Princeton, 1946.

[6.52] Theil, H. "Economics and information theory". North Holland, Amsterdam, 1967.

[6.53] Dowson, D. C. and Wragg, A. "Maximum entropy distributions having prescribed first and second moments". IEEE IT-16, pp. 689-693, 1973.

[6.54] Wragg , A. and Dowson, D.C. "Fitting continuous probability density functions over [0,∞) using information theory ideas". IEEE IT-16, pp. 220-230, 1970.

[6.55] Barrow, D. F. and Cohen, A. C. "On some functions involving Mill's ratio". Ann. Math. Statistics, 25, pp. 405-408, 1954.

[6.56] Tagliani, A. "On the application of maximum entropy to the problem of moments". Jour. Math. Phys. 34, pp. 326-337, 1993.

[6.57] Tagliani, A. "Maximum entropy the Hamburger moments problem". Submitted to Jour. Math. Phys. 1993.

[6.58] Powles, J. G. and Carranza, B. "An information theory of nuclear magnetic resonance". In Magnetic Resonance. Coogan C. K. Ed., pp. 133-161, 1970.

[6.59] Siddall, J. N. and Diab, Y. "The use in probabilistic design of probability curves generated by maximizing the Shannon entropy function constrained by moments". Jour. of Engin. for Industry. A. S. M. E. 97 , pp. 843-852, 1975.

[6.60] Baker, R. "Probability estimation and information principles" Structural Stability. 9, pp. 97-116, 1990.

Chapter 7

ENTROPY AND LARGE DEVIATIONS

The following is taken almost literally from [7.1]. It comprises the very basic results in the theory of large deviations. In that reference you will find quite a lot about the subject and its applications to statistical mechanics. Also, check with [7.2] for more.

We shall consider a probability space $(\Omega,\mathcal{T},\mathrm{P})$ on which we have defined a family of independent, identically distributed random variables $\{X_n : n \geq 1\}$ taking values on a finite set $S=\{x_1,\ldots,x_N\}$.

It is clear that any measure ρ on S (equipped with the σ-algebra P(S) the class of all subsets of S) can be written as

$$\rho(A) = \sum_1^N \rho_i \delta_{x_i}(A)$$

where $\delta_{x_i}(A)$ is 1 or 0 depending on whether x_i is in A or not. We shall set

$$S_n/n = \sum_1^n X_k/n, \quad n \geq 1$$

and define the empirical frequencies $L_{n,i}$ by

$$L_{n,i}(\omega) = \frac{1}{n} \sum_{j=1}^n \delta_{X_j(\omega)}(\{x_i\}) \tag{7.1}$$

where the $\omega \in \Omega$ is written to emphasize that the $L_{n,i}$ are random variables which count, for each realization $X_n(\omega)$ of the process, the frequency with which the sequence $X_n(\omega)$ takes the value x_i.

$$S_n/n = \sum_{i=1}^n x_i L_{n,i} \tag{7.2}$$

The $L_{n,i}$ also define the empirical measures

$$L_n(A) = \sum_{i \in \wedge} L_{n,i} = \frac{1}{n} \sum_{j=1}^n \delta_{X_j}(A).$$

If for the law P on $(\Omega,\mathcal{F})$, $P(X_k=i)= \rho_i$ then

$$EX_k = \Sigma x_i \rho_i = m_\rho$$

and the summands in $L_{n,1}$ are independent, identically distributed random variables, taking values in the set of all probability measures on S.

According to the law of large numbers, for any $\varepsilon>0$ the following limits hold true

$$\lim_{n\to\infty} P\{|S_n/n - m_\rho| \geq \varepsilon\} = 0$$

$$\lim_{n\to\infty} P\left\{\max_{1\leq i\leq N} |L_{n,i} - \rho_i| \geq \varepsilon\right\} = 0$$

where the vector $(\rho_1, ..., \rho_N)$ is the limit of the random vector $(L_{n,1},...L_{n,N})$.

To begin with we shall consider the fluctuations of $L_{n,1}$ about their means when S={0,1}, or if you will, for the head and tail game. All the basic results and techniques already appear in this case, in which counting is simpler.

For the time being set S={0,1}, $\rho=1/2(\delta_0 +\delta_1)$, $\rho_0 = \rho_1 = 1/2$.$L_{n,0}= 1-S_n/n$, $L_{n,1}= S_n/n$. Therefore, $|L_{n,0} - \rho_0| = |S_n/n - m_\rho|$ From this

$$P(|S_n/n - m_\rho| \geq \varepsilon) = P\left(\max_{i=0,1} |L_{n,i} - \rho_i| \geq \varepsilon\right)$$

Let $Q_n^{(1)}$ denote the distribution of S_n/n as an $\mathfrak{R}$-valued variable and set $A = \{t \in \mathfrak{R} : |t - m_\rho| \geq \varepsilon\}$ with $0<\varepsilon<1/2$.

Certainly, $A\cap[0,1] \neq \varnothing$ and $Q_n^{(1)}(A) = P\{|S_n/n - m_\rho| \geq \varepsilon\}$ is positive for large enough n. Since $m_\delta \notin A, Q_n^{(1)}(A) \to 0$ *as* $n \to \infty$.

Let us define

$$I^{(1)}(z) = \begin{cases} z\ln 2z + (1-z)\ln 2(1-z) & z \in [0,1] \\ \infty & \textit{otherwise} \end{cases} \tag{7.3}$$

with 0ln0=0 as usual. Note that $I^{(1)}(z)$ is symmetric about 1/2 and has its minimum there. The following result relates the decay of $Q_n^{(1)}(A)$ to $I^{(1)}(z)$.

Theorem 7.4. With the notations introduced above

$$\lim_{n\to\infty} \tfrac{1}{n}\ln Q_n^{(1)}(A) = \lim_{n\to\infty} \tfrac{1}{n}\ln P\left\{\left|\tfrac{S_n}{n} - m_\rho\right| \geq \varepsilon\right\} = -\min_{z\in A} I^{(1)}(z)$$

Comment. Since A is a closed set and $m_\rho \notin A$, $\min_{z\in A} I^{(1)}(z)$ is strictly larger than $I^{(1)}(m_\rho)=0$. Therefore $Q_n^{(1)}(A)$ tends to zero exponentially fast as $n \to \infty$.

Proof: S_n ranges over the set $\{0,1,...,n\}$ and

$$P(S_n=k) = \binom{n}{k}/2^n.$$

If we put $A_n=\{k:|\ k/n-1/2|\geq \varepsilon\}$ we have

$$Q_n^{(1)}(A) = \sum_{k\in A_n} P(S_n=k) = \sum_{k\in A_n} \binom{n}{k}/2^n$$

Since there are (n+1) terms in the sum

$$\max_{k\in A_n} \binom{n}{k}/2^n \leq Q_n^{(1)}(A) \leq (n+1) \max_{k\in A_n} \binom{n}{k}/2^n$$

and since lnz is an increasing function, we have

$$\max_{k\in A_n} \left\{ \tfrac{1}{n} \ln \frac{\binom{n}{k}}{2^n} \right\} \leq \tfrac{1}{n} \ln Q_n^{(1)}(A) \leq \tfrac{\ln(1+n)}{n} + \max_{k\in A_n} \left\{ \tfrac{1}{n} \ln \frac{\binom{n}{k}}{2^n} \right\}$$

To conclude the proof we need the following.

Lemma 7.5. The following estimate is uniform in $k \leq n$.

$$\tfrac{1}{n} \ln \binom{n}{k} = -\tfrac{k}{n} \ln \tfrac{k}{n} - \left(1-\tfrac{k}{n}\right) \ln \left(1-\tfrac{k}{n}\right) + O\left(\tfrac{\lg n}{n}\right) \quad as\ n \to \infty.$$

Proof: For k=0 or k=n it is obviously true. From Stirting's theorem, ln n!=n ln n-n+O(ln n). Therefore

$$\tfrac{1}{n} \ln \binom{n}{k} = \ln n - \tfrac{k}{n} \ln k - \tfrac{n-k}{n} \ln (n-k) + \tfrac{1}{n} O(\ln n).$$

Since $\ln n = -\tfrac{k}{n} \ln \tfrac{1}{n} - \tfrac{(n-k)}{n} \ln \tfrac{1}{n}$ with which we obtain the estimate

$$\tfrac{1}{n}\ln\binom{n}{k}\tfrac{1}{2^n} = \ln\tfrac{1}{2} - \tfrac{k}{n}\ln\tfrac{k}{n} - \left(1-\tfrac{k}{n}\right)\ln\left(1-\tfrac{k}{n}\right) + O\left(\tfrac{\ln n}{n}\right)$$

$$= -I^{(1)}\left(\tfrac{k}{n}\right) + O\left(\tfrac{\ln n}{n}\right).$$

Back to the theorem. As both ln n/n and ln(n+1)/n are O(ln n/n) we have

$$\tfrac{1}{n}\ln Q_n^{(1)}\left(\left\{\tfrac{k}{n}\right\}\right) = -I^{(1)}\left(\tfrac{k}{n}\right) + O\left(\tfrac{\ln n}{n}\right)$$

from which

$$\lim \tfrac{1}{n}\ln Q_n^{(1)}(A) = \lim_{n\to\infty} \max_{k\in A_n}\left(-I^{(1)}\left(\tfrac{k}{n}\right)\right) = -\lim_{n\to\infty} \min_{k\in A_n} I^{(1)}\left(\tfrac{k}{n}\right).$$

We are almost there. Note to finish that

$$\left\{z \in [0,1] : z = \tfrac{k}{n} \quad \textit{for some} \quad k \quad \textit{in} \quad A_n\right\} \subset A \cap [0,1]$$

and since $I^{(1)}(z) = \infty$ whenever z is not in A_n, then

$$\lim \tfrac{1}{n}\ln Q_n^{(1)}(A) = -\min_{z\in A\cap[0,1]} I^{(1)}(z) = -\min_{z\in A} I^{(1)}(z).$$

The missing steps are contained in

Lemma 7.6. Let $A \subset \Re$ be a compact set, $f: A \to \Re$ a continuous function and $A_n \subset A$ closed sets such that for any $a \in A$ there exists a sequence $a_n \in A_n$ and $a_n \to a$. Then

$$\lim_{n\to\infty} \min_{A_n} f(x) = \min_{A} f(a).$$

Proof: Let a be such that

$$\min_{A} f(x) = f(a)$$

and a_n be as above. Then

(7.7) $$f(a) = \lim f(a_n) \geq \lim_{n} \min_{A_n} f(x) \geq \min_{A} f(x).$$

Let us now consider the general case: $S=\{x_1,\ldots x_n\}$, (and let the x_i be real numbers such that $x_1<\ldots<x_n$). As mentioned above

$$\mathbf{P}\,(S)= \{(p_1,\ldots,p_n) : \sum_1^N p_i = 1, ; p_i \geq 0\}$$

is a compact subset of $\mathfrak{R}^n$. Recall that the entropy of $\nu = \sum_1^n \nu_i \delta_{x_i}$ relative to $p = \Sigma p_i \delta_{x_i}$ is

$$S_p(\nu) = -\Sigma\, \nu_i \ln (\nu_i / p_i).$$

Assume that on $(\Omega,\mathcal{F})$ we have a probability P_p such that $P(X_n=x_i)=p_i$ for all n. Let us denote by $Q_n^{(1)}$ and $Q_n^{(2)}$ the distributions of S_n/n and L_n with respect to P.

Let A_1 and A_2 be the Borel sets defined by

$$A_1 = \{t \in \mathfrak{R} : |t - m_p| \geq \varepsilon\},\ 0 < \varepsilon < \min\{m_p - x_1, x_N - m_p\}$$

$$A_2 = \left\{\nu \in \mathbf{P}(S) : \max_{i=1,2,\ldots} |\nu_i - p_i| \geq \varepsilon\right\},\ 0 < \varepsilon < \min_i \{p_i, 1 - p_i\}$$

and define

$$S(p,z) = \begin{cases} \max\{S_p(\nu) : \nu \in \mathbf{P}(S),\ \Sigma \nu_i x_i = z\} & z \in [x_1, x_n] \\ -\infty & z \notin [x_1, x_n] \end{cases} \tag{7.8}$$

We are now ready for

Theorem 7.9. With all the jargon introduced above

$$(i) \qquad \lim_{n\to\infty} \tfrac{1}{n} Q_n^1(A_1) = \max_{z \in A_1} S(p,z)$$

$$(ii) \qquad \lim_{n\to\infty} \tfrac{1}{n} Q_n^2(A_2) = \max_{\nu \in A_2} S_p(\nu).$$

Proof: Let us take care of (ii) to begin with. For each n and $\omega \in \Omega$ fixed, let $1 \leq i \leq N$ and $k_i = \#\{x_i$ appears in the sequence $X_1,\ldots,X_N\}$. Then $L_{n,i}=k_i/n$ and $L_n(.)$ is in A_2 if and only if $\mathbf{k}=(k_1,\ldots k_N)$ is in the set

$$A_{2,n} = \left\{\mathbf{k} : 0 \leq k_i \leq N,\ \Sigma k_i = n,\ \max_{1\leq i\leq n} \left|\frac{k_i}{n} - p_i\right| \geq \varepsilon\right\}.$$

Introduce the (standard) symbols

$$C(n, \mathbf{k}) = n!/n_1!, \ldots, n_N!, \quad \mathbf{p}^{\mathbf{k}} = p_1^{k_1}, \ldots, p_N^{k_N}$$

with the aid of which we have

$$Q_n^2(A_2) = \sum_{\mathbf{k} \in A_2} P\left\{L_{n,i} = \frac{k_i}{n}, \quad 1 \le i \le N\right\} = \sum_{\mathbf{k} \in A_{n,2}} C(n, \mathbf{k}) \mathbf{p}^{\mathbf{k}}$$

An obvious variation on the theme of Lemma 7.5 yields

$$\frac{1}{n} \ln C(n, \mathbf{k}) = -\Sigma \frac{k_i}{n} \ln \frac{k_i}{n} + O\left(\frac{\ln n}{n}\right)$$

as n becomes large. From this it is clear that

$$\frac{1}{n} \ln C(n, k) \mathbf{p}^{\mathbf{k}} = -\Sigma \frac{k_i}{n} \ln \frac{k_i}{n p_i} + O\left(\frac{\ln n}{n}\right).$$

Setting, for any $\mathbf{k}$,

$$\nu_{\mathbf{k}/n} = \Sigma \frac{k_i}{n} \delta_{x_i} \in \mathbf{P}(S),$$

it is clear that $L_{n,i}=k_i/n$ if and only if $L_n=\nu_{k/n}$. From this we have

(7.10) $$\frac{1}{n} \ln Q_n^2(\{\nu_{\mathbf{k}/n}\}) = S_p(\nu_{\mathbf{k}/n}) + O\left(\frac{\ln n}{n}\right).$$

Noticing now, that in the sum defining $Q_n^2(A_2)$ there are less than $(n+1)^N$ terms, we can proceed as in the proof of Theorem 7.4 to obtain

$$\frac{1}{n} \ln Q_n^2(A_2) = \max_{\mathbf{k} \in A_{2n}} \left\{S_p(\nu_{k|n}) + O\left(\frac{\ln n}{n}\right)\right\}.$$

Again, for each n, the set

$$\{\nu \in \mathbf{P}(S) : \nu = \nu_{\mathbf{k}/n} \quad \textit{for some} \quad \mathbf{k} \in A_{2,n}\}$$

is a subset of A_2. Invoking Lemma 7.6 the following is clear

$$\lim \tfrac{1}{n} \ln Q_n^2(A_2) = \max_{\nu \in A_2} S_p(\nu).$$

To obtain (i) from (ii) we proceed as follows:
Since $S_{n/n} = \Sigma x_i L_{n,i}$ we see that $S_{n/n} \in A_1$ iff $L_n \in B_2 = \{\nu \in \mathbf{P}(S) : |\Sigma x_i \nu_i - m_p| \geq \varepsilon\}$
Therefore $Q_n^{(1)}(A_1) = Q_n^{(2)}(B_2)$. A variation on the theme developed above yields

$$\lim \tfrac{1}{n} \ln Q_n^{(1)}(A_1) = \lim \tfrac{1}{n} \ln Q_n^{(2)}(B_2) = \max_{\nu \in B_2} S_p(\nu).$$

To get the result we want we need to compute the right hand side. Note to begin with that

$$\max_{\nu \in B_2} S_p(\nu) = \max_{z \in A_1 \cap [x_1, x_N]} \max \{S_p(\nu) : \nu \in \mathbf{P}(S), \quad \Sigma x_i \nu_i = z\}$$

$$= \max_{z \in A_1 \cap [x_1, x_N]} S(p, z)$$

note now, that we defined $S(p,z) = -\infty$ whenever it is not in $[x_1, x_N]$. Therefore

$$\lim_{n \to \infty} \tfrac{1}{n} \ln Q_n^{(1)}(A_1) = \max_{z \in A_1} S(p, z).$$

Comments. Since $\max \{S(p,z) : z \in A_1\}$ is negative, this theorem asserts that for N large, the probability that a microscopic configuration is such that the empirical mean $\int x L_N(dx)$ differs a little bit from the actual mean of X_1 is exponentially small. See chapter of [7.1] for more on this and [7.3] for a variation on the theme.

REFERENCES

[7.1] Ellis, R. S. "*Entropy, large deviations and statistical mechanics*". Springer Verlag, Berlin, 1985.

[7.2] Bucklew, J. A. "*Large deviation techniques in decision, simulation, and estimation*" John Wiley & Sons, New York, 1990.

[7.3] Robert, C. "*An entropy concentration theorems an application in artificial intelligence and descriptive statistics*". Jour. Appl. Prob., **27**, pp. 303-313, 1990.

Chapter 8

MAXIMUM ENTROPY AND CONDITIONAL PROBABILITIES

In this chapter we present a very interesting way of "understanding" how do maxentropic distributions arise. We go through the simplest situation, lifting the results from [8.1-3] but direct the reader to [6.41], [8.2], [8.3] and especially [8.4] for history, different and more comprehensive results. It should also become apparent that the results here and in the previous chapter are deeply connected.

We shall see that under some obvious regularity conditions, the maxentropic distributions appear as limits of conditional distributions. But before we go into that, let us go through an interlude on martingales, and changes of measure that, on one hand illuminates the maximum entropy method and, on the other explains some of the changes of measure that are made below.

Consider a probability space $(\Omega,\mathcal{F},P)$ and an increasing family $\mathcal{F}_n$ of sub-σ-algebras of $\mathcal{F}$. We shall say that a process (an indexed family of random variables) M_n is a martingale if : (i) $E|M_n|<\infty \quad \forall n$, (ii) $M_n\in\mathcal{F}_n \quad \forall n$, and (iii) $E[M_{n+1}|\mathcal{F}_n]=M_n \quad \forall n$.

Assume now that M_n is positive, and define

$$E_M[H] = E[HM_n] \tag{8.1}$$

for every bounded H in $\mathcal{F}_n$. The martingale property provides us with consistency for (8.1). To have (8.1) for any bounded H in $\mathcal{F}$, just approximate H by an appropriate sequence H_n.

Consider now a collection $\{X_n: n\geq 1\}$ of independent, identically distributed random variables such that

$$D(X) = \{\lambda \in \Re | E[\exp -\lambda X_1] < \infty\}$$

has a nonempty interior. A standard convexity argument implies that D(X) is an interval containing 0, and the comments in the appendix to chapter 5 tell us that $\mathcal{T}(\lambda)= -d\ln Z(\lambda)/d\lambda$ is a differentiable bijection between the int (D(X)) and int (conv(range X_1)). That is for each $a\in$ int(conv(range X_1)) there is $\lambda_a\in$ int (D(X)) such that $a= -d\ln Z/d\lambda(\lambda_a)$.

Notice now that $M_n=\exp-\lambda S_n/Z(\lambda)^n$ is a positive martingale, where $S_n=\sum_1^n X_K$, and that

$$E_M(X_1) = E[X_1 e^{-\lambda X_1}]/Z(\lambda)$$

In the probabilistic literature M_n is called Wald's martingale and the change of measure in (8.1) is the discrete time analogue of the Cameron-Martin-Girsanov transformation employed in sections 6.10 and 6.11. This set up can be greatly generalized, but let us not do it here. Let us just prove

Lemma 8.2. Let G be an integrable function. Let M_n be any positive martingale and P_M be defined as in (8.1). Then, for G in I_n

$$E_M[G|S_n] = E[GM_n|S_n]/M_n.$$

Proof: It is a rather simple consequence of the defining identities

Corollary 8.3. Let $M_n=\exp-\lambda S_n/Z(\lambda)^n$. Then if G is integrable and I_n-measurable

$$E[G|S_n] = E_M[G|S_n]$$

The first result we quote from [8.1] is contained in

Proposition 8.4. Assume that $\{X_n: n\geq 1\}$ are independent, finitely valued, identically distributed with $P(X_1=x_i)=p_i$. Let $\min\{x_i\}<a<\max\{x_i\}$, then

$$P\left(X_1 = j|\tfrac{1}{n}S_n = a\right) = e^{-\bar{\lambda}x_j}p_j/Z\left(\bar{\lambda}\right)$$

where $\bar{\lambda}$ is such that $-d\ln Z(\bar{\lambda})/d\lambda = \alpha$.

Proof: Let us denote P_M by $\hat{P}$ Then, the corollary above implies that

$$P(X_1 = j|S_n = an) = \hat{P}(X_1 = j|S_n = \alpha n)$$

$$= \hat{P}(X_1 = j)\hat{P}(X_2 + \ldots + X_n = n\alpha - j)/\hat{P}(S_n = n\alpha).$$

Since $\hat{P}\left(\left|\frac{S}{n} - \alpha\right| > \varepsilon\right)$ behaves as $\exp-nK(\varepsilon)$ for appropriate $K(\varepsilon)$, see Theorem 7.4, the result we want follows.

This result is extended in

Theorem 8.5. Let $\{X_n: n\geq 1\}$ denote a sequence of real valued, independent, identically distributed random variables. Let $U: \Re \to \Re$ be a bounded measurable function and $h: \Re \to \Re$ be such that $D(X)=\{\lambda\in \Re: Z(\lambda)=E[\exp-\lambda h(X)]<\infty\}$ has a nonempty interior. Let C denote the closure of the convex set generated by the range of X_1. Choose $\lambda\in$ int D(X) such that $E_M[h(X_1)]=a$ for $a\in$ int C, and put $S_n=\sum_{j=1}^{n} h(X_j)$. Then

$$\lim_{n\to\infty} E_M\left[U(X_1)g\left(\frac{S_n}{n}\right)\right] = E_M[U(X_1)]g(a)$$

for any smooth function g of rapid decay.

Proof: Write $g(x)=\int e^{-ikx}\hat{g}(k)dk/2\pi$. We have set things up so that Fubini's theorem can be applied to justify exchanging integrals to obtain

$$E_M\left[U(X_1)g\left(\frac{S_n}{n}\right)\right] = \int \frac{dk}{2\pi}\hat{g}(k)E_M[U(X_1)\exp-(iS_n/n)]$$

Now, the $h(X_n)$ are still independent relative to P_M and the expectation under the integral sign can be further computed as

$$E_M[U(X_1)\exp-i(kS_n/n)]$$

$$= E_M\left[U(X_1)\exp-\left(\tfrac{1}{n}X_1\right)\right]E\left[\exp-\left(\lambda+\tfrac{ik}{n}\right)\sum_2^n X_j\right]\exp-(n-1)\kappa(\lambda)$$

where we are again using $\kappa(\lambda)=\ln Z(\lambda)$. The way we choose λ insures us with the approximation

$$\exp-(n-1)\left(\kappa\left(\lambda+\tfrac{ik}{n}\right)-\kappa(\lambda)\right) = \exp-(n-1)\left(\tfrac{ik}{n}\kappa'(\lambda)+0\left(\tfrac{1}{n}\right)\right)$$

which tends to exp-(ika) as n goes to infinity.

Therefore

$$\lim E_M\left[U(X_1)g\left(\frac{S_n}{n}\right)\right] = \int \frac{dk}{2\pi}\hat{g}(k)e^{-ika}E_M[U(X_1)]$$

$$= E_M[U(X_1)]g(a).$$

And we are ready to state

Corollary 8.6. With the notations above

$$\lim_{n\to\infty} E[U(X_1)|S_n/n = a] = E_M[U(X_1)].$$

Proof: From corollary 8.3 it follows that

$$E[U(X_1)|S_n/n = a] = E_M[U(X_1)|S_n/n = a].$$

To conclude the proof, all you need is to know that when regular conditional probabilities exist, then

$$E_M[U(X_1)|S_n/n = a] = \lim_k E\left[U(X_1)g_k\left(\frac{S_n}{n}\right)\right] / E\left[g_k\left(\frac{S_n}{n}\right)\right]$$

where $\left\{g_k(x)\right\}$ is any sequence (like $\left(\frac{k}{2\pi}\right)^{1/2} \exp -k(x-a)^2/2$) peaking about a.

Nice huh? The natural, and obvious, interpretation is that conditioning with respect to $\{S_n/n=a\}$ concentrates the probability on the distribution yielding value a to h(X).

REFERENCES

[8.1] Van Campenhout, J. M. and Cover, T. M. "Maximum entropy and conditional probability". IEEE, IT-27, pp. 483-489, 1981.

[8.2] Zabell, S. L. "Rates of convergence for conditional expectations"

[8.3] Natarajan, S. "Large deviations, hypothesis testing and source coding for finite Markov chains". IEEE, IT-31, pp. 360-365, 1985.

[8.4] Csizar, I., Cover, T. M. and Choi, B. S. "Conditional limits theorems under Markov conditioning". IEEE, IT-33, pp. 788-801, 1983.

Chapter 9

MAXIMUM ENTROPY AND STATISTICS

This chapter consists of sections not bearing a relation with each other, but all related to statistics.

9.1 Gauss principle and minimum discrimination.

Let Φ_i: $\mathfrak{R} \to \mathfrak{R}$, i=1, K be measurable functions and $F_0(x)$ be the guess we make about the unknown distribution function F(x) of a random variable x. Assume that

$$Z(\boldsymbol{\lambda}) = \int e^{-(\lambda, \Phi(x))} dF_0(x)$$

is finite on $\mathfrak{R}^k$ The density $dF/dF_0(x)$ minimizing $K(F_1,F_0)=-SF_{F_1}(F)$ over the set of distribution functions $\{F: dF<<dF_0, E_F(\Phi)=\mathbf{c}\}$ is given by

(9.1.1) $$\frac{dF}{dF_0}(x) = (Z(\boldsymbol{\lambda}))^{-1} \exp-(\boldsymbol{\lambda}(c), \Phi(x))$$

where $\boldsymbol{\lambda}(\mathbf{c})$ is such that

(9.1.2) $$\int \Phi(x)(Z(\boldsymbol{\lambda}))^{-1} \exp-(\boldsymbol{\lambda}(\mathbf{c}), \Phi(x)) dF_0(x) = -\nabla_{\boldsymbol{\lambda}} \ln Z(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}(\mathbf{c})} = \mathbf{c}$$

That much is old news. In [9.1] from which we are quoting, Campbell observed the following. Assume there is a measure m(dx) on $\mathfrak{R}$ with respect to which F and dF_0 have densities $f(x,\mathbf{c})$ and $f_0(x)$ respectively.

If we put $\Phi_0(x)\equiv 1$ and define $\lambda_0=\ln Z(\boldsymbol{\lambda})$, then if n measurements $x_1,\ldots,x_n$ of a random variable X distributed according to (9.2), i.e.

(9.1.3) $$f(x, \mathbf{c}) = f_0(x) \exp-\sum_0^K \lambda_k \Phi_k(x)$$

then we have

Lemma 9.1.4. With the notations introduced above, the maximum likelihood estimators of the c_k are

(9.1.5) $$\hat{c}_k = \frac{1}{n}\sum_{i=1}^{n} \Phi_k(x_i)$$

Proof: Just form $\ln \prod_{i=1}^{n} f(x_i, \mathbf{c})$ and differentiate (9.1.3) with respect to c_k, equate the derivative to zero, use (9.1.2) to obtain (9.1.5).

Before going to the converse, we shall recall (the generalization of) Gauss' method. Again let us assume that the unknown distribution with respect to a measure m(dx)) of a real valued random variable is f(x,**c**). The experimenter knows a sample $x_1,\ldots,x_n$ and $f_0(x)=f(x,\mathbf{c}_0)$ for some $\mathbf{c}_0$.

Gauss' methods is based on assuming that

i) The right f(x,**c**) corresponds to the value of **c** maximizing the likelihood

(9.1.6) $$\ln \prod_{1}^{n} f(x_i, \mathbf{e})$$

ii) That the correct value of **c** is given by the estimator

$$\hat{c}_K = \frac{1}{n}\sum_{i=1}^{n} \Phi_k(x_i) \qquad k = 1, 2, \ldots, K$$

for some appropriate family $\Phi_1(x),\ldots,\Phi_K(x)$. We are ready now for

Lemma 9.1.7. Assume that the functions $1,\Phi_1(x),\ldots,\Phi_K(x)$ are independent (none of them is a function of the others) and that they are continuously differentiable. Assume also that f(x,**c**) is continuously differentiable, once with respect to x and twice with respect to **c**. If f(x,**c**) is obtained via Gauss' method, then it is of the form

$$f(x, \mathbf{c}) = f_0(x)\exp-(\lambda(\mathbf{c}), \Phi(x))$$

for some appropriate $\lambda(\mathbf{c})$.

Proof: The correct value of **c** has to satisfy both

(9.1.8) $$\frac{\partial}{\partial c_k}\sum_{j=1}^{n} \ln f(x_j, \mathbf{c}) = 0 \qquad k = 1, 2, \ldots, K$$

and (9.1.5). If we think of $\mathbf{x}=(x_1,\ldots,x_n)$ as the coordinates of a point in $\mathfrak{R}^n$, then both (9.1.8) and (9.1.5) rewritten as

(9.1.9) $$\frac{1}{n}\sum_{j=1}^{n}(\Phi_k(x_j)-c_k)=0 \qquad k=1,2,\ldots,K$$

determine the same value of **c**. This is the key determine the same k-dimensional surface. This means that each normal to (9.1.5) is a linear combinations of the normals to (9.1.9). That is, for each $1\le i\le n$

$$\frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial C_k}\sum_{j=1}^{n}\ln f(x_j,\mathbf{c})\right)=\sum_{l=0}^{K}a_{kl}\frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n}\Phi_l(x_j)-c_l\right)$$

where the a_k depend on **c**. Obviously, for every i,

$$\frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial C_k}\ln f(x_i,\mathbf{c})\right)=\sum_{l=0}^{K}a_{Kl}\frac{\partial}{\partial x_i}\Phi_l(x_i)$$

Since the functions involved depend only on one coordinate x_i, we can give any generic name to it, let it be x.

Now, integrating both sides of the last identity with respect to x and noticing that (9.1.8) and (9.1.9) have to hold, it is clear that the integration constants have to be such that

(9.1.10) $$\frac{\partial}{\partial c_k}\ln f(x,\mathbf{c})=\sum_{l=0}^{K}a_{Kl}(\Phi_l(x)-c_l)$$

Since $\partial^2\ln f(x,\mathbf{c})/\partial c_k\,\partial c_l=\partial^2\ln f(x,\mathbf{c})/\partial c_l\,\partial c_k$, from that last identity it follows that

$$\sum_{l=1}^{K}\left(\frac{\partial a_{l}}{\partial c_l}-\frac{\partial a_{ll}}{\partial c_k}\right)(\Phi_l(x)-c_l)-a_{kl}+a_{lk}=0$$

and bringing in the assumption of the independence of the set $1,\Phi_1(x),\ldots,\Phi_K(x)$, we obtain that

$$\partial a_{kl}/\partial c_l=\partial a_{ll}/\partial c_k, \qquad \mu_{kl}=\mu_{lk}$$

The first implies that for each $1\le i\le K$ there exists a $\lambda_l(\mathbf{c})$ such that $a_{ki}=-\partial\lambda_i/\partial c_k$. The second implies that there is a function $V(\mathbf{c})$ such that $\lambda_i=\partial V/\partial c_i$. Now, we can rewrite (9.1.10) as

$$\frac{\partial\ln}{\partial C_k}f(x,\mathbf{c})=-\frac{\partial}{\partial C_k}\sum_{l=1}^{K}[\lambda_l(\Phi_l-C_l)]+V \quad .$$

Integrating both sides along any curve joining $\mathbf{c}_0$ to **C** and since $f_0(x)=f(x,\mathbf{c}_0)$ we obtain

$$f(x,\mathbf{c})=f_0(x)\exp-[(\boldsymbol{\lambda}(c),\boldsymbol{\Phi}(x))+\lambda_0]$$

where the identification of λ_0 with all remaining constants in the exponent is clear. Since the left hand side is to be normalized, it is clear that $\exp\lambda_0 = Z(\lambda)$ as above

This concludes Campbell's proof.

Notice one interesting consequence of the extension of Gauss' method. When **c** is given according to (9.1.5) and when f(x,**c**) is the right distribution, and when n tends to infinity then

$$\frac{1}{n}\sum_{i=1}^{n} \ln f(x_i, \mathbf{c})/f_0(x_i) \to \int \ln \frac{f(x,\mathbf{c})}{f_0(x)} dF(x)$$

$$\frac{1}{n}\sum_{i=1}^{n} \Phi_k(x_i) \to \int \Phi_k(x) dF(x)$$

where of course dF=f(x,**c**)m(dx). Certainly we would also obtain

$$-\frac{1}{n}\sum_{i=1}^{n} \ln f(x_i, \mathbf{c}) \to S_m(X) = -\int f(x, \mathbf{c}) \ln f(x, \mathbf{c}) m(dx)$$

Thus, maximum likelihood makes the entropy functional easy to accept, at least for statisticians. And for them, the maximum entropy method for looking for distribution functions must also be natural. The gist or the crux of the problem of characterizing distribution becomes identical with the issue of choosing a family $\{\Phi_k(x), k=1,...,K\}$ to characterize the parameters of the distribution.

In the next section we shall see how the notion of sufficiency anticipated Campbell results from a different point of view.

9.2 *Sufficiency.*

Here we shall refer to the second chapter of [4.1] and from the second and third chapters of [0.2]. Recall that if the probability P on $(\Omega,\mathcal{F})$ has a density f with respect to a measure μ on $(\Omega,\mathcal{F})$ then the restriction on P to a sub-σ-algebra G has a density $E_\mu[f|G]$ with respect to the restriction of μ to G.

We defined for $P,Q \in \mathbf{P}(\Omega)$ and $\mu \in M(\Omega)$

$$K_\mu^G(P,Q) = I_\mu^G(P,Q) = \int E_\mu[f|g] \ln \frac{E_\mu[f|g]}{E_\mu[f|g]} d\mu$$

and a variation of the proof of Lemma 4.9 establishes that

(9.2.1) $$K_\mu(P,Q) \geq K_{\bar{\mu}}^{G}\left(\bar{P},\bar{Q}\right)$$

where $\bar{P},\bar{Q},\bar{\mu}$ denote restrictions to G.

Definition 9.2.2. G is sufficient for P,Q (relative to μ) whenever the identity holds in (9.2.1). That is whenever

$$\frac{dP/d\mu}{E_\mu[dP/d\mu|G]} = \frac{dQ/d\mu}{E_\mu[dQ/d\mu|G]}$$

holds a. e. μ.

In the setup of Lemma 4.9 $G = \sigma(\Phi)$ for Φ: $(\Omega,\mathcal{F}) \to (\Omega',\mathcal{F}')$, $P' = P \circ \Phi^{-1}$ and $\mu' = \mu \circ \Phi^{-1}$, we note that $E_\mu\,[dP/d\mu|\Phi] = (dP'/d\mu') \circ \Phi$, etc., therefore the sufficiency condition becomes

$$((dP'/d\mu') \circ \Phi)^{-1} dP/d\mu = ((dQ/d\mu) \circ \Phi)^{-1} dQ/d\mu$$

Consider now the following setup. Let X: $\Omega \to (E,\mathcal{E})$ be an E-valued random variable. Let Q be a probability on $(\Omega,\mathcal{F})$ and m be a measure on $\mathfrak{R}$ such that

$$Q(X \in A) = \int_A P(\xi) m(d\xi)$$

(or the Q-distribution of X has a density f with respect to m).

Let Φ: $E \to \mathfrak{R}^k$ be an $\mathfrak{R}^k$-valued Borel measurable function such that

$$Z(\lambda) = E_Q[e^{-(\lambda,\Phi(X))}] = \int_E e - (\lambda, \Phi(\xi)) p(\xi) m(d\xi)$$

is defined for λ in some convex set $D \in \mathfrak{R}^k$ Denote by $P(\lambda)$ the measure on Ω with density $dP(\lambda)/dQ = \exp-(\lambda,\Phi(X))/Z(\lambda)$. Let $\theta(\lambda)$ be given by

(9.2.3) $$\theta(\lambda) = E_{p(\lambda)}(\Phi(X))$$

Suppose you take N independent copies $X_1,\ldots,X_N$ distributed according to $P_N = P(\lambda) \otimes \ldots \otimes P(\lambda)$ and let $Q \otimes \ldots \otimes Q$ be the product of N copies of Q. Then

$$\theta = \frac{1}{N} \sum_{i=1}^{N} \Phi(x_i)$$

is a sufficient statistics for $K(P_N,Q_N) = NK(P(\lambda),Q)$.

Thus supplementing this result with the uniqueness in the correspondence $\lambda \to \theta(\lambda)$ contained in (9.2.3) neatly rounds up a bunch of ideas.

9.3 *Some very elementary Bayesian statistics.*

We shall present some very elementary results about the Bayesian approach to density estimation. For more the reader should take a look at [9.2] or [9.3].

When considering a parametric family $\rho(x,\theta)$ it proves convenient to think of θ as the values of a random variable about which we know an a priori distribution $G(\theta)$. We want to estimate θ given the values $X_1,\ldots,X_n$, of a random variable X having distribution function $\rho(x|\theta)$.

To produce the Bayesian estimator of θ the following procedure is applied

i) The posterior distribution of θ given the observations $x_1,\ldots,x_n$ is

$$h(\theta|x_1,\ldots,x_n) = \rho(x_1,\ldots,x_n|\theta)p(\theta)\left(\int f(x_1,\ldots,x_n|\theta)p(\theta)\nu(d\theta)\right)^{-1}$$

The measure ν is preassigned on the range T of θ.

ii) A utility (or loss) function $L(\hat{\theta},\theta)$ is chosen in advance. It is assumed to describe the gain or penalty incurred in choosing the estimator $\hat{\theta}$ when the correct value of parameter happens to be θ. For example; $(\hat{\theta}-\theta)=L(\hat{\theta},\theta)$.

iii) The risk measure for $\hat{\theta}=\hat{\theta}(x_1,\ldots,x_n)$ is defined by

$$R\left(\hat{\theta},\theta\right) = \int L\left(\hat{\theta},\theta\right)\rho(x_1,\ldots,x_n|\theta)dx_1\ldots dx_n$$

and the expected risk is

$$R\left(\hat{\theta}\right) = \int R\left(\hat{\theta},\theta\right)g(\theta)\nu(d\theta)\,.$$

The Bayes estimator $\hat{\theta}_B$ is the estimator $\hat{\theta}$ which minimizes $R(\hat{\theta})$. When $L(\hat{\theta},\theta)$ is convex in the first variable, then $R(\hat{\theta})$ is a convex functional of $\hat{\theta}$ and we can do variational analysis. For full details see [9.2].

It is reasonably easy to verify that $\hat{\theta}_B$ is the estimator at which the a posteriori risk

$$R\left(\hat{\theta}|x_1,\ldots,x_n\right) = \int L\left(\hat{\theta},\theta\right)h(\theta|x_1,\ldots,x_n)\nu(d\theta)$$

reaches its minimum.

The problem is then how to choose $g(\theta)$. In [9.2] there are a few examples in which $g(\theta)$ is chosen as the density (with respect to $\nu(d\theta)$) which maximizes

$$S_\nu(\theta) = -\int g(\theta)\ln g(\theta)\nu(d\theta)$$

subject to

$$\int g(\theta)\nu(d\theta) = 1$$

$$\int f_K(\theta)g(\theta)\nu(d\theta) = \mu_K \qquad K = 1, \ldots, m.$$

This is a level 1 maxentropic reconstruction problem about which you have heard enough by now. We present instead an example missing in [9.2] which is a (minor) variation on the theme of section 9 of [9.3].

Assume that we somehow know that $P(\theta)$ is a member of a family of densities $p(\theta,a)$ on TxA, a being some countable set on which we have another a priori density $w(\theta,a)$ chosen perhaps according to some invariance principle (or god given to cut short the regression). Assume that on the basis of an observation of a we decide to look for the distribution $P(\theta,a)$ concentrated on an a maximizing the relative entropy (or Kullback distance)

$$S = -K(P, W) = -\sum_a \int P(\theta, a)\ln \frac{P(\theta,a)}{W(\theta,a)}\nu(d\theta)$$

subject to the obvious constraint

$$\int P(\theta, a_0)\nu(d\theta) = \sum_a \int P(\theta, a)\delta_{aa_0}\nu(d\theta) = 1 .$$

The maxentropic $P_{ME}(\theta,a)$ which is our candidate for $P(\theta)$ is

$$\mathbf{P}_{\mathrm{ME}}(\theta, a) = \begin{cases} \left(\int W(\theta, a_0)\nu(d\theta)\right)^{-1} W(\theta, a_0) \equiv P(\theta) & \text{if } a = a_0 \\ 0 & \text{otherwise} \end{cases}$$

And this is a good point to stop.

REFERENCES

[9.1] Campbell, O. L. "Equivalences of Gauss' principle and minimum discrimination information estimation of probabilities" Ann. of Math. Stat. 41, pp. 1011-1015, 1970.

[9.2] Berger, J. O. "Statistical Decision Theory and Bayesian Analysis". Springer-Verlag, Berlin, 1985.

[9.3] Akaike, H. "Prediction and entropy" in "A Celebration of Statistics". Atkinson, A. C. and Fienberg, S. E., Eds. Springer-Verlag, Berlin, 1985.