ENTROPY IN PROBABILITY AND STATISTICS

by

Wolfgang André Rolke

A Dissertation Presented to the

FACULTY OF THE GRADUATE SCHOOL

UNIVERSITY OF SOUTHERN CALIFORNIA

In Partial Fulfillment of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

(Applied Mathematics)

May 1992

UNIVERSITY OF SOUTHERN CALIFORNIA
THE GRADUATE SCHOOL
UNIVERSITY PARK
LOS ANGELES, CALIFORNIA 90007

This dissertation, written by

Wolfgang Andre Rolke

*under the direction of h*is........ *Dissertation Committee, and approved by all its members, has been presented to and accepted by The Graduate School, in partial fulfillment of requirements for the degree of*

DOCTOR OF PHILOSOPHY

Barbara Solomon
Dean of Graduate Studies

Date April 16, 1992

DISSERTATION COMMITTEE

Joseph C. Watkins
Chairperson

S. Tavaré

Robert W. Ashcome

Dedication

For Linda

Acknowledgements

I would like to thank Professor J. Watkins for being more than just a thesis advisor. His friendship has made a long journey much more enjoyable.

I would also like to thank Professor S. Tavaré and Professor R. Osborne for serving on my committee.

I am grateful for the Department of Mathematics for supporting me through the years. The teaching assistantship not only paid my bills, it also taught me how to be a teacher. This research was supported in part by National Science Foundation grant DMS 90-11480.

Finally I would like to thank my cat Lolita for being there when by blood pressure went up to far.

Abstract

We develop a theory of entropy, where entropy is defined as the Legendre - Fenchel transform of the logarithmic moment generating function of a probability measure on a Banach space. A variety of properties relating the probability measure and its entropy are proven. It is shown that the entropy of a large class of stochastic processes can be approximated by the entropies of the finite - dimensional distributions of the process. For several types of measures we find explicit formulas for the entropy, for example for stochastic processes with independent increments and for Gaussian processes. For the entropy of Markov chains, evaluated at the observations of the process, we prove a central limit theorem. Theorems relating weak convergence of probability measures on a finite dimensional space and pointwise convergence of their entropies are developed and then used to give a new proof of Donsker's theorem. Finally the use of entropy in statistics is discussed. We show the connection between entropy and Kullback's minimum discrimination information. A central limit theorem yields a test for the independence of a sequence of observations.

Contents

Chapter 0

Introduction

In the 1930's Harold Cramér became interested in the following question : Given a sequence of independent and identically distributed random variables $\{X_i\}_{i=1}^{\infty}$ with $EX_1 = 0$ and $VarX_1 = 1$ we know from the law of large numbers that $\frac{1}{n}\sum_{i=1}^{n} X_i \to 0$ in probability. What can be said about the rate of convergence?

If the random variable X_1 has a finite moment generating function, then a simple calculation using Chebyshev's inequality shows the following :

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} X_i > x\right) &= P\left(\exp[\lambda\frac{1}{n}\sum_{i=1}^{n} X_i] > \exp[\lambda x]\right) \\ &\leq \frac{E\exp[\lambda\frac{1}{n}\sum_{i=1}^{n} X_i]}{\exp[\lambda x]} = \frac{\left(Ee^{\frac{\lambda}{n}X_1}\right)^n}{e^{\lambda x}} \\ &= \exp\left\{-\left(\lambda x - n\log Ee^{\frac{\lambda}{n}X_1}\right)\right\} \end{aligned}$$

for all $\lambda, x > 0$. Taking the infimum over all $\lambda > 0$ on the right hand side gives

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} X_i > x\right) &\leq \exp\left\{-\sup_{\lambda>0}\left(\lambda x - n\log Ee^{\frac{\lambda}{n}X_1}\right)\right\} \\ &= \exp\left\{-n\sup_{\lambda>0}\left(\lambda x - \log Ee^{\lambda X_1}\right)\right\} \end{aligned}$$

or

$$\frac{1}{n}\log P\left(\frac{1}{n}\sum_{i=1}^{n} X_i > x\right) \leq -I(x)$$

where

$$I(x) \;=\; \sup_{\lambda>0}\left\{\lambda x - \log Ee^{\lambda X_1}\right\}$$

and so we find that the convergence in probability is exponentially fast.

It is easy to show that the sup in the definition of I above can be taken over all $\lambda \in R$, see for example Deuschel and Stroock [2], Lemma 1.2.3 i, and so we are led to the definition of the entropy of a probability measure on the real line :

Let $\mu \in \mathcal{P}(R)$, $x \in R$, then

$$\Lambda^*_\mu(x) \;=\; \sup\left\{\lambda x - \log\int e^{\lambda t}\mu(dt) \;\;:\;\; \lambda \in R\right\}$$

This functional is also called the large deviation rate function or the Legendre - Fenchel transform of the log moment generating function.

In this dissertation we provide an indepth study of the entropy of probability measures in Banach spaces.

Entropies arise not only in large deviation theory as described above but also in statistics as the Kullback's minimum discrimination information, in signaling theory and in other fields.

In Chapter 1 we define the entropy of a probability measure on a Banach space as the convex-conjugate or Legendre-Fenchel transform of the log moment generating function of the measure. We prove a variety of properties of the entropy, such as convexity, lower-semicontinuity, non-negativity and that it is zero at the mean of the measure. We also prove that the entropy is additive for independent measures, that is, if $\mu, \nu \in \mathcal{P}(E)$ are independent, $x, y \in E$ then

$$\Lambda^*_{\mu\times\nu}(x,y) \;=\; \Lambda^*_\mu(x) \;+\; \Lambda^*_\nu(y)$$

This property will be used extensively in later Chapters.

In Chapter 2 we establish the connection between entropy and relative entropy defined as follows: For $\mu, \nu \in \mathcal{P}(E)$ the relative entropy of ν with respect to μ is defined by

$$H(\nu|\mu) \;=\; \begin{cases} \int_E f\log f d\mu & \text{if } \nu << \mu \text{ and } f = \frac{d\nu}{d\mu} \\ \infty & \text{otherwise} \end{cases}$$

Using techniques from large deviation theory we show that

$$\Lambda_\mu^*(x) = \inf\left\{H(\nu|\mu) \,:\, \nu \in \mathcal{P}(E)\,,\ \int_E t\nu(dt) = x\right\}$$

This identity gives an alternative method of computing the entropy Λ^*, or at least an upper bound.

In Chapter 3 we turn to the study of the entropy of measures on R^d. First we prove a series of structure theorems, showing a strong connection between the support of the measure and the set of values for which the entropy is finite. A useful inequality between the entropy of a measure and the entropies of its marginals is established. Finally a theorem is proven relating weak convergence of measures with pointwise convergence of the entropies. The completeness of these theorems is shown with some counterexamples.

In Chapter 4 we show that the entropy of a stochastic process with sample paths in the space of right continuous functions with left hand limits can be approximated by the entropies of the finite-dimensional distributions of the process. As examples we use this approximation to find the entropy of Brownian motion and of the Poisson process. This approximation can also be used to estimate the entropy of a process from a finite number of observations.

In Chapter 5 we use the approximation from Chapter 4 to give a new proof of Donsker's theorem, namely that if $\{X_n\}_{n=1}^{\infty}$ is a sequence of independent and identically distributed random variables with mean 0 and variance 1 and if

$$S_n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{[nt]} X_i$$

then

$$S_n \Rightarrow B$$

where B is standard Brownian Motion.

In Chapter 6 we turn to the study of entropies of Gaussian processes. We show the connection between the entropy of a Gaussian process and the eigenvalues and eigenfunctions of the covariance function of the process, and also between the entropy and the norm of the reproducing kernel Hilbert space of the covariance function.

In Chapter 7 we consider the entropy of finite-state Markov processes. An explicit formula is found for two-state chains and a central limit theorem is proven for n-state chains.

In Chapter 8 we will study the usefulness of the entropy in statistics. We will show the close connection between entropy and Kullback's minimum discrimination information. We will derive a nonparametric test for a distribution on a finite state space, and finally we will develop a test for independence in a sequence of observations.

Chapter 1

Entropy

In this Chapter we will define the entropy of a probability measure on a Banach space as the complex-conjugate of the log moment generating function of the probability measure. We will state some results from the theory of convex functions and use these to derive several important properties of the entropy.

Let (E, ρ) be a separable Banach space with metric ρ induced by the norm, and let $\mu \in \mathcal{P}(E)$, the space of probability measures on E. We will denote by $\langle \lambda, x \rangle$ the natural pairing of the Banach space E and its dual E^*. The log moment generating function Λ_μ of μ is defined by

$$\Lambda_\mu(\lambda) = \log \int_E \exp[\langle \lambda, t \rangle] \mu(dt) \quad \lambda \in E^*$$

Throughout this dissertation we will assume that

$$0 < \int_E \exp[\alpha \parallel t \parallel] \mu(dt) < \infty \quad \forall \alpha \geq 0$$

and we will denote the space of all probability measures on E satisfying this condition by $\mathcal{P}(E)$. Note that this assumption also implies $\Lambda_\mu(\lambda) < \infty$ because by the Schwartz inequality:

$$| \langle \lambda, t \rangle | \leq \parallel \lambda \parallel \cdot \parallel t \parallel \quad \forall t \in E \quad \forall \lambda \in E^*$$

Let $L(E)$ be the class of lower-semicontinuous and convex functions on E.

Definition 1.1 Let $f \in L(E^*)$. Then the Legendre-Fenchel transform $TLF : L(E^*) \longmapsto L(E)$ is defined by :

$$f^*(x) := TLF(f)(x) = \sup\{\langle\lambda, x\rangle - f(\lambda) \ : \ \lambda \in E^*\} \quad x \in E$$

Note

f^* is also called the convex-conjugate of f. The Legendre-Fenchel transform has been studied extensively in the theory of convex functions, especially for the case of d-dimensional Euclidean space. For further information, see Rockafellar [9].

Lemma 1.1 *We will need the following properties of the Legendre-Fenchel transform:*

1. *Let $f, g \in L(E^*)$ with $f \geq g$. Then $f^* \leq g^*$.*

2. *Let $\{f_n, f\} \subseteq L(E^*)$ and $f_n \to f$ pointwise. Then*

$$\liminf_{n\to\infty} f_n^*(x) \geq f^*(x) \ \ \forall x \in E$$

   that is, f^* is lower-semicontinous, or l.s.c.

3. *If f is not identically infinite, then*

$$f(\lambda) = \sup\{\langle\lambda, x\rangle - f^*(x) \ : \ x \in E\}$$

   for all $\lambda \in E^*$. In other words, $(f^*)^* \equiv f$.

Proof.

1. Let $\lambda \in E^*$. Then

$$\begin{aligned} & f(\lambda) \geq g(\lambda) \\ \Rightarrow \quad \langle\lambda, x\rangle - f(\lambda) &\leq \langle\lambda, x\rangle - g(\lambda) \\ \Rightarrow \quad f^*(x) &\leq g^*(x) \end{aligned}$$

2.

$$\liminf_{n\to\infty} f_n^*(x)$$

$$= \liminf_{n\to\infty} \sup \{\langle\lambda, x\rangle - f_n(\lambda) \; : \; \lambda \in E^*\}$$

$$\geq \liminf_{n\to\infty} \{\langle\lambda, x\rangle - f_n(\lambda)\} \quad \forall\lambda \in E^*$$

$$= \quad \langle\lambda, x\rangle - f(\lambda) \quad \forall\lambda \in E^*$$

3. See Deuschel, Stroock, Theorem 2.2.15 [2].

□

Now we can define the entropy of μ, Λ_μ^*, as the Legendre-Fenchel transform of the log moment generating function of the measure :

$$\Lambda_\mu^* \; : E \longmapsto [0, \infty]$$

and

$$\Lambda_\mu^*(x) := TLF(\Lambda_\mu)(x) = \sup \{\langle\lambda, x\rangle - \Lambda_\mu(\lambda); \; \lambda \in E^*\}$$

Lemma 1.2 *$\forall\, \mu, \nu \in \mathcal{P}(E)$ we have*

1. *If we let $S = \{x \in E \; : \; \Lambda_\mu^*(x) < \infty\}$, then S is convex.*
2. *$\Lambda_\mu^* \geq 0$.*
3. *Λ_μ^* is l.s.c and convex.*
4. *$\int_E t\mu(dt) = x \;\Rightarrow\; \Lambda_\mu^*(x) = 0$.*
5. *$\Lambda_\mu^*(x) = \Lambda_\nu^*(x) \;\; \forall x \in E \iff \mu \equiv \nu$.*

Proof.

1. Assume $x, y \in S$, $\alpha \in (0,1)$ and $z = \alpha x + (1-\alpha)y$. Then

$$\Lambda_\mu^*(z) = \sup\{\langle\lambda, z\rangle - \log \int_E \exp[\langle\lambda, t\rangle]\mu(dt) \; : \; \lambda \in E^*\}$$

$$= \sup\{\alpha\langle\lambda, x\rangle - \alpha \log \int_E \exp[\langle\lambda, t\rangle]\mu(dt)$$

$$+ (1-\alpha)\langle\lambda, y\rangle - (1-\alpha)\log\int_E \exp[\langle\lambda, t\rangle]\mu(dt) \ : \ \lambda \in E^*\}$$

$$\leq \alpha\Lambda_\mu^*(x) + (1-\alpha)\Lambda_\mu^*(y) \ < \ \infty$$

and so $z \in S$.

2.

$$\Lambda_\mu^*(x) \geq \langle 0, x\rangle - \log\int_E e^{\langle 0,t\rangle}\mu(dt) = 0$$

3. Assume $\lim_{n\to\infty} \| x_n - x \| = 0$ with $x \in E$. Then for each $\lambda \in E^*$

$$\lim_{n\to\infty} \Lambda_\mu^*(x_n) \geq \lim_{n\to\infty}\{\langle\lambda, x_n\rangle - \log\int_E \exp[\langle\lambda, t\rangle]\mu(dt)\}$$

$$= \langle\lambda, x\rangle - \log\int_E \exp[\langle\lambda, t\rangle]\mu(dt) \ \ \forall\lambda \in E$$

Hence

$$\lim_{n\to\infty} \Lambda_\mu^*(x_n) \geq \Lambda_\mu^*(x)$$

The convexity of Λ_μ^* follows from an application of Hölder's inequality.

4. By Jensen's inequality we have

$$\langle\lambda, x\rangle - \log\int_E e^{\langle\lambda,t\rangle}\mu(dt) \leq \langle\lambda, x\rangle - \int_E \langle\lambda, t\rangle\mu(dt)$$

$$= \langle\lambda, x\rangle - \langle\lambda, \int_E t\mu(dt)\rangle = 0$$

5. " $\Rightarrow$ " By Lemma 1.1 we have

$$\Lambda_\mu(\lambda) = \Lambda_\nu(\lambda) \quad \forall\lambda \in E^*$$

"$\Leftarrow$" Trivial.

□

It will be convenient to define the entropy of a random variable on the Banach space E. Let X be a random variable on E, and assume that μ is the law of X. Then we can define the entropy of X as the entropy of its law μ :

$$\Lambda_X^* \equiv \Lambda_\mu^*$$

Next we will show that the entropy is additive for independent random variables. This property will play a central role throughout this dissertation :

Lemma 1.3 *Let X, Y be independent random variables on E. Then for $x,y \in E$*

$$\Lambda_{(X,Y)}^*((x,y)) = \Lambda_X^*(x) + \Lambda_Y^*(y)$$

Proof.

From the independence of X and Y it follows that

$$E[\exp(\langle(\lambda_1,\lambda_2),(X,Y)\rangle)] = E[\exp(\langle\lambda_1, X\rangle)] \cdot E[\exp(\langle\lambda_2, Y\rangle)]$$

and so

$$\sup\{\langle(\lambda_1,\lambda_2),(x,y)\rangle - \log E[\exp(\langle(\lambda_1,\lambda_2),(X,Y)\rangle)] \;:\; (\lambda_1,\lambda_2) \in E^* \times E^*\}$$

$$= \sup\{\langle\lambda_1, x\rangle - \log E[\exp(\langle\lambda_1, X\rangle)]\} + \sup\{\langle\lambda_2, y\rangle - \log E[\exp(\langle\lambda_2, Y\rangle)]\}$$

□

Example 1.1 As a first example of entropies we will compute the entropy of a d-dimensional normal random vector. Let X be normal with mean vector $m \in R^d$ and covariance matrix $\Sigma \in R^{dxd}$. We assume X to be nondegenerate, i.e. Σ is positive definite. Hence there exists a unitary matrix A such that

$$A \cdot \Sigma \cdot A^T = D = \begin{pmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \lambda_d^2 \end{pmatrix}$$

where $\lambda_1^2, \cdots, \lambda_d^2$ are the eigenvalues of Σ. The Banach space pairing is now the scalar inner product, and so

$$\langle \lambda, X \rangle = \lambda' X = \lambda' A^T A X = \langle A\lambda, AX \rangle$$

and AX is a multivariate normal random vector with mean Am and covariance matrix D, i.e. independent components. Hence

$$E[\exp\{\langle \lambda, X \rangle\}] = E[\exp\{\sum_{i=1}^{d} (A\lambda)_i (AX)_i\}]$$

$$= \prod_{i=1}^{d} E[\exp\{(A\lambda)_i (AX)_i\}]$$

Because A has full rank, we have

$$\{A\lambda \, ; \ \lambda \in R^d\} = R^d$$

and so

$$\sup\{\langle \lambda, x \rangle - \log E[\exp\{\lambda' X\}]; \ \ \lambda \in R^d\}$$

$$= \sup\{\langle A\lambda, AX \rangle - \sum_{i=1}^{d} \log E[\exp\{(A\lambda)_i (AX)_i\}]; \ \ A\lambda \in R^d\}$$

$$= \sum_{i=1}^{d} \Lambda^*_{(AX)_i}((Ax)_i)$$

So we see that it suffices to find the entropy of a normal random variable on R^1. Suppose $Y \sim N(m, \sigma^2)$. Then

$$E[e^{t \cdot Y}] = e^{tm + \frac{1}{2}\sigma^2 t^2}$$

$$\Lambda^*_Y(y) = \sup\{ty - \log E[e^{tY}] \ : \ t \in R\} = \sup\{ty - tm - \frac{1}{2}\sigma^2 t^2 \ : \ t \in R\}.$$

$$\frac{d}{dt}\{ty - tm - \frac{1}{2}\sigma^2 t^2\} = y - m - \sigma^2 t = 0$$

$$\Rightarrow \quad t = \frac{y - m}{\sigma^2}$$

and so

$$\Lambda_Y^*(y) = \frac{1}{2} \cdot \frac{(y-m)^2}{\sigma^2}$$

Note that $(AX)_i \sim N((Am)_i, \lambda_i^2)$ and so we get :

$$\Lambda_X^*(x) = \frac{1}{2}\sum_{i=1}^{d} \frac{(A(x-m))_i^2}{\lambda_i^2} = \frac{1}{2}[A(x-m)]^T D^{-1} A(x-m)$$

$$= \frac{1}{2}(x-m)^T A^T D^{-1} A(x-m) = \frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)$$

Example 1.2 As a second example we will compute the entropy of standard Brownian motion. By Wiener's theorem there exists a unique measure $\mathcal{W}$ on

$$\Theta = \{\theta \in C([0,\infty); \mathbf{R}^d) : \theta(0) = 0 \ \ and \ \ \lim_{t\to\infty} \frac{|\,\theta(t)\,|}{t} = 0\}$$

with the property that

$$\Lambda_{\mathcal{W}}(\lambda) \equiv \log[\int_\Theta \exp[\langle \lambda, x\rangle]\mathcal{W}(dx)] = 1/2 \int_0^\infty \int_0^\infty s \wedge t \ \lambda(ds) \cdot \lambda(dt) \ \ \lambda \in \Theta^*$$

In the following we will consider only those $\lambda \in \Theta^*$ which are non-atomic and have compact support. These are dense in Θ^* and so the supremum will be the same.

Let H^1 be the space of functions $x \in \Theta$ with the property that

$$x(t) = \int_0^t \dot{x}(s)ds \ \ \forall t \geq 0$$

where $\dot{x} \in L^2$ and $L^2 = L^2([0,\infty); R^d)$. Note that

$$\int_0^\infty \int_0^\infty s \wedge t \ \lambda(ds) \cdot \lambda(dt) = \int_0^\infty |\,\lambda((t,\infty))|^2 dt$$

First assume that $x \in H^1$. Then we get

$$\Lambda_{\mathcal{W}}^*(x) = \sup\{\int_0^\infty \dot{x}(t)\lambda((t,\infty))dt - 1/2\int_0^\infty |\lambda((t,\infty))|^2 dt \ : \ \lambda \in \Theta^*\}$$

$$= \sup\{(\dot{x},\varphi)_{L^2} - \frac{1}{2} \parallel \varphi\parallel_{L^2}^2 : \varphi \in L^2\}$$

$$= 1/2 \parallel \dot{x}\parallel_{L^2}^2$$

Here the first equation holds because $\lambda \in \Theta^*$ with compact support implies $\lambda(\cdot, \infty) \in L^2$, and because each $\varphi \in L^2$ gives rise to a $\lambda \in \Theta^*$ by $\lambda(\cdot, \infty) = \varphi$.

Now assume that $\Lambda^*_{\mathcal{W}}(x) < \infty$. Let $\varphi \in C_c^\infty((0,\infty)$ and define $\lambda \in \Theta^*$ by $\varphi(t) = \lambda((t,\infty)), t \geq 0$. Then

$$-\int_0^\infty x(t)\dot{\varphi}(t)dt - 1/2 \parallel \varphi\|^2_{L^2} = \langle \lambda, x\rangle - \Lambda_{\mathcal{W}}(\lambda) \leq \Lambda^*_{\mathcal{W}}(x)$$

and so there exists a unique $\dot{x} \in L^2$ such that

$$-\int_0^\infty x(t)\dot{\varphi}(t)dt = \int_0^\infty \dot{x}(t)\varphi(t)dt.$$

Hence $x \in H^1$ and we have

$$\Lambda^*_{\mathcal{W}}(x) = \begin{cases} \frac{1}{2}\int_0^\infty \dot{x}(t)^2 dt & \text{if } x \in H^1 \\ \infty & \text{otherwise.} \end{cases}$$

Chapter 2

Entropy and relative entropy

In this chapter we will introduce the concept of relative entropy and show its connection with the entropy as it was defined in the previous chapter. Assume that there is a signed measure λ defined on E. Further assume we have probability measures $\mu, \nu \in \mathcal{P}(E)$ such that $\mu \ll \lambda$, $\nu \ll \lambda$ and we let g, f be the Radon-Nikodym derivatives of μ, ν, respectively, with respect to λ. Then we define the relative entropy of μ with respect to ν, denoted by $H(\nu \mid \mu)$ as follows:

$$H(\nu \mid \mu) = \begin{cases} \int_E f(t) \log \frac{f(t)}{g(t)} \lambda(dt) & \text{if } \nu \ll \mu \\ \infty & \text{otherwise} \end{cases}$$

Lemma 2.1 $\forall \mu, \nu \in \mathcal{P}(E)$ *we have*

$$H(\nu \mid \mu) \geq 0$$

with equality if and only if $f = g$ *a.e.* $[\lambda]$.

Proof.

If $H(\nu \mid \mu) = \infty$, there is nothing to prove. So assume $H(\nu \mid \mu) < \infty$, that is, $\nu \ll \mu$. Let $h(t) = \frac{f(t)}{g(t)}$. Then

$$\int_E f(t) \log \frac{f(t)}{g(t)} \lambda(dt) = \int_E h(t) \log h(t) \mu(dt)$$

Let $\varphi(t) = t \log t, t > 0$. Since $0 < h(x) < \infty$ *a.e.* $[\mu]$, a Taylor series expansion of φ gives :

$$\varphi(h(x)) = \varphi(1) + \dot{\varphi}(1)(h(x) - 1) + \frac{1}{2}\ddot{\varphi}(\xi_x)(h(x) - 1)^2$$

where ξ_x lies between $h(x)$ and 1, which implies $0 < \xi_x < \infty$. Since $\varphi(1) = 0$, $\dot{\varphi}(1) = 1$ and $\ddot{\varphi}(t) = t^{-1}$, we get

$$\int_E \varphi(h(x))\mu(dx) = \int_E (h(x) - 1)\mu(dx) + \frac{1}{2}\int_E \xi_x^{-1}(h(x) - 1)^2 \mu(dx)$$

$$= \int_E f(x)\lambda(dx) - 1 + \frac{1}{2}\int_E \frac{(h(x) - 1)^2}{\xi_x}\mu(dx)$$

$$= \frac{1}{2}\int_E \frac{(h(x) - 1)^2}{\xi_x}\mu(dx) \ \geq \ 0$$

and equality holds iff $h = \frac{f}{g} \equiv 1$ *a.e.* $[\lambda]$.

□

There is a variational formula for H as shown in the next lemma:

Lemma 2.2 $\forall\, \nu, \mu \in \mathcal{P}(E)$ *we have*

$$H(\nu \mid \mu) = \sup\{\int_E \varphi d\nu - \log \int_E e^{\varphi} d\mu \ ;: \ \varphi \in C_b(E, R)\}$$

Proof.

See Lemma 3.2.13 in Deuschel & Stroock [2].

□

Definition 2.1 We define the I-functional of μ on E as follows:

$$I_\mu(x) = \inf\{\ H(\nu \mid \mu)\ ;\ \nu \in \mathcal{P}(E),\ f = \frac{d\nu}{d\lambda} \text{ and } \int_E t d\nu = x\}$$

Now we can prove the following relationship between entropy and relative entropy:

Theorem 2.1 *Let E be a separable Banach space and assume $\mu \in \mathcal{P}(E)$. Let $x \in E$. Assume that* $\forall \alpha \in [0,\infty)$

$$\int_E e^{\alpha \|t\|_E} \mu(dt) < \infty$$

Then we have :

$$\Lambda_\mu^*(x) = I_\mu(x)$$

Proof.

The proof of this theorem relies heavily on the theory of large deviations. Let $\Omega \equiv E^{\mathbf{N}}$ with the product topology. Next, for $n \in \mathbf{N}$, we use $X_n : \Omega \longmapsto E$ to denote the n^{th} coordinate map . Let $\bar{S}_n = \frac{1}{n}\sum_{i=1}^n X_i$. For $\mu \in \mathcal{P}(E)$ let $P \equiv \mu^{\mathbf{N}}$ and let μ_n be the distribution of $\bar{S}_n$ under P. Now by a theorem of Donsker and Varadhan, (Theorem 3.3.11 in Deuschel & Stroock [2],) we know that Λ_μ^* is a good rate function for the large deviations of $\{\mu_n; n \geq 1\}$. But by Lemma 2.2.1 in Deuschel & Stroock [2] rate functions are unique, hence it suffices to show that I_μ is also a good rate function that governs the large deviations of $\{\mu_n, n \geq 1\}$. To this end let $\tilde{\mu}_n \in \mathcal{P}(\mathcal{P}(E))$ be the distribution of L_n under μ^n, where L_n is defined as follows:

$$\bar{\sigma} = (\sigma_1, .., \sigma_n) \in E^n \longmapsto L_n(\bar{\sigma}) \equiv \frac{1}{n}\sum_{i=1}^n \delta_{\sigma_i}$$

Then by Sanov's theorem, Theorem 3.2.17 in Deuschel & Stroock [2], $H(\cdot \mid \mu)$ is a good, convex rate function on $\mathcal{P}(E)$ and $\{\tilde{\mu}_n : n \geq 0\}$ satisfies the full large deviation principle with rate function $H(\cdot \mid \mu)$. Next define the mapping m : $\mathcal{P}(E) \longmapsto E$ by

$$m(\nu) = \int_E t\nu(dt)$$

In the next step we will construct an increasing sequence of closed sets Γ_L such that m is continuous on Γ_L $\forall L \geq 0$ and we will use this to prove that $\{\mu_n, n \geq 1\}$ satisfies the large deviation principle with rate function I_μ. To this end define $f : E \longmapsto R^+$ by

$$f(x) = \sup\{\alpha \parallel x \parallel_E - \log \int_E \exp[\alpha \parallel t \parallel_E]\mu(dt) \ : \ \alpha \geq 0\}$$

and $\forall\, L \geq 0$ let

$$\Gamma_L = \{\nu \in \mathcal{P}(E) \ : \ \int_E f(x)\nu(dx) \leq L\}$$

Then by Lemma 3.3.8 and Lemma 3.3.9 in Deuschel & Stroock [2] we have that Γ_L is closed $\forall\, L \geq 0$ and that m is continuous on Γ_L $\forall\, L \ \geq 0$. By the definition of the measures $\{\tilde{\mu}_n, n \geq 0\}$ we see that they are concentrated on elements of $\mathcal{P}(E)$ of the form $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{\sigma_i}$. For those we have for some $L \geq 0$:

$$\int_E f(x)\nu(dx) = \frac{1}{n}\sum_{i=1}^{n} f(\sigma_i) \leq L$$

Hence if we define $\Gamma_\infty = \bigcup_{L\geq 0}\Gamma_L$ we have

$$\tilde{\mu}_n(\Gamma_\infty) = 1$$

Let $X_1,\ X_2,\ \ldots$ be i.i.d. μ , then $f(X_1),\ f(X_2),\ \ldots$ is an i.i.d sequence of real-valued random variables with law $\mu \circ f^{-1}$. Now for $t \in R$ and $X \sim \mu$ we get :

$$\Lambda_{f(X)}(t) := \log E[e^{t \cdot f(X)}] \leq -\log(1-t) \ \ 0 \leq t < 1$$

where the inequality follows from Lemma 3.3.10 in Deuschel & Stroock [2].

Because $f \geq 0$, we have that $\Lambda_{f(X)}$ is finite in a neighborhood of 0. Hence by Lemma 1.2.3 (ii) in Deuschel & Stroock [2]

$$\Lambda^*_{f(X)}(x) \to \infty \text{ as } \mid x \mid \to \infty$$

For $t > 0$ and using Chebyshev's inequality with the exponential function we get:

$$P\{\frac{1}{n}\sum_{i=1}^{n} f(X_i) > L\} \leq \frac{E[\exp(t\sum_{i=1}^{n} f(X_i))]}{e^{tnL}}$$

Using the independence of the X_i's we get

$$\frac{1}{n}\log P\{\frac{1}{n}\sum_{i=1}^{n} f(X_i) > L\}$$

$$\leq \ \ -\sup\{tL - \log E[\exp(tf(X_1))] \ \ : \ t > 0\}$$

$$= -\Lambda^*_{f(X)}(L)$$

for L big enough so that the sup over R^+ is the same as the sup over R by Lemma 1.2.3 in Deuschel & Stroock [2], where the last equality follows from the fact that $f \geq 0$. So now we have

$$\limsup_{n\to\infty} \frac{1}{n} \log \tilde{\mu}_n(\Gamma_L^c)$$

$$= \limsup_{n\to\infty} \frac{1}{n} \log P\{\frac{1}{n}\sum_{i=1}^{n} f(X_i) > L\}$$

$$\leq -\Lambda^*_{f(X)}(L)$$

Now, there exists a sequence $\{m_n\}_{n=1}^{\infty} \subseteq C(\mathcal{P}(E), E)$ such that $m_n \mid_{\Gamma_n} = m \mid_{\Gamma_n}$. $\{\tilde{\mu}_n, n \geq 0\}$ satisfies the large deviation principle with rate function H, and so:

$$-\inf_{\Gamma_L^c} H(\cdot \mid \mu) \leq \limsup_{n\to\infty} \frac{1}{n} \log \tilde{\mu}_n(\Gamma_L^c) \leq -\Lambda^*_{f(X)}(L)$$

$$\Rightarrow \quad H(\nu \mid \mu) \;\geq\; \Lambda^*_{f(X)}(L) \quad \forall\, \nu \in \Gamma_L^c$$

We want to apply Lemma 2.1.4 in Deuschel & Stroock [2] to prove our assertion. To this end, we need the following :

$$\lim_{n\to\infty} \sup \;\{\; \rho(m_n(\nu), m(\nu)) \;:\; \nu \in \mathcal{P}(E),\ and\ H(\nu \mid \mu) \leq L\}$$

$$= \lim_{n\to\infty} \sup \;\{\; \rho(m_n(\nu), m(\nu)) \;:\; \nu \in \Gamma_n^c\ and\ H(\nu \mid \mu) \leq L\}$$

$$= 0 \quad \forall\, L \geq 0$$

The first equation follows from the choice of m_n because $m_n \mid_{\Gamma_n} = m \mid_{\Gamma_n}$ implies $\rho(m_n(\nu), m(\nu)) = 0$ for $\nu \in \Gamma_n$. The second equation holds because $\nu \in \Gamma_n^c$ implies

$$\Lambda^*_{f(X)}(n) \;\leq\; H(\nu \mid \mu) \;\leq\; L$$

Since $\Lambda^*_{f(X)}(n) \;\to \infty$ as $n \to \infty$ the set eventually becomes empty for a finite n. We also have that

$$\lim_{L\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log\tilde{\mu}_n(\{\nu\in\mathcal{P}(E)\ :\ \rho(m_L(\nu),(m(\nu))\geq\delta\})$$

$$\leq\lim_{L\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log\tilde{\mu}_n(\Gamma_L^c)=-\infty\quad\forall\delta>0$$

So the conditions of Lemma 2.1.4 are satisfied and we get that $\{\tilde{\mu}_n\circ m^{-1}\ ;\ n\geq 1\}$ satisfies the large deviation principle with rate function I_μ. But

$$m(\nu)=\int_E t\frac{1}{n}\sum_{i=1}^{n}\delta_{\sigma_i}(dt)=\frac{1}{n}\sum_{i=1}^{n}\sigma_i$$

and so $\tilde{\mu}_n\circ m^{-1}=\mu_n$ and we get that

$$\Lambda_\mu^*=I_\mu$$

□

Example 2.1 In this example we will use the I-functional described in the above theorem to find the entropy of a normal random variable on R with mean m and variance σ^2 . That is we have to find

$$\inf\{\int_{-\infty}^{\infty}f(t)\log\frac{f(t)}{g(t)}dt;\int_{-\infty}^{\infty}f(t)dt=1,\int_{-\infty}^{\infty}t\cdot f(t)dt=x\ ,\ f\geq 0\ a.e.\}$$

where

$$g(t)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2}\frac{(t-m)^2}{\sigma^2}\}$$

Let $T\ :\ C_0^\infty(R)\longmapsto R$, where C_0^∞ is the space of infinitely differentiable functions that vanish at infinity, be defined by

$$T(f)=\int_{-\infty}^{\infty}f(t)\log\frac{f(t)}{g(t)}dt$$

First we will find the Gateaux differential of T:

$$\lim_{\alpha\to 0}\frac{1}{\alpha}\{T(f+\alpha\psi)-T(f)\}=\int_{-\infty}^{\infty}[\log\frac{f(t)}{g(t)}+1]\cdot\psi(t)dt$$

So using Lagrange multipliers we get a necessary condition for the infimum:

$$\int_{-\infty}^{\infty}[\log\frac{f(t)}{g(t)}+1+\lambda x+\tau]\cdot\psi(t)dt=0$$

for all $\psi\in C_0^\infty(R)$. Hence using ψ's close to point masses we get

$$\log\frac{f(t)}{g(t)}+1+\lambda x+\tau=0\quad\forall t\in R$$

or

$$f(t)=g(t)\cdot e^{-(1+\lambda x+\tau)}=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}[(t-m)^2+2\sigma^2(1+\lambda x+\tau)]\}$$

$$=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(t-(m-\lambda\sigma^2))^2\}\cdot\exp\{\frac{1}{2}\lambda^2\sigma^2-\ \lambda m-\tau-1\}$$

But f is the density of a random variable with mean x, and so we get the following conditions for λ and τ :

$$\frac{1}{2}\lambda^2\sigma^2-\lambda m-\tau-1=0$$

$$m-\lambda\sigma^2=x$$

and then the infimum is attained for f_0, where f_0 is the density of a normal random variable with mean x and variance σ^2 . So

$$\Lambda^*_{N(m,\sigma^2)}(x)=\int_{-\infty}^{\infty}f_0(t)\log\frac{f_0(t)}{g(t)}dt$$

$$=\frac{1}{2\sigma^2}\int_{-\infty}^{\infty}[(t-m)^2-(t-x)^2]f_0(t)dt$$

$$=\frac{1}{2}\frac{(x-m)^2}{\sigma^2}$$

Example 2.2 In many cases it is impossible to find a closed expression for Λ^*_μ, but it is relatively easy to find upper bounds using the relative entropy. As an example, consider the uniform distribution on the points $\{1,2,..,n\}$ and let μ be the measure associated with this distribution. Then

$$\Lambda_\mu(\lambda)\ =\ \frac{1}{n}\sum_{k=1}^{n}e^{\lambda k}\quad\lambda\in R$$

Hence

$$\Lambda_\mu^*(x) = \sup\left\{\lambda x - \log(\sum_{k=1}^n e^{\lambda k}) + \log n \ : \ \lambda \in R\right\}$$

$$\frac{d}{d\lambda}\left\{\lambda x - \log(\sum_{k=1}^n e^{\lambda k}) + \log n\right\} = x - \frac{\sum_{k=1}^n k e^{\lambda k}}{\sum_{k=1}^n e^{\lambda k}} = 0$$

$$\Rightarrow \sum_{k=1}^n (x-k)e^{\lambda k} = 0$$

and this equation in general can not be solved explicitly if $n \geq 5$. But if Y has a distribution on $\{1,..,n\}$, say $P\{Y=k\} = p_k$ and if $EY = x$, then denoting by ν the measure associated with Y we get

$$\Lambda_\mu^*(x) \leq H(\nu \mid \mu) = \log n + \sum_{k=1}^n p_k \log p_k$$

and any choice of $\{p_k\}_{k=1}^n$ with $\sum_{k=1}^n p_k = 1$ and $\sum_{k=1}^n k \cdot p_k = x$ will give an upper bound for $\Lambda_\mu^*(x)$.

Example 2.3 Let $X \sim N(0,1)$ and $Y \sim N(x, \sigma^2)$. Let $H(Y|X)$ be the relative entropy of the measures of X and Y. Then

$$\Lambda_X^*(x) \leq H(Y \mid X)$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(t-x)^2] \log \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}(t-x)^2]}{\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}t^2]} dt$$

$$= E[\frac{1}{2\sigma^2}((\sigma^2-1)Y^2 + 2xY - x^2) - \log\sigma]$$

$$= \frac{1}{2}(\sigma^2 + x^2 - 1) - \log\sigma$$

Taking the infimum over all σ we get

$$\Lambda_X^*(x) \leq \frac{1}{2}x^2$$

We already computed $\Lambda_X^*(x) = \frac{1}{2}x^2$, so in this case the bound is sharp. We can also see that in some cases it suffices to take the infimum of H over a much smaller class of functions.

Chapter 3

Entropy of measures on finite dimensional spaces

In this chapter we will study the entropy of random variables on R^d. Note that on R^d the natural Banach space pairing amounts to the inner product :

$$\langle \lambda, x \rangle \; = \; \lambda' x \; = \; \sum_{i=1}^{d} \lambda_i x_i$$

Throughout this chapter we assume that X is a random variable on R^d, that it has law μ and that $E[e^{\alpha \|X\|}] < \infty \; \forall \alpha \geq 0$.

First we will study the relationship between the set of values for which the entropy Λ_μ^* is finite and the support of the measure μ , that is, the set of all values $x \in R^d$ with the property that for any open set G with $x \in G$ we have $\mu(G) > 0$.

In the following we denote by C_μ the convex hull of the support of μ.

Lemma 3.1 *If $x \notin \bar{C}_\mu$, the closure of C_μ, then $\Lambda_\mu^*(x) = \infty$*

Proof.

By the Hahn-Banach theorem there exists $\lambda_0 \in E^*$ and an $a \in R$ such that

$$\langle \lambda_0, z \rangle \leq a \quad \forall \, z \in \bar{C}_\mu$$

and

$$\langle \lambda_0, x \rangle > a$$

So we get

$$\Lambda_\mu^*(x) \geq \langle n\cdot\lambda_0, x\rangle - \log\int_{\bar{C_\mu}} e^{\langle n\cdot\lambda_0, z\rangle}\mu(dz)$$

$$\geq \; n\cdot(\langle\lambda_0, x\rangle - a) \;\rightarrow\; \infty \;\; as \;\; n \;\rightarrow\; \infty$$

because $\mu(C_\mu^c) = 0$. Hence

$$\Lambda_\mu^*(x) \; = \; \infty$$

□

Note that the proof did not depened on the finite dimensionality of the space E. So this lemma is also true in the infinite-dimensional case. Let riC_μ denote the interior of C_μ relative to the smallest linear manifold on which μ is concentrated. Then we have the following lemma:

Lemma 3.2 *If $x \in riC_\mu$, then $\Lambda_\mu^*(x) < \infty$*

Proof.

Let $H_{\lambda,x} = \{z \in R^d \; : \; \langle\lambda, z\rangle \; > \; \langle\lambda, x\rangle\}$. Then

$$\langle\lambda, x\rangle - \log\int_{R^d} e^{\langle\lambda,z\rangle}\mu(dz)$$

$$\leq \;\; \langle\lambda, x\rangle - \log\int_{H_{\lambda,x}} e^{\langle\lambda,z\rangle}\mu(dz)$$

$$\leq \;\; \langle\lambda, x\rangle - \log\{e^{\langle\lambda,x\rangle}\cdot\mu(H_{\lambda,x})\} \;\; = \;\; -\log\{\mu(H_{\lambda,x})\}$$

Let $\{y\}^\delta \;\equiv\; \{x \;\in\; R^d \;\; : \; \parallel \; x \;-\; y \; \parallel < \; \delta\}$ be an open set in R^d. Then for $x \in riC_\mu \;\; \exists \; \{y_\alpha\}_{\alpha\in A}$ and a $\delta > 0$ such that $\mu(\{y_\alpha\}^\delta) > 0 \;\; \forall\alpha \in A$, and $\forall \; \lambda \in R^d \;\; \exists\alpha \in A$ such that $y_\alpha \in H_\lambda$. As a matter of fact, in R^d we can find a finite A with the above property, and so

$$\mu(H_{\lambda,x}) \geq \min\{\mu(\{y_\alpha\}^\delta) \; > \; 0 \;\; \forall\,\lambda \in R^d$$

Hence

$$\Lambda_\mu^*(x) \;\; \leq \sup\{-\log\mu(H_{\lambda,x})\} < \infty$$

□

Now we know that for points in the interior of the convex hull of the support of the measure the entropy is finite. We can also say something about where the supremum might be attained:

Lemma 3.3 *Let $\{\mu_\alpha\}_{\alpha\in I}$, where I is some index set, be a collection of probability measures on R^d, and let $x \in R^d$. Assume that there exists a finite number of points $a_1, \cdots, a_m$ in R^d and a $\delta > 0$ such that x lies in the interior of the convex hull of the a_i's and, for all $t \in R^d$ and all a_i's, we have*

$$\mu_\alpha(\{y : t'y > t'a_i\}) \geq \delta > 0.$$

Then there exists an $L > 0$, independent of α, with

$$\Lambda^*_{\mu_\alpha}(x) = \sup\left\{t'x - \log \int_{R^d} e^{t'y}\mu_\alpha(dy) \;:\; ||t|| < L\right\}$$

Proof.

Let $t \in R^d$ with $||t|| = 1$. The assumption that x be in the interior of the convex hull of the a_i implies that $t'x < t'a_i$ for some i. Then

$$\begin{aligned} & t'x - \log \int_{R^d} e^{t'y}\mu_\alpha(dy) \\ \leq\; & t'x - \log \int_{\{y:t'y>t'a_i\}} e^{t'y}\mu_\alpha(dy) \\ \leq\; & t'(x - a_i) - \log \mu_\alpha(\{y : t'y > t'a_i\}) \\ \leq\; & t'(x - a_i) - \log \delta \end{aligned}$$

but then

$$(rt)'x - \log \int_{R^d} e^{(rt)'y}\mu_\alpha(dy) \;\leq\; rt'(x - a_i) - \log \delta \;<\; 0$$

for $r \in R$ big enough. Here the right side does not depend on α, and so r does not depend on α. Of course the entropy is positive, and so all values of t that make the expression inside the supremum negative can be ignored. By the above derivation that is true for all t with $||t|| > r$.

□

Let X be a random variable on R^d. By Lemma 1.1 we know that $EX = m$ implies $\Lambda^*_X(m) = 0$. Now we will show that in finite dimensions the converse is also true.

Lemma 3.4 *Let $X = (X_1, .., X_d)$ be a random vector on R^d and suppose $\Lambda^*_{X_k}(m_k) = 0$. Then $(EX)_k = m_k$*

Proof.

First note that the overall assumption $E[e^{\alpha\|X\|}] < \infty \; \forall\alpha \geq 0$ implies the existence of EX and that

$$EX = \frac{d}{dt}E[e^{t'X}] \mid_{t=(0,\cdots,0)}$$

Then $\Lambda^*_{X_k}(m_k) = 0$

$$\Rightarrow \quad tm_k - \log E[e^{tX_k}] \; \leq 0 \;\; \forall t \in R$$

$$\Rightarrow \quad E[e^{tX_k}] \; \geq \; e^{tm_k} \;\; \forall t \in R$$

Let $t > 0$, then

$$E[\frac{e^{tX_k} - 1}{t}] \; \geq \; \frac{e^{tm_k} - 1}{t} \quad \forall t > 0$$

$$(EX)_k \geq \frac{d}{dt}(e^{tm_k}) \mid_{t=0} = m_k$$

Analogously with $t < 0$ we get $(EX)_k \leq m_k$, and so we have

$$(EX)_k = m_k$$

□

Remark

If $\Lambda^*_X(m) = 0$ for some $m \in R^d$, then $EX = m$.

In the next lemma we show that this relationship between means of random variables and zeros of their entropies also holds for a sequence of random variables whose means converge.

Lemma 3.5 *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables on R^d such that $f(x) := \lim_{n\to\infty} \Lambda^*_{X_n}(x)$ exists for all $x \in R^d$, where infinity is admissible as a limit. Assume there exists an $\bar{x} \in R^d$ with*

$$f(\bar{x}) = 0 \quad and \quad f(x) > 0 \ \ \forall x \neq \bar{x}$$

Then

$$x_n \ := \ E[X_n] \to \bar{x}$$

Proof.

Using Jensen's inequality with e^x we see that the assumption $Ee^{\alpha \|X_n\|} < \infty \ \forall \alpha \in R$ implies that $E \parallel X_n \parallel < \infty$. Now let $\Gamma = \{x \ : \ f(x) < \infty\}$, then f is continuous on Γ because it is lower semicontinuous and convex.

Next note that $\Lambda^*_{X_n}(x_n) = 0 \ \forall n \geq 1$ by Lemma 1.2 part 4. Now assume that $x_n \not\to \bar{x}$. This implies that there exists a subsequence $\{n_k\}$ and a $\delta > 0$ such that $\parallel x_{n_k} - \bar{x} \parallel \geq \delta$ for all $k \geq 1$. Let

$$m = \inf\{f(x) \ : \ x \in \Gamma \ and \ \parallel x - \bar{x} \parallel = \delta\}$$

Then $m > 0$ by compactness and continuity, and using the convexity of f we have $f(x_{n_k}) \geq m$ for all $k \geq 1$. Note that this is trivially true if $x_{n_k} \notin \Gamma$. But now there exists an $M \geq 1$ such that

$$\Lambda^*_{X_m}(x_{n_k}) \geq \frac{m}{2} > 0 \quad \forall m \geq M$$

which is a contradiction to $\Lambda^*_{X_{n_k}}(x_{n_k}) = 0 \ \forall k \geq 1$.

□

In the next lemma we will study the relationship between the entropy of a measure and the entropy of its marginals. The inequality proven in the next lemma will also be true for measures on an infinite dimensional space and will be useful in Chapter 4 when we consider the entropy of stochastic processes.

Lemma 3.6 *Let $\mu^{(k)}$ be the k'th marginal of μ and let $p_1, .., p_d$ be positive numbers with $\frac{1}{p_1} + ... + \frac{1}{p_d} = 1$. Then*

$$\Lambda^*_\mu(x) \geq \sum_{k=1}^{d} \frac{1}{p_k} \Lambda^*_{\mu^{(k)}}(x_k)$$

where $x = (x_1, .., x_d)^T$. We also have $\Lambda^*_\mu(x) \geq \Lambda^*_{\mu^{(k)}}(x_k)$ for all $1 \leq k \leq d$.

Proof.

For the first part, let $z = (z_1, .., z_d)^T$. Then using Hölder's inequality, we get

$$\int_{R^d} \exp[t'z]\mu(dz) \leq \prod_{k=1}^{d} (\int_R \exp[p_k t_k z_k]\mu^{(k)}(dz_k))^{\frac{1}{p_k}}$$

$$\begin{aligned} \Rightarrow \quad & t'x - \log \int_{R^d} \exp[t'z]\mu(dz) \\ \geq \quad & \sum_{i=1}^{d} \left\{ t_i x_i - \frac{1}{p_i} \log \int_R \exp[p_i t_i z_i]\mu^{(i)}(dz_i) \right\} \\ = \quad & \sum_{i=1}^{d} \frac{1}{p_i} \left\{ p_i t_i x_i - \log \int_R \exp[p_i t_i z_i]\mu^{(i)}(dz_i) \right\} \end{aligned}$$

and the first assertion follows by taking the sup over t on both sides. The second follows from the first if we let $p_k \to 1$, because then $p_j \to \infty \; \forall j \neq k$.

□

The rest of this chapter will be devoted to the study of the relationship between weak convergence of measures and pointwise convergence of entropies.

Lemma 3.7 *Let $\{\mu_n, \mu\} \in \mathcal{P}(R^d)$ and $\mu_n \Rightarrow \mu$. Assume that $x \in C^0_\mu$, the interior of C_μ in R^d. Then $\exists\, N \geq 1$ such that $x \in C^0_{\mu_n}$ $\forall n \geq N$.*

Proof.

$x \in C^0_\mu \;\Rightarrow\; \exists\; \{y_k\}_{k=1}^m\; ; \epsilon, \delta\; > 0$ such that $\mu(\{y_k\}^\delta) \geq \epsilon > 0$, $1 \leq k \leq m$ and x lies in the interior of the convex hull of the y_k's. But by the weak convergence of $\{\mu_n\}$ and the Portmanteau lemma we have

$$\liminf_{n\to\infty} \mu_n(\{y_k\}^\delta) \;\geq\; \mu(\{y_k\}^\delta) \;\geq \epsilon > 0$$

$$\Rightarrow \quad \{y_k\}^\delta \in C^0_{\mu_n} \quad n \geq N \quad 1 \leq k \leq m$$

$$\Rightarrow \quad x \in C^0_{\mu_n}$$

□

In the next lemma we will show that our assumptions on the measures are enough so that weak convergence implies convergence in mean of some functions of the measures. This will be needed in the later theorems.

Lemma 3.8 *Let $\{\mu_n, \mu\} \in \mathcal{P}(R^d)$ with $\mu_n \Rightarrow \mu$. Let ψ be a measurable function such that there exists $\beta > 0$ with*

$$\psi(t)e^{-\beta|t|} \to 0 \ \ as \ \ t \to \infty$$

boundedly and pointwise, and assume that our basic assumption holds for all $\mu_n \circ \psi^{-1}$ and for $\mu \circ \psi^{-1}$, namely

$$\int_{R^d} \exp(\alpha \parallel t \parallel)\, (\mu_n \circ \psi^{-1})(dt) \quad < \infty \ \ \forall n \geq 1$$

and

$$\int_{R^d} \exp(\alpha \parallel t \parallel)\, (\mu \circ \psi^{-1})(dt) \quad < \infty$$

Then $\{\mu_n \circ \psi^{-1}\}$ is uniformly integrable and

$$\int_{R^d} \psi(t)\mu_n(dt) \ \to \ \int_{R^d} \psi(t)\mu(dt)$$

Proof.

It suffices to show this lemma for $\psi(t) = e^{\beta t}$. By Propositions 2.2 and 2.3, Appendix 3 in in Ethier, Kurtz [4] it sufficies to find an increasing convex function φ on $[0,\infty)$ with $\frac{\varphi(t)}{t} \to \infty$ as $t \to \infty$ and

$$\sup_n \int \varphi(|t|)\mu_n(dt) \ < \ \infty$$

Let $\varphi(t) = t^{\frac{\alpha}{\beta}}$ with $\alpha > \beta$. Then φ clearly fulfills the first three conditions. Let $L > 0$. By weak convergence and the usual approximation with continuous functions we have

$$\begin{aligned} & \int \varphi(|t|) I_{[-L,L]}(t)(\mu_n \circ \psi^{-1})(dt) \\ = & \int \varphi(e^{\beta t}) I_{[-L,L]}(e^{\beta t}) \mu_n(dt) \\ \rightarrow & \int \varphi(e^{\beta t}) I_{[-L,L]}(e^{\beta t}) \mu(dt) \\ \leq & \int e^{\alpha |t|} \mu(dt) \; =: C \; < \; \infty \end{aligned}$$

Hence by the monotone convergence theorem we get

$$\limsup_{n \to \infty} \int \varphi(|t|) \mu_n(dt) \;\leq\; 2C \;<\; \infty$$

□

Note that with $\psi(t) = t$ the last lemma implies the convergence of the means and with $\psi(t) = e^{\alpha t}$ it implies convergence of the moment generating functions.

Theorem 3.1 *Let $\{\mu_n, \mu\} \in \mathcal{P}(R^d)$ and denote by ρ_σ the d-dimensional normal random variable with mean vector 0 and covariance matrix $\sigma^2 I_d$. Then if $\mu_n \Rightarrow \mu$ we have*

$$\lim_{\sigma \to 0} \lim_{n \to \infty} \Lambda^*_{\mu_n * \rho_\sigma}(x) = \Lambda^*_\mu(x) \;\; \forall x \in R^d$$

where $\mu_n * \rho_\sigma$ is the convolution of μ_n and ρ_σ

Proof.

By the previous lemma $\eta_n := \int z \mu_n(dz) \rightarrow \int z \mu(dz) =: \eta$. Let $X_n \sim \mu_n$ and $Z_\sigma \sim N(0, \sigma^2 I_d)$ with X_n independent of Z_σ $\forall n \geq 1$, $\sigma > 0$. Note that

$$\begin{aligned} 0 \;=\; \Lambda^*_{\mu_n}(\eta_n) \;&=\; \sup\left\{ t'\eta_n \;-\; \Lambda_{\mu_n}(t) \;:\; t \in R^d \right\} \\ &\geq\; t'\eta_n \;-\; \Lambda_{\mu_n}(t) \;\;\; \forall t \in R^d \end{aligned}$$

and so

$$\Lambda_{\mu_n * \rho_\sigma}(t) = \log E[\exp\{t'(X_n + Z_\sigma)\}]$$

$$= \log E[\exp\{t'X_n\}] + \log E[\exp\{t'Z_\sigma\}]$$

$$= \Lambda_{\mu_n}(t) + \frac{1}{2}\sigma^2 \parallel t \parallel^2$$

$$\geq t'\eta_n + \frac{1}{2}\sigma^2 \parallel t \parallel^2$$

where we also used the result of Example 1.1. Then

$$\frac{\Lambda_{\mu_n * \rho_\sigma}(t)}{\parallel t \parallel} \rightarrow \infty \text{ as } \parallel t \parallel \rightarrow \infty$$

We also see that $\Lambda_{\mu_n * \rho_\sigma}$ is strictly convex, and so there exists a unique $t_{n,\sigma} \in R^d$ such that for fixed $x \in R^d$

$$\bigtriangledown\{t'_{n,\sigma}\, x - \Lambda_{\mu_n * \rho_\sigma}(t_{n,\sigma})\} = x - \bigtriangledown\Lambda_{\mu_n * \rho_\sigma}(t_{n,\sigma}) = 0$$

By weak convergence and Lemma 3.8

$$\lim_{n\rightarrow\infty} \Lambda_{\mu_n * \rho_\sigma}(t) = \Lambda_{\mu * \rho_\sigma}(t)$$

and by strict convexity plus differentiability

$$\lim_{n\rightarrow\infty} \bigtriangledown\Lambda_{\mu_n * \rho_\sigma}(t) = \bigtriangledown\Lambda_{\mu * \rho_\sigma}(t) \quad \forall t \in R^d$$

Also by strict convexity the inverse function of $\bigtriangledown\Lambda_{\mu_n * \rho_\sigma}$, denoted by $(\bigtriangledown\Lambda_{\mu_n * \rho_\sigma})^{-1}$ exists, is continuous and converges pointwise to the inverse function of $\bigtriangledown\Lambda_{\mu * \rho_\sigma}$, denoted by $(\bigtriangledown\Lambda_{\mu * \rho_\sigma})^{-1}$. But

$$t_{n,\sigma} = (\bigtriangledown\Lambda_{\mu_n * \rho_\sigma})^{-1}(x) \rightarrow (\bigtriangledown\Lambda_{\mu * \rho_\sigma})^{-1}(x) =: t_\sigma$$

as $n \rightarrow \infty$ and $\forall\, \sigma > 0$. Hence

$$\lim_{n\rightarrow\infty} \Lambda^*_{\mu_n * \rho_\sigma}(x)$$

$$= \lim_{n\rightarrow\infty} \{t'_{n,\sigma}\, x - \Lambda_{\mu_n * \rho_\sigma}(t_{n,\sigma})\}$$

$$= \quad t_\sigma' \, x - \Lambda_{\mu*\rho_\sigma}(t_\sigma) = \Lambda^*_{\mu*\rho_\sigma}(x)$$

Now

$$\Lambda^*_{\mu*\rho_\sigma}(x) \ \geq \ t'x - \Lambda_\mu(t) - \frac{1}{2}\sigma^2 \parallel t \parallel^2 \quad \forall\, t \in R^d$$

$$\Rightarrow \quad \liminf_{\sigma \to 0} \Lambda^*_{\mu*\rho_\sigma}(x) \ \geq \ \Lambda^*_\mu(x)$$

but $\frac{1}{2}\sigma^2 \parallel t \parallel^2 \geq 0$ and so

$$\Lambda^*_{\mu*\rho_\sigma}(x) \ \leq \ \Lambda^*_\mu(x)$$

and we have

$$\lim_{\sigma \to 0} \Lambda^*_{\mu*\rho_\sigma}(x) = \Lambda^*_\mu(x)$$

□

Remark

The statement of the previous theorem is not true without the convolution with ρ_σ. As an example, consider $\mu_n = \delta_{\frac{1}{n}}$, the point mass at $\frac{1}{n}$ on R. Then $\mu_n \Rightarrow \delta_0$, but

$$\Lambda^*_{\mu_n}(0) = \sup\{-\frac{t}{n} \ : \ t \in R\} \ = \ \infty \quad \forall n \geq 1$$

On the other hand, $\Lambda^*_{\delta_0}(0) = 0$, and so

$$\Lambda^*_{\mu_n}(0) \not\to \Lambda^*_{\delta_0}(0)$$

Theorem 3.2 *Let $\{\mu_n, \mu\} \in \mathcal{P}(R^d)$ with $\mu_n \Rightarrow \mu$. Let $x \in R^d$ be such that $x \notin \delta C_\mu$, the boundary of C_μ. Then*

$$\Lambda^*_{\mu_n}(x) \to \Lambda^*_\mu(x) \quad \textit{as} \ \ n \to \infty$$

Proof.

If $x \notin \bar{C}_\mu$, then $\Lambda^*_\mu(x) = \infty$. But also $\exists \lambda \in R^d$, $a \in R$ such that

$$\mu(H_\lambda) = 1$$

where $H_\lambda = \{z \in R^d : \lambda' z < a \le \lambda' x\}$ is an open set, and by the Portmonteau lemma the weak convergence implies that

$$\limsup_{n\to\infty} \mu_n(\{z : \lambda' z \ge a\}) \le \mu(H_\lambda^c) = 0$$

Now for any $m \ge 1$

$$\begin{aligned} &(m\lambda)'x - \log E[e^{(m\lambda)'X_n}] \\ &= m\lambda' x - \log\left(\int_{H_\lambda} e^{m\lambda' z}\mu_n(dz) + \int_{H_\lambda^c} e^{m\lambda' z}\mu_n(dz)\right) \\ &\ge m\lambda' x - \log\left(e^{ma}\mu_n(H_\lambda) + \int_{H_\lambda^c} e^{m\lambda' z}\mu_n(dz)\right) \\ &\ge m\lambda' x - ma - \log(\mu_n(H_\lambda)) - \frac{\int_{H_\lambda^c} e^{m\lambda' z}\mu_n(dz)}{e^{ma}\mu_n(H_\lambda)} \end{aligned}$$

because by Taylor's theorem

$$\log(x+y) = \log x + \frac{1}{x}y - \frac{1}{2}\frac{1}{\xi^2}y^2 \le \log x + \frac{y}{x}$$

where ξ lies between x and $x+y$. Pick $M > 0$. Then there exists $m \ge 1$ such that

$$m(\lambda' x - a) > 2M$$

and there exists $N \ge 1$ such that for all $n \ge N$

$$\frac{\int_{H_\lambda^c} e^{m\lambda' z}\mu_n(dz)}{e^{ma}\mu_n(H_\lambda)} < M$$

This is possible because $\int_{H_\lambda^c} e^{m\lambda' z}\mu_n(dz) \to 0$ as $n \to \infty$ by weak convergence.

$$\Lambda_{\mu_n}^*(x) \ge M \quad \forall n \ge N$$

But M was arbitrary, and so $\Lambda_{\mu_n}^*(x) \to \infty$ as $n \to \infty$.

So now assume $x \in C_\mu^0$. This means there exist $a_1, \cdots, a_m$ and $\delta > 0$ such that x lies in the interior of the convex hull of the a_i's and for all $t \in R^d$ we have

$\mu(\{y : t'y > t'a_i\}) \geq \delta$. By the weak convergence and the Portmanteau lemma we have

$$\liminf_{n\to\infty} \mu_n(\{y : t'y > t'a_i\} \geq \delta$$

Hence by Lemma 3.3 we know there exists $N \geq 1$ and an $L > 0$ such that

$$\Lambda^*_{\mu_n}(x) = t'_n x - \log \int_{R^d} e^{t'_n y} \mu_n(dy) \quad ||t_n|| < L$$

then there exists a convergent subsequence $t_{n_k} \to t_0$ such that

$$\limsup_{k\to\infty} \Lambda^*_{\mu_{n_k}}(x) = \limsup_{k\to\infty} \left\{ t'_{n_k} x - \log \int_{R^d} e^{t'_{n_k} y} \mu_{n_k}(dy) \right\}$$

$$= \; t'_0 x - \log \int_{R^d} e^{t'_0 y} \mu(dy) \; \leq \; \Lambda^*_\mu(x)$$

where the convergence of the moment generating functions follows from Lemma 3.8. The opposite inequality follows from the usual estimate with the sup and so the theorem is proven.

□

Lemma 3.9 *Let $\{\mu_n\} \subset \mathcal{P}(R^d)$ and define*

$$f(x) = \liminf_{n\to\infty} \Lambda^*_{\mu_n}(x) \;\; \forall x \in R^d$$

If there exists a bounded set $\Gamma \subseteq R^d$ such that $f(x) = 0 \; \forall x \in \Gamma$ and $f(x) > 0 \; \forall x \notin \Gamma$, then $\{\mu_n\}$ is tight.

Proof.

First we note that as the liminf of convex functions, f is convex, and because Γ is bounded, we have by convexity $f(x) \to \infty$ as $\| x \| \to \infty$. Denote by $\mu_n^{(k)}$ the k^{th} marginal of μ_n. Then for $x \in \Gamma$

$$0 \;\leq\; f^{(k)}(x_k) \equiv \liminf_{n\to\infty} \Lambda^*_{\mu_n^{(k)}}(x_k) \;\leq\; \liminf_{n\to\infty} \Lambda^*_{\mu_n}(x) \;=\; f(x) \;=\; 0$$

$$\Rightarrow \quad f^{(k)}(x_k) = 0 \;\; \forall \, x \in \Gamma \; ; \; 1 \leq k \leq d$$

Next we will show that $f^{(k)}(x) \to \infty$ as $| x | \to \infty$. Assume this is wrong. Then without loss of generality there exists a sequence $\{x_m\}_{m=1}^{\infty}$ with $x_m \to \infty$ such that $f^{(k)}(x_m) \leq M < \infty$. Then

$$\liminf_{n\to\infty} \Lambda^*_{X_n^{(k)}}(x) \leq M \quad \forall x_0 < x < x_m$$

where $x_0 = \bar{x}_k$ for some $\bar{x} \in \Gamma$ and $\forall m \geq 1$. That means there exists a subsequence $\{n_j\}$ such that

$$\Lambda^*_{X_{n_j}^{(k)}}(x) \leq 2M \quad \forall x_0 \leq x \leq x_m \; \forall j \geq J.$$

$$\Rightarrow \; tx - \log E[\exp\{tX_{n_j}^{(k)}\}] \leq 2M$$

$$\Rightarrow \; E[\exp\{tX_{n_j}^{(k)}\}] \geq e^{tx-2M}$$

$\forall x_0 \leq x \leq x_m$, $\forall j \geq J$, $\forall t \in R$. But $x_m \to \infty$, and so

$$E[\exp\{tX_{n_j}^{(k)}\}] = \infty$$

$\forall t > 0$ and $j \geq J$. But by our overall assumption $E[e^{t'X}] < \infty$, so using $t = (0,..,0,t_k,0..,0)^T$ with $t_k > 0$ we get a contradiction. Next we will show that all the moment generating functions of the marginals are bounded uniformly in n on a neighborhood of 0 : $f^{(k)}(x) \to \infty$ as $| x | \to \infty$ implies that there exists an $\tilde{x} > 0$ such that $f(\tilde{x}) > 3$. Hence there exists an N such that $\forall n \geq N$ $\Lambda^*_{\mu_n}(\tilde{x}) > 2$, and that means there exist a $\tilde{t}_n$ such that

$$\tilde{t}_n\tilde{x} - \log E[e^{\tilde{t}_n X_n^{(k)}}] \; \geq \; 1$$

$$\Rightarrow \quad E[e^{\tilde{t}_n X_n^{(k)}}] \; \leq \; e^{\tilde{t}_n\tilde{x}-1} \; < \infty$$

Note that $\liminf_{n\to} \tilde{t}_n \geq 0$ because

$$1 \leq \tilde{t}_n\tilde{x} - \log E[e^{\tilde{t}_n X_n^{(k)}}]\tilde{x} - \log E[e^{\tilde{t}_n X_n^{(k)}}] \leq \tilde{t}_n(\tilde{x} - E[X_n^{(k)}])$$

Now let $\tilde{t} = \liminf_{n\to} \tilde{t}_n$. By Lemma 1.2.3 in Deuschel & Stroock [2] we know that $\tilde{t} > 0$. Let $0 < t < \tilde{t}$. Then $\tilde{t}/t > 1$, and with Jensen's inequality we get

$$E[e^{tX_n^{(k)}}] \leq (E[e^{\tilde{t}X_n^{(k)}}])^{\frac{t}{\tilde{t}}} < \infty$$

$\forall t \in (0, \tilde{t})$ and $\forall n \geq N$. Using $\tilde{x} < 0$, we can get the same result for $-\delta < t < 0$ and so the claim is proven. Now we are in a position to prove the tightness of the marginals $\{\mu_n^{(k)}\}$. Using Chebyshev's inequality with the exponential function we get

$$\mu_n^{(k)}([R,\infty)) \leq \frac{E[e^{t_1 X_n^{(k)}}]}{e^{t_1 R}} \leq Me^{-t_1 R}$$

for some $t_1 > 0$ and

$$\mu_n^{(k)}((-\infty,-R]) \leq \frac{E[e^{-t_2 X_n^{(k)}}]}{e^{-t_2 R}} \leq Me^{t_2 R}$$

for some $t_2 < 0$.

$$\Rightarrow \quad \limsup_{n\to\infty} \mu_n^{(k)}([-R,R]^c) \;\leq\; M(e^{-t_1 R} + e^{t_2 R}) \;\to 0$$

as $R \to \infty$. This proves the tightness of $\{\mu_n^{(k)}\}$ for $1 \leq k \leq d$, and the tightness of $\{\mu_n\}$ follows by Ethier, Kurtz [4].

□

Example 3.1 Let μ_n be the law of a bivariate normal with mean 0 and covariance matrix

$$\Sigma = \begin{bmatrix} n & 1 \\ 1 & 1 \end{bmatrix}$$

Then using example 1.1 we get

$$\Lambda_{\mu_n}^*((x,y)^T) = \frac{1}{2(n-1)}\left(x^2 - 2xy + ny^2\right) \to \frac{1}{2}y^2 \neq 0$$

$\forall (x,y)^T$ with $x = 0$. Here of course $\{\mu_n\}$ is not tight and the conditions of the lemma are not fulfilled.

Theorem 3.3 *Let $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(R^d)$ and assume that the limit $f(x) = \lim_n \Lambda^*_{\mu_n}(x)$ exists. Here infinity is allowed as a limit. Assume there exists $\bar{x} \in R^d$ with $f(\bar{x}) = 0$ and $f(x) > 0 \;\; \forall x \neq \bar{x}$. Then there exists a $\mu \in \mathcal{P}(R^d)$ such that $f(x) = \Lambda^*_\mu(x) \;\; \forall x \in R^d$ and $\mu_n \Rightarrow \mu$.*

Proof.

As the limit of convex functions, f is convex, and the assumptions imply that $f(x) \to \infty$ as $\| x \| \to \infty$. Hence by Lemma 3.6, $\{\mu_n\}$ is tight. That means we can find a convergent subsequence, say $\{\mu_{n_k}\}$ with $\mu_{n_k} \Rightarrow \mu$ for some $\mu \in \mathcal{P}(R^d)$. Let $\rho_\sigma = N(0, \sigma^2 I)$. Then by Theorem 1.1:

$$\lim_{\sigma \to 0} \lim_{k \to \infty} \Lambda^*_{\mu_{n_k} * \rho_\sigma}(x) = \Lambda^*_\mu(x) \quad \forall x \in R^d$$

But $\lim_{k\to\infty} \Lambda^*_{\mu_{n_k}}(x) = f(x)$ and so we have $f(x) = \Lambda^*_\mu(x)$.

If $\{\mu_{n_j}\}$ is another subsequence with $\mu_{n_j} \Rightarrow \nu$, then

$$\Lambda^*_\mu(x) = f(x) = \Lambda^*_\nu(x) \;\; \forall x \in R^d$$

and so $\mu = \nu$, and the theorem is proven.

□

Example 3.2 Let $X, X_1, X_2, ..$ be independent and identically distributed real valued random variables with mean m. Let $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Then

$$E[e^{\lambda S_n}] \;=\; (E[e^{\frac{\lambda}{n} X}])^n$$

and so

$$\left\{\lambda x - \log E[e^{\lambda S_n}]\right\} \;=\; n \cdot \left\{\frac{\lambda}{n} x - \log E[e^{\frac{\lambda}{n} X}]\right\} \quad \forall \lambda \in R$$

$$\Rightarrow \quad \Lambda^*_{S_n}(x) \;=\; n \cdot \Lambda^*_X(x) \;\to\; \Lambda^*_{\delta_m}(x)$$

where

$$\Lambda^*_{\delta_m}(x) = \begin{cases} 0 & \text{if } x = m \\ \infty & \text{otherwise} \end{cases}$$

$$\Rightarrow \quad \frac{1}{n}\sum_{k=1}^{n} X_k \;\Rightarrow\; \delta_m$$

In the next example we will prove the central limit theorem for independent and identically distributed random variables.

Example 3.3 Let $X, X_1, X_2, \ldots$ be independent and identically distributed real valued random variables with mean 0 and variance 1, and let

$$S_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k$$

Then

$$\begin{aligned} \Lambda_{S_n}(t) &= \log E[e^{\frac{t}{\sqrt{n}} \sum_{i=1}^{n} X_i}] \\ &= n \log E[e^{\frac{t}{\sqrt{n}} X}] = n \cdot \Lambda_X(\frac{t}{\sqrt{n}}) \end{aligned}$$

Note : $\Lambda_X(0) = 0$, $\dot{\Lambda}_X(0) = EX = 0$ and $\ddot{\Lambda}_X(0) = EX^2 = 1$. Then by Taylor's theorem we have

$$\Lambda_X(\frac{t}{\sqrt{n}}) = \frac{1}{2} \ddot{\Lambda}_X(\zeta \frac{t}{\sqrt{n}}) \frac{t^2}{n} , \ \zeta \in (0,1)$$

Let Z_σ be a normal random variable with mean 0, variance σ^2, and independent of the X_i's. Then

$$\begin{aligned} \lim_{n \to \infty} \Lambda^*_{S_n + Z_\sigma}(x) &= \sup \left\{ tx - \frac{1}{2} \ddot{\Lambda}_X(\zeta \frac{t}{\sqrt{n}}) t^2 - \frac{1}{2} \sigma^2 t^2 \right\} \\ &\geq \lim_{n \to \infty} \left\{ tx - \frac{1}{2} \ddot{\Lambda}_X(\zeta \frac{t}{\sqrt{n}}) t^2 - \frac{1}{2} \sigma^2 t^2 \right\} \\ &= tx - \frac{1}{2} t^2 - \frac{1}{2} \sigma^2 t^2 = tx - \frac{1}{2} (\sigma^2 + 1) t^2 \end{aligned}$$

for all $t \in R$, and so taking the sup over $t \in R$ on the right side gives

$$\lim_{n \to \infty} \Lambda^*_{S_n + Z_\sigma}(x) \geq \frac{1}{2} \frac{x^2}{\sigma^2 + 1} \quad \forall x \in R$$

To proof the inequality in the opposite direction, we need to use Lemma 3.3 for $S_n + Z_\sigma$. First we need to show that S_n can be both positive and negative for all n:

For X we know that $EX = 0$ and $VarX = 1$. Therefore there exists an $\epsilon_1 > 0$ and a $p > 0$ such that $P(X > \epsilon_1) = p$. Then by independence we get

$$P(S_n \geq \epsilon_1\sqrt{n}) = P(\sum_{i=1}^{n} X_i \geq \epsilon_1 n)$$

$$\geq P(X_1 \geq \epsilon_1, \cdots, X_n \geq \epsilon_1) = p^n > 0$$

Analogously there is an $\epsilon_2 > 0$ and a $q > 0$ such that $P(X \leq -\epsilon_2) = q > 0$ and

$$P(S_n \leq -\epsilon_2\sqrt{n}) \geq q^n > 0$$

This first of all implies that $\Lambda^*_{S_n}(x) < \infty$ for all $x \in R$ and n big enough. It also implies that for all $x \in R$ and $\sigma > 0$ there is a $p_\sigma > 0$ such that $P(S_n + Z_\sigma > x + 1) \geq p_\sigma$ and a $q_\sigma > 0$ such that $P(S_n + Z_\sigma < x - 1) \geq q_\sigma$, where p_σ and q_σ do not depend on n. Hence we can use the last lemma, and we get that there exists an $L_\sigma > 0$, independent of n, such that

$$\Lambda^*_{S_n+Z_\sigma}(x) = t_{n,\sigma} \cdot x - \log E[e^{t_{n,\sigma}S_n}] - \frac{1}{2}\sigma^2 t^2_{n,\sigma} \quad |t_{n,\sigma}| \leq L_\sigma$$

Now any bounded sequence has a convergent subsequence, say $t_{n_k,\sigma} \to t_\sigma$, and so again using Taylor's theorem we get

$$\lim_{k\to\infty} \Lambda^*_{S_{n_k}+Z_\sigma}(x)$$

$$= \lim_{k\to\infty}\left\{t_{n_k,\sigma}x - \frac{1}{2}\ddot{\Lambda}_X(\zeta\frac{t_{n_k,\sigma}}{\sqrt{n}})t^2_{n_k,\sigma} - \frac{1}{2}\sigma^2 t^2_{n_k,\sigma}\right\}$$

$$= t_\sigma x - \frac{1}{2}(\sigma^2+1)t^2_\sigma \leq \frac{1}{2}\frac{x^2}{\sigma^2+1}$$

but the right hand side is independent of the choice of subsequence, and so we have

$$\lim_{n\to\infty} \Lambda^*_{S_n+Z_\sigma}(x) = \frac{1}{2}\frac{x^2}{\sigma^2+1}$$

for all $\sigma > 0$. Therefore we see that as a function of σ $\lim_{n\to\infty} \Lambda^*_{S_n+Z_\sigma}(x)$ is right-continuous at 0 and so

$$\lim_{n\to\infty} \Lambda^*_{S_n}(x)$$

$$\lim_{n\to\infty} \Lambda^*_{S_n+Z_0}(x)$$

$$= \lim_{\sigma\downarrow 0} \lim_{n\to\infty} \Lambda^*_{S_n+Z_\sigma}(x) = \frac{1}{2}x^2$$

$\forall x \in R$. Hence it follows from the last theorem that

$$S_n \Rightarrow Z_1$$

Chapter 4

Approximation theorems

In this chapter we will show that for a large class of continuous time stochastic processes there is an approximation of their entropies in terms of the entropies of their finite - dimensional distributions.

We will denote by $D_R[0,T]$ the space of right - continuous functions from $[0,T]$ into R having left limits, and we endow this space with the Skorohod topology. Then $D_R[0,T]$ is not a Banach space, but this will not matter in the following.

First we will find the dual of $D_R[0,T]$. Obviously, $C_R[0,T] \subseteq D_R[0,T]$, where $C_R[0,T]$ is the space of continuous functions from $[0,T]$ into R with the sup-norm. The pairing $\langle z, x\rangle$ defines a bounded linear functional on $C_R[0,T]$ and by the Riesz Representation Theorem we know that every bounded linear functional φ on $C_R[0,T]$ is of the form

$$\varphi(x) \;=\; \langle \mu, x\rangle_{C_R[0,T]} \;=\; \int_0^T x(t)d\mu(t)$$

where $x \in C_R[0,T]$ and μ is a regular complex Borel measure with $\| \varphi \| = |\mu|([0,T])$. Here $|\mu|$ is the total variation of μ.

Noting that every bounded linear functional is continuous, we can extend this representation to $D_R[0,T]$: If $x \in D_R[0,T]$, then there exists a sequence $\{x_n\} \in C_R[0,T]$ such that $x_n(t) \leq x(t)\ \forall t \in [0,T], \forall n \geq 1$ and $\| x_n - x \| \rightarrow 0$. Then

$$\varphi(x) \;=\; \lim_{n\to\infty} \varphi(x_n) \;=\; \int_0^T x_n(t)\mu(dt) \;=\; \int_0^T x(t)\mu(dt)$$

by Lebesgue's dominated convergence theorem.

Next we will investigate what kind of measures μ can arise in this way. First note that φ is assumed to be real-valued. Let $0 \leq a < b \leq T$. Then the indicator function $I_{[a,b)} \in D_R[0,T]$ and

$$\varphi(I_{[a,b)}) = \mu([a,b)) \in R$$

But intervals of the form $[a,b)$ are dense in the Borel sets of $[0,T]$, and so μ has to be real-valued.

Denote by $I_{[a,T]}$ the indicator function of the interval $[a,T]$, then $I_{[a,T]} \in D_R[0,T]$ for all $a \geq 0$. Now bounded linear operators are continuous, and so

$$\mu([a,T]) = \int_a^T \mu(dt)$$

$$= \varphi(I_{[a,T]}) = \lim_{h \downarrow 0} \varphi(I_{[a+h,T]})$$

$$= \lim_{h \downarrow 0} \int_{a+h}^T \mu(dt) = \lim_{h \downarrow 0} \mu([a+h,T])$$

because $I_{[a+h,T]} \rightarrow I_{[a,T]} \in D$, and so μ has no atom at a.

From this it follows that $\lambda(t) := \mu([0,t])$ is continuous. Therefore the dual of $D_R[0,T]$ is contained in $C_R[0,T]$ and

$$\langle \lambda, x \rangle = \int_0^T x(t) d\lambda(t)$$

with $x \in D_R[0,T]$ and $\lambda \in C_R[0,T]$. Note that λ is also of bounded variation because μ has a finite total variation. In the following we will denote the space of real-valued continuous functions of bounded variation on $[0,T]$ by $CB_R[0,T]$.

Now define $\pi_{nT} : D_R[0,T] \longmapsto R^{[nT]}$ by

$$\pi_{nT} x := \left(x(\frac{1}{n}), \ldots, x(\frac{[nT]}{n}) \right)^T$$

with $x \in D_R[0,T]$. With these preliminaries we get:

$$\Lambda^*_{X_T}(x) = \sup \left\{ \int_0^T x(t) d\lambda(t) - \log E[\exp\{\int_0^T X(t) d\lambda(t)\}] \; : \; \lambda \in CB_R[0,T] \right\}$$

and

$$\Lambda^*_{\pi_{nT}X}(\pi_{nT}x) = \sup\left\{\sum_{k=1}^{[nT]}(\pi_{nT}x)_k\lambda_k - \log E[\exp\{\sum_{k=1}^{[nT]}(\pi_{nT}X)_k\lambda_k\}] \ : \ \lambda \in R^{[nT]}\right\}$$

$$= \quad \sup\left\{\sum_{k=1}^{[nT]} x(\frac{k}{n})\lambda_k \ - \ \log E[\exp\{\sum_{k=1}^{[nT]} X(\frac{k}{n})\lambda_k\}] \ : \ \lambda \in R^{[nT]}\right\}$$

Lemma 4.1 $\forall\, n \geq 1$*:*

$$0 \ \leq \ \Lambda^*_{\pi_{nT}X}(\pi_{nT}x) \ \leq \ \Lambda^*_{X_T}(x)$$

Proof.

We always have $\Lambda^*_\mu \geq 0$.

$$\Lambda^*_{\pi_{nT}X}(\pi_{nT}x)$$

$$= \ \sup\left\{\int_0^T x(t)d\lambda(t) - \log E[\exp\{\int_0^T X(t)d\lambda(t)\}] \ : \ \lambda = \sum_{k=1}^{[nT]}\alpha_k I_{[\frac{k-1}{n},\frac{k}{n})}\right\}$$

$$\leq \ \Lambda^*_{X_T}(x)$$

where the sup is taken over all $\alpha \in R^{[nT]}$, because every continuous function can be approximated by step functions.

□

Theorem 4.1 $\forall x \in D_R[0,T]$ *we have*

$$\lim_{n\to\infty} \Lambda^*_{\pi_{nT}X}(\pi_{nT}x) \ = \ \Lambda^*_{X_T}(x)$$

Proof.

First note that if $\lambda \in CB_R[0,T]$ and if we let

$$f_n(x) := \sum_{k=1}^{[nT]} x(\frac{k}{n})\left(\lambda(\frac{k}{n}) - \lambda(\frac{k-1}{n})\right)$$

then

$$\exp\{f_n(X)\} \;\leq\; \exp\left\{\sum_{k=1}^{[nT]} \mid X(\frac{k}{n}) \mid \cdot \mid \lambda(\frac{k}{n}) - \lambda(\frac{k-1}{n}) \mid \right\}$$

$$\leq\; \exp\{\| X \| \cdot \| \lambda \|\}$$

So with our basic assumption and the dominated convergence theorem we have

$$\lim_{n\to\infty} E[e^{f_n(X)}] \;=\; E[e^{\int_0^T X(t)d\lambda(t)}].$$

First assume that $\Lambda^*_{X_T}(x) = \infty$. Then there exists a sequence $\{\lambda_m\} \subseteq CB_R[0,T]$ such that

$$\int_0^T x(t)d\lambda_m(t) - \log E[\exp\{\int_0^T X(t)d\lambda_m(t)\}] \;\to\; \infty$$

as $m \to \infty$. Let $K > 0$, then there exists $M \geq 1$ such that for all $m \geq M$

$$\int_0^T x(t)d\lambda_m(t) - \log E[\exp\{\int_0^T X(t)d\lambda_m(t)\}] \;\geq\; 2K$$

and there exists $N \geq 1$ such that for all $n \geq N$

$$\sum_{k=1}^{[nT]} x(\frac{k}{n})\left(\lambda_m(\frac{k}{n}) - \lambda_m(\frac{k-1}{n})\right) - \log E\left[\exp\{\sum_{k=1}^{[nT]} X(\frac{k}{n})\left(\lambda_m(\frac{k}{n}) - \lambda_m(\frac{k-1}{n})\right)\}\right]$$

$$\geq\; K$$

But K was arbitrary, and so

$$\lim_{n\to\infty} \Lambda^*_{\pi_{nT}X}(\pi_{nT}x) \;=\; \infty$$

Now assume $\Lambda^*_{X_T}(x) = L < \infty$. Let $\epsilon > 0$, then there exists $\lambda \in CB_R[0,T]$ such that

$$\int_0^T x(t)d\lambda(t) - \log E[\exp\{\int_0^T X(t)d\lambda(t)\}] \;\geq L - \epsilon$$

Then there exists $N \geq 1$ such that for all $n \geq N$

$$\sum_{k=1}^{[nT]} x(\frac{k}{n})\left(\lambda(\frac{k}{n}) - \lambda(\frac{k-1}{n})\right) - \log E\left[\exp\{\sum_{k=1}^{[nT]} X(\frac{k}{n})\left(\lambda(\frac{k}{n}) - \lambda(\frac{k-1}{n})\right)\}\right]$$

$$\geq L - 2\epsilon$$

$$\Rightarrow \quad \Lambda^*_{\pi_{nT}X}(\pi_{nT}x) \ \geq \ L - 2\epsilon$$

and using Lemma 4.1 we get

$$\lim_{n\to\infty} \Lambda^*_{\pi_{nT}X}(\pi_{nT}x) \ = \ L \ = \ \Lambda^*_{X_T}(x)$$

□

Example 4.1 In this example we will use the technique of the last theorem to find the entropy of Brownian motion $\{B(t), t \in [0,T]\}$ with initial distribution $B(0) = 0$ a.s. In this case $\pi_{nT}B$ has a $[nT]$ - dimensional normal distribution with mean vector 0 and covariance matrix Σ given by $\Sigma_{ij} = \frac{i \wedge j}{n}$ for $i, j = 1, \cdots, [nT]$, or

$$\Sigma = \frac{1}{n} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \cdots & [nT] \end{pmatrix}$$

By example 1.1 we know that

$$\Lambda^*_{\pi_{nT}B}(\pi_{nT}x) = \frac{1}{2}(\pi_{nT}x)^T \Sigma^{-1} (\pi_{nT}x)$$

Σ^{-1} is found to be

$$\Sigma^{-1} = n \cdot \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & 0 & -1 & 1 \end{pmatrix}$$

and so we get

$$\Lambda^*_{\pi_{nT}B}(\pi_{nT}x) = \frac{n}{2}(\pi_{nT}x)^T \cdot \begin{pmatrix} 2x(\frac{1}{n}) - x(\frac{2}{n}) \\ -x(\frac{1}{n}) + 2x(\frac{2}{n}) - x(\frac{3}{n}) \\ -x(\frac{2}{n}) + 2x(\frac{3}{n}) - x(\frac{4}{n}) \\ \vdots \\ -x(\frac{[nT]-2}{n}) + 2x(\frac{[nT]-1}{n}) - x(\frac{[nT]}{n}) \\ -x(\frac{[nT]-1}{n}) + x(\frac{[nT]}{n}) \end{pmatrix}$$

$$= \frac{n}{2}\left[2\sum_{k=1}^{[nT]-1} x^2(\frac{k}{n}) - 2\sum_{k=1}^{[nT]} x(\frac{k}{n})x(\frac{k-1}{n}) + x^2(\frac{[nT]}{n})\right]$$

$$= \frac{n}{2}\sum_{k=1}^{[nT]}\left(x(\frac{k}{n}) - x(\frac{k-1}{n})\right)^2$$

$$= \frac{1}{2}\sum_{k=1}^{[nT]}\left(\frac{x(\frac{k}{n}) - x(\frac{k-1}{n})}{\frac{1}{n}}\right)^2\left(\frac{k}{n} - \frac{k-1}{n}\right)$$

$$\to \begin{cases} \frac{1}{2}\int_0^T \dot{x}(t)^2 dt & \text{if } \dot{x} \text{ exists a.e.} \\ \infty & \text{otherwise} \end{cases}$$

Theorem 4.2 *Let $\{X(t); t \in [0,T]\}$ be a time-homogeneous process with independent increments and assume $X(0) = 0$ a.s. Then*

$$\Lambda^*_{\pi_{nT}X}(\pi_{nT}x) = \sum_{k=1}^{[nT]} n \cdot \Lambda^*_{X(\frac{k}{n})-X(\frac{k-1}{n})}(x(\frac{k}{n}) - x(\frac{k-1}{n})) \cdot (\frac{k}{n} - \frac{k-1}{n})$$

Proof.

$$E\left[\exp\{\lambda'(\pi_{nT}X)\}\right] = \prod_{k=1}^{[nT]} E\left[\exp\{(A^{-1^T}\lambda)_k(X(\frac{k}{n}) - X(\frac{k-1}{n}))\}\right]$$

where A is the $[nT] \times [nT]$ matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & . & 0 \\ 0 & \cdots & . & . & -1 & 1 \end{pmatrix}$$

Now

$$\Lambda^*_{\pi_{nT}X}(\pi_{nT}x)$$

$$= \sup\left\{\lambda'(\pi_{nT}x) - \sum_{k=1}^{[nT]} \log E\left[\exp\{(A^{-1^T}\lambda)_k(X(\frac{k}{n}) - X(\frac{k-1}{n}))\}\right] \ : \ \lambda \in R^{[nT]}\right\}$$

$$= \sup\{\sum_{k=1}^{[nT]}(A^{-1^T}\lambda)_k(x(\frac{k}{n}) - x(\frac{k-1}{n}))$$

$$- \sum_{k=1}^{[nT]} \log E\left[\exp\{(A^{-1^T}\lambda)_k(X(\frac{k}{n}) - X(\frac{k-1}{n}))\}\right] \ : \ \lambda \in R^{[nT]}\}$$

$$= \sum_{k=1}^{[nT]} \sup\left\{\theta(x(\frac{k}{n}) - x(\frac{k-1}{n})) - \log E\left[\exp\{\theta(X(\frac{k}{n}) - X(\frac{k-1}{n}))\}\right] \ : \ \theta \in R\right\}$$

$$= \sum_{k=1}^{[nT]} n \cdot \Lambda^*_{X(\frac{k}{n})-X(\frac{k-1}{n})}(x(\frac{k}{n}) - x(\frac{k-1}{n})) \cdot (\frac{k}{n} - \frac{k-1}{n})$$

□

Example 4.2 Let $\{N_t; t \geq 0\}$ be a Poisson process with rate function η, i.e N has independent increments and $N(t) - N(s) \sim P(\eta(t) - \eta(s))$, a Poisson random variable with parameter $\eta(t) - \eta(s)$. We will assume that η is differentiable in $[0, T]$. Using the last theorem and the fact that the log moment generating function of a Poisson random variable X with parameter s is given by

$$\log E[e^{\lambda X}] = s(e^{\lambda} - 1)$$

and noting that $N(\frac{k}{n})-N(\frac{k-1}{n})$ is a Poisson random variable with parameter $\eta(\frac{k}{n})-\eta(\frac{k-1}{n})$ we get

$$\begin{aligned}
&\Lambda^*_{N(\frac{k}{n})-N(\frac{k-1}{n})}(u) \\
&= \sup\left\{\alpha u \;-\; (\eta(\frac{k}{n})-\eta(\frac{k-1}{n}))(e^\alpha-1)\;:\;\alpha\in R\right\}
\end{aligned}$$

But

$$\begin{aligned}
&\frac{d}{d\alpha}\left\{\alpha u \;-\; (\eta(\frac{k}{n})-\eta(\frac{k-1}{n}))(e^\alpha-1)\right\} \\
&= u-(\eta(\frac{k}{n})-\eta(\frac{k-1}{n}))e^\alpha \;=\; 0
\end{aligned}$$

implies

$$\alpha \;=\; \log\frac{u}{\eta(\frac{k}{n})-\eta(\frac{k-1}{n})}$$

and so

$$\Lambda^*_{N(\frac{k}{n})-N(\frac{k-1}{n})}(u) \;=\; u\log\frac{u}{\eta(\frac{k}{n})-\eta(\frac{k-1}{n})}-u+\eta(\frac{k}{n})-\eta(\frac{k-1}{n})$$

where $u\in R$ and $\eta(\frac{k}{n})-\eta(\frac{k-1}{n})\in R$. Then

$$\begin{aligned}
&\Lambda^*_{\pi_{nT}N}(\pi_{nT}x) \\
&= \sum_{k=1}^{[nT]} n\cdot\Lambda^*_{N(\frac{k}{n})-N(\frac{k-1}{n})}(x(\frac{k}{n})-x(\frac{k-1}{n}))\cdot(\frac{k}{n}-\frac{k-1}{n}) \\
&= \sum_{k=1}^{[nT]}((x(\frac{k}{n})-x(\frac{k-1}{n}))\log\frac{x(\frac{k}{n})-x(\frac{k-1}{n})}{\eta(\frac{k}{n})-\eta(\frac{k-1}{n})} \\
&\quad -x(\frac{k}{n})-x(\frac{k-1}{n})+\eta(\frac{k}{n})-\eta(\frac{k-1}{n})) \\
&= \sum_{k=1}^{[nT]}\left(\frac{x(\frac{k}{n})-x(\frac{k-1}{n})}{\frac{1}{n}}\log\frac{x(\frac{k}{n})-x(\frac{k-1}{n})}{\eta(\frac{k}{n})-\eta(\frac{k-1}{n})}\right)(\frac{k}{n}-\frac{k-1}{n}) \;-\; x(\frac{[nT]}{n})+\eta(\frac{[nT]}{n}) \\
&\to\begin{cases}\int_0^T \dot{x}(t)\log\frac{\dot{x}(t)}{\dot{\eta}(t)}dt \;-\; x(T)\;+\;\eta(T) & \text{if } \dot{x} \text{ exists a.e.}\\ \infty & \text{otherwise}\end{cases}
\end{aligned}$$

Assume that we have a process X with independent increments as in the theorem above. Assume that $\mathcal{A}$ is the generator of X and that the function $f_\lambda(x) = e^{\lambda x}$ is in the domain of $\mathcal{A}$. Then

$$\lim_{h\downarrow 0} \frac{E[e^{\lambda(X(t+h)-X(t))}] - 1}{h}$$

$$\lim_{h\downarrow 0} E\left[e^{-\lambda X(t)} \frac{E[(e^{\lambda X(t+h)} \mid X(t)] - e^{\lambda X(t)}}{h})\right]$$

$$= E\left[e^{-\lambda X(t)}(\mathcal{A}f_\lambda)(X(t))\right]$$

and

$$\lim_{h\downarrow 0} \frac{\log E\left[e^{\lambda(X(t+h)-X(t))}\right]}{h} = E\left[e^{-\lambda X(t)}(\mathcal{A}f_\lambda)(X(t))\right]$$

Hence it follows from the theorems in this chapter that

$$\Lambda^*_{X_T}(x) \geq \int_0^T \left(E\left[e^{-\lambda X(t)}(\mathcal{A}f_\lambda)(X(t))\right]\right)^* (\dot{x}(t))dt$$

if $\dot{x}$ exists a.e., and ∞ otherwise. If the interchange of lim and sup can be justified, then equality holds.

Example 4.3 Consider standard Brownian motion. Then $(\mathcal{A}e^{\lambda\cdot})(x) = \frac{1}{2}\lambda^2 e^{\lambda x}$. Hence

$$E\left[(\mathcal{A}e^{\lambda\cdot}(B_t)e^{-\lambda B_t}\right]$$

$$= E\left[\frac{1}{2}\lambda^2 e^{\lambda B_t} e^{-\lambda B_t}\right] = \frac{1}{2}\lambda^2$$

$$\Rightarrow \quad \Lambda^*_{B_T}(x) \geq \int_0^T \sup\left\{\lambda\dot{x}(t) - \frac{1}{2}\lambda^2 \ : \ \lambda \in R\right\} dt$$

$$= \frac{1}{2}\int_0^T \dot{x}(t)^2 dt$$

Chapter 5

Donsker's theorem

In this chapter we will use the techniques developed in the previous chapters to give a new proof of a famous theorem by Donsker. For a discussion of Donsker's theorem see Durrett [3].

Theorem 5.1 *Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables on R with mean 0 and variance 1 and assume that X_k has a moment generating function. Let*

$$S_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} X_k$$

Then

$$S_n \Rightarrow B$$

where B is Brownian motion on R.

Proof.

We will use Theorem 7.2. and Theorem 7.8. in Chapter 3 of Ethier, Kurtz [4]. That is, we have to verify the following two conditions:

1) The finite-dimensional distributions of S_n converge weakly to those of B, that is there exists a $D \subset [0, \infty)$ dense such that

$$(S_n(t_1), \cdots, S_n(t_k)) \Rightarrow (B(t_1), \cdots, B(t_k))$$

for all finite collections $\{t_1, \cdots, t_k\} \subset D$.

2) The sequence $\{S_n\}$ is relatively compact in $D_R[0,\infty)$.

Let $D = \left\{(\frac{j_1}{N},\cdots,\frac{j_k}{N}) \;:\; k,N \geq 1 \;,\; 0 < j_1 < \cdots < j_k\right\}$. Then D is a dense subset of $[0,\infty)$. For notational convenience only we will restrict ourselves to S_{nN}. Now fix $N, k, j_1, \cdots, j_k$. Then

$$
\begin{aligned}
S^n &\equiv \left(S_{nN}(\frac{j_1}{N}),\cdots,S_{nN}(\frac{j_k}{N})\right) \\
&= \left(\frac{1}{\sqrt{nN}}\sum_{i=1}^{nj_1} X_i,\cdots,\frac{1}{\sqrt{nN}}\sum_{i=1}^{nj_k} X_i\right) \\
&= \frac{1}{\sqrt{nN}} M \cdot \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{nj_k} \end{pmatrix}
\end{aligned}
$$

where

$$
M = \begin{pmatrix}
1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & 0 \\
1 & \cdots & \cdots & \cdots & 1 & 0 & \cdots & 0 \\
\vdots & & & & \vdots & \ddots & & 0 \\
1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1
\end{pmatrix}
$$

where the i^{th} row of M has a one in the first nj_i positions and zeroes thereafter.

Let $x \in R^k$ and define $\bar{x} \in R^{nj_k}$ by

$$
(\bar{x})_i = \frac{1}{nj_l - nj_{l-1}}(x_l - x_{l-1})
$$

for $nj_{l-1} < i \leq nj_l$, $l = 1,\cdots,k$ where we set $x_0 = 0$ and $j_0 = 0$. Then

$$
M \cdot \bar{x} = x
$$

and

$$
\begin{aligned}
\Lambda^*_{S^n}(x) &= \sup\left\{\lambda^T x - \log E\left[e^{\lambda^T S^n}\right] \;:\; \lambda \in R^k\right\} \\
&= \sup\left\{(M^T\lambda)^T\bar{x} - \sum_{i=1}^{nj_k}\log E\left[e^{\frac{1}{\sqrt{nN}}(M^T\lambda)_i X_1}\right] \;:\; \lambda \in R^k\right\}
\end{aligned}
$$

$$\le \sup\left\{\eta^T\bar{x} - \sum_{i=1}^{nj_k}\log E\left[e^{\frac{1}{\sqrt{nN}}\eta_i X_1}\right] \;:\; \eta \in R^{nj_k}\right\}$$

$$= \sum_{i=1}^{nj_k}\sup\left\{\alpha\bar{x}_i - \log E\left[e^{\frac{1}{\sqrt{nN}}\alpha X_1}\right] \;:\; \alpha\in R\right\}$$

$$= \sum_{i=1}^{k}\sum_{l=nj_{i-1}+1}^{nj_i}\sup\left\{\frac{\alpha}{\sqrt{N}}(\sqrt{N}\frac{x_i - x_{i-1}}{nj_i - nj_{i-1}}) - \log E\left[e^{\frac{1}{\sqrt{n}}\frac{\alpha}{\sqrt{N}}X_1}\right] \;:\; \alpha\in R\right\}$$

$$= \sum_{i=1}^{k}(j_i - j_{i-1})\sup\left\{\beta\sqrt{N}\frac{x_i - x_{i-1}}{j_i - j_{i-1}} - n\log E\left[e^{\frac{1}{\sqrt{n}}\beta X_1}\right] \;:\; \beta\in R\right\}$$

$$= \sum_{i=1}^{k}(j_i - j_{i-1})\sup\left\{\beta\sqrt{N}\frac{x_i - x_{i-1}}{j_i - j_{i-1}} - \log E\left[e^{\beta\frac{1}{\sqrt{n}}\sum_{l=1}^{n}X_l}\right] \;:\; \beta\in R\right\}$$

$$= \sum_{i=1}^{k}(j_i - j_{i-1})\Lambda^*_{S_n(1)}\left(\frac{\sqrt{N}(x_i - x_{i-1})}{j_i - j_{i-1}}\right)$$

and using Example 3.3 we get

$$\lim_{n\to\infty}\Lambda^*_{S^n}(x) \;=\; \frac{1}{2}\sum_{i=1}^{k}\frac{(x_i - x_{i-1})^2}{\frac{j_i}{N} - \frac{j_{i-1}}{N}}.$$

Now $\left(B(\frac{j_1}{N}),\cdots,B(\frac{j_k}{N})\right)$ has a multivariate normal distribution with mean vector 0 and covariance matrix Σ with

$$\Sigma = \frac{1}{N}\cdot\begin{pmatrix} j_1 & j_1 & j_1 & j_1 & \cdots & j_1 \\ j_1 & j_2 & j_2 & j_2 & \cdots & j_2 \\ j_1 & j_2 & j_3 & j_3 & \cdots & j_3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ j_1 & j_2 & j_3 & j_4 & \cdots & j_k \end{pmatrix}$$

Then

$$\Sigma^{-1} = N\cdot\begin{pmatrix} \frac{1}{j_1}+\frac{1}{j_2-j_1} & -\frac{1}{j_2-j_1} & 0 & 0 & \cdots & 0 \\ -\frac{1}{j_2-j_1} & \frac{1}{j_2-j_1}+\frac{1}{j_3-j_2} & -\frac{1}{j_3-j_2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & -\frac{1}{j_k-j_{k-1}} & \frac{1}{j_k-j_{k-1}} \end{pmatrix}$$

From Example 1.1 we know that the entropy of a multivariate normal random variable with mean vector 0 and covariance Σ is given by

$$\frac{1}{2}x^T\Sigma^{-1}x$$

$$= \frac{1}{2}\sum_{i=1}^{k}\left(\left(\frac{N}{j_i - j_{i-1}} + \frac{N}{j_{i+1} - j_i}\right)x_i^2 - 2\frac{N}{j_i - j_{i-1}}x_i x_{i-1}\right) - \frac{N}{j_k - j_{k-1}}x_k^2$$

$$= \frac{1}{2}\sum_{i=1}^{k}\frac{(x_i - x_{i-1})^2}{\frac{j_i}{N} - \frac{j_{i-1}}{N}}.$$

and so we have

$$\lim_{n\to\infty} \Lambda^*_{S^n}(x) \leq \Lambda^*_{(B(\frac{j_1}{N}),\cdots,B(\frac{j_k}{N}))}(x)$$

for all $x \in R^k$. On the other hand

$$\begin{aligned}
\Lambda^*_{S^n}(x) &= \sup\left\{\lambda^T x - \log E\left[e^{\lambda^T S^n}\right] \;:\; \lambda \in R^k\right\} \\
&= \sup\left\{\lambda^T x - \sum_{i=1}^{nj_k}\log E\left[e^{\frac{1}{\sqrt{nN}}(M^T\lambda)_i X_1}\right] \;:\; \lambda \in R^k\right\} \\
&\geq \lambda^T x - \sum_{i=1}^{nj_k}\log E\left[e^{\frac{1}{\sqrt{nN}}(M^T\lambda)_i X_1}\right] \\
&= \lambda^T x - \sum_{i=1}^{k}\sum_{l=nj_{i-1}+1}^{nj_i}\log E\left[e^{\frac{1}{\sqrt{nN}}(\sum_{s=i}^{k}\lambda_s)X_1}\right] \\
&= \lambda^T x - \sum_{i=1}^{k}(j_i - j_{i-1})n\log E\left[e^{\frac{1}{\sqrt{n}}(\frac{1}{\sqrt{N}}\sum_{s=i}^{k}\lambda_s)X_1}\right] \\
&= \lambda^T x - \sum_{i=1}^{k}(j_i - j_{i-1})\log E\left[e^{(\frac{1}{\sqrt{N}}\sum_{s=i}^{k}\lambda_s)S_n(1)}\right]
\end{aligned}$$

for all $\lambda \in R^k$, and again using Example 3.3 we get

$$\lim_{n\to\infty} \Lambda^*_{S^n}(x) \geq \lambda^T x - \sum_{i=1}^{k}(j_i - j_{i-1})\frac{1}{2}\left(\frac{1}{\sqrt{N}}\sum_{s=i}^{k}\lambda_s\right)^2$$

$$= \lambda^T x - \sum_{i=1}^{k}(j_i - j_{i-1})\frac{1}{2N}\left(\sum_{s=i}^{k}\lambda_s\right)^2$$

Taking the supremum on the right side we get

$$\lim_{n\to\infty} \Lambda^*_{S^n}(x) \geq \sup\left\{\lambda^T x - \sum_{i=1}^{k}(j_i - j_{i-1})\frac{1}{2N}\left(\sum_{s=i}^{k}\lambda_s\right)^2 \;:\; \lambda \in R^k\right\}$$

$$= \sup\left\{\lambda^T x - \frac{1}{2}\sum_{i=1}^{k}(T\lambda)_i^2 \;:\; \lambda \in R^k\right\}$$

$$= \sup\left\{\lambda^T x - \frac{1}{2}\lambda^T T^T T\lambda \;:\; \lambda \in R^k\right\}$$

$$= \frac{1}{2}x^T(T^T T)^{-1}x$$

where

$$T = \begin{pmatrix} \sqrt{\frac{j_1}{N}} & \sqrt{\frac{j_1}{N}} & \sqrt{\frac{j_1}{N}} & \cdots & \sqrt{\frac{j_1}{N}} \\ 0 & \sqrt{\frac{j_2}{N}-\frac{j_1}{N}} & \sqrt{\frac{j_2}{N}-\frac{j_1}{N}} & \cdots & \sqrt{\frac{j_2}{N}-\frac{j_1}{N}} \\ 0 & 0 & \sqrt{\frac{j_3}{N}-\frac{j_2}{N}} & \cdots & \sqrt{\frac{j_3}{N}-\frac{j_2}{N}} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \sqrt{\frac{j_k}{N}-\frac{j_{k-1}}{N}} \end{pmatrix}$$

It is easy to compute $T^T \cdot T = \Sigma$, and so

$$\lim_{n\to\infty} \Lambda^*_{S^n}(x) \geq \frac{1}{2}x^T\Sigma^{-1}x$$

and we can conclude that

$$\lim_{n\to\infty} \Lambda^*_{S^n}(x) = \Lambda^*_{(B(\frac{j_1}{N}),\cdots,B(\frac{j_k}{N})}(x)$$

for all $x \in R^k$. Note that

$$\Lambda^*_{(B(\frac{j_1}{N}),\cdots,B(\frac{j_k}{N}))}(x) = \frac{1}{2}\left(\frac{x_1^2}{\frac{j_1}{N}} + \sum_{i=2}^{k}\frac{(x_i - x_{i-1})^2}{\frac{j_i}{N}-\frac{j_{i-1}}{N}}\right)$$

is equal to 0 if and only if $x = 0$, and so using theorem 4.3 we get

$$S^n \equiv \left(S_{nN}(\frac{j_1}{N}), \cdots, S_{nN}(\frac{j_k}{N})\right) \Rightarrow \left(B(\frac{j_1}{N}), \cdots, B(\frac{j_k}{N})\right)$$

as $n \to \infty$, and so the first condition is fulfilled.

For the second condition we will use Theorem 8.8 in Chapter 3 of Ethier-Kurtz [4], that is, we will show that for all $n \geq 1$ and $h > 0$

$$E\left[|S_n(t+h) - S_n(t)|^2 \cdot |S_n(t) - S_n(t-h)|^2\right] \leq 16h^2$$

First note that if $h < \frac{1}{2n}$ then one of the two terms in the expectation is 0, and the inequality follows trivially. So now assume $h \geq \frac{1}{2n}$. Then

$$\begin{aligned}
& E\left[|S_n(t+h) - S_n(t)|^2 \cdot |S_n(t) - S_n(t-h)|^2\right] \\
= \;& E\left[|\frac{1}{\sqrt{n}} \sum_{k=[nt]+1}^{[n(t+h)]} X_k|^2 \cdot |\frac{1}{\sqrt{n}} \sum_{k=[n(t-h)]+1}^{[nt]} X_k|^2\right] \\
= \;& E\left[|\frac{1}{\sqrt{n}} \sum_{k=[nt]+1}^{[n(t+h)]} X_k|^2\right] \cdot E\left[|\frac{1}{\sqrt{n}} \sum_{k=[n(t-h)]+1}^{[nt]} X_k|^2\right] \\
\leq \;& \frac{1}{n^2}\left(\frac{1}{\sqrt{n}} \sum_{k=[nt]+1}^{[n(t+h)]} EX_k^2\right) \cdot \left(\frac{1}{\sqrt{n}} \sum_{k=[n(t-h)]+1}^{[nt]} EX_k^2\right) \\
= \;& \frac{1}{n^2}([n(t+h)] - [nt]) \cdot ([nt] - [n(t-h)]) \\
\leq \;& \left(\frac{[n(t+h)] - [n(t-h)]}{n}\right)^2
\end{aligned}$$

Of course we have

$$[n(t+h)] \leq n(t+h) \;\; and \;\; [n(t-h)] \geq n(t-h) - 1$$

and so we get

$$E\left[|S_n(t+h) - S_n(t)|^2 \cdot |S_n(t) - S_n(t-h)|^2\right]$$

$$\leq \left(\frac{[n(t+h)] - [n(t-h)]}{n}\right)^2$$

$$\leq \left(\frac{n(t+h) - n(t-h) + 1}{n}\right)^2 = \left(\frac{2nh+1}{n}\right)^2 \leq (4h)^2$$

because $h \geq \frac{1}{2n}$ implies $1 \leq 2nh$, and so the theorem is proven.

□

Chapter 6

Entropy of Gaussian processes

In this chapter, we will study the entropy of Gaussian processes. We already computed the entropy of Brownian motion, and the techniques used, especially the approximation using the entropy of the finite-dimensional distributions, are still valid here.

We start by giving a short introduction into the theory of Gaussian processes on the space $D_E[0,T]$, $T \geq 0$. This exposition follows closely the article by N. Jain and M. Marcus in Kuelbs [6].

A process $\{X(t), t \in [0,T]\}$ is called Gaussian if the finite-dimensional vectors $(X(t_1), \cdots, X(t_n))$ have a normal distribution for all finite collections $t_1, \cdots, t_n \in T$. If the mean of the vector is zero, than $\{X(t)\}$ is called a centered Gaussian process.

In the following we will also assume that $\{X(t), t \in [0,T]\}$ is square-integrable, i.e. $EX(t)^2 < \infty, t \in [0,T]$.

The mean function m and the covariance function Γ of a Gaussian process are defined as follows :

$$m(t) = EX(t) \quad , \quad t \in [0,T]$$

$$\Gamma(t,s) = E(X(t) - m(t))(X(s) - m(s)) \quad , \quad s,t \in [0,T]$$

Definition 6.1 Let $T > 0$. A real-valued function Γ on $[0,T]^2$ is called a covariance kernel if

$$\Gamma(s,t) = \Gamma(t,s)$$

and Γ is nonnegative definite, i.e. given $t_1, \cdots, t_n$ and $a_1, \cdots, a_n \in R$

$$\sum_{j,k} a_j a_k \Gamma(t_j, t_k) \geq 0$$

In the remainder of this chapter we will assume that Γ is a continuous mapping from $[0,T]^2 \to R$. To the covariance kernel Γ we can attach a Hilbert space $H(\Gamma)$ of real-valued functions as follows : Let

$$S = \left\{ \sum_{j=1}^{n} a_j \Gamma(t_j, \cdot) : a_1, \cdots, a_n; t_1, \cdots, t_n; n \geq 1 \right\}$$

On S define an inner product by

$$\left(\sum_{j=1}^{n} a_j \Gamma(t_j, \cdot), \sum_{k=1}^{m} b_k \Gamma(s_k, \cdot) \right) = \sum_{j=1}^{n} \sum_{k=1}^{m} a_j b_k \Gamma(t_j, s_k)$$

Let $f \in S$, then we have

$$\begin{aligned} f(t) &= \sum_{j=1}^{n} a_j \Gamma(t_j, t) \\ &= (f, \Gamma(t, \cdot)) \end{aligned}$$

Next

$$(f, f) = \sum_{i,j} a_i a_j \Gamma(t_i, t_j) \geq 0$$

because Γ is positive definite. Assume $(f, f) = 0$, then

$$| f(t) |^2 = | (f, \Gamma(t, \cdot)) |^2 \leq (f, f) \cdot \Gamma(t, t) = 0$$

where the inequality is the Schwartz inequality for semi-inner products. Therefore we have an inner product on S. Furthermore, if $\{f_n\} \in S$, then

$$| f_n(t) - f_m(t) |^2 = | (f_n - f_m, \Gamma(t, \cdot)) |^2$$

$$\leq \| f_n - f_m \|^2 \cdot \Gamma(t, t)$$

Hence if $\{f_n\}$ is Cauchy with respect to the inner product norm on S, then it is pointwise Cauchy. We close S under this norm and identify the limit elements with the pointwise limits. The closure will be denoted by $H(\Gamma)$ and is called the reproducing kernel Hilbert space (rkhs) of Γ.

Theorem 6.1 *Let* Γ *be a covariance kernel on* $[0,T]^2$. *Then there exists a Hilbert space* $H(\Gamma)$ *such that*

1.

$$\Gamma(t,\cdot) \in H(\Gamma) \quad \forall t \in [0,T]$$

2.

$$(f, \Gamma(t,\cdot)) = f(t) \quad \forall f \in H(\Gamma),\ t \in [0,T]$$

Theorem 6.2 *Let* Γ *be a covariance kernel on* $[0,T]^2$. *Let* H *be a Hilbert space of real-valued functions on* $[0,T]$ *with inner product* $(\cdot,\cdot)_1$. *Suppose*

1.

$$\Gamma(t,\cdot) \in H \ \ \forall t \in [0,T]$$

2.

$$(f, \Gamma(t,\cdot))_1 = f(t) \forall f \in H(\Gamma) \ \ , \ \ t \in [0,T]$$

then $H = H(\Gamma)$.

Proofs for these two theorems can be found in Kuelbs [6].

Next we will find a complete orthonormal system (CONS) for $H(\Gamma)$ in terms of the eigenvalues and eigenfunctions of Γ. If Γ is a continuous covariance kernel on $[0,T]^2$ and $f \in L^2([0,T])$, then we define the operator K by :

$$Kf(s) \ = \ \int_0^T \Gamma(s,t) f(t) dt$$

K is a compact operator and hence has a countable number of eigenvalues, where the eigenvalues are repeated according to their multiplicities. The eigenspaces are all finite-dimensional, and so the multiplicities are also finite.

K is self-adjoint because

$$(Kf, g) = \int_0^T (\Gamma(\cdot, s), g)\, f(s) ds$$
$$= \int_0^T g(s) f(s) ds = (f, Kg)$$

Definition 6.2 The eigenvalues $\{\lambda_n^2\}$ and eigenfunctions $\{\psi_n\}$ of a covariance kernel Γ are defined by

$$\lambda_n^2 \cdot \psi_n(s) = \int_0^T \Gamma(s,t)\psi_n(t)dt$$

We see that all the eigenvalues are positive because if μ is an eigenvalue with eigenfunction φ then

$$\mu \parallel \varphi \parallel^2 = (\mu\varphi, \varphi) = (K\varphi, \varphi)$$
$$= \int_0^T (\Gamma(\cdot, s), \varphi)\, \varphi(s) ds = \int_0^T \varphi^2(s) ds > 0$$

Note that the continuity of Γ implies the continuity of ψ_n for all $n \geq 1$. We assume the system $\{\psi_n\}$ to be normalized, i.e.

$$\int_0^T \psi_n(t)\psi_m(t)dt = \delta_{nm}$$

Remark

Note that this implies Mercer's theorem, i.e.

$$\Gamma(t, s) = \sum_n \lambda_n^2 \psi_n(t)\psi_n(s) \quad \forall 0 \leq s, t, \leq T$$

To see this, fix $s \in [0, T]$. Then $\Gamma(s, \cdot) \in H(\Gamma)$ and

$$\Gamma(s,t) = \sum_n \left(\int_0^T \Gamma(s,u)\psi_n(u)du \right) \psi_n(t)$$
$$= \sum_n \lambda_n^2 \psi_n(s)\psi_n(t)$$

Theorem 6.3 *Let $\{X(t)\ ,\ t \in [0,T]\}$ be a Gaussian process with covariance kernel Γ, and let the eigenvalues and eigenfunctions of Γ be denoted by $\{\lambda_n^2\}$ and $\{\psi_n\}$, respectively. Then*

$$X(t) \;=\; \sum_{n=1}^{\infty} \lambda_n \psi_n(t) Y_n \;+\; m(t)$$

where $\{Y_n\}$ are independent $N(0,1)$ random variables.

The proof of this theorem can be found in Kuelbs [6].

Remark

The above is called the Karhunen-Loeve expansion.

We will now use this expansion to find a formula for the entropy of a Gaussian process $\{X(t), t \in [0,T]\}$ in terms of eigenvalues and the eigenfunctions of Γ. Without loss of generality we assume the process to be centered.

Theorem 6.4 *Let $\{X(t), t \in [0,T]\}$ be a centered Gaussian process with covariance function Γ. Let $\{\lambda_n^2\}$ and $\{\psi_n\}$ be the eigenvalues and eigenfunctions of Γ, respectively. Then*

1. *If $x \in H(\Gamma)$ and $x(t) \;=\; \sum_1^{\infty} a_n \psi_n(t)$, $t \in [0,T]$, then*

$$\Lambda_X^*(x) \;=\; \sum_{n=1}^{\infty} \frac{a_n^2}{\lambda_n^2}$$

2. *If $x \notin H(\Gamma)$, then $\Lambda_X^*(x) \;=\; \infty$.*

Proof.

First let $x \in H(\Gamma)$ and note that

$$\log E\left[\exp(\int_0^T X(s) d\lambda(s))\right]$$

$$= \;\log E\left[\exp(\int_0^T \sum_{j=1}^{\infty} \lambda_j \psi_j(s) Y_j d\lambda(s))\right]$$

$$= \;\sum_{j=1}^{\infty} \log E\left[\exp\left((\int_0^T \lambda_j \psi_j(s) d\lambda(s)) Y_j\right)\right]$$

$$= \sum_{j=1}^{\infty} \frac{1}{2} \left(\int_0^T \lambda_j \psi_j(s) d\lambda(s) \right)$$

Hence

$$\Lambda_X^*(x) = \sup \left\{ \int_0^T x(t) d\lambda(t) - \sum_{j=1}^{\infty} \frac{1}{2} \left(\int_0^T \lambda_j \psi_j(s) d\lambda(s) \right)^2 : \lambda \in C[0,T] \right\}$$

The Frechet-derivative is found to be

$$\int_0^T x(s) dh(s) - \sum_{j=1}^{\infty} \lambda_j^2 \left(\int_0^T \psi_j(s) d\lambda(s) \right) \left(\int_0^T \psi_j(s) dh(s) \right)$$

where $h \in C^\infty[0,T]$, and so because of concavity, a necessary and sufficient condition for an optimum is

$$\int_0^T \left[x(s) - \sum_1^{\infty} \lambda_j^2 \left(\int_0^T \psi_j(t) d\lambda(t) \right) \psi_j(s) \right] dh(s) = 0$$

for all $h \in C^\infty[0,T]$, or

$$x(s) = \sum_1^{\infty} \lambda_j^2 \left(\int_0^T \psi_j(t) d\lambda(t) \right) \psi_j(s)$$

Let $x \in H(\Gamma)$. Then there exists an expansion of x in terms of the eigenfunctions of Γ, i.e. there exists $\{a_j\}$ with

$$x(s) = \sum_1^{\infty} a_j \psi_j(s) \quad \forall s \in [0,T]$$

Let λ be defined by

$$d\lambda(s) = \sum_{j=1}^{\infty} \frac{a_j}{\lambda_j^2} \psi_j(s) ds$$

Then λ satisfies the equation above and so it is an optimizing element in $C[0,T]$. Hence

$$\Lambda_X^*(x)$$

$$= \int_0^T \sum_{j=1}^{\infty} a_j\psi_j(s) \cdot \sum_{i=1}^{\infty} \frac{a_i}{\lambda_i^2}\psi_i(s)ds - \frac{1}{2}\sum_{j=1}^{\infty}\lambda_j^2\left(\int_0^T \psi_j(s)\sum_{i=1}^{\infty}\frac{a_i}{\lambda_i^2}\psi_i(s)ds\right)^2$$

$$= \sum_{i,j}\frac{a_ia_j}{\lambda_i^2}\int_0^T \psi_i(s)\psi_j(s)ds - \frac{1}{2}\sum_i \lambda_i^2\left(\sum_j \frac{a_j}{\lambda_j^2}\int_0^T \psi_i(s)\psi_j(s)ds\right)^2$$

$$= \frac{1}{2}\sum_j^{\infty}\frac{a_j^2}{\lambda_j^2}$$

It remains to show part 2). To do that, assume that $\Lambda_X^*(x) < \infty$. We will show that this implies $x \in H(\Gamma)$. Let

$$a_n = \int_0^T x(t)\psi_n(t)dt$$

Then it sufficies to show that $\sum_1^\infty a_n^2 < \infty$. But we have

$$\int_0^T x(t)\lambda(t)dt - \frac{1}{2}\sum_1^{\infty}\lambda_n^2\left(\int_0^T \psi_n(t)\lambda(t)dt\right)^2 \leq \Lambda_X^*(x) < \infty$$

for all $\lambda \in C[0,T]$.

Now let

$$\lambda^N(t) = \sum_{n=1}^{N}\frac{a_n}{\lambda_n^2}\cdot\psi_n(t) \quad t \in [0,T]$$

then $\lambda_N \in C[0,T] \ \forall N \geq 1$ and

$$\int_0^T x(t)\lambda^N(t)dt - \frac{1}{2}\sum_1^{\infty}\lambda_n^2\left(\int_0^T \psi_n(t)\lambda^N(t)dt\right)^2 \leq \Lambda_X^*(x) < \infty$$

$$\Rightarrow \int_0^T x(t)\sum_{n=1}^{N}\frac{a_n}{\lambda_n^2}\cdot\psi_n(t)dt - \frac{1}{2}\sum_{i=1}^{\infty}\lambda_i^2\left(\int_0^T \psi_i(t)\sum_{n=1}^{N}\frac{a_n}{\lambda_n^2}\cdot\psi_n(t)dt\right)^2$$

$$= \sum_{n=1}^{N}\frac{a_n^2}{\lambda_n^2} - \frac{1}{2}\sum_{n=1}^{N}\frac{a_n^2}{\lambda_n^2}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\frac{a_n^2}{\lambda_n^2} \leq \Lambda_X^*(x) < \infty$$

for all $N \geq 1$, and so

$$\sum_{n=1}^{\infty} \frac{a_n^2}{\lambda_n^2} < \infty$$

but K is a bounded linear operator and so

$$\lambda_n^2 \leq \| K \| < \infty$$

$$\Rightarrow \sum_{n=1}^{\infty} a_n^2 \leq \| K \| \cdot \sum_{n=1}^{\infty} \frac{a_n^2}{\lambda_n^2} < \infty$$

□

Remark

For a Gaussian random variable X on R^d with mean zero and positive definite covariance matrix Σ we found the entropy to be

$$\Lambda_X^*(x) = \frac{1}{2} x^T \Sigma^{-1} x$$

where $x \in R^d$. Σ positive definite implies the existence of a unitary matrix U and positive eigenvalues $\lambda_1^2, \cdots, \lambda_d^2$ such that

$$U^T \Sigma U = D$$

where D is a diagonal matrix with entries the eigenvalues of Σ. The eigenvectors form a complete basis of R^d and so there exist $a = (a_1, \cdots, a_d)^T$ such that $x = Ua$. Hence

$$\begin{aligned} \Lambda_X^*(x) &= \frac{1}{2} x^T \left(UDU^T\right)^{-1} x \\ &= \frac{1}{2} (U^T x)^T D^{-1} U^T x \\ &= \frac{1}{2} a^T D^{-1} a = \frac{1}{2} \sum_{j=1}^{d} \frac{a_j^2}{\lambda_j^2} \end{aligned}$$

So we see that the formula above is the infinite analog of the formula for the entropy of a Gaussian random variable in finite dimensions.

Example 6.1 Now we use the formula above to find the entropy of the Ornstein-Uhlenbeck process, i.e. of a mean zero Gaussian process with covariance function

$\Gamma(s,t) = e^{-\beta|s-t|}$, $\beta > 0$. First we have to find an equation for the eigenvalues and eigenfunctions of Γ.

$$\lambda_n^2 \psi_n(s) = \int_0^T \Gamma(s,t)\psi_n(t)dt$$

$$= \int_0^s e^{-\beta s+\beta t}\psi_n(t)dt + \int_s^T e^{\beta s-\beta t}\psi_n(t)dt$$

Differentiating with respect to s gives

$$\lambda_n^2 \dot{\psi}_n(s) = -\beta \int_0^s e^{-\beta s+\beta t}\psi_n(t)dt + \beta \int_s^T e^{\beta s-\beta t}\psi_n(t)dt$$

and

$$\lambda_n^2 \ddot{\psi}_n(s)$$

$$= \beta^2 \int_0^s e^{-\beta s+\beta t}\psi_n(t)dt - \beta\psi_n(s) + \beta^2 \int_s^T e^{\beta s-\beta t}\psi_n(t)dt - \beta\psi_n(s)$$

So we get

$$\lambda_n^2 \ddot{\psi}_n(s) = \left(\lambda_n^2\beta^2 - 2\beta\right)\psi_n(s)$$

with boundary condition

$$\lambda_n^2 \psi_n(0) = \int_0^T e^{-\beta t}\psi_n(t)dt$$

or

$$\lambda_n^2 \left(\ddot{\psi}_n(s) - \beta\psi_n(s)\right) = -2\beta\psi_n(s)$$

Now let $x \in H(\Gamma)$. To make the computations somewhat easier, we will assume that x is twice differentiable. Define $a_n = \int_0^T x(t)\psi_n(t)dt$. Then

$$2 \cdot \Lambda_X^*(x) = \sum_{n=1}^{\infty} \frac{a_n^2}{\lambda_n^2}$$

$$= \sum_{n=1}^{\infty} \lambda_n^{-2} \left(\int_0^T x(t)\psi_n(t)dt\right)^2$$

$$= \sum_{n=1}^{\infty} \lambda_n^{-2} \left(\int_0^T x(t)\frac{\lambda_n^2}{2\beta}\left[\ddot{\psi}_n(t) - \beta^2\psi_n(t)\right] dt\right)^2$$

$$\sum_{n=1}^{\infty} \frac{\lambda_n^2}{4\beta^2} \left(\int_0^T x(t)\ddot{\psi}_n(t)dt \; - \; \beta^2 \int_0^T x(t)\psi_n(t)dt \right)^2$$

Now

$$\int_0^T x(t)\ddot{\psi}_n(t) \;\; = \;\; C_n \; + \; \int_0^T \ddot{x}(t)\psi_n(t)dt$$

where

$$C_n \; = \; x(T)\dot{\psi}_n(T) - x(0)\dot{\psi}_n(0) - \dot{x}(T)\psi_n(T) + \dot{x}(0)\psi_n(0)$$

Note that using the continuity of $\dot{\psi}_n$ and the differential equation above we have

$$\begin{aligned} C_n = \; & x(T)\frac{1}{\lambda_n^2}\left(-\beta \int_0^T \Gamma(T,t)\psi_n(t)dt\right) \; - \; x(0)\frac{1}{\lambda_n^2}\left(\beta \int_0^T \Gamma(0,t)\psi_n(t)dt\right) \\ & - \dot{x}(T)\psi_n(T) \; + \; \dot{x}(0)\psi_n(0) \\ = \; & -\beta x(T)\psi_n(T) - \beta x(0)\psi_n(0) - \dot{x}(T)\psi_n(T) + \dot{x}(0)\psi_n(0) \\ = \; & -\left(\beta x(T) + \dot{x}(T)\right)\psi_n(T) \; - \; \left(\beta x(0) - \dot{x}(0)\right)\psi_n(0) \\ =: & -\left(\alpha(T)\psi_n(T) \; + \; \alpha(0)\psi_n(0)\right) \end{aligned}$$

where we define

$$\alpha(t) \; = \; \beta x(t) \; - \; \dot{x}(t)$$

for all $t \in [0,T]$. Then we get

$$\begin{aligned} 2\Lambda_X^*(x) \; & = \; \sum_n \frac{a_n^2}{\lambda_n^2} \\ & = \; \frac{1}{4\beta^2}\sum_n \lambda_n^2 \left(C_n \; + \; \int_0^T (\ddot{x}(t) - \; \beta^2 x(t))\psi_n(t)dt \right)^2 \end{aligned}$$

Fix a u with $0 \le u \le T$. Then integrating by parts gives the following :

$$\begin{aligned} & \int_0^T (\ddot{x}(t) - \beta^2 x(t))\Gamma(u,t)dt \\ = \; & \int_0^u \ddot{x}(t)e^{\beta(t-u)}dt \; - \beta^2 \int_0^u x(t)e^{\beta(t-u)}dt \\ + \; & \int_u^T \ddot{x}(t)e^{\beta(u-t)}dt \; - \; \beta^2 \int_u^T x(t)e^{\beta(u-t)}dt \end{aligned}$$

$$= \dot{x}(u) - \dot{x}(0)e^{-\beta u} - \beta \int_0^u \dot{x}(t)e^{\beta(t-u)}dt$$

$$- \beta x(u) + \beta x(0)e^{-\beta u} + \beta \int_0^u \dot{x}(t)e^{\beta(t-u)}dt$$

$$+ \dot{x}(T)e^{\beta(u-T)} - \dot{x}(u) + \beta \int_u^T \dot{x}(t)e^{\beta(u-t)}dt$$

$$+ \beta x(T)e^{\beta(u-T)} - \beta x(u) - \beta \int_u^T \dot{x}(t)e^{\beta(u-t)}dt$$

$$= \alpha(0)\Gamma(0,u) - 2\beta x(u) + \alpha(T)\Gamma(T,u)$$

Hence

$$8\beta^2\Lambda_X^*(x)$$

$$= \sum_n \lambda_n^2[\alpha^2(T)\psi_n^2(T) + \alpha^2(0)\psi_n^2(0) + 2\alpha(T)\alpha(0)\psi_n(T)\psi_n(0)$$

$$- 2\left(\alpha(T)\psi_n(T) + \alpha(0)\psi_n(0)\right)\int_0^T (\ddot{x}(t) - \beta^2 x(t))\psi_n(t)dt$$

$$+ \int_0^T\int_0^T (\ddot{x}(t) - \beta^2 x(t))\psi_n(t)(\ddot{x}(s) - \beta^2 x(s))\psi_n(s)dsdt]$$

$$= \alpha^2(T) + 2\alpha(T)\alpha(0)e^{-\beta T} + \alpha^2(0)$$

$$-2\alpha(T)\int_0^T (\ddot{x}(t) - \beta^2 x(t))\Gamma(T,t)dt - 2\alpha(0)\int_0^T (\ddot{x}(t) - \beta^2 x(t))\Gamma(0,t)dt$$

$$+ \int_0^T\int_0^T (\ddot{x}(t) - \beta^2 x(t))(\ddot{x}(s) - \beta^2 x(s))\Gamma(s,t)dsdt$$

$$= \alpha^2(T) + 2\alpha(T)\alpha(0)e^{-\beta T} + \alpha^2(0)$$

$$- 2\alpha(T)\left(\alpha(T) - 2\beta x(T) + \alpha(0)e^{-\beta T}\right)$$

$$- 2\alpha(0)\left(\alpha(0) - 2\beta x(0) + \alpha(T)e^{-\beta T}\right)$$

$$+ \int_0^T (\ddot{x}(t) - \beta^2 x(t))\left(\alpha(T)e^{-\beta(T-t)} - 2\beta x(t) + \alpha(0)e^{-\beta T}\right)dt$$

$$= -\alpha^2(T) - 2\alpha(T)\alpha(0)e^{-\beta T} - \alpha^2(0)$$

$$+ 4\beta\alpha(T)x(T) + 4\beta\alpha(0)x(0)$$

$$+ \alpha(T)\left(\alpha(T) - 2\beta x(T) + \alpha(0)e^{-\beta T}\right)$$

$$+ \alpha(0)\left(\alpha(0) - 2\beta x(0) + \alpha(T)e^{-\beta T}\right)$$

$$- 2\beta \int_0^T (\ddot{x}(t) - \beta^2 x(t))x(t)dt$$

$$= \; 2\beta\alpha(T)x(T) \; + \; 2\beta\alpha(0)x(0)$$

$$- 2\beta\left(\dot{x}(T)x(T) - \dot{x}(0)x(0) - \int_0^T \dot{x}^2(t)dt - \int_0^T \beta^2 x^2(t)x(t)dt\right)$$

$$= \; 2\beta^2\left(\int_0^T \left(\frac{1}{\beta}\dot{x}^2(t) + \beta x^2(t)\right) dt \; + \; x^2(0) + x^2(T)\right)$$

and so finally

$$\Lambda_X^*(x) \; = \; \frac{1}{4}\left(\int_0^T \left(\frac{1}{\beta}\dot{x}^2(t) + \beta x^2(t)\right) dt \; + \; x^2(0) + x^2(T)\right)$$

if $x \in H$ and infinity otherwise.

Note that in this example we were able to compute the entropy without finding explicit expressions for the eigenvalues and eigenfunctions of Γ.

Theorem 6.5 *Let $X(t), t \in [0,T]$ be a centered Gaussian process covariance kernel Γ. Let $H(\Gamma)$ be the reproducing kernel Hilbert space of Γ with inner product norm $\| \cdot \|_{H(\Gamma)}$. Then*

$$\Lambda_X^*(x) \; = \; \begin{cases} \frac{1}{2} \| x \|_{H(\Gamma)}^2 & \text{if } x \in H(\Gamma) \\ \infty & \text{otherwise} \end{cases}$$

Proof.

By the previous theorem, it suffices to show that if $x \in H(\Gamma)$ and

$$x(t) \; = \; \sum_n a_n \psi_n(t)$$

then

$$\sum_n \frac{a_n^2}{\lambda_n^2} \; = \| x \|^2$$

But

$$\| x \|^2 = (x, x)$$

$$= \left(\sum_n a_n \psi_n \, , \, \sum_k a_k \psi_k \right)$$

$$= \sum_{n,k} a_n a_k \, (\psi_n, \psi_k)$$

$$= \sum_{n,k} \frac{a_n a_k}{\lambda_n^2 \lambda_k^2} \left(\int_0^T \Gamma(t, \cdot)\psi_n(t)dt \, , \, \int_0^T \Gamma(s, \cdot)\psi_k(s)ds \right)$$

$$= \sum_{n,k} \frac{a_n a_k}{\lambda_n^2 \lambda_k^2} \int_0^T \int_0^T \psi_n(t)\psi_k(s) \; (\Gamma(t, \cdot) \, , \, \Gamma(s, \cdot) \; dt \; ds)$$

$$= \sum_{n,k} \frac{a_n a_k}{\lambda_n^2 \lambda_k^2} \int_0^T \int_0^T \psi_n(t)\psi_k(s)\Gamma(t, s) \; dt \; ds$$

$$= \sum_{n,k} \frac{a_n a_k}{\lambda_n^2 \lambda_k^2} \int_0^T \psi_n(t) \left(\int_0^T \psi_k(s)\Gamma(t, s)ds \right) dt$$

$$= \sum_{n,k} \frac{a_n a_k}{\lambda_n^2 \lambda_k^2} \int_0^T \psi_n(t)\lambda_k^2 \psi_k(t)dt$$

$$= \sum_n \frac{a_n^2}{\lambda_n^2}$$

□

Example 6.2 Let $\varphi \in C[0, T]$, $\varphi > 0$ and let

$$X(t) = \int_0^t \varphi(s)dB(s) \, , \; 0 \leq t \leq T$$

where B is standard Brownian Motion. Let

$$H = \left\{ x \in C[0, T] \; : \; \exists \dot{x} \in L^2([0, T]) \;\; with \;\; x(t) = \int_0^t \dot{x}(s)ds \right\}$$

with inner product

$$(x, y) = \int_0^T \frac{\dot{x}(s)\dot{y}(s)}{\varphi^2(s)} ds$$

Now

$$\Gamma(t, s) = E \left(\int_0^t \varphi(u)dB(u) \cdot \int_0^s \varphi(v)dB(v) \right)$$

$$= \int_0^{t \wedge s} \varphi^2(u) du$$

The derivative of $\Gamma(t, \cdot)$ exists everywhere but at t and is

$$\frac{d}{ds}\Gamma(t,s) = \varphi^2(s) I_{[0,t]}(s)$$

Hence $\Gamma(t, \cdot) \in H$ and

$$(x, \Gamma(t, \cdot)) = \int_0^T \frac{\dot{x}(s)\varphi^2(s) I_{[0,t]}(s)}{\varphi^2(s)} ds = x(t)$$

So H is the reproducing kernel Hilbert space of Γ and

$$\Lambda_X^*(x) = \begin{cases} \frac{1}{2}\int_0^T \frac{\dot{x}^2(s)}{\varphi^2(s)} ds & \text{if } x \in H \\ \infty & \text{otherwise} \end{cases}$$

This technique is especially useful to verify that a certain expression holds for Λ^* because the conditions of Theorem 6.2 are usually easy to check.

Chapter 7

Entropy of finite-state Markov chains

In this chapter we will study the entropy of Markov chains on a finite state space $S = \{0, \cdots, N\}, N \geq 1$. For the case of two states we will derive an explicit formula for the entropy. In the general case we will prove a central limit theorem for $\Lambda^*_{X_1,\cdots,X_n}(X_1, \cdots, X_n)$, i.e. for the entropy at the observed path. We will denote the initial distribution by ν and the transition probabilities by $P = (p_{ij})$. Let $X^n = (X_1, \cdots, X_n)^T$. Then we have the following theorem:

Theorem 7.1 *Let $S = \{0,1\}$. Then for $x \in S^n$ we have*

$$\Lambda^*_{X^n}(x) = -\log P(X_1 = x_1) - \sum_{k=1}^{n-1} \log p_{x_k x_{k+1}}$$

Proof.

For $\lambda \in R^n$ we have

$$\begin{aligned} & E\left[\exp\left(\langle \lambda, X^n \rangle\right)\right] \\ = \; & E\left[\exp\left(\sum_{k=1}^{n} \lambda_k X_k\right)\right] \\ = \; & \sum_{\zeta \in S^{n+1}} \nu(\zeta_0) p_{\zeta_0 \zeta_1} e^{\lambda_1 \zeta_1} \cdots p_{\zeta_{n-1}\zeta_n} e^{\lambda_n \zeta_n} \end{aligned}$$

Let $\varphi \in C(R^n)$ be defined by

$$\varphi(\lambda) = \sum_{k=1}^{n} \lambda_k x_k - \log E\left[\exp\left(\sum_{k=1}^{n} \lambda_k X_k\right)\right]$$

Case 1 : $x_j = 0$

$$\Rightarrow \frac{d\varphi}{d\lambda_j} = -\frac{E\left[X_j \exp\left(\sum_{k=1}^{n} \lambda_k X_k\right)\right]}{E\left[\exp\left(\sum_{k=1}^{n} \lambda_k X_k\right)\right]} \leq 0$$

for all $\lambda \in R^n$. So φ is monotonically decreasing in λ_j.

Case 2 : $x_j = 1$

$$\Rightarrow \frac{d\varphi}{d\lambda_j} = 1 - \frac{E\left[X_j \exp\left(\sum_{k=1}^{n} \lambda_k X_k\right)\right]}{E\left[\exp\left(\sum_{k=1}^{n} \lambda_k X_k\right)\right]}$$

$$= \frac{E\left[(1-X_j) \exp\left(\sum_{k=1}^{n} \lambda_k X_k\right)\right]}{E\left[\exp\left(\sum_{k=1}^{n} \lambda_k X_k\right)\right]} \geq 0$$

for all $\lambda \in R^n$. So φ is monotonically increasing in λ_j.

$$\Lambda^*_{X^n}(x) = \sup\{\varphi(\lambda) : \lambda \in R^n\}$$

$$\lim_{\lambda_k \to (-1)^{1-x_k}\infty} \varphi(\lambda)$$

$$= -\log\left(\nu_0 p_{0x_1} p_{x_1x_2} \cdots p_{x_{n-1}x_n} + \nu_1 p_{1x_1} p_{x_1x_2} \cdots p_{x_{n-1}x_n}\right)$$

$$= -\log P(X_1 = x_1) - \sum_{k=1}^{n-1} \log p_{x_k x_{k+1}}$$

□

Example 7.1 Let $\nu = \delta_0$ and

$$P = \begin{bmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{bmatrix}$$

$$\Lambda^*_{X^5}((0,0,0,0,0)^T) = -\log P(X_1 = 0) - \sum_{k=1}^{4} \log p_{00}$$

$$= -5 \log p_{00} = 8.05$$

In order to prove the next result, we first have to introduce the notion of an r-block. Let I be the space of all subsets of $\bigcup_{n=0}^{\infty} S^n \cup S^{\infty}$, that is, the space of all paths of finite or infinite length and give it the smallest σ -algebra that contains all paths. Let $r \in S$ and define an r-block B as an element of I that starts with r but does not contain any further r. The space of r-blocks is a measurable subset of I, and we give it the relative topology. Let $\tau_1, \tau_2, \cdots$ be the consecutive times with $X_n = r$. The m'th r-block B_m is then the sample sequence X from τ_m to just before τ_{m+1}, shifted to the left so as to start at time 0. Formally, let $\tau_1 = \inf\{n : X_n = r\}$ if $X_n = r$ for some finite n, otherwise $\tau_1 = \infty$. If $\tau_k < \infty$, then let $\tau_{k+1} = \inf\{n > \tau_k : X_n = r\}$, otherwise $\tau_{k+1} = \infty$. On $\tau_m < \infty$, let B_m be the sequence of length $\tau_{m+1} - \tau_m$, whose n'th term is $X_{\tau_m + n}$ for $0 \leq n < \tau_{m+1} - \tau_m$. On $\tau_m = \infty$, let $B_m = \emptyset$. Let $\mu = P_r B_1^{-1}$. Then we have the following result due to Doeblin (1938):

Theorem 7.2 *The sequence $B_1, B_2, \ldots$ is independent and identically distributed.*

Theorem 7.3 *Let $\{X_n, n \geq 0\}$ be an irreducible Markov chain on $S = \{0, \cdots, N\}$ with stationary distribution π. Let $r \in S$ and $\{B_m\}$ be the sequence of r-blocks. Assume $P(X_1 = r) = 1$. Further let*

$$\mu = E\left[\Lambda^*_{B_1}(B_1)\right]$$

$$\sigma^2 = Var\left[\Lambda^*_{B_1}(B_1)\right]$$

Then

$$\frac{\Lambda^*_{X^n}(X^n) - \pi_r n\mu}{\sqrt{\pi_r n \sigma^2}} \Rightarrow Z$$

where Z is a standard normal random variable.

Proof.

Let $\zeta_n = \sum_{k=1}^n I_r(X_k)$ be the number of visits to r up to time n. Then by the additivity of the entropy for independent random variables we have

$$\Lambda^*_{X^n}(X^n) = \sum_{k=1}^{\zeta_n - 1} \Lambda^*_{B_1}(B_k) + \Lambda^*_{X^n}\left((r, X_{\zeta_n+1}, \cdots, X_n)\right)$$

Now by the standard central limit theorem for iid random variables we get

$$\frac{\sum_{k=1}^{[\pi_r n]} \Lambda^*_{B^1}(B_k) - \pi_r n\mu}{\sqrt{\pi_r n \sigma^2}} \Rightarrow Z$$

So it remains to show that $\frac{1}{n}\zeta_n \to \pi_r$ in probability and that

$$\frac{1}{\sqrt{n}}\Lambda^*_{X^n}\left((r, X_{\zeta_n+1}, \cdots, X_n)\right) \to 0$$

in probability. But the first is well a known fact from the theory of Markov chains, namely the ergodic theorem, and the second follows from

$$\Lambda^*_{X^n}\left((r, X_{\zeta_n+1}, \cdots, X_n)^T\right) \leq \Lambda^*_{X_n}(B_{\zeta_n}) = \Lambda^*_{B_1}(B_1) < \infty \ a.s.$$

□

Chapter 8

Entropy and Statistics

In 1957 Solomon Kullback published his now classical book "Information Theory and Statistics." In this book he derives major portions of modern Statistics such as parameter estimation and hypothesis testing. His starting point is a functional now known as the Kullback-Liebler infomation, which is the same as the relative entropy discussed in Chapter 2. Using the notation from Kullback [7] we have the following : Let f and g be densities of a dominated set of probability measures on a measurable space and let λ be the dominating measure. Then for any Borel set A we let

$$\mu(A) := \int_A g(t)\lambda(dt)$$

and

$$\nu(A) := \int_A f(t)\lambda(dt)$$

Then the Kullback-Leibler information is defined by

$$I(f:g) \;=\; \int f(t)\log\frac{f(t)}{g(t)}\lambda(dt) \;=\; H(\nu|\mu)$$

An important notion in Kullback [7] is the minimum discrimination information $I(*:g)$, which is obtained by minimizing $I(f:g)$ over all members of the dominated set of probability measures subject to the constraint

$$\int T(x)\nu(dx) \;=\; \Theta$$

where T is a measurable statistic and Θ is a constant, that is, we find the minimum over the set of measures for which T is an unbiased estimator of the parameter Θ.

Letting $T(x) = x$ we get

$$
\begin{aligned}
I(* : g) &= \inf\left\{I(f : g) : \int T(x)f(x)\lambda(dx) = \Theta\right\} \\
&= \inf\left\{H(\nu|\mu) : \int x\nu(dx) = \Theta\right\} = \Lambda^*_\mu(\Theta)
\end{aligned}
$$

where the last inequality follows from Theorem 2.1. So we see that for the case $T(x) = x$ Kullback's minimum discrimination information and entropy coincide. It is therefore no surprise that entropy should be of interest to the statistician.

As was shown by Kullback, using the minimum discrimination information and therefore entropy to do statistics leads in many cases to well known procedures. As an example, consider the following :

We want to study the following simple hypothesis testing problem. Assume $X_1, \cdots, X_n$ are independent samples from a normal population with unknown mean μ and standard deviation 1. We want to test the null hypothesis $H_0 : \mu = 0$ vs. the alternative $H_1 : \mu = 1$. Note that under the null hypothesis the entropy is zero at 0 and becomes bigger as we move farther away from zero. Therefore we choose the critical region of the test as follows:

$$
\{\Lambda^*_{H_0}(X_1, \cdots, X_n) - \Lambda^*_{H_1}(X_1, \cdots, X_n) \geq c\}
$$

where $\Lambda^*_{H_0}$ is the entropy computed under the null hypothesis and $\Lambda^*_{H_1}$ is the entropy under the alternative.

Using the independence of the observations and Example 1.1 we get

$$
\begin{aligned}
&\{\Lambda^*_{H_0}(X_1, \cdots, X_n) - \Lambda^*_{H_1}(X_1, \cdots, X_n) \geq c\} \\
&= \{\frac{1}{2}\sum_{i=1}^{n} X_i^2 - \frac{1}{2}\sum_{i=1}^{n}(X_i - 1)^2 \geq c\} \\
&= \{\sum_{i=1}^{n} X_i - \frac{n}{2} \geq c\}
\end{aligned}
$$

$$= \{\sum_{i=1}^{n} X_i \geq c_1\}$$

which is of course the same critical region one gets using the likelihood ratio statistic.

Next we will look at another hypothesis testing procedure, but this time the resulting test will be different from those commonly employed.

Say we have an independent sample $X_1, \cdots, X_n$ on the set $\{s_1, \cdots, s_N\}$, that is, $P(X_i = s_j) = \alpha_j \;\; j = 1, \cdots, N$ and $i = 1, \cdots, n$, with the α_i's unknown and we want to test $H_0 : p = \beta$ vs. $H_1 : p \neq \beta$. Now again denoting by $\Lambda^*_{H_0}$ the entropy under the null hypothesis and letting f_k be the number of observations equal to s_k we get

$$\begin{aligned} Z \;&:=\; \frac{\sum_{i=1}^{n} \Lambda^*_{H_0}(X_i) \;-\; n \cdot E[\Lambda^*_{H_0}(X_1)]}{\sqrt{n \cdot Var[\Lambda^*_{H_0}(X_1)]}} \\ &= \frac{\sum_{k=1}^{N} f_k \Lambda^*_{H_0}(s_k) \;-\; n \cdot \sum_{k=1}^{N} \beta_k \Lambda^*_{H_0}(s_k)}{\sqrt{n \cdot \left(\sum_{k=1}^{N} \beta_k \Lambda^{*2}_{H_0}(s_k) - (\sum_{k=1}^{N} \beta_k \Lambda^*_{H_0}(s_k))^2\right)}} \\ &= \frac{\sum_{k=1}^{N} (f_k - n\beta_k) \Lambda^*_{H_0}(s_k)}{\sqrt{n \cdot \left(\sum_{k=1}^{N} \beta_k \Lambda^{*2}_{H_0}(s_k) - (\sum_{k=1}^{N} \beta_k \Lambda^*_{H_0}(s_k))^2\right)}} \\ &\Rightarrow \;\; N(0,1) \end{aligned}$$

that is, the test statistic Z converges to a standard normal random variable by the central limit theorem.

The test statistic usually employed here is Pearson's χ^2 statistic, which is given by

$$\chi^2 \;=\; \sum_{k=1}^{N} \frac{(f_k - n\beta_k)^2}{n\beta_k}$$

Under the null hypothesis χ^2 has asymptotically a χ^2-distribution with $N - 1$ degrees of freedom.

Evidence from simulation suggests that the entropy test based on Z has a slightly larger type I error than the χ^2 test but that it has a lower type II error. In Table 1 we have the results of some simulations of the type I error for $s_k = k - 1$, $\alpha_k = 0.2 \forall k$ and $N = 5$:

Type I		
n	Z	χ^2
50	0.102	0.042
100	0.121	0.045
150	0.114	0.038
500	0.102	0.046

In Table 2 we collected the results of a simulation for the type II error. Again we have $s_k = k-1$, $\alpha_k = 0.2 \forall k$ and $N = 5$ but now the null hypothesis specifies the following probabilities : $H_0 : \beta_1 = 0.24$, $\beta_2 := 0.16$, $\beta_3 = 0.2$, $\beta_4 = 0.16$, $\beta_5 = 0.24$

Type II		
n	Z	χ^2
50	0.653	0.773
100	0.539	0.701
150	0.435	0.572
500	0.053	0.072

Note also that the test statistic Z depends explicitly on the state space, whereas the χ^2 statistic is the same no matter what the state space looks like. This might be useful in some applications. As an example how the type I error changes depending on the state space, consider the results of the simlation shown in the next table. Here we use the uniform distribution:

s_0	s_1	s_2	s_3	s_4	z
-2	-1	0	1	2	0.940
-5	-1	0	1	5	0.106
-10	-1	0	1	10	0.100
-20	-1	0	1	20	0.062
-30	-1	0	1	30	0.041

In the last example we will use the additivity property of entropies of independent random variables to derive a test whether a given sequence $\{X_1, \cdots, X_n\}$ is an independent sequence or not. Moreover, if it is not independent, we will be able to tell whether it depends linearly on an independent sequence.

As usual we will need the assumption that the population has a moment generating function. Denoting by $\mu^{(n)}$ the law of $\{X_1, \cdots, X_n\}$ and by μ the law of X_1 we can state the null hypothesis of independence in the following way:

$$H_0 : \Lambda^*_{\mu^{(n)}}(X_1, \cdots, X_n) = \sum_{i=1}^{n} \Lambda^*_\mu(X_i)$$

Of course we don't know $\mu^{(n)}$, but we can estimate it using the empirical distribution function.

Under the null hypothesis it is admissible to split the sequence up into pieces of any length. We will split the sequence $\{X_1, \cdots, X_n\}$ into pieces of length $k > 1$. The choice of the right k will be of importance, and in general the procedure should be repeated with different values of k.

If we would use the whole sequence to estimate the moment generating function and to evaluate the entropy, than the resulting sum would not be the sum of independent random variables. It can be shown that we still have

$$\frac{1}{n} \sum_{i=1}^{n} \Lambda^*_{\mu^{(n)}}(X_i) \to E[\Lambda^*_\mu(X_1)]$$

where the convergence is in probability, but evidence from simulations shows the convergence to be very slow. Therefore we will use one part of the sequence to estimate the moment generating function and the other to evaluate the test statistic.

This leads to the following procedure : Let $0 < r < 1$ and $s = [rn]$. Let $k > 1$ and assume that $k|s$ and that $k|n$. This last assumption is for convenience only and of no importance to the theory.

Next we split up the sequence in two pieces of lengths s and $n - s$, and then split these pieces up further into sequences of length k :

$$w_j = (X_{jk+1}, \cdots, X_{(j+1)k}) \quad j = 0, \cdots, \frac{s}{k} - 1$$

$$v_j = (X_{jk+1}, \cdots, X_{(j+1)k}) \quad j = \frac{s}{k}, \cdots, \frac{n}{k} - 1$$

Denote by δ_W the probability measure in R^k which puts probability $\frac{k}{s}$ at each point w_j and denote by δ_X the probability measure in R that puts probability $\frac{1}{n}$ at each of the observations X_i.

Now the estimate for $\Lambda^*_{(X_1,\cdots,X_n)}(X_1,\cdots,X_n)$ is given by

$$\hat{\Lambda}^* := \sum_{j=\frac{s}{k}}^{\frac{n}{k}-1} \Lambda^*_{\delta_W}(v_j)$$

$$= \sum_{j=\frac{s}{k}}^{\frac{n}{k}-1} \sup\left\{\lambda' v_j - \log\frac{k}{s}\sum_{i=0}^{\frac{s}{k}-1}\exp(\lambda' w_i) \;:\; \lambda \in R^k\right\}$$

The estimates for the expectation and the variance of Λ^* are computed using the respective estimates of the moments as follows :

$$\hat{E}[(\Lambda^*_{X_1}(X_1))^m] = \frac{1}{n}\sum_{i=1}^{n}(\Lambda^*_{\delta_X}(X_i))^m$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sup\left\{\lambda \cdot X_i - \log\frac{1}{n}\sum_{j=1}^{n} e^{\lambda X_j} \;:\; \lambda \in R\right\}\right)^m \qquad m = 1,2$$

and

$$\hat{\mu} = \hat{E}[\Lambda^*_{X_1}(X_1)]$$

$$\hat{\sigma}^2 = \hat{E}[(\Lambda^*_{X_1}(X_1))^2] - (\hat{\mu})^2$$

Then it follows from the central limit theorem that

$$Z := \frac{\hat{\Lambda}^* - (n-s)\hat{\mu}}{\sqrt{(n-s)\hat{\sigma}^2}} \Rightarrow N(0,1)$$

where $N(0,1)$ stands for the law of a standard normal random variable.

The next table shows the results of a simulation study for this test statistic. In this simulation we generated sequences with 750 observations each. The first 500 of these were used to estimate the moment generating function, the last 250 to compute the test statistic Z, that is, we used $r = \frac{2}{3}$. We used $k = 2$ for $\hat{\Lambda}^*$.

We generated three sequences of independent standard normal random variables $u_1, v_1, w_1, \cdots, u_{375}, v_{375}, w_{375}$. Then for the column marked "Independent" we computed the $X_i^{\prime}s$ as follows :

$$X_{2i-1} = u_i \ , \ X_{2i} = v_i$$

For the column with " Linearly Dependent" we used

$$X_{2i-1} = u_i + v_i \ , \ X_{2i} = u_i - v_i$$

and for the column with "Nonlinearly Dependent" we let

$$X_{2i-1} = u_i * v_i * w_i \ , \ X_{2i} = (u_i + v_i) * w_i$$

Number	Independent	Linearly Dependent	Nonlinearly Dependent
1	-0.909	-1.283	-1.098
2	-0.625	-0.314	1.198
3	-0.633	-1.738	-1.741
4	-1.025	-1.007	-0.364
5	-0.638	0.116	-1.466
6	0.126	-1.251	-0.618
7	0.662	0.033	-2.307
8	0.065	-1.017	-1.697
9	-1.011	-1.252	-1.194
10	-1.901	-0.064	-0.240
11	0.410	-0.682	-1.523
12	0.368	-0.056	-1.416
13	0.226	-0.063	-2.142
14	0.051	0.650	0.621
15	-0.492	0.164	

This simulation shows that this test can distinguish clearly between sequences that are linearly dependent on an independent sequence and those that are not.

On the other hand, it does not work very well to separate independent sequences and those that are linear transformations of an independent sequence. The reason for this lies in the following :

Let $U = (u_1, \cdots, u_n)^T$ be a sequence of independent random variables, and let T be a nonsingular $n \times n$ matrix. Suppose we observe $X = T \cdot U$. Then

$$\Lambda^*_{\mu^{(n)}}(X)$$

$$= \sup\left\{\lambda^T X \; - \; \log E[e^{\lambda^T X}] \; : \; \lambda \in R^n\right\}$$

$$= \sup\left\{(\lambda^T T)U \; - \; \log E[e^{(\lambda^T T)U}] \; : \; \lambda \in R^n\right\}$$

$$= \sum_{i=1}^{n} \Lambda^*_{u_1}(u_i) \; = \; \sum_{i=1}^{n} \Lambda^*_{u_1}((T^{-1}X)_i)$$

and so the entropy of the sequence is the sum of entropies just as it would be for an independent sequence.

This can be used in the following manner: first one tests for independence as described above. Should the test reject the null hypothesis, one can conclude that the sequence does not even depend linearly on an independent sequence. Should the test fail to reject H_0 one repeats the test with a nonlinear transformation of the data that preserves the independence, for example $Y_i = X_i^2$. Should the test again fail to reject H_0 one can conclude that the sequence is independent, otherwise one has a sequence that depends linearly on an independent sequence.

In a second simulation study, we generated an independent sequence on the set $\{0, 1, 2\}$. Here the distribution was the uniform, and the length N of the sequence was 9000. Using different values of r, we found the following values for the type I error:

r	α
0.222	0.35
0.333	0.25
0.444	0.16
0.555	0.13
0.667	0.12
0.777	0.07
0.888	0.11

This simulation shows that the type I error decreases when r increases. That implies that it is important to estimate the moment generating function well.

Simulations of this test have been seriously hampered by the lack of an appropriate optimization procedure to compute the various Λ^*'s. The usual Newton algorithm where one finds improved estimates of the maximizing λ using the formula

$$\lambda_{n+1} = \lambda_n - \frac{f'(\lambda_n)}{f''(\lambda_n)}$$

does not work very well here because the functions f under consideration, namely

$$f(\lambda) = \lambda \cdot x - \log E[e^{\lambda X}]$$

are often nearly linear everywhere but at a neighborhood of the maximum. That of course leads to f'' being nearly zero and renders the Newton algorithm useless. As an example, consider the two-point distribution $P(X = 0) = 0.99$ and $P(X = 10) = 0.01$. Then

$$\Lambda_X^*(2) = \sup\left\{\lambda \cdot 2 - \log(0.99 + 0.01 e^{\lambda \cdot 10})\right\}$$

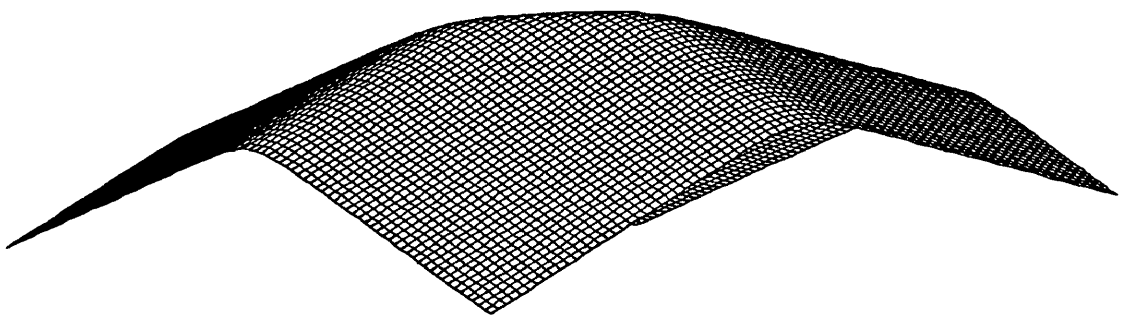
Analytically one finds the sup at $\lambda = 0.32$, but the Newton algorithm fails to find this value for any starting value outside of $(0.13, 0.63)$.

The problem becomes even more serious if one tries to optimize in higher dimensions.

Example 8.1 We used Matlab to illustrate this optimization problem. Consider the probability measure in R^2 giving probability 0.001 to the points $(0,0)$, $(0,1)$, (1,0) and probability 0.997 to the point (1,1). Then finding $\Lambda^*(1,1)$ means optimizing the following function:

$$\varphi(\lambda_1, \lambda_2) = \lambda_1 + \lambda_2 - \log\left(0.001 + 0.001e^{\lambda_1} + 0.001e^{\lambda_2} + 0.997 * e^{\lambda_1+\lambda_2}\right)$$

This function is shown in the picture for $-20 \le \lambda_1 \le 20$ and $-20 \le \lambda_2 \le 20$. It is easy to see that the function is flat nearly everywhere.



The simulations that could be done supported the theory above, but to draw a final conclusion about the usefulness of this test further simulations using improved numerical algorithms will be necessary.

Chapter 9

References

[1] P. Billingsley. *Statistical Inference for Markov Processes,* The University of Chicago Press, 1961

[2] J.D. Deuschel and D. Stroock. *Large Deviations,* Academic Press, Mathematics 137, 1989

[3] R. Durrett. *Probability: Theory and Examples,* Wadsworth & Brooks/Cole, 1991

[4] S.N. Ethier and T. Kurtz. *Markov Processes,* John Wiley & Sons, 1986

[5] D. Freedman. *Markov Chains,* Springer Verlag, New York, 1983

[6] J. Kuelbs. *Probability on Banach Spaces,* Dekker, Advances in Probability, Vol. 4, 1978

[7] S. Kullback. *Information Theory and Statistics,* John Wiley & Sons, 1957

[8] Proceedings. *Probability in Banach Spaces V,* Springer Verlag, Lecture Notes in Mathematics Vol. 1153, 1984

[9] R.T. Rockafellar. *Convex Functions,*
Princeton University Press, 1970

[10] W. Rudin. *Real and Complex Analysis,*
McGraw Hill, 1966