READING NOTES

Information Theory and the Central Limit Theorem

Oliver Johnson (2004)

Xuanzhe Hou

October 2026

Contents

1 Introduction to Information Theory **1**
1.1 Entropy and relative entropy . . . 1
1.2 Link to thermodynamic entropy . . . 5
1.3 Fisher information . . . 6
1.4 Previous information-theoretic proofs . . . 8

2 Convergence in Relative Entropy **10**
2.1 Motivation . . . 10
2.2 Generalised bounds on projection eigenvalues . . . 12
2.3 Rates of convergence . . . 14

3 Non-Identical Variables and Random Vectors **17**
3.1 Non-identical random variables . . . 17
3.2 Random vectors . . . 20

4 Dependent Random Variables **24**
4.1 Introduction and notation . . . 24
4.2 Fisher information and approximate convolution . . . 26
4.3 Proof of the approximate subadditive relation . . . 28

5 Convergence to Stable Laws **31**
5.1 Introduction to stable laws . . . 31
5.2 Parameter estimation for stable distributions . . . 33
5.3 Extending de Bruijn's identity . . . 35
5.4 Relationships between forms of convergence . . . 36
5.5 Steps toward a Brown inequality . . . 37

6 Convergence on Compact Groups **38**
6.1 Probability on compact groups . . . 38
6.2 Convergence in relative entropy . . . 41
6.3 Comparison of forms of convergence . . . 43
6.4 Proofs of the convergence in relative entropy theorems . . . 44

7 Convergence to the Poisson Distribution **46**
7.1 Entropy and the Poisson distribution . . . 46
7.2 Fisher information . . . 48
7.3 Strength of the bounds . . . 51
7.4 A Poisson de Bruijn identity . . . 52
7.5 L^2 bounds on Poisson distance . . . 53

8 Free Random Variables **56**
8.1 Operators, traces, and spectral distributions . . . 56

8.2 Cauchy and R-transforms 56
8.3 Freeness and conjugate variables 57
8.4 Free Fisher information and free entropy 58
8.5 Projection inequality and information-theoretic free CLT 58

9 Appendices A–E **60**
9.1 Appendix A: calculating entropies 60
9.2 Appendix B: Poincaré inequalities 61
9.3 Appendix C: de Bruijn's identity 62
9.4 Appendix D: entropy power inequality 62
9.5 Appendix E: comparing modes of convergence 63

1 Introduction to Information Theory

All logarithms are natural logarithms.

1.1 Entropy and relative entropy

1.1.1 Discrete entropy

The starting point is that a rare event is more informative than a common one, and independent pieces of information should add.

Definition 1.1 (Information and discrete entropy; Book Definitions 1.1–1.2). If A has positive probability, the information gained by observing A is

$$I(A) = -\log \mathbb{P}(A).$$

If a discrete random variable X has mass function $p(x)$, its Shannon entropy is

$$H(X) = H(p) = -\sum_x p(x) \log p(x),$$

with $0 \log 0 := 0$.

For a Bernoulli variable with success probability p,

$$H(X) = -p \log p - (1-p) \log(1-p).$$

This is symmetric about $p = 1/2$, vanishes at 0 and 1, and is largest at $1/2$.

Lemma 1.2 (Elementary entropy properties; Book Lemma 1.1). For random variables with finite ranges:

1. $H(X) \geq 0$, with equality exactly when X is deterministic;
2. relabeling the states does not change entropy;
3. if X and Y are independent, then $H(X, Y) = H(X) + H(Y)$;
4. if X takes at most m values, then $H(X) \leq \log m$, with equality for the uniform distribution.

The Kraft inequality characterizes the possible lengths of binary codewords.

Lemma 1.3 (Kraft inequality; Book Lemma 1.2). There is a uniquely decipherable binary code with integer codeword lengths $r_1, \ldots, r_m$ if and only if

$$\sum_{i=1}^{m} 2^{-r_i} \leq 1.$$

An ideal code would use length $-\log_2 p_i$ for state i, so the expected optimal code length is governed by Shannon entropy.

1.1.2 Differential entropy

For a continuous variable, the entropy of finer and finer quantizations diverges. The finite part that remains after subtracting the grid contribution is differential entropy.

Definition 1.4 (Differential entropy; Book Definition 1.4). If X has density f on $\mathbb{R}$, define

$$H(X) = H(f) = -\int_{\mathbb{R}} f(x) \log f(x)\, \mathrm{d}x,$$

whenever the integral is well-defined.

If X_Δ records the cell of a grid of width Δ containing X, then under standard regularity conditions

$$H(X_\Delta) + \log \Delta \longrightarrow H(X) \qquad (\Delta \downarrow 0).$$

Unlike discrete entropy, differential entropy may be negative and is not invariant under changes of scale. If $a \neq 0$, then

$$H(aX + b) = H(X) + \log |a|. \tag{1.1}$$

Translation leaves it unchanged, while scaling contributes the logarithm of the Jacobian.

Useful reference values are

$$H(\mathrm{Unif}[0, L]) = \log L, \qquad H(\mathrm{Exp}(\text{mean } \mu)) = 1 + \log \mu,$$

and

$$H(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{2} \log(2\pi e \sigma^2).$$

Differential entropy changes under rescaling. Relative entropy compares two laws and is invariant when both undergo the same change of variables.

1.1.3 Relative entropy

Definition 1.5 (Relative entropy; Book Definition 1.5). For mass functions p, q or densities f, g, respectively,

$$\mathrm{D}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}, \qquad \mathrm{D}(f\|g) = \int f(x) \log \frac{f(x)}{g(x)}\, \mathrm{d}x.$$

If the first law is not absolutely continuous with respect to the second, the divergence is $+\infty$.

Relative entropy is nonnegative, but it is neither symmetric nor a metric.

Theorem 1.6 (Gibbs inequality; Book Lemma 1.4). For probability laws P, Q,

$$\mathrm{D}(P\|Q) \geq 0,$$

with equality if and only if $P = Q$ almost everywhere.

Proof idea. Apply $\log u \leq u - 1$ to $u = q/p$ and average under p. Equality in the scalar inequality forces $q/p = 1$ wherever $p > 0$. ⊛

The log-sum inequality follows from the same argument and implies the data processing inequality. Pinsker's inequality gives a quantitative comparison with total variation:

$$\mathrm{D}(P\|Q) \geq \frac{1}{2}\|P-Q\|_1^2 . \tag{1.2}$$

Thus convergence in relative entropy implies convergence in total variation, hence weak convergence. Appendix E shows that the converse implications fail.

Relative entropy also has operational interpretations.

- In coding, $\mathrm{D}(P\|Q)$ is the expected excess description length incurred by coding as if the law were Q when the true law is P.
- In hypothesis testing, sums of log-likelihood ratios grow at an asymptotic rate given by a relative entropy. The Bernoulli calculation in Book Lemma 1.7 is an elementary large-deviation instance.
- If Q is uniform on m points, then $\mathrm{D}(P\|Q) = \log m - H(P)$.

Theorem 1.7 (Information projection; Book Lemma 1.9). Let $\mathcal{C}$ be a closed convex family of probability laws and let P^* minimize $\mathrm{D}(P\|Q)$ over $P \in \mathcal{C}$. Then

$$\mathrm{D}(P\|Q) \geq \mathrm{D}(P\|P^*) + \mathrm{D}(P^*\|Q), \qquad P \in \mathcal{C}.$$

Consequently, any minimizing sequence converges to P^* in relative entropy when the minimum is finite.

The most important use of relative entropy in the book is the Gaussian variational principle.

Theorem 1.8 (Maximum entropy at fixed variance; Book Lemma 1.11). Let X have mean μ, variance σ^2, and density f. If $Z \sim \mathcal{N}(\mu,\sigma^2)$ has density ϕ_{μ,σ^2}, then

$$\mathrm{D}(f\|\phi_{\mu,\sigma^2}) = H(Z) - H(X) \geq 0. \tag{1.3}$$

Equality holds exactly for the Gaussian law.

Proof. The logarithm of a Gaussian density is a quadratic polynomial. Therefore $\int f \log \phi_{\mu,\sigma^2}$ depends only on the first two moments of f and equals the same integral under the matching Gaussian. Expanding $\mathrm{D}(f\|\phi)$ leaves precisely $H(\phi) - H(f)$. ⊛

Similarly, the exponential law maximizes entropy among distributions supported on $[0,\infty)$ with a prescribed mean (Book Lemma 1.12). Among all Gaussians, the closest one to a given density in relative entropy matches its mean and variance (Book Lemma 1.13).

1.1.4 Mutual information and conditional entropy

Definition 1.9 (Mutual and conditional information; Book Definition 1.6). For discrete random variables X, Y,

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = \mathrm{D}(P_{X,Y} \| P_X P_Y),$$

and

$$H(Y \mid X) = H(X,Y) - H(X).$$

Gibbs' inequality immediately yields

$$H(X,Y) \leq H(X) + H(Y),$$

with equality exactly under independence. Conditional entropy is an average:

$$H(Y \mid X) = \sum_x \mathbb{P}(X = x) H(Y \mid X = x).$$

The chain rule is $H(X,Y) = H(X) + H(Y \mid X)$.

For independent discrete variables, entropy cannot decrease under convolution:

$$H(X+Y) \geq \max\{H(X), H(Y)\}. \tag{1.4}$$

Indeed, $H(X+Y \mid X) = H(Y)$ and conditioning cannot increase entropy. For normalized sums of continuous variables, the change in entropy under scaling is accounted for by the entropy power inequality in Appendix D.

1.1.5 Axiomatic definition of entropy

Johnson reviews axiomatic characterizations due to Rényi and Faddeev. Symmetry, continuity, normalization, additivity for independent experiments, and a decomposition rule force a narrow class of information measures.

Definition 1.10 (Rényi entropy and divergence; Book Definitions 1.8–1.9). For $\alpha > 0$, $\alpha \neq 1$,

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \sum_x p(x)^\alpha,$$

and

$$\mathrm{D}_\alpha(p \| q) = \frac{1}{\alpha - 1} \log \sum_x p(x)^\alpha q(x)^{1-\alpha}.$$

As $\alpha \to 1$, l'Hôpital's rule recovers Shannon entropy and Kullback–Leibler divergence.

Shannon entropy is obtained when the decomposition rule uses ordinary probabilistic averaging.

1.2 Link to thermodynamic entropy

1.2.1 Definition of thermodynamic entropy

If a macrostate is compatible with R microscopic configurations, Boltzmann entropy is

$$S = k \log R.$$

For a large ensemble of replicas with empirical frequencies p_i, the number of compatible arrangements is multinomial. Stirling's formula gives

$$\log \frac{n!}{\prod_i (np_i)!} = -n \sum_i p_i \log p_i + o(n).$$

After dividing by the number of replicas, the thermodynamic expression becomes Shannon entropy, up to Boltzmann's constant.

Relative entropy corresponds naturally to free energy. If a Gibbs density is

$$g_\beta(x) = \frac{e^{-\beta E(x)}}{Z_\beta}, \qquad Z_\beta = \int e^{-\beta E(x)} \, \mathrm{d}x,$$

then

$$\mathrm{D}(f \| g_\beta) = -H(f) + \beta \mathbb{E}_f E(X) + \log Z_\beta. \tag{1.5}$$

At fixed mean energy, minimizing relative entropy to equilibrium is the same as maximizing entropy, or minimizing Helmholtz free energy after physical constants are restored.

1.2.2 Maximum entropy and the Second Law

Theorem 1.11 (Gibbs state under an energy constraint; Book Lemma 1.16). Among probability vectors satisfying

$$\sum_i p_i = 1, \qquad \sum_i p_i E_i = \overline{E},$$

the entropy is maximized by

$$p_i = \frac{e^{-\beta E_i}}{Z_\beta}, \qquad Z_\beta = \sum_j e^{-\beta E_j},$$

where β is chosen to match $\overline{E}$.

Under an energy constraint, the Gibbs distribution maximizes entropy. Under a variance constraint, the Gaussian does. This gives the connection between the Second Law and the entropic central limit theorem discussed in the book.

1.3 Fisher information

1.3.1 Definition and properties

Definition 1.12 (Parameter Fisher information; Book Definition 1.11). For a differentiable parametric density $p(x;\theta)$, $\theta \in \mathbb{R}^d$, the Fisher information matrix is

$$\mathcal{I}(\theta)_{ij} = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i}\log p(X;\theta)\frac{\partial}{\partial\theta_j}\log p(X;\theta)\right].$$

Most of the book uses the location score, differentiated in the observation variable.

Definition 1.13 (Score and Fisher information; Book Definition 1.12). If X has differentiable density f, define

$$\rho_X(x) = \frac{f'(x)}{f(x)} = \frac{\mathrm{d}}{\mathrm{d}x}\log f(x), \qquad J(X) = \mathbb{E}[\rho_X(X)^2].$$

If the derivative or the integral does not exist, set $J(X) = +\infty$.

For $X \sim \mathcal{N}(\mu,\sigma^2)$,

$$\rho_X(x) = -\frac{x-\mu}{\sigma^2}, \qquad J(X) = \frac{1}{\sigma^2}.$$

The score is affine if and only if the density is Gaussian. Under an affine change,

$$J(aX+b) = \frac{1}{a^2}J(X). \tag{1.6}$$

The score is also an operator defined by integration by parts.

Lemma 1.14 (Stein identity; Book Lemma 1.18). For every smooth test function g with suitable boundary decay,

$$\mathbb{E}[\rho_X(X)g(X)] = -\mathbb{E}[g'(X)]. \tag{1.7}$$

In particular, $\mathbb{E}\rho_X(X) = 0$.

Theorem 1.15 (Cramér–Rao inequality; Book Lemma 1.19). If X has mean μ and variance σ^2, then

$$J(X) \geq \frac{1}{\sigma^2}.$$

Equality holds exactly for $\mathcal{N}(\mu,\sigma^2)$.

Proof from the score. Equation (1.7) with $g(x) = x-\mu$ gives $\mathbb{E}[(X-\mu)\rho_X(X)] = -1$. Cauchy–Schwarz yields $1 \leq \sigma^2 J(X)$. Equality forces the score to be proportional to $X-\mu$, hence affine; integrating the score gives a Gaussian density. ⊛

Definition 1.16 (Relative and standardized Fisher information; Book Definitions 1.13–1.14). For densities f, g,

$$J(f\|g) = \int f(x)\left(\frac{f'(x)}{f(x)} - \frac{g'(x)}{g(x)}\right)^2 \mathrm{d}x.$$

If X has mean μ and variance σ^2, then

$$J_{\mathrm{st}}(X) = \sigma^2 J(X) - 1 \tag{1.8}$$

$$= \sigma^2 \mathbb{E}\left[\rho_X(X) + \frac{X-\mu}{\sigma^2}\right]^2 \geq 0. \tag{1.9}$$

The standardized quantity is scale invariant and vanishes exactly at the matching Gaussian. It is the local counterpart of the entropy deficit in Equation (1.3).

For a centered density f of variance one, de Bruijn's identity relates relative entropy to standardized Fisher information:

$$\mathrm{D}(f\|\phi) = \frac{1}{2}\int_0^\infty \left[J(X+Z_t) - \frac{1}{1+t}\right] \mathrm{d}t = \frac{1}{2}\int_0^\infty \frac{J_{\mathrm{st}}(X+Z_t)}{1+t}\,\mathrm{d}t. \tag{1.10}$$

Here $Z_t \sim \mathcal{N}(0,t)$ is independent of X. Appendix C proves the identity from the heat equation.

1.3.2 Behaviour on convolution

Conditional expectation gives the score of a sum and hence bounds its Fisher information.

Lemma 1.17 (Score of a sum; Book Lemma 1.20). If U, V are independent and $W = U + V$, then

$$\rho_W(W) = \mathbb{E}[\rho_U(U) \mid W] = \mathbb{E}[\rho_V(V) \mid W]. \tag{1.11}$$

Differentiate the convolution. If $r(w) = \int f(u)g(w-u)\,\mathrm{d}u$, then differentiating either g or, after integration by parts, f gives

$$\frac{r'(w)}{r(w)} = \int \frac{f'(u)}{f(u)}\frac{f(u)g(w-u)}{r(w)}\,\mathrm{d}u.$$

The final factor is the conditional density of U given $U+V=w$. ⊛

Taking a convex combination of the two conditional representations and applying conditional Jensen gives the Fisher convolution inequalities.

Theorem 1.18 (Fisher information inequalities; Book Lemmas 1.21–1.22). For independent U, V and $0 \le \beta \le 1$,

$$J(U+V) \le \beta^2 J(U) + (1-\beta)^2 J(V),$$

and

$$J\Big(\sqrt{\beta}\,U + \sqrt{1-\beta}\,V\Big) \le \beta J(U) + (1-\beta) J(V).$$

Optimizing the first inequality gives Stam's reciprocal form

$$\frac{1}{J(U+V)} \ge \frac{1}{J(U)} + \frac{1}{J(V)}. \tag{1.12}$$

Equality in the optimized inequality requires Gaussian variables.

For IID centered variables with variance σ^2, let

$$U_n = \frac{X_1 + \cdots + X_n}{\sqrt{n\sigma^2}}.$$

The convolution inequality implies

$$(n+m) J(U_{n+m}) \le n J(U_n) + m J(U_m),$$

and the same subadditivity holds for J_{st}. In particular, $J_{\mathrm{st}}(U_{2n}) \le J_{\mathrm{st}}(U_n)$. If Fisher information is finite at some convolution level, the arguments in Chapter 2 show that $J_{\mathrm{st}}(U_n)$ tends to zero.

1.4 Previous information-theoretic proofs

1.4.1 Rényi's method

Rényi's method uses monotonicity and compactness to study homogeneous convolution and Markov evolution.

Rényi's three steps; Book Definition 1.15. Let Y_n be an iterated convolution or Markov evolution and let F be a nonnegative functional.

1. Prove $F(Y_{n+1}) \le F(Y_n)$, with a rigid equality case.
2. Use compactness to extract a convergent subsequence $Y_{n_k} \to Y$.
3. Use continuity to compare $F(Y)$ and $F(Y * X)$. Equality identifies Y as equilibrium, and subadditivity or uniqueness fills the gaps between subsequences.

Theorem 1.19 (Finite-state Markov convergence; Book Theorem 1.1). Let P be a transition matrix on a finite state space with strictly positive entries and invariant law π. If $P_i^{(n)}$ is the ith row of P^n, then

$$\mathrm{D}(P_i^{(n)} \| \pi) \longrightarrow 0 \qquad \text{for every } i.$$

The monotonicity step applies Jensen's inequality to the time-reversed transition kernel. Compactness is automatic on the finite simplex. Equality forces every subsequential limit row to equal π. Spectral methods also give an exponential rate.

The argument uses identical distributions. Chapter 3 treats independent summands with different laws.

1.4.2 Convergence of Fisher information

Theorem 1.20 (Classical central limit theorem; Book Theorem 1.2). If $X_1, X_2, \dots$ are IID with finite, positive variance, then their sum, centered and divided by its standard deviation, converges weakly to the standard normal distribution.

Linnik initiated the information approach to the central limit theorem. Shimizu and Brown studied convergence of Fisher information, and Barron later proved convergence in relative entropy under the optimal finiteness condition.

Theorem 1.21 (Shimizu's smoothed Fisher convergence; Book Theorem 1.3). Let X_i be IID, centered, with variance σ^2, and put

$$U_n = \frac{X_1 + \cdots + X_n}{\sqrt{n\sigma^2}}, \qquad Y_n = U_n + Z_t,$$

where $t > 0$ and $Z_t \sim \mathcal{N}(0, t)$ is independent. Then

$$J_{\mathrm{st}}(Y_n) \longrightarrow 0,$$

so Y_n converges weakly to $\mathcal{N}(0, 1 + t)$.

Shimizu's compactness argument. Fisher information decreases under normalized IID convolution. Gaussian smoothing gives smooth densities and enough compactness to select a weakly convergent subsequence. Continuity of the smoothed Fisher information and the equality case of the convolution inequality force the subsequential limit to be Gaussian. Subadditivity then identifies the limit along the full sequence.

Chapter 2 develops Brown's projection argument. It bounds the projection deficit, identifies the limiting score as linear, and uses de Bruijn's identity to obtain convergence in relative entropy.

2 Convergence in Relative Entropy

Let $X_1, X_2, \dots$ be IID with mean zero and variance $\sigma^2 > 0$, and write

$$U_n = \frac{X_1 + \cdots + X_n}{\sqrt{n\sigma^2}}.$$

When no confusion is possible, $\mathrm{D}(U_n)$ denotes the divergence from the standard normal density ϕ.

2.1 Motivation

2.1.1 The Fisher information deficit

Relative entropy contains a logarithm of the convolution density and is difficult to control directly. Fisher information has a quadratic structure and admits an exact conditional expectation identity.

Lemma 2.1 (Exact Fisher deficit; Book Lemma 2.1). Let U, V be IID with score functions ρ_U, ρ_V, and let ρ_* be the score of $(U+V)/\sqrt{2}$. Then

$$J(U) - J\left(\frac{U+V}{\sqrt{2}}\right) = \mathbb{E}\left[\rho_*\left(\frac{U+V}{\sqrt{2}}\right) - \frac{\rho_U(U) + \rho_V(V)}{\sqrt{2}}\right]^2. \tag{2.1}$$

Expansion. The score of the sum is the conditional expectation of either component score. Expand the square in Equation (2.1); the cross term is evaluated by conditioning on the sum, while independence and $\mathbb{E}\rho_U(U) = 0$ remove the mixed component term. What remains is exactly the difference of the two Fisher informations. ⊛

Apply the identity to two independent copies of U_n. The nonnegative difference

$$J_{\mathrm{st}}(U_n) - J_{\mathrm{st}}(U_{2n})$$

is the squared distance between a score depending only on the sum and an additive function of the two blocks. If this difference tends to zero, then a function of $x + y$ is becoming close to a sum of a function of x and a function of y. For Gaussian variables, exact equality forces affine functions. A quantitative projection inequality, combined with a Poincaré inequality, bounds the distance of the score from an affine function by this error. Since Gaussian scores are affine, the resulting bound controls standardized Fisher information.

2.1.2 Projections and adjoints

Let U, V be independent with densities p, q, let $W = U + V$ have density r, and use the natural weighted L^2 inner products. Define

$$(Mh)(w) = \mathbb{E}[h(U) \mid U + V = w] = \int h(x) \frac{p(x)q(w-x)}{r(w)}\,\mathrm{d}x, \tag{2.2}$$

$$(Lk)(x) = \mathbb{E}[k(x+V)] = \int k(x+y)q(y)\,\mathrm{d}y. \tag{2.3}$$

Lemma 2.2 (Adjoint relation; Book Lemma 2.3). The operator $L : L^2(r) \to L^2(p)$ is the adjoint of $M : L^2(p) \to L^2(r)$:

$$\langle g, Mh\rangle_{L^2(r)} = \langle Lg, h\rangle_{L^2(p)}\,.$$

Moreover, differentiation commutes with L, and $M\rho_U = \rho_W$.

The two relevant closed subspaces are

$$\{f(U+V)\} \quad \text{and} \quad \{g(U)+h(V)\}.$$

Constants and linear functions belong to both subspaces. The self-adjoint operator LM describes successive projection onto these two subspaces.

Lemma 2.3 (Eigenvalue bound from a projection inequality; Book Lemma 2.4). Suppose U, V are IID and centered, with finite positive variance, and write $W = U + V$. Assume that for some $K \geq 0$ and every f satisfying $\mathbb{E}f(W) = \mathbb{E}[Wf(W)] = 0$,

$$\mathbb{E}\big[f(U+V) - Lf(U) - Lf(V)\big]^2 \geq K\,\mathbb{E}[Lf(U)^2].$$

Then every nonzero g satisfying $\mathbb{E}g(U) = \mathbb{E}[Ug(U)] = 0$ satisfies

$$\frac{\mathbb{E}[Mg(U+V)^2]}{\mathbb{E}[g(U)^2]} \leq \frac{1}{2+K}.$$

The orthogonal complements of the affine functions are preserved by L and M. Since $L = M^*$, the operators LM and ML have the same norm on these subspaces. The assumed residual inequality is equivalent to $\mathbb{E}f(W)^2 \geq (2+K)\mathbb{E}[Lf(U)^2]$, which gives the displayed bound. Thus a lower bound on the projection residual bounds the projection eigenvalues after the affine modes have been removed.

2.1.3 The normal case

For Gaussian variables, the projection operators can be diagonalized explicitly. Let $H_k(x;\sigma^2)$ be the Hermite polynomials defined by

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} H_k(x;\sigma^2) = \exp\left(\frac{tx}{\sigma^2} - \frac{t^2}{2\sigma^2}\right). \tag{2.4}$$

They form an orthogonal basis of $L^2(\mathcal{N}(0,\sigma^2))$.

After the normalization that maps the variance of the sum back to one, both Gaussian projection maps act on the kth Hermite mode by the factor $2^{-k/2}$. Hence LM has eigenvalue 2^{-k}. The constant and linear modes correspond to $k = 0, 1$; after the affine part is removed, the largest remaining eigenvalue is $1/4$, attained by the quadratic mode.

Lemma 2.4 (Brown's Gaussian projection inequality; Book Lemma 2.5). If U, V are independent IID Gaussian variables, then for any admissible f, g there exist constants a, b such that

$$\mathbb{E}\big[g(U) - aU - b\big]^2 \leq \mathbb{E}\big[f(U+V) - g(U) - g(V)\big]^2.$$

Equivalently, the nonlinear component has a strict spectral gap from the affine modes.

Spectral proof. Expand g in Hermite polynomials. Subtract the H_0 and H_1 components. On the remaining span, the largest eigenvalue of LM is $1/4$. Pythagoras converts this eigenvalue bound into the displayed residual inequality.

2.1.4 Results of Brown and Barron

Brown adds independent Gaussian noise to obtain a Gaussian lower bound for the density. If $Y = X + Z_t$ with $\mathbb{E}X = 0$, $\operatorname{Var} X = 1$, and $t > 0$, a Hermite expansion of the smoothed score bounds its non-affine coefficients by the decrease in Fisher information under doubling.

Barron turns Brown's smoothed Fisher convergence into the strong entropic CLT.

Theorem 2.5 (Barron's entropic CLT; Book Theorem 2.1). Let $X_1, X_2, \ldots$ be IID with densities and finite, positive variance. Let g_n be the density of U_n, the centered sum divided by its standard deviation. Then

$$\mathrm{D}(g_n \| \phi) \longrightarrow 0$$

if and only if $\mathrm{D}(g_m \| \phi) < \infty$ for at least one m.

Proof of Barron's theorem. Brown's sandwich inequality and a uniform integrability argument show that, for every $t > 0$,

$$J(U_n + Z_t) \longrightarrow \frac{1}{1+t}.$$

The integral de Bruijn identity then transfers this pointwise convergence of smoothed Fisher information to relative entropy. The integral is controlled using the finite entropy assumption. Necessity follows because a sequence converging to zero is finite from some index onward.

Projection inequalities for general independent variables give quantitative versions of this argument.

2.2 Generalised bounds on projection eigenvalues

2.2.1 Projection of functions in L^2

Let Y_1, Y_2 be independent, let $f(Y_1 + Y_2)$ be centered, and define its coordinate projections

$$g_1(u) = \mathbb{E}[f(u + Y_2)], \qquad g_2(v) = \mathbb{E}[f(Y_1 + v)].$$

Proposition 2.6 (General projection inequality; Book Proposition 2.2). For every $\beta \in [0, 1]$, there is a constant μ such that

$$\mathbb{E}\big[f(Y_1 + Y_2) - g_1(Y_1) - g_2(Y_2)\big]^2 \tag{2.5}$$

$$\geq \frac{1}{\overline{J}} \left\{ \beta \mathbb{E}[g_1'(Y_1) - \mu]^2 + (1-\beta)\mathbb{E}[g_2'(Y_2) - \mu]^2 \right\}, \tag{2.6}$$

where

$$\overline{J} = (1-\beta)J(Y_1) + \beta J(Y_2).$$

Core calculation. Multiply the projection residual by the score of the other variable and condition. Cauchy–Schwarz bounds the resulting correlation by the residual norm times the relevant Fisher information. The Stein identity converts the correlation into $g_i' - \mu$, where $\mu = \mathbb{E}f'(Y_1 + Y_2)$. Weight the two bounds by β and $1 - \beta$. ⊛

The proposition says that if a function of a sum is nearly additive, then each projected component has derivative close to a common constant. A Poincaré inequality is now needed to conclude that the component itself is close to a common linear function.

2.2.2 Restricted Poincaré constants

Definition 2.7 (Restricted Poincaré constant; Book Definition 2.2). For a random variable Y, define

$$R_Y^* = \sup_g \frac{\mathbb{E}[g(Y)^2]}{\mathbb{E}[g'(Y)^2]},$$

where the supremum is over absolutely continuous g satisfying

$$\mathbb{E}g(Y) = 0, \qquad \mathbb{E}g'(Y) = 0, \qquad 0 < \mathbb{E}g(Y)^2 < \infty.$$

The condition $\mathbb{E}g'(Y) = 0$ excludes nonconstant linear functions.

For a centered variable of variance one,

$$\frac{\mathbb{E}Y^4 - 1}{4} \le R_Y^* \le R_Y,$$

where R_Y is the ordinary Poincaré constant. If $Z \sim \mathcal{N}(0, \sigma^2)$, then

$$R_Z^* = \frac{\sigma^2}{2},$$

attained by $g(x) = x^2 - \sigma^2$.

Proposition 2.8 (Contraction under a restricted Poincaré inequality; Book Proposition 2.3). Let Y_1, Y_2 be IID and centered, with finite Fisher information J and finite restricted Poincaré constant R^*. For every centered $f(Y_1 + Y_2)$ there is μ such that

$$\mathbb{E}\big[f(Y_1 + Y_2) - Lf(Y_1) - Lf(Y_2)\big]^2 \ge \frac{1}{JR^*}\mathbb{E}[Lf(Y_1) - \mu Y_1]^2.$$

Consequently,

$$J_{\mathrm{st}}\left(\frac{Y_1 + Y_2}{\sqrt{2}}\right) \le J_{\mathrm{st}}(Y_1)\frac{2JR^*}{1 + 2JR^*}. \tag{2.7}$$

The factor is scale invariant because $J(aY)R_{aY}^* = J(Y)R_Y^*$. For a Gaussian, $JR^* = 1/2$, and the result recovers Brown's eigenvalue bound.

2.2.3 Convergence of restricted Poincaré constants

The contraction for two variables alone initially gives a rate along powers of two whose constant depends poorly on the smoothing level. Johnson controls this by showing that restricted Poincaré constants themselves converge to the Gaussian value.

Lemma 2.9 (Restricted constant under doubling; Book Lemma 2.9). Let Y_1, Y_2 be IID and centered, with variance one, finite ordinary Poincaré constant R, and restricted constant R^*. If $Z = (Y_1 + Y_2)/\sqrt{2}$, then

$$R^*_Z - \frac{1}{2} \leq \frac{R^* - 1/2}{2} + \sqrt{2RJ_{\mathrm{st}}(Y_1)}. \tag{2.8}$$

The first term contracts the excess restricted constant; the second is an error that disappears as the Fisher deficit decreases. Summing this recursion along dyadic convolution powers makes the excess constants summable.

Theorem 2.10 ($O(1/n)$ Fisher rate, implicit constant; Book Theorem 2.2). Let $X_1, X_2, \ldots$ be IID with variance σ^2, finite Fisher information $I = J(X_1)$, and finite ordinary Poincaré constant R. Then there is a finite constant $C = C(I, R)$ such that

$$J_{\mathrm{st}}(U_n) \leq \frac{C}{n}, \qquad n \geq 1. \tag{2.9}$$

Convergence along powers of two. Equation (2.7) first gives convergence along $n = 2^k$. Lemma 2.9 makes the product of contraction factors comparable with 2^{-k}. The bound for every n also follows from the several-summand inequality in Theorem 2.11, since $R^* \leq R$.

2.3 Rates of convergence

2.3.1 Explicit $O(1/n)$ bounds

A projection argument for several summands gives an explicit constant and sufficient control under Gaussian smoothing to apply de Bruijn's identity.

Theorem 2.11 (Explicit Fisher rate; Book Theorem 2.3). Let $X_1, X_2, \ldots$ be IID with variance σ^2 and finite restricted Poincaré constant R^*. Then

$$J_{\mathrm{st}}(U_n) \leq \frac{2R^*}{2R^* + (n-1)\sigma^2}\, J_{\mathrm{st}}(X_1), \qquad n \geq 1. \tag{2.10}$$

Theorem 2.12 (Explicit rate of convergence in relative entropy; Book Theorem 2.4). Under the same IID and variance assumptions, suppose the ordinary Poincaré constant R is finite. Then

$$\mathrm{D}(U_n \| \phi) \leq \frac{2R}{2R + (n-1)\sigma^2}\, \mathrm{D}(X_1/\sigma \| \phi) \leq \frac{2R}{n\sigma^2}\, \mathrm{D}(X_1/\sigma \| \phi). \tag{2.11}$$

From Fisher to entropy. Apply Theorem 2.11 to each smoothed summand $X_1 + Z_t$. Gaussian convolution does not increase the ordinary Poincaré constant by more than the added variance, so the coefficient can be bounded uniformly in t by the expression involving R. Insert the resulting Fisher bound into de Bruijn's identity and integrate. ⊛

Qualitative convergence also holds without a finite Poincaré constant.

Theorem 2.13 (Fisher convergence from one finite level; Book Theorem 2.5). Let X_i be IID with finite variance. If $J_{\mathrm{st}}(U_m) < \infty$ for at least one m, then

$$J_{\mathrm{st}}(U_n) \longrightarrow 0.$$

Truncation. Split the real line into a bounded interval and its complement. On each fixed bounded interval, the convolution density is eventually bounded away from zero, giving a finite Poincaré constant for its restriction. On the tails, uniform integrability controls the Fisher contribution. Let the interval $[-B_n, B_n]$ expand slowly enough to balance the two errors.

2.3.2 Comparison with other forms of convergence

A $\Gamma(n, \theta)$ variable is a sum of n independent exponential variables. After centering and rescaling, the book obtains

$$J_{\mathrm{st}}(\Gamma(n,\theta)) = \frac{2}{n-2} \quad (n > 2), \qquad \mathrm{D}(\Gamma(n,\theta)) \sim \frac{1}{3n}.$$

Both quantities are of order $1/n$.

The classical Berry–Esseen theorem gives a bound of order $n^{-1/2}$ for distribution functions.

Theorem 2.14 (Berry–Esseen bound; Book Section 2.3.2). Let X_i be IID with mean zero, variance $\sigma^2 > 0$, and finite absolute third moment. If F_n is the distribution function of U_n and Φ is the standard normal distribution function, then for an absolute constant A,

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \le A \frac{\mathbb{E}|X_1|^3}{\sigma^3 \sqrt{n}}. \tag{2.12}$$

For a centered variable X with variance one and density f, analytic inequalities bound distances between densities by the square root of the Fisher deficit:

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

and

$$\int |f - \phi| \le \sqrt{2 J_{\mathrm{st}}(X)}.$$

Thus an $O(1/n)$ Fisher bound gives an $O(n^{-1/2})$ bound on the distance between densities.

2.3.3 Extending the Cramér–Rao lower bound

Lower bounds follow by choosing test functions in the Cramér–Rao argument and completing the square.

Lemma 2.15 (Generalized Cramér–Rao; Book Lemma 2.10). Let U, V have densities f, g, and let $\rho_g = g'/g$. For every admissible test function k,

$$J(f\|g) \geq \frac{\big(\mathbb{E}k'(U) + \mathbb{E}[\rho_g(U)k(U)]\big)^2}{\mathbb{E}[k(U)^2]}. \tag{2.13}$$

If, in addition, U is centered with variance σ^2 and $V \sim \mathcal{N}(0, \sigma^2)$, this becomes

$$J_{\mathrm{st}}(U) \geq \sigma^2 \frac{\big(\mathbb{E}k'(U) - \mathbb{E}[Uk(U)]/\sigma^2\big)^2}{\mathbb{E}[k(U)^2]}.$$

Polynomial choices make the numerator depend on moments that can be followed under convolution.

Theorem 2.16 (Lower bound in terms of skewness; Book Lemma 2.11). Assume $\mathbb{E}X_1^4 < \infty$ and let

$$s = \frac{\mathbb{E}[(X_1 - \mathbb{E}X_1)^3]}{\sigma^3}$$

be the skewness. Then

$$\liminf_{n\to\infty} nJ_{\mathrm{st}}(U_n) \geq \frac{s^2}{2}. \tag{2.14}$$

More precisely, if γ is excess kurtosis, the proof gives

$$J_{\mathrm{st}}(U_n) \geq \frac{s^2}{n(2+\gamma/n)} = \frac{s^2}{2n+\gamma}.$$

For the variance-one sum U_n, choose $k(u) = u^2 - 1$. The resulting numerator involves the third moment, which scales as $n^{-1/2}$. Squaring gives a lower bound of order $1/n$.

Theorem 2.17 (First unmatched Hermite mode; Book Lemma 2.12). Suppose X_1 has a finite $2r$th moment and its moments through order r agree with those of the matching Gaussian Z. Then

$$\liminf_{n\to\infty} n^{r-1} J_{\mathrm{st}}(U_n) \geq \frac{\big(m_{r+1}(X_1) - m_{r+1}(Z)\big)^2}{r!\,\sigma^{2r+2}}, \tag{2.15}$$

where m_j denotes the centered jth moment.

The proof chooses the rth Hermite polynomial as k. Hermite orthogonality cancels all matched lower moments and isolates the first unmatched one. The first moment that differs from the Gaussian moment determines this lower bound.

3 Non-Identical Variables and Random Vectors

3.1 Non-identical random variables

3.1.1 Lindeberg's condition

Let $X_1, X_2, \ldots$ be independent, centered, real-valued random variables with

$$\mathrm{Var}(X_i) = \sigma_i^2, \qquad v_n = \sum_{i=1}^{n} \sigma_i^2, \qquad U_n = \frac{X_1 + \cdots + X_n}{\sqrt{v_n}}.$$

The variance weights

$$\alpha_i = \frac{\sigma_i^2}{v_n}$$

replace the equal weights $1/n$ of the IID problem.

Definition 3.1 (Lindeberg condition; Book Condition 1). For $\varepsilon > 0$, define

$$\Lambda_\varepsilon(n) = \frac{1}{v_n} \sum_{i=1}^{n} \mathbb{E}\left[X_i^2 \mathbf{1}_{\{|X_i| \geq \varepsilon \sqrt{v_n}\}}\right]. \tag{3.1}$$

The Lindeberg condition is $\Lambda_\varepsilon(n) \to 0$ for every $\varepsilon > 0$.

For a fixed sequence of independent variables with positive variances, this condition implies $v_n \to \infty$ and bounds the aggregate contribution of large summands.

Definition 3.2 (Individual smallness; Book Condition 2). The array is individually small if

$$\max_{1 \leq i \leq n} \frac{\sigma_i^2}{v_n} \longrightarrow 0. \tag{3.2}$$

Lindeberg implies individual smallness. The converse need not hold without a distributional condition on the tails.

Theorem 3.3 (Lindeberg–Feller; Book Theorem 3.1). For independent centered variables, the normalized sum U_n converges weakly to $\mathcal{N}(0,1)$ and individual smallness holds if and only if the Lindeberg condition holds.

Convergence in relative entropy requires additional assumptions. Barron's counterexample in Appendix E has finite variance and weak convergence, but every normalized sum has infinite divergence from the Gaussian.

Linnik proved weak convergence by truncating the variables and adding small Gaussian perturbations. He also introduced a tail condition for individual summands.

Definition 3.4 (Individual Lindeberg bound; Book Condition 3). Given $\varepsilon > 0$ and $\zeta_1 > 0$, the variable X_j, $j \le n$, obeys the individual Lindeberg bound if

$$\mathbb{E}\left[X_j^2 \mathbf{1}_{\{|X_j| \ge \varepsilon \sqrt{v_n}\}}\right] \le \zeta_1 \sigma_j^2. \tag{3.3}$$

Linnik's entropy increment estimate relates the decrease in relative entropy after adding a summand to the difference between Fisher information and its Cramér–Rao lower bound. Projection inequalities give a quantitative estimate for several independent summands.

3.1.2 Weighted projection inequalities

Let $S_n = X_1 + \cdots + X_n$, and let $f(S_n)$ be centered. Define the coordinate projections

$$f_i(x) = \mathbb{E}[f(X_1 + \cdots + X_{i-1} + x + X_{i+1} + \cdots + X_n)]. \tag{3.4}$$

If $J_i = J(X_i)$, the amount of information in all variables other than X_i determines the weight assigned to the ith projected derivative.

Proposition 3.5 (Projection inequality for independent summands; Book Proposition 3.1). There exists a constant μ such that

$$\mathbb{E}\left[f(S_n) - \sum_{i=1}^{n} f_i(X_i)\right]^2 \ge \frac{1}{2} \sum_{i=1}^{n} \left(\sum_{j \ne i} \frac{1}{J_j}\right) \mathbb{E}[f_i'(X_i) - \mu]^2. \tag{3.5}$$

Successive projections. Set

$$g_i(x) = \mathbb{E}[f(x + X_{i+1} + \cdots + X_n)].$$

Project $g_i(X_1 + \cdots + X_i)$ onto the sum of a function of $X_1 + \cdots + X_{i-1}$ and the coordinate function $f_i(X_i)$. The squared increments of these successive projections telescope. The Stein identity and Cauchy–Schwarz inequality bound the ith increment below by

$$\left(\sum_{j<i} \frac{1}{J_j}\right) \mathbb{E}[f_i'(X_i) - \mu]^2.$$

Repeat the argument in the reverse order and average the two bounds. Every complementary information sum $\sum_{j \ne i} 1/J_j$ then appears with factor $1/2$.

The reciprocal Fisher weights are optimal in the Cauchy–Schwarz step: for fixed $\sum_j u_j$, the quadratic form $\sum_j u_j^2 J_j$ is minimized by $u_j \propto 1/J_j$.

Assume now that each variable has a finite restricted Poincaré constant R_i^*. Applying the restricted inequality to $f_i - \mu x$ converts derivative control into distance from a linear function.

Lemma 3.6 (Projection inequality with restricted Poincaré constants; Book Lemma 3.2). Under the preceding assumptions, there is a constant μ such that

$$\mathbb{E}\left[f(S_n) - \sum_{i=1}^{n} f_i(X_i)\right]^2 \geq \sum_{i=1}^{n} \frac{\sum_{j\neq i} 1/J_j}{2R_i^*} \mathbb{E}[f_i(X_i) - \mu X_i]^2. \tag{3.6}$$

In the IID case, the coefficient becomes $n(n-1)/(2R^*J)$ after summing identical terms, recovering the strengthening behind Chapter 2.

The score of S_n can now be inserted for f. Decompose each coordinate projection of the sum score into a Gaussian linear part and a non-Gaussian residual. The quadratic inequality

$$ax^2 + by^2 \geq \frac{ab}{a+b}(x-y)^2$$

combines the projection residual with the coordinate residual.

Proposition 3.7 (Non-identical Fisher contraction; Book Proposition 3.2). Let X_i be independent and centered, with variance σ_i^2, Fisher information J_i, and finite restricted Poincaré constant R_i^*. Define

$$c_i = \frac{\sum_{j\neq i} 1/J_j}{2R_i^*}, \qquad \alpha_i = \frac{\sigma_i^2}{v_n}.$$

Then

$$J_{\mathrm{st}}(U_n) \leq \sum_{i=1}^{n} \frac{\alpha_i}{1+c_i} J_{\mathrm{st}}(X_i). \tag{3.7}$$

Here $J_{\mathrm{st}}(X_i) = \sigma_i^2 J_i - 1$.

For IID variables, $\alpha_i = 1/n$ and $c_i = (n-1)/(2R^*J)$, so Equation (3.7) reduces to the Chapter 2 contraction.

Definition 3.8 (Average boundedness; Book Condition 4). There exist constants C, D with $C > 0$ and $D < \infty$ such that

$$\inf_n \frac{\sum_{i=1}^n 1/J_i}{v_n} \geq C, \tag{3.8}$$

$$\sup_n \frac{\sum_{i=1}^n \sigma_i^2 J_{\mathrm{st}}(X_i) R_i^*}{v_n} \leq D. \tag{3.9}$$

The first inequality bounds the sum of reciprocal Fisher informations below by a fixed fraction of the total variance. The second bounds the average weighted by the variances of $J_{\mathrm{st}}(X_i)R_i^*$.

Theorem 3.9 (Non-identical CLT in Fisher information; Book Theorem 3.2). Assume the Lindeberg condition and Condition 4. Then

$$J_{\mathrm{st}}\left(\frac{X_1+\cdots+X_n}{\sqrt{v_n}}\right)\longrightarrow 0. \tag{3.10}$$

Put $m_n=\max_{1\le i\le n}\sigma_i^2$. With the constants in Condition 4, whenever $Cv_n>m_n$ the proof gives

$$J_{\mathrm{st}}(U_n)\le\frac{2D}{Cv_n-m_n}. \tag{3.11}$$

Since $m_n/v_n\to 0$, this is at most $4D/(Cv_n)$ for all sufficiently large n.

Substitution into the weighted contraction. Cramér–Rao gives $1/J_i\le\sigma_i^2$. Condition 4 therefore implies

$$\sum_{j\neq i}\frac{1}{J_j}\ge Cv_n-m_n,\qquad 1+c_i\ge\frac{Cv_n-m_n}{2R_i^*}.$$

Insert this into Equation (3.7) and apply Equation (3.9). The Lindeberg condition implies $m_n/v_n\to 0$ and $v_n\to\infty$, so the bound vanishes. ⊛

The moment implications of a restricted Poincaré bound can also help verify Lindeberg. For a centered Y, applying the definition to $g(y)=y^2-\mathrm{Var}(Y)$ gives the estimate

$$\mathbb{E}Y^4\le\mathrm{Var}(Y)^2+4\,\mathrm{Var}(Y)R_Y^*.$$

Thus Chebyshev reduces the Lindeberg tails to

$$\frac{\sum_{i=1}^n(\sigma_i^4+4\sigma_i^2R_i^*)}{\varepsilon^2v_n^2}.$$

In particular, Lindeberg follows when this expression tends to zero for every $\varepsilon>0$; individual smallness controls the term involving squared variances, while the book's average bound on $\sum_i\sigma_i^2R_i^*$ controls the other term.

3.2 Random vectors

3.2.1 Definitions

Throughout the vector subsection, random vectors are centered. Let Y be a d-dimensional random vector with density f.

Definition 3.10 (Vector entropy and relative entropy; Book Definition 3.1).

$$H(Y)=-\int_{\mathbb{R}^d}f(y)\log f(y)\,\mathrm{d}y,\qquad \mathrm{D}(f\|g)=\int_{\mathbb{R}^d}f(x)\log\frac{f(x)}{g(x)}\,\mathrm{d}x.$$

Definition 3.11 (Score vector and Fisher information matrix; Book Definition 3.2). Let Y have differentiable density f and positive definite covariance matrix C. Define

$$\rho_Y(x) = \nabla \log f(x) = \frac{\nabla f(x)}{f(x)},$$

and

$$J(Y) = \mathbb{E}[\rho_Y(Y)\rho_Y(Y)^T], \tag{3.12}$$
$$J(Y\|Z) = J(Y) - C^{-1}, \tag{3.13}$$
$$J_{\mathrm{st}}(Y) = C(J(Y) - C^{-1}), \tag{3.14}$$

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

For the Gaussian,

$$\rho_Z(x) = -C^{-1}x, \qquad J(Z) = C^{-1}.$$

The vector Stein identity is the componentwise formula obtained by integration by parts.

Lemma 3.12 (Vector Stein identity; Book Lemma 3.3). For every smooth scalar test function h with suitable decay,

$$\mathbb{E}[h(Y)\rho_Y(Y)] = -\mathbb{E}[\nabla h(Y)]. \tag{3.15}$$

In particular,

$$\mathbb{E}[Y\rho_Y(Y)^T] = -I_d.$$

Consequently,

$$J(Y\|Z) = \mathbb{E}\big[(\rho_Y(Y) + C^{-1}Y)(\rho_Y(Y) + C^{-1}Y)^T\big] \succeq 0.$$

Thus the matrix relative Fisher information is positive semidefinite and vanishes exactly when $\rho_Y(x) = -C^{-1}x$.

3.2.2 Behaviour on convolution

Lemma 3.13 (Score vector of a sum; Book Lemma 3.4). If U, V are independent random vectors and $W = U + V$, then

$$\rho_W(W) = \mathbb{E}[\rho_U(U) \mid W] = \mathbb{E}[\rho_V(V) \mid W]. \tag{3.16}$$

The proof differentiates the convolution density in each coordinate and divides by the density of the sum. It is exactly the scalar argument, now applied componentwise.

The entropy deficit from a Gaussian with covariance C can be written as

$$\mathrm{D}(f\|\phi_C) = H(\phi_C) - H(f) + \frac{1}{2}(\mathrm{tr}(C^{-1}B) - d),$$

when f has covariance B. If $B = C$, this reduces to the entropy difference.

Let $Y_t = X + Z_{Ct}$, where $Z_{Ct} \sim \mathcal{N}(0, tC)$ is independent and X has covariance C. Along this path, write

$$J_{\mathrm{st},C}(Y_t) := C\left(J(Y_t) - \frac{1}{1+t}C^{-1}\right).$$

This is the standardized Fisher deficit expressed in the original covariance scale; equivalently, it is $(1+t)^{-1}$ times the usual deficit standardized by the covariance $(1+t)C$ of Y_t. The matrix de Bruijn identity in Appendix C simplifies when the covariances agree to

$$\mathrm{D}(f\|\phi_C) = \frac{1}{2}\int_0^\infty \mathrm{tr}\big(J_{\mathrm{st},C}(Y_t)\big)\,\mathrm{d}t. \tag{3.17}$$

Integrating a bound on the trace of the Fisher information matrix gives a bound on relative entropy.

3.2.3 Projection inequalities

For independent vectors Y_1, Y_2 and centered $f(Y_1+Y_2)$, define

$$g_1(u) = \mathbb{E}[f(u+Y_2)], \qquad g_2(v) = \mathbb{E}[f(Y_1+v)].$$

Proposition 3.14 (Vector projection inequality; Book Proposition 3.3). For every $\beta \in [0,1]$, there is a vector μ such that

$$\mathbb{E}\big[f(Y_1+Y_2) - g_1(Y_1) - g_2(Y_2)\big]^2 \tag{3.18}$$

$$\geq \frac{1}{\overline{J}}\left\{\beta\mathbb{E}\,\|\nabla g_1(Y_1)-\mu\|^2 + (1-\beta)\mathbb{E}\,\|\nabla g_2(Y_2)-\mu\|^2\right\}, \tag{3.19}$$

where

$$\overline{J} = (1-\beta)\,\mathrm{tr}\,J(Y_1) + \beta\,\mathrm{tr}\,J(Y_2).$$

The proof defines vector-valued correlations between the projection residual and the two score vectors. Cauchy–Schwarz introduces the traces of the Fisher matrices; the vector Stein identity identifies those correlations with gradient residuals.

Definition 3.15 (Vector restricted Poincaré constant; Book Definition 3.3).

$$R_Y^* = \sup_g \frac{\mathbb{E}[g(Y)^2]}{\mathbb{E}\,\|\nabla g(Y)\|^2},$$

where g is absolutely continuous and satisfies

$$\mathbb{E}g(Y) = 0, \qquad \mathbb{E}\nabla g(Y) = 0, \qquad 0 < \mathbb{E}g(Y)^2 < \infty.$$

The condition $\mathbb{E}\nabla g(Y) = 0$ removes all affine directions. To obtain a scalar contraction from a matrix Fisher quantity, test the score through a positive definite matrix.

Theorem 3.16 (Vector Fisher contraction; Book Theorem 3.3). Let Y_1, Y_2 be IID centered random vectors with positive definite covariance C, finite Fisher information matrix, and finite restricted Poincaré constant R^*. Let $Z \sim \mathcal{N}(0, C)$ and define

$$J_0 = \operatorname{tr} J(Y_1).$$

For every positive definite matrix D,

$$\operatorname{tr}\left[DJ\left(\frac{Y_1 + Y_2}{\sqrt{2}} \middle\| Z\right)\right] \le \operatorname{tr}[DJ(Y_1 \| Z)] \frac{2J_0R^*}{1 + 2J_0R^*}. \tag{3.20}$$

Proof. Factor $D = B^T B$. Then

$$\operatorname{tr}(DJ(Y\|Z)) = \mathbb{E}\left\|B(\rho_Y(Y) + C^{-1}Y)\right\|^2$$

for Y with covariance C. Apply the vector projection inequality componentwise to B times the relative score. Its denominator is J_0. The relative score is orthogonal to affine functions, and so are its coordinate projections. The restricted Poincaré inequality applies after subtracting their mean gradients, while orthogonality ensures that subtracting a linear function cannot reduce their squared norms. The resulting projection bound has coefficient $1/(J_0R^*)$; the calculation in Chapter 2 gives the factor $2J_0R^*/(1 + 2J_0R^*)$. ⊛

Iterating Theorem 3.16 over IID vector convolution powers gives convergence of the standardized Fisher matrix to zero whenever the required Fisher and Poincaré quantities are finite. Equation (3.17) then yields convergence in relative entropy to the matching multivariate Gaussian. Combining the weighted scalar argument of Section 3.1 with the vector argument gives the corresponding argument for vectors with different distributions.

4 Dependent Random Variables

For dependent variables, the score identity involves joint scores as well as marginal scores. Gaussian smoothing gives bounds on their difference, while Rosenblatt mixing controls dependence between separated blocks. These estimates yield an approximate subadditivity inequality for Fisher information.

Standardized Fisher information converges to zero at each positive smoothing level. Under an integrability assumption, de Bruijn's identity gives convergence in relative entropy for the original sums after normalization by their variance.

4.1 Introduction and notation

Let $(X_i)_{i\in\mathbb{Z}}$ be a doubly infinite, strictly stationary sequence with $\mathbb{E}X_i = 0$. Write

$$S_n = \sum_{i=1}^{n} X_i, \qquad v_n = \mathrm{Var}(S_n), \qquad U_n = \frac{S_n}{\sqrt{n}}.$$

For $\tau > 0$, let $Z_i^{(\tau)} \sim \mathcal{N}(0,\tau)$ be IID and independent of the entire process, and define the smoothed normalized sum

$$V_n^{(\tau)} = \frac{\sum_{i=1}^{n}(X_i + Z_i^{(\tau)})}{\sqrt{n}} \stackrel{d}{=} U_n + Z^{(\tau)}.$$

Its variance is $v_n/n + \tau$. When it exists, the long-run variance is

$$v = \sum_{j=-\infty}^{\infty} \mathrm{Cov}(X_0, X_j) = \lim_{n\to\infty} \frac{v_n}{n}.$$

For a smooth density p, retain the notation

$$\rho_p(x) = \frac{p'(x)}{p(x)}, \qquad J(X) = \mathbb{E}\rho_p(X)^2, \qquad J_{\mathrm{st}}(X) = \mathrm{Var}(X)J(X) - 1.$$

The Cramér–Rao inequality gives $J_{\mathrm{st}}(X) \geq 0$, with equality precisely for a Gaussian law.

4.1.1 Mixing coefficients

Mixing coefficients measure dependence between the sigma fields generated by separated sets of variables.

Definition 4.1 (Rosenblatt coefficients for two random variables). For random variables S, T, define

$$\alpha(S,T) = \sup_{A,B} \left|\mathbb{P}(S \in A, T \in B) - \mathbb{P}(S \in A)\mathbb{P}(T \in B)\right|,$$

$$\phi(S,T) = \sup_{\substack{A,B \\ \mathbb{P}(S\in A)>0}} \left|\mathbb{P}(T \in B \mid S \in A) - \mathbb{P}(T \in B)\right|,$$

$$\psi(S,T) = \sup_{\substack{A,B \\ \mathbb{P}(S\in A)\mathbb{P}(T\in B)>0}} \frac{\left|\mathbb{P}(S \in A, T \in B) - \mathbb{P}(S \in A)\mathbb{P}(T \in B)\right|}{\mathbb{P}(S \in A)\mathbb{P}(T \in B)}.$$

The suprema run over measurable A and B.

Let $\mathcal{F}_a^b = \sigma(X_a, \dots, X_b)$, allowing infinite endpoints. Define, for example,

$$\alpha(t) = \sup\{\alpha(S,T) : S \text{ is } \mathcal{F}_{-\infty}^0\text{-measurable},\ T \text{ is } \mathcal{F}_t^\infty\text{-measurable}\},$$

and define $\phi(t)$ and $\psi(t)$ similarly. The process is α-mixing, ϕ-mixing, or ψ-mixing when the corresponding coefficient tends to zero. Since

$$\psi(S,T) \geq \phi(S,T) \geq \alpha(S,T),$$

ψ-mixing implies ϕ-mixing, which implies α-mixing.

Every m-dependent process is α-mixing. A finite moving average

$$X_k = \sum_{j=0}^{m-1} a_j V_{k-j}$$

of independent variables is therefore a basic example. For infinite moving averages, mixing also depends on the law of the innovations. Stationary aperiodic Harris chains satisfy strong mixing.

Remark 4.2 (Mixing is stronger than covariance decay). The coefficient $\alpha(t)$ bounds dependence between all pairs of events in sigma fields separated by t time units. In positively associated processes, such as FKG systems, covariance bounds can also be used to establish limit theorems.

4.1.2 The main Fisher and entropy limits

Theorem 4.3 (Book Theorem 4.1: smoothed Fisher convergence). Suppose (X_i) is stationary, $\mathbb{E}\,|X_0|^{2+\delta} < \infty$ for some $\delta > 0$, and

$$\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty.$$

Then, for every $\tau > 0$,

$$J_{\mathrm{st}}\left(V_n^{(\tau)}\right) \longrightarrow 0.$$

The same summability hypothesis implies $v_n/n \to v < \infty$.

The conclusion also covers the degenerate case $v = 0$. For instance, if $X_j = Y_j - Y_{j+1}$ with (Y_j) IID, then $S_n = Y_1 - Y_{n+1}$ and $U_n \to 0$ in probability. Gaussian smoothing still gives

$$J_{\mathrm{st}}(V_n^{(\tau)}) \le \left(\frac{v_n}{n} + \tau\right) J(Z^{(\tau)}) - 1 = \frac{v_n}{n\tau} \longrightarrow 0.$$

Theorem 4.4 (Book Theorem 4.2: convergence in relative entropy). In addition to the hypotheses of Theorem 4.3, assume

$$v = \sum_{j=-\infty}^{\infty} \mathrm{Cov}(X_0, X_j) > 0.$$

For $N \ge 1$, set

$$f_N(\tau) = \sup_{n \ge N} \frac{n J_{\mathrm{st}}(V_n^{(\tau)})}{v_n + n\tau}.$$

If $\int_0^\infty f_N(\tau)\,\mathrm{d}\tau < \infty$ for some N, and if g_n is the density of $S_n/\sqrt{v_n}$, then

$$\mathrm{D}(g_n \| \phi) \longrightarrow 0,$$

where ϕ is the standard Gaussian density.

De Bruijn's identity expresses relative entropy as an integral of the smoothed Fisher deficit. Theorem 4.3 gives pointwise convergence, and f_N is an integrable dominating function. The result follows by dominated convergence.

4.2 Fisher information and approximate convolution

For independent variables, the score of a sum is an orthogonal projection and Fisher information is subadditive. Dependence changes this into an inequality with an error term.

Definition 4.5 (Score projection deficit). Let X, Y have marginal scores ρ_X, ρ_Y, and let $\widetilde{\rho}$ be the score of $\sqrt{\beta}X + \sqrt{1-\beta}Y$, where $0 < \beta < 1$. Define

$$\begin{aligned}\Delta(X, Y, \beta) = \mathbb{E}\big[&\sqrt{\beta}\,\rho_X(X) + \sqrt{1-\beta}\,\rho_Y(Y) \\ &- \widetilde{\rho}(\sqrt{\beta}X + \sqrt{1-\beta}Y)\big]^2.\end{aligned}$$

For independent X, Y, this is exactly the drop in the weighted Fisher information. It vanishes precisely when both variables are Gaussian with equal variances. Under dependence, the identity includes additional terms involving joint scores.

Theorem 4.6 (Approximate Fisher inequality). Let $\max\{\mathrm{Var}(S), \mathrm{Var}(T)\} \le K\tau$ and

$$X = S + Z_S^{(\tau)}, \qquad Y = T + Z_T^{(\tau)},$$

where the two Gaussian noises are mutually independent and independent of (S, T). For every $0 < \eta < 1/3$ there is a constant $C = C(K, \tau, \eta)$ such that

$$\beta J(X) + (1-\beta) J(Y) - J(\sqrt{\beta} X + \sqrt{1-\beta} Y) + C\alpha(S,T)^{1/3-\eta} \ge \Delta(X, Y, \beta).$$

If S, T have a uniformly bounded kth moment, the error may be taken with any exponent strictly below $k/(k+4)$.

The error is bounded by a power of $\alpha(S, T)$. When this coefficient is zero, the independent projection identity is recovered.

4.2.1 Takano's stronger density coefficient

For comparison, if (S, T) has joint density $p_{S,T}$ and marginal densities p_S, p_T, define

$$\delta_r(S,T) = \left(\iint p_S(s) p_T(t) \left| \frac{p_{S,T}(s,t)}{p_S(s) p_T(t)} - 1 \right|^r \, \mathrm{d}s \, \mathrm{d}t \right)^{1/r}.$$

For continuous densities,

$$4\alpha(S,T) \le \delta_4(S,T) \le \delta_\infty(S,T) = \psi(S,T).$$

Takano's density condition is stronger than the α-mixing assumption. Johnson obtains convergence for the full normalized sum by controlling the terms between separated blocks.

4.2.2 Smoothing is continuous for the mixing coefficient

Theorem 4.7 (Continuity of the mixing coefficient under smoothing). Let $X = S + Z_S^{(\tau)}$ and $Y = T + Z_T^{(\tau)}$, with $\max\{\mathrm{Var}(S), \mathrm{Var}(T)\} \le K\tau$. The Gaussian noises are mutually independent and independent of (S, T, Z). Suppose $\mathbb{E}Z = 0$ and $\mathrm{Var}(Z) \le \varepsilon$; Z may depend on S, T. There is a function $r_K(\varepsilon)$, depending also on the fixed smoothing scale τ, such that

$$\alpha(X + Z, Y) \le \alpha(X, Y) + r_K(\varepsilon), \qquad r_K(\varepsilon) \longrightarrow 0 \quad (\varepsilon \downarrow 0).$$

This estimate controls the change in the mixing coefficient when a short separating interval is added to a block.

4.2.3 A bound on standardized Fisher information

For a tail profile $\chi(R) \downarrow 0$, let $\mathcal{C}_\chi$ consist of variables satisfying

$$\mathbb{E}\left[X^2 \mathbf{1}_{\{|X| \geq R\sqrt{\mathrm{Var}(X)}\}}\right] \leq \mathrm{Var}(X)\chi(R).$$

The nonlinear part of a score is measured by its distance from affine functions under a Gaussian reference law. A lower bound on the smoothed joint density controls this distance in terms of $\Delta(X, Y, \beta)$. If X, Y remain in a common class $\mathcal{C}_\chi$ and β stays away from 0 and 1, then $\Delta(X, Y, \beta) \to 0$ implies $J_{\mathrm{st}}(X) \to 0$.

Ibragimov's moment theorem gives uniform control of the tails. Under the hypotheses of Theorem 4.3, the block sums belong uniformly to one $\mathcal{C}_\chi$, and $v_n/n \to v$.

4.2.4 Separated blocks

Divide the index set into long blocks separated by shorter intervals. The separating intervals grow in length but remain small relative to the blocks. This construction ensures that

- each block accounts for a positive fraction of the total variance;
- the separating intervals contribute negligibly to the normalized sum;
- the mixing coefficient between distinct blocks tends to zero.

With separating intervals of order $\sqrt{m}$, the book obtains an error of the form

$$c(m) = C\left[\alpha(\sqrt{m}) + r_K(m^{-1/2})\right]^{1/3-\eta} \longrightarrow 0.$$

Consequently, for $m \geq n$, one has the approximate subadditivity relation

$$J_{\mathrm{st}}(V_{m+n}^{(\tau)}) \leq \frac{m}{m+n} J_{\mathrm{st}}(V_m^{(\tau)}) + \frac{n}{m+n} J_{\mathrm{st}}(V_n^{(\tau)}) + c(m) - \Delta\left(V_m^{(\tau)}, \widetilde{V}_n^{(\tau)}, \frac{m}{m+n}\right).$$

Along $n = 2^k$, either the deficit is large enough to make the Fisher information decrease, or it is small enough to force the score to be nearly linear. A telescoping argument therefore gives $J_{\mathrm{st}}(V_{2^k}^{(\tau)}) \to 0$; subadditivity then gives convergence for every n.

4.3 Proof of the approximate subadditive relation

4.3.1 Joint scores and the exact decomposition

Let $p_{X,Y}$ be a smooth joint density and define its coordinate scores by

$$\rho_{X,Y}^{(1)}(x, y) = \partial_x \log p_{X,Y}(x, y), \qquad \rho_{X,Y}^{(2)}(x, y) = \partial_y \log p_{X,Y}(x, y).$$

If $\widetilde{\rho}$ is the score of $X + Y$, differentiation under the integral gives

$$\widetilde{\rho}(z) = \mathbb{E}[\rho_{X,Y}^{(1)}(X, Y) \mid X + Y = z] = \mathbb{E}[\rho_{X,Y}^{(2)}(X, Y) \mid X + Y = z].$$

Thus conditional expectation still describes the score of the sum, but it projects a *joint* score rather than a marginal score.

For constants a, b, define the dependence correction

$$M_{a,b}(x,y) = a(\rho^{(1)}_{X,Y}(x,y) - \rho_X(x)) + b(\rho^{(2)}_{X,Y}(x,y) - \rho_Y(y)).$$

It is identically zero under independence. The two-dimensional Stein identity gives the following decomposition. Put $W = \sqrt{\beta}X + \sqrt{1-\beta}Y$ and now let $\widetilde{\rho}$ denote the score of W:

$$\begin{aligned}&\beta J(X) + (1-\beta)J(Y) - J(\sqrt{\beta}X + \sqrt{1-\beta}Y)\\ &\quad + 2\sqrt{\beta(1-\beta)}\mathbb{E}[\rho_X(X)\rho_Y(Y)] + 2\mathbb{E}[M_{\sqrt{\beta},\sqrt{1-\beta}}(X,Y)\widetilde{\rho}(W)] = \Delta(X,Y,\beta).\end{aligned}$$

The proof problem has now been reduced to bounding the final two dependence terms.

4.3.2 Bounds on smoothed densities and scores

Gaussian convolution supplies all of the local regularity used later. Assume $\mathbb{E}S = \mathbb{E}T = 0$ in the following density estimates.

1. If $\max\{\mathrm{Var}(S), \mathrm{Var}(T)\} \le K\tau$, then Chebyshev's inequality and the Gaussian kernel give the product lower bound
$$p_{X,Y}(x,y) \ge \frac{e^{-4K}}{4}\phi_{\tau/2}(x)\phi_{\tau/2}(y).$$
In particular, the smoothed joint density is positive.

2. Score moments are bounded by a wider Gaussian convolution. In particular, for $r \ge 1$,
$$\left(\mathbb{E}\left|\rho_X(X)\right|^r\right)^{1/r} \le \left(\frac{2^{1/r}2r}{\tau e}\right)^{1/2},$$
with analogous bounds for each coordinate of the joint score.

3. On the central interval, the score has the sharp cubic growth bound
$$\int_{-B\sqrt{\tau}}^{B\sqrt{\tau}} \rho_X(u)^2\,\mathrm{d}u \le \frac{8B^3}{\sqrt{\tau}}(3+2K), \qquad B > 1.$$

4. Ibragimov's covariance inequality and bounded Gaussian kernels give
$$|p_{X,Y}(x,y) - p_X(x)p_Y(y)| \le \frac{2}{\pi\tau}\alpha(S,T),$$
together with analogous first-derivative bounds.

Let
$$L_B = \{(x,y) : |x| \le B\sqrt{\tau},\ |y| \le B\sqrt{\tau}\}.$$
Combining these estimates on L_B bounds the correction involving the joint score by
$$|\mathbb{E}[M_{a,b}(X,Y)\widetilde{\rho}\,\mathbf{1}_{L_B}]| \le C(K,\tau)(|a|+|b|)\alpha(S,T)B^4.$$

4.3.3 Bounds on tails

Outside L_B, Chebyshev gives probability $O(B^{-2})$. Hölder's inequality, the uniform bound on score moments, and a choice of exponent close to one give

$$\left|\mathbb{E}[M_{a,b}\widetilde{\rho}\,\mathbf{1}_{L_B^c}]\right| \le \frac{C(K,\tau,\eta)(|a|+|b|)}{B^{2-\eta}}.$$

If a kth moment is uniformly bounded, the exponent improves to every value below k. For the remaining cross term, the estimates on L_B and its complement give

$$|\mathbb{E}[\rho_X(X)\rho_Y(Y)]| \le C_1\alpha(S,T)B^4 + C_2B^{-2}.$$

Balancing the two generic errors,

$$C_1\alpha B^4 + C_2B^{-(2-\eta)},$$

with B of order $\alpha^{-1/6}$ yields $C\alpha^{1/3-\eta}$. With a bounded kth moment, the same calculation gives any exponent below $k/(k+4)$, as in Theorem 4.6.

4.3.4 Control of the mixing coefficient after smoothing

To prove Theorem 4.7, first truncate to L_B. Chebyshev makes the complement uniformly small. On the box, repeated triangle inequalities separate three errors:

- the original dependence, bounded by $\alpha(X,Y)$;
- the change from p_X to p_{X+Z};
- the corresponding change in the joint density.

Local L^2 score control bounds density ratios under a small translation. A representative estimate is

$$\int_{|w|\le B\sqrt{\tau}} |p_{X+Z}(w) - p_X(w)|\,\mathrm{d}w \le e^{C\varepsilon^{1/5}} - 1 + 2\varepsilon^{1/5}.$$

The bivariate term is handled in the same way. First choose B large to control the tails and then let $\varepsilon \downarrow 0$ on the central box. This produces $r_K(\varepsilon) \to 0$.

5 Convergence to Stable Laws

5.1 Introduction to stable laws

5.1.1 Definitions

The Gaussian is only one member of the family of stable laws. The defining feature is invariance, up to centering and scaling, under repeated addition.

Definition 5.1 (Stable limit). A nondegenerate random variable Z is stable if there are IID random variables $X_1, X_2, \ldots$, positive constants a_n, and real constants b_n such that

$$\frac{1}{a_n}\left(\sum_{i=1}^{n} X_i - b_n\right) \Longrightarrow Z. \tag{5.1}$$

The normalizing constants have the form $a_n = n^{1/\alpha}L(n)$ for some $0 < \alpha \le 2$, where L is slowly varying. Thus $L(\lambda n)/L(n) \to 1$ for every fixed $\lambda > 0$.

The parameter α is the stability index. The cases $\alpha = 2$ and $\alpha = 1$ include the Gaussian and Cauchy limits, respectively. When $\alpha = 1$ and the summands have a finite mean, the ordinary law of large numbers instead produces a deterministic limit under n-scaling.

Theorem 5.2 (Characteristic functions of stable laws). A stable characteristic function has the form

$$\Psi(t) = \exp\Big(i\gamma t - c|t|^{\alpha}\big(1 + i\beta\,\mathrm{sgn}(t)\tan\tfrac{\pi\alpha}{2}\big)\Big), \qquad \alpha \neq 1, \tag{5.2}$$

$$\Psi(t) = \exp\Big(i\gamma t - c|t|\big(1 - i\beta\,\mathrm{sgn}(t)\tfrac{2}{\pi}\log|t|\big)\Big), \qquad \alpha = 1, \tag{5.3}$$

where $\gamma \in \mathbb{R}$, $c > 0$, $0 < \alpha \le 2$, and $-1 \le \beta \le 1$.

Here γ is a location parameter, c is a scale parameter, and β describes skewness. The value $c = 0$ in the characteristic function gives a point mass at γ. For $c > 0$, stable densities are smooth and bell-shaped, but they are not generally log-concave. Closed density formulas are available only for a few families. In Johnson's parametrization,

$$\phi_{c,\gamma}(x) = \frac{1}{\sqrt{4\pi c}}\exp\left(-\frac{(x-\gamma)^2}{4c}\right), \qquad (\alpha,\beta) = (2,0), \tag{5.4}$$

$$p_{c,\gamma}(x) = \frac{c}{\pi\{c^2 + (x-\gamma)^2\}}, \qquad (\alpha,\beta) = (1,0), \tag{5.5}$$

$$\ell_{c,\gamma}(x) = \frac{c}{\sqrt{2\pi|x-\gamma|^3}}\exp\left(-\frac{c^2}{2|x-\gamma|}\right)\mathbf{1}_{\{x>\gamma\}}, \qquad (\alpha,\beta) = \left(\frac{1}{2},-1\right). \tag{5.6}$$

For $\beta = 1$, the Lévy density is reflected onto $(-\infty, \gamma)$. Except for the asymmetric $\alpha = 1$ case, scale and location act by

$$g_{\gamma,c}(x) = c^{-1/\alpha}\, g_{0,1}\left(\frac{x-\gamma}{c^{1/\alpha}}\right). \tag{5.7}$$

5.1.2 Domains of attraction

Definition 5.3 (Domain of normal attraction). A law X belongs to the domain of normal attraction of a stable law Z of index α if (5.1) holds with $a_n \sim n^{1/\alpha}$.

For $\alpha = 2$, finite variance is the classical condition. For $0 < \alpha < 2$, the tails must be regularly varying with exponent $-\alpha$.

Theorem 5.4 (Tail characterization). A random variable X is in the domain of normal attraction of an α-stable law, $0 < \alpha < 2$, if and only if there are constants k_1, k_2, not both zero, such that

$$F_X(x)|x|^\alpha \longrightarrow k_1, \qquad x \to -\infty, \tag{5.8}$$

$$(1 - F_X(x))x^\alpha \longrightarrow k_2, \qquad x \to +\infty. \tag{5.9}$$

Consequently, a non-Gaussian stable variable Z has

$$\mathbb{E}|Z|^t < \infty \quad (0 < t < \alpha), \qquad \mathbb{E}|Z|^t = \infty \quad (t \geq \alpha). \tag{5.10}$$

The book also recalls an L^1 rate theorem. Let g be an α-stable density with $1 \leq \alpha \leq 2$. Choose b_n so that $n^{-1/\alpha}(\sum_{i=1}^n Z_i - b_n)$ has density g when the Z_i are IID with density g. Suppose the summands X_i are IID with density p and

$$\int_{\mathbb{R}} x^m \{p(x) - g(x)\}\,\mathrm{d}x = 0, \qquad m = 0, \ldots, 1 + \lfloor \alpha \rfloor,$$

and

$$\int_{\mathbb{R}} |x|^{1+\alpha} |p(x) - g(x)|\,\mathrm{d}x < \infty.$$

Then the density p_n of $U_n = n^{-1/\alpha}(\sum_{i=1}^n X_i - b_n)$ satisfies

$$\int_{\mathbb{R}} |p_n(x) - g(x)|\,\mathrm{d}x = O(n^{-1/\alpha}). \tag{5.11}$$

5.1.3 Entropy of stable laws

The entropy power inequality gives a restriction on the possible stability index.

Lemma 5.5 (Upper bound on the stability index). Suppose X, Y are IID, $H(X)$ is finite, and $(X + Y)/2^{1/\alpha}$ has the same law as X. Then $\alpha \leq 2$.

Proof. The entropy power inequality gives $H(X + Y) \geq H(X) + \frac{1}{2}\log 2$. Stability gives

$$H(X) = H\left(\frac{X + Y}{2^{1/\alpha}}\right) = H(X + Y) - \frac{\log 2}{\alpha}.$$

Thus $1/\alpha \geq 1/2$. ⊛

The Cauchy and Lévy entropies are

$$H(p_{c,\gamma}) = \log(4\pi c), \qquad H(\ell_{c,\gamma}) = \frac{1}{2}\log(16\pi e c^4) + \frac{3}{2}\kappa, \tag{5.12}$$

where κ is Euler's constant.

5.2 Parameter estimation for stable distributions

The scale and location of a stable approximation can be selected by minimizing relative entropy, minimizing Fisher information distance, or matching the expected logarithm of the density.

5.2.1 Minimizing relative entropy

For a density p,

$$D(p\|g_{c,\gamma}) = -H(p) - \int p(x)\log g_{c,\gamma}(x)\,\mathrm{d}x. \tag{5.13}$$

At an interior optimum, the scale and location scores have zero expectation under p:

$$\int p(x)\,\partial_c \log g_{c,\gamma}(x)\,\mathrm{d}x = 0, \qquad \int p(x)\,\partial_\gamma \log g_{c,\gamma}(x)\,\mathrm{d}x = 0. \tag{5.14}$$

Lemma 5.6 (Closest Cauchy law in relative entropy). If the minimizing Cauchy parameters are (C, G) and $f_X(z) = \mathbb{E}(z - X)^{-1}$ is the Cauchy transform, then

$$f_X(G + iC) = -\frac{i}{2C}. \tag{5.15}$$

Proof. Expand $D(p\|p_{c,\gamma})$ using $\log\{c^2 + (x-\gamma)^2\} - \log(c/\pi)$. The two score equations give

$$\int p(x)\frac{G - x}{C^2 + (x-G)^2}\,\mathrm{d}x = 0, \qquad \int p(x)\frac{C}{C^2 + (x-G)^2}\,\mathrm{d}x = \frac{1}{2C}.$$

Combining them as a complex fraction yields (5.15).

Lemma 5.7 (Closest Lévy scale at a fixed location). Fix γ such that $X > \gamma$ almost surely. Suppose $0 < \mathbb{E}[1/(X-\gamma)] < \infty$ and $\mathbb{E}|\log(X-\gamma)| < \infty$. The scale minimizing the cross entropy with a Lévy target is

$$c = \frac{1}{\sqrt{\mathbb{E}[1/(X-\gamma)]}}. \tag{5.16}$$

When $H(X)$ is finite, it also minimizes relative entropy.

The cross entropy is

$$\frac{1}{2}\log(2\pi) - \log c + \frac{3}{2}\mathbb{E}\log(X-\gamma) + \frac{c^2}{2}\mathbb{E}\frac{1}{X-\gamma}.$$

Differentiation in c gives (5.16). After substituting this scale, the location is chosen by minimizing

$$\frac{1}{2}\log\mathbb{E}\frac{1}{X-\gamma} + \frac{3}{2}\mathbb{E}\log(X-\gamma)$$

over the admissible values of γ.

5.2.2 Minimizing Fisher information distance

For location scores $\rho_p = p'/p$ and $\rho_g = g'/g$, define

$$J(p\|g) = \int p(x)\{\rho_p(x) - \rho_g(x)\}^2 \,\mathrm{d}x. \tag{5.17}$$

When the boundary terms vanish, integration by parts rewrites the part depending on the target as

$$\int p(x)\left(2\rho_g'(x) + \rho_g(x)^2\right) \,\mathrm{d}x. \tag{5.18}$$

For the Gaussian family, minimizing (5.17) gives the same parameters as KL minimization:

$$\mu = \mathbb{E}X, \qquad \sigma^2 = \operatorname{Var} X. \tag{5.19}$$

For a Lévy target with location fixed at 0,

$$\rho_{c,0}(x) = -\frac{3}{2x} + \frac{c^2}{2x^2}, \tag{5.20}$$

and Fisher fitting gives a different scale rule,

$$c^2 = 7\frac{\mathbb{E}(X^{-3})}{\mathbb{E}(X^{-4})}. \tag{5.21}$$

Thus the two minimization criteria select different Lévy scales.

5.2.3 Matching the logarithm of the density

Definition 5.8 (Cross-entropy functional). For densities f, g, set

$$\Lambda_g(f) = -\int_{\mathbb{R}} f(x)\log g(x)\,\mathrm{d}x, \tag{5.22}$$

and set $\Lambda_g(f) = \infty$ if the support of f is not contained in that of g.

For a symmetric stable family $p_c(x) = c^{-1/\alpha}p_1(x/c^{1/\alpha})$ with mode zero, Johnson proposes choosing c by the matching equation

$$\Lambda_{p_c}(h) = H(p_c). \tag{5.23}$$

Lemma 5.9 (Uniqueness of the matched scale). If $\Lambda_{p_1}(h) < \infty$, equation (5.23) has a unique solution $c > 0$.

Proof. Use the scaling $p_c(x) = c^{-1/\alpha}p_1(x/c^{1/\alpha})$ and put $u = c^{-1/\alpha}$. The relevant function is $K(u) = \int h(y)\log p_1(uy)\,\mathrm{d}y$. Unimodality makes K strictly decreasing; tail behavior makes $K(u) \to -\infty$ as $u \to \infty$; and its endpoint values straddle $\int p_1 \log p_1$. Hence there is exactly one crossing.

For a distribution function F and score $\rho_g = g'/g$, integration by parts yields, for any t,

$$\Lambda_g(f) = -\log g(t) + \left[\int_{-\infty}^{t} \rho_g(y)F(y)\,\mathrm{d}y - \int_t^{\infty} \rho_g(y)(1-F(y))\,\mathrm{d}y\right]. \tag{5.24}$$

For the Cauchy score $-2y/(1+y^2)$, the tail conditions for the domain of attraction make this quantity finite.

5.3 Extending de Bruijn's identity

5.3.1 Partial differential equations

Gaussian smoothing satisfies the heat equation. Stable densities also satisfy partial differential equations, including

$$\frac{\partial^2 \ell_c}{\partial c^2} = 2\frac{\partial \ell_c}{\partial x}, \tag{5.25}$$

$$\frac{\partial^2 p_c}{\partial c^2} + \frac{\partial^2 p_c}{\partial x^2} = 0. \tag{5.26}$$

Thus Lévy smoothing is governed by a first spatial derivative and second scale derivative, whereas the Cauchy family satisfies Laplace's equation in (x, c).

For rational $\alpha = m/n$, with m, n positive and coprime, the book records Medgyessy's general constant-coefficient relation. Put $B = \beta \tan(\pi\alpha/2)$ when $\alpha \neq 1$. For $\alpha = 1$, require $\beta = 0$ and put $B = 0$. If $k \in \mathbb{Z}$ and a positive integer M satisfy

$$B = \tan\left(\frac{\pi k}{Mn} - \frac{\pi\alpha}{2}\right), \tag{5.27}$$

then the stable density satisfies

$$(-1)^{Mn+k}(1+B^2)^{Mn/2}\frac{\partial^{Mm} g}{\partial x^{Mm}} - \frac{\partial^{Mn} g}{\partial c^{Mn}} = 0, \tag{5.28}$$

If g_c satisfies this equation, differentiation under the convolution shows that $f * g_c$ does as well.

5.3.2 Derivatives of relative entropy

Assume $D(f\|g_1) < \infty$. Let $h_c = f * g_c$, let $k_c = g_1 * g_c$, and define

$$D^c = D(h_c\|k_c), \qquad J^c(h_c\|k_c) = \int h_c(x)\left(\partial_c \log h_c(x) - \partial_c \log k_c(x)\right)^2 \mathrm{d}x. \tag{5.29}$$

Here J^c measures the difference between the scale scores. Differentiating relative entropy twice gives

$$\frac{\partial^2 D^c}{\partial c^2} = \int \frac{\partial^2 h_c}{\partial c^2} \log \frac{h_c}{k_c}\,\mathrm{d}x - \int \frac{h_c}{k_c}\frac{\partial^2 k_c}{\partial c^2}\,\mathrm{d}x + J^c(h_c\|k_c). \tag{5.30}$$

The PDE converts the first two integrals into spatial score terms.

Theorem 5.10 (Lévy de Bruijn analogue). If h_c is the density of $f * \ell_{c,0}$ and k_c that of $\ell_{1,0} * \ell_{c,0}$, then

$$\frac{\partial^2}{\partial c^2} D(h_c\|k_c) = J^c(h_c\|k_c). \tag{5.31}$$

Theorem 5.11 (Cauchy de Bruijn analogue). If h_c is the density of $f * p_{c,0}$ and k_c that of $p_{1,0} * p_{c,0}$, then

$$\frac{\partial^2}{\partial c^2} D(h_c\|k_c) = J(h_c\|k_c) + J^c(h_c\|k_c), \tag{5.32}$$

where J compares location scores and J^c compares scale scores.

The book also derives a mixed expression for the $(\alpha, \beta) = (3/2, -1)$ family. The integrand contains mixed derivative terms and need not be nonnegative.

5.3.3 Integral form of the identities

For the Cauchy semigroup, Taylor's formula with integral remainder gives, for $s < t$,

$$D^s = D^t - (t-s)(D^t)' + \int_s^t (u-s)(D^u)'' \, \mathrm{d}u. \tag{5.33}$$

Data processing under additional Cauchy noise makes D^t decreasing. A log-sum argument proves the contraction, and convergence of D^t rules out a nonzero limit of $(t-s)(D^t)'$. Scale invariance, entropy increase under convolution, uniform integrability of the cross-entropy term, and entropy semicontinuity then identify the endpoint limits. The resulting identity is

$$D(f\|p_{1,0}) = \int_0^\infty u\,(D^u)'' \, \mathrm{d}u, \tag{5.34}$$

with $(D^u)''$ supplied by (5.32).

5.4 Relationships between forms of convergence

For a symmetric log-concave target h with score $\rho_h = h'/h$, relative Fisher information controls both L^1 and uniform errors.

Proposition 5.12 (Fisher distance controls strong distributional distances). There is a constant $K = K(h)$ such that, for every density f,

$$\int_{\mathbb{R}} |f(x) - h(x)| \, \mathrm{d}x \le 2K\sqrt{J(f\|h)}, \tag{5.35}$$

$$\sup_x |f(x) - h(x)| \le (1 + Kh(0))\sqrt{J(f\|h)}. \tag{5.36}$$

Proof. Write $p(x) = h(x)(f(x)/h(x))' = f'(x) - f(x)\rho_h(x)$. Cauchy–Schwarz gives $\int_L |p| \le \sqrt{J(f\|h)}$ on every set L. Integrating the identity $f - Ch = h\int_0^x p/h$ separately near the mode and in the tails, then choosing the splitting point optimally, yields (5.35). The same representation gives (5.36).

Stable densities need not be log-concave. For a stable target Z of index α, Johnson proposes the scale invariant quantity

$$J_{\mathrm{st}}(X\|Z) = \left(\mathbb{E}|X|^k\right)^{2/k} J(X\|Z), \qquad 0 < k < \alpha. \tag{5.37}$$

For sums normalized and centered for the target Z, Johnson conjectures that this quantity tends to zero if and only if X_1 is in the domain of normal attraction of Z and $J_{\mathrm{st}}(X_1\|Z) < \infty$.

5.5 Steps toward a Brown inequality

For stable laws, orthogonal expansions and conditional expectation operators provide partial analogues of Brown's Gaussian projection inequality.

- A change of variables transports the even Hermite polynomials to an orthogonal basis for the Lévy density. For the Cauchy density, rational transforms of Chebyshev polynomials provide orthogonal bases.
- The stable analogues of the two projection maps are adjoint.
- If p is an α-stable density satisfying

$$p(u) = 2^{1/\alpha} \int_{\mathbb{R}} p(x)p(2^{1/\alpha}u - x)\,\mathrm{d}x,$$

  then

$$u_k(x) = \frac{p^{(k)}(x)}{p(x)} \tag{5.38}$$

  is an eigenfunction of the relevant conditional expectation operator with eigenvalue $2^{-k/\alpha}$.
- Unlike the Gaussian case, the two projection maps are not generally equal. Hence knowing eigenfunctions of one composition does not immediately identify the full spectrum needed for Brown's inequality.

Stable convolution also gives a lower bound for the smoothed density.

Lemma 5.13 (Stable smoothing lower bound). Let Z_c have an α-stable density g_c with $\beta \neq \pm 1$, and let p_c be the density of $U + Z_c$. For every fixed $A > 0$, there is a constant $R = R(g_c, A) > 0$ such that

$$p_c(x) \geq R\,\mathbb{P}(|U| \leq A)\, g_{c/2}(x) \qquad \text{for every } x. \tag{5.39}$$

The proof uses unimodality, the scaling $g_{c/2}(x) = 2^{1/\alpha} g_c(2^{1/\alpha}x)$, and the contribution from $|U| \leq A$. For normalized sums, a uniform bound $\sup_n \mathbb{E}|U_n|^k \leq M$ with $0 < k < \alpha$ gives

$$\mathbb{P}(|U_n| \leq A) \geq 1 - \frac{M}{A^k}.$$

Choose $A > M^{1/k}$ to obtain a positive lower bound independent of n. For one-sided stable targets, lower bounds are restricted to the support of the convolution.

6 Convergence on Compact Groups

On $\mathbb{R}$, normalized convolution approaches a Gaussian and entropy approaches the Gaussian maximum under a variance constraint. On a compact group there is no rescaling and no variance constraint. The natural equilibrium is Haar measure itself. Its density is the constant function 1, it is unchanged by convolution, and it maximizes entropy among all densities with respect to Haar measure.

6.1 Probability on compact groups

6.1.1 Topological groups and Haar probability

Definition 6.1 (Topological and compact groups). A topological group is a Hausdorff topological space G equipped with a group operation such that

$$(g, h) \longmapsto gh^{-1}$$

is continuous. It is a compact group when the underlying topological space is compact.

Finite groups with the discrete topology, the torus $\mathbb{T}^d$, the orthogonal group $O(d)$, and the special orthogonal group $SO(d)$ are standard examples. The real line under addition is a topological group but is not compact.

Compactness guarantees a unique normalized Haar probability measure μ. It is invariant under left translation, right translation, and inversion: for every bounded continuous F and every $t \in G$,

$$\int_G F(s)\,\mathrm{d}\mu(s) = \int_G F(st)\,\mathrm{d}\mu(s) = \int_G F(ts)\,\mathrm{d}\mu(s) = \int_G F(s^{-1})\,\mathrm{d}\mu(s).$$

The book often abbreviates $\mathrm{d}\mu(s)$ to $\mathrm{d}s$.

If f is a density with respect to μ, define

$$H(f) = -\int_G f \log f \;\mathrm{d}\mu, \qquad \mathrm{D}(f\|g) = \int_G f \log \frac{f}{g}\;\mathrm{d}\mu.$$

Writing 1_G for the uniform density,

$$\mathrm{D}(f\|1_G) = \int_G f \log f \;\mathrm{d}\mu = -H(f) \geq 0.$$

Thus

$$H(f) \leq 0, \qquad H(f) = 0 \Longleftrightarrow f = 1_G \quad \mu\text{-a.e.}$$

This is the maximum entropy property on compact groups.

For probability measures P, Q on G, their convolution is the law of the group product XY for independent $X \sim P$ and $Y \sim Q$. Equivalently,

$$\int_G F(u)\,(P * Q)(\,\mathrm{d}u) = \int_G \int_G F(st)\,P(\,\mathrm{d}s)Q(\,\mathrm{d}t).$$

If P, Q have densities f, g, then one convenient convention is

$$(f * g)(u) = \int_G f(ut^{-1})g(t)\,\mathrm{d}\mu(t).$$

Haar invariance makes the equivalent left- and right-shift formulas available even when G is noncommutative.

6.1.2 Weak convergence of convolution products

Let $\nu_1, \nu_2, \dots$ be probability measures on G, and set

$$\mu_n = \nu_1 * \cdots * \nu_n.$$

When μ_n has a density, denote it by g_n. If the increments are supported on a coset of a proper normal subgroup, their products can remain confined to cosets or vary periodically. On $\mathbb{Z}_2$, for example, repeated convolution of a point mass at the nonidentity element alternates between the two point masses.

Theorem 6.2 (Book Theorem 6.1: Ito–Kawada–Stromberg). In the IID case $\nu_i = \nu$, suppose that the smallest closed subgroup containing $\operatorname{supp}(\nu)$ is G. Suppose also that $\operatorname{supp}(\nu)$ is not contained in a coset of any proper closed normal subgroup. Then

$$\nu^{*n} \Longrightarrow \mu$$

weakly.

Theorem 6.3 (Book Theorem 6.2: a minorization bound). Suppose that for every measurable $A \subseteq G$,

$$\nu_i(A) \geq c_i \mu(A), \qquad 0 \leq c_i \leq 1.$$

Then

$$\sup_A |\mu_n(A) - \mu(A)| \leq \prod_{i=1}^{n} (1 - c_i).$$

In particular, $\sum_i c_i = \infty$ implies convergence in total variation.

Write $\nu_i = c_i \mu + (1 - c_i)\widetilde{\nu}_i$ and use $\mu * \lambda = \lambda * \mu = \mu$. The coefficient of the remaining convolution $\widetilde{\nu}_1 * \cdots * \widetilde{\nu}_n$ is $\prod_i (1 - c_i)$.

Weak compactness of probability measures is automatic on compact G, but compactness of densities in a stronger norm is not. On the circle,

$$f_m(u) = m \mathbf{1}_{[0,1/m]}(u)$$

has no L^1-Cauchy subsequence. Additional regularity must therefore come from convolution.

Definition 6.4 (Uniform equicontinuity). A family $\mathcal{F} \subset C(G)$ is uniformly equicontinuous if, for every $\varepsilon > 0$, there is a neighborhood V of the identity such that

$$|f(s) - f(t)| \leq \varepsilon \quad \text{for every } f \in \mathcal{F} \text{ whenever } ts^{-1} \in V.$$

Arzelà–Ascoli turns uniform equicontinuity and uniform boundedness on compact G into total boundedness in the supremum norm.

Theorem 6.5 (Book Theorem 6.4: a uniformly convergent subsequence). If ν_m has a continuous density for some m, then the densities g_n of $\nu_1 * \cdots * \nu_n$, $n \geq m$, have a uniformly convergent subsequence.

Indeed, convolution preserves the modulus of continuity of g_m: translation of the argument before averaging changes every later g_n by no more than it changed g_m. Also, $\|g_n\|_\infty \le \|g_m\|_\infty$. Arzelà–Ascoli then supplies the subsequence.

6.1.3 Conditions for uniform convergence

The results on uniform convergence in this subsection concern connected compact groups. They use harmonic analysis and the Peter–Weyl theorem.

For a probability measure ν and $x \ge 0$, let $H^+(\nu - x\mu)$ be a Hahn positive set for the signed measure $\nu - x\mu$, and define

$$M_\nu(x) = \mu(H^+(\nu - x\mu)), \qquad N_\nu(y) = \inf\{x \ge 0 : M_\nu(x) < y\}.$$

If the absolutely continuous part of ν has density p, then $M_\nu(x) = \mu\{s : p(s) \ge x\}$. The bounds below use

$$\widetilde{S}(\nu) = \int_0^\infty M_\nu(x)^3 \, \mathrm{d}x, \qquad S(\nu) = \int_0^{\pi/4} y^2 N_\nu(y) \, \mathrm{d}y.$$

Theorem 6.6 (Book Theorem 6.5: uniformly bounded increment densities). Let G be connected. Suppose ν_i has a density p_i satisfying $p_i \le c_i < \infty$. For $n \ge 3$, the convolution density obeys an estimate of the form

$$\sup_{s \in G} |g_n(s) - 1| \le C c_1 c_2 \left(\sum_{i=3}^{n} c_i^{-2} \right)^{-1}.$$

Consequently, uniformly bounded IID densities give an $O(n^{-1})$ uniform rate.

The proof on the circle bounds Fourier coefficients and then uses Parseval. Peter–Weyl extends the argument from characters to matrix coefficients of irreducible unitary representations on a general compact group.

The next two results weaken boundedness.

- **Book Theorem 6.6** assumes that the first six increment measures have $L^2(G, \mu)$ densities. Its bound has the form

  $$\|g_n - 1\|_\infty \le C \prod_{i=1}^{6} \|p_i\|_2 \left(\sum_{i=7}^{n} \widetilde{S}(\nu_i) \right)^{-1}.$$

  Thus divergence of the sum gives uniform convergence, with an $O(n^{-1})$ IID rate. The proof decomposes measures into pieces with bounded densities.

- **Book Theorem 6.7** requires only two increment measures to possess L^2 densities, say p_i, p_j. It gives

  $$\|g_n - 1\|_\infty \le \|p_i - 1\|_2 \, \|p_j - 1\|_2 \prod_{\substack{k \le n \\ k \ne i,j}} \left(1 - \frac{11}{24} S(\nu_k) \right).$$

  The densities converge uniformly when this product tends to zero; $\sum_k S(\nu_k) = \infty$ is sufficient. The proof controls invariant measures on unit spheres arising from irreducible representations.

Theorem 6.7 (Book Theorem 6.8: exponential IID uniform convergence). Let G be connected. If $\nu_i = \nu$ and ν has a density $f \in L^{1+\varepsilon}(G, \mu)$ for some $\varepsilon > 0$, then there are constants $a, b > 0$ such that, for all sufficiently large n,

$$\sup_{s \in G} |g_n(s) - 1| \le a e^{-bn}.$$

Hölder's inequality first improves integrability under convolution until an L^2 density is available; **Book Theorem 6.7** then applies. Stronger bounds are available for certain non-Abelian groups through estimates on their irreducible representations.

6.2 Convergence in relative entropy

6.2.1 Two coefficients for density lower bounds

The entropy argument works on any compact group, for measures that need not initially have densities. Using M_ν and N_ν defined above, set

$$c(\nu) = N_\nu(1), \qquad d(\nu) = \left[\int_0^\infty y^{-2} (1 - M_\nu(y)) \, \mathrm{d}y \right]^{-1}, \tag{6.1}$$

with value zero when the integral diverges. Write $c_n = c(\nu_n)$ and $d_n = d(\nu_n)$. If ν_n has density f_n, then

$$c_n = \operatorname{ess\,inf} f_n, \qquad d_n = \left(\int_G \frac{1}{f_n(s)} \, \mathrm{d}\mu(s) \right)^{-1}.$$

Thus c_n is the largest coefficient for which $\nu_n \ge c_n \mu$, and d_n is the harmonic mean of the density. One has $d_n \ge c_n$, and d_n can be positive when $c_n = 0$.

6.2.2 The theorem with an explicit rate

Theorem 6.8 (Book Theorem 6.9). Suppose g_m exists and $\mathrm{D}(g_m \| 1_G) < \infty$. For every $n \ge m + 1$,

$$\mathrm{D}(g_n \| 1_G) \le (1 - c_n) \, \mathrm{D}(g_{n-1} \| 1_G). \tag{6.2}$$

Hence, if $\sum_{n > m} c_n = \infty$, then

$$\mathrm{D}(g_n \| 1_G) \longrightarrow 0.$$

Iteration gives the explicit estimate

$$\mathrm{D}(g_n \| 1_G) \le \mathrm{D}(g_m \| 1_G) \prod_{i=m+1}^{n} (1 - c_i) \le \mathrm{D}(g_m \| 1_G) \exp\left(- \sum_{i=m+1}^{n} c_i \right).$$

In the IID case, a positive essential infimum therefore gives exponential convergence in relative entropy.

6.2.3 A broader theorem without an explicit rate

Theorem 6.9 (Book Theorem 6.10). Suppose that, for some m and some $\varepsilon > 0$,

$$\|g_m\|_{1+\varepsilon} < \infty, \qquad \sum_{n>m} d_n = \infty.$$

Then

$$\mathrm{D}(g_n \| 1_G) \longrightarrow 0.$$

The condition on d_n permits more general increments than the condition on c_n. The proof first obtains L^1 convergence from $\sum d_n = \infty$, then uses the $L^{1+\varepsilon}$ bound to deduce convergence in relative entropy.

6.2.4 Entropy monotonicity and equality

Let $R_u f(v) = f(vu)$ denote right translation. If g_n is the density of $\nu_1 * \cdots * \nu_n$ and $\mathrm{D}(g_{n-1} \| 1_G) < \infty$, the entropy increment has the exact identity

$$\mathrm{D}(g_{n-1} \| 1_G) - \mathrm{D}(g_n \| 1_G) = \int_G \mathrm{D}(g_{n-1} \| R_u g_n)\, \nu_n(\,\mathrm{d}u). \tag{6.3}$$

The right side is nonnegative. Therefore

$$H(g_n) \geq H(g_{n-1}), \qquad \mathrm{D}(g_n \| 1_G) \leq \mathrm{D}(g_{n-1} \| 1_G).$$

The identity is a formula for mutual information. If X and Y are independent and Y has finite entropy, then

$$H(YX) - H(Y) = I(X; YX).$$

Conditioning on $X = u$ merely translates the law of Y, so the conditional entropy remains $H(Y)$. The increase in entropy is exactly the information that the product retains about the new increment.

The equality case has a subgroup characterization.

Proposition 6.10 (Book Proposition 6.1: equality case). Let X, Y be IID with density f and $\mathrm{D}(f \| 1_G) < \infty$. If

$$\mathrm{D}(f * f \| 1_G) = \mathrm{D}(f \| 1_G),$$

then $\mathrm{supp}(f)$ is a coset Hu of a closed subgroup H with $\mu(H) > 0$, and

$$uHu^{-1} = H, \qquad f = \frac{\mathbf{1}_{Hu}}{\mu(H)} \quad \mu\text{-a.e.}$$

The subgroup H is open, and u belongs to its normalizer.

Indeed, equality in (6.3) says that X and YX are independent. The resulting translation identities make f constant on a coset Hu. Products of elements in the support imply $uHu^{-1} = H$. Conversely, the uniform law on such a coset has a convolution square uniform on Hu^2, so its entropy is unchanged.

6.3 Comparison of forms of convergence

Convergence in relative entropy and uniform convergence are incomparable for a general target density. If the target vanishes, even uniform convergence can leave relative entropy infinite: on $\mathbb{Z}_2$, a sequence assigning mass $1/n$ where the target assigns zero converges uniformly but has infinite KL distance. Conversely, KL convergence can coexist with narrow density spikes, so it need not imply uniform or $L^{1+\varepsilon}$ convergence.

For convergence to the Haar density,

$$\begin{aligned} L^{1+\varepsilon} \text{ convergence for some } \varepsilon > 0 &\Longrightarrow \mathrm{D}(\cdot \,\|1_G) \text{ convergence} \\ &\Longrightarrow L^1 \text{ convergence} \\ &\Longrightarrow \text{weak convergence.} \end{aligned}$$

The first and second implications are strict. Uniform convergence implies $L^{1+\varepsilon}$ convergence on compact G, so it is stronger still.

If a general target h is bounded away from zero, the elementary chi-square bound

$$\mathrm{D}(f\|h) \le \int_G \frac{(f-h)^2}{h}\,\mathrm{d}\mu \le \frac{1}{c}\,\|f-h\|_2^2\,, \qquad h \ge c > 0,$$

implies convergence in relative entropy from convergence in L^2.

Theorem 6.11 (Book Theorem 6.11). Let f_n, h be probability densities, with $h \ge c > 0$. If, for some $\varepsilon > 0$,

$$\int_G |f_n - h|^{1+\varepsilon}\,\mathrm{d}\mu \longrightarrow 0,$$

then

$$\mathrm{D}(f_n\|h) \longrightarrow 0.$$

The proof truncates at a height K. On $\{f_n/h \le K\}$, the logarithm is controlled by L^1 convergence; on the set where the density is large, a small power of f_n/h dominates its logarithm, and $L^{1+\varepsilon}$ convergence makes the tail uniformly negligible.

Convolution is an L^p contraction:

$$\|f * \nu\|_p \le \|f\|_p\,, \qquad p \ge 1.$$

This follows from Jensen or Hölder and Haar invariance.

Theorem 6.12 (Book Theorem 6.12). Let g_n be the densities of $\nu_1 * \cdots * \nu_n$. Suppose a density exists from some index onward and

$$\|g_m\|_{1+\varepsilon} < \infty$$

for some m and $\varepsilon > 0$. If

$$\|g_n - 1_G\|_1 \longrightarrow 0,$$

then

$$\mathrm{D}(g_n\|1_G) \longrightarrow 0.$$

The $L^{1+\varepsilon}$ norm cannot increase after time m. Interpolation gives $L^{1+\varepsilon'}$ convergence for some $0 < \varepsilon' < \varepsilon$, after which Theorem 6.11 applies.

6.4 Proofs of the convergence in relative entropy theorems

6.4.1 Proof of Book Theorem 6.9: explicit rate

Two lemmas isolate the argument.

Lemma 6.13 (Shift averaging). For probability densities p, q on G,

$$\int_G \mathrm{D}(p\|R_u q)\,\mathrm{d}\mu(u) \geq \mathrm{D}(p\|1_G).$$

To see this, average the cross-entropy term over u. Haar invariance makes $\int q(vu)\,\mathrm{d}\mu(u) = 1$, and Jensen gives $\int \log q(vu)\,\mathrm{d}\mu(u) \leq 0$.

Lemma 6.14 (A uniform component survives convolution). If ν_1 has a density and $c_2 = c(\nu_2)$, then $\nu_1 * \nu_2$ has a density g_2 satisfying

$$g_2 \geq c_2 \quad \mu\text{-a.e.}$$

The definition of c_n says, up to arbitrarily small loss, that $\nu_n \geq c_n\mu$. Apply this lower bound to the identity for the entropy increment (6.3), and then apply shift averaging:

$$\begin{aligned}\mathrm{D}(g_{n-1}\|1_G) - \mathrm{D}(g_n\|1_G) &\geq c_n \int_G \mathrm{D}(g_{n-1}\|R_u g_n)\,\mathrm{d}\mu(u)\\ &\geq c_n\,\mathrm{D}(g_{n-1}\|1_G).\end{aligned}$$

Rearrangement gives (6.2).

6.4.2 Proof without an explicit rate

The coefficient $d(\nu)$ in (6.1) is designed for a weighted Cauchy–Schwarz inequality.

Lemma 6.15 (Book Lemma 6.9: weighted integral bound). For every bounded measurable F,

$$\int_G F(u)^2 \nu(\,\mathrm{d}u) \geq d(\nu)\left(\int_G |F(u)|\,\mathrm{d}\mu(u)\right)^2.$$

When ν has density f, this is simply Cauchy–Schwarz with weights f and $1/f$.

Theorem 6.16 (Book Theorem 6.13). Assume a density g_m exists and $\mathrm{D}(g_m\|1_G) < \infty$. If

$$\sum_{n>m} d_n = \infty,$$

then

$$\|g_n - 1_G\|_1 \longrightarrow 0.$$

Kullback–Pinsker and (6.3) give, using natural logarithms,

$$\mathrm{D}(g_{n-1}\|1_G) - \mathrm{D}(g_n\|1_G) \geq \frac{1}{2}\int_G \|g_{n-1} - R_u g_n\|_1^2\,\nu_n(\,\mathrm{d}u).$$

Applying Book Lemma 6.9 replaces ν_n by Haar measure and introduces the factor d_n. Moreover, averaging all shifts of g_n gives the constant density:

$$\int_G R_u g_n \, \mathrm{d}\mu(u) = 1_G.$$

Minkowski's inequality therefore yields

$$\int_G \|g_{n-1} - R_u g_n\|_1 \, \mathrm{d}\mu(u) \geq \|g_{n-1} - 1_G\|_1 .$$

Combining the estimates,

$$\mathrm{D}(g_{n-1}\|1_G) - \mathrm{D}(g_n\|1_G) \geq \frac{d_n}{2} \|g_{n-1} - 1_G\|_1^2 . \tag{6.4}$$

Convolution contracts the L^1 distance to 1_G, so this distance is nonincreasing. If it stayed above a positive constant while $\sum d_n = \infty$, summing (6.4) would force the nonnegative relative entropy below zero. This proves Theorem 6.16.

Theorem 6.16 gives $\|g_n - 1_G\|_1 \to 0$. Under the additional assumption $\|g_m\|_{1+\varepsilon} < \infty$, Theorem 6.12 then proves Theorem 6.9.

7 Convergence to the Poisson Distribution

The law of small numbers concerns sums of rare events whose total mean remains bounded. Their Poisson approximation can be studied through discrete scores, finite differences, and convolution with Poisson distributions.

Throughout this section the random variables take values in $\{0, 1, 2, \ldots\}$ and have finite means. The Poisson parameters are positive, and $P_\lambda(x) = e^{-\lambda}\lambda^x/x!$ denotes the Poisson mass function. In sums of nonnegative terms, a quotient $0/0$ is interpreted as zero and a positive numerator divided by zero as $+\infty$.

Write $(x)_k = x(x-1)\cdots(x-k+1)$ for the falling factorial, with $(x)_0 = 1$.

7.1 Entropy and the Poisson distribution

7.1.1 The law of small numbers

Let $X_i \sim \text{Bernoulli}(p_i)$ and

$$S_n = \sum_{i=1}^n X_i, \qquad \lambda = \sum_{i=1}^n p_i. \tag{7.1}$$

For independent Bernoulli variables, a small value of $\max_i p_i/\lambda$ gives a useful Poisson approximation.

Theorem 7.1 (Le Cam upper bound). For independent Bernoulli summands,

$$d_{\mathrm{TV}}(S_n, \mathrm{Po}(\lambda)) \le \frac{8}{\lambda}\sum_{i=1}^n p_i^2. \tag{7.2}$$

When $p_i = \lambda/n$, the bound is $8\lambda/n$.

The order cannot generally be improved. If $\ell_i = -\log(1-p_i)$ and $\sum_i \ell_i \le 1$, the book records the lower bound

$$d_{\mathrm{TV}}(S_n, \mathrm{Po}(\mu)) \ge \frac{1}{2}\left(\sum_{i=1}^n \ell_i^2\right)\exp\left(-\sum_{i=1}^n \ell_i\right) \quad \text{for every } \mu. \tag{7.3}$$

For $p_i = \lambda/n$ and $\lambda < 1$, this is asymptotic to a positive constant times $1/n$.

The Poisson law also has an extremal entropy characterization.

Definition 7.2 (Class of Bernoulli sums). Let

$$\mathcal{B}(\lambda) = \left\{\sum_{i=1}^n X_i : X_i \text{ independent Bernoulli}, \ \mathbb{E}\sum_i X_i = \lambda\right\}. \tag{7.4}$$

Theorem 7.3 (Poisson entropy supremum).

$$\sup_{X \in \mathcal{B}(\lambda)} H(X) = H(\mathrm{Po}(\lambda)). \tag{7.5}$$

The supremum is not attained within $\mathcal{B}(\lambda)$, because a Poisson variable is not a finite Bernoulli sum.

There is no elementary closed form for Poisson entropy. If $Z \sim \mathrm{Po}(\lambda)$, then

$$H(Z) = \lambda - \lambda \log \lambda + \mathbb{E} \log(Z!), \tag{7.6}$$

and Stirling's formula gives $H(Z) \sim \frac{1}{2} \log(2\pi e \lambda)$ as $\lambda \to \infty$.

Lemma 7.4 (The closest Poisson law matches the mean). Let X have mean $\mu = \mathbb{E}X$ with $0 < \mu < \infty$. Then

$$\inf_{\lambda > 0} D(X \| \mathrm{Po}(\lambda)) = D(X \| \mathrm{Po}(\mathbb{E}X)). \tag{7.7}$$

Proof. The divergence satisfies

$$D(X \| \mathrm{Po}(\lambda)) = D(X \| \mathrm{Po}(\mu)) + \lambda - \mu - \mu \log(\lambda/\mu).$$

The last three terms are $D(\mathrm{Po}(\mu) \| \mathrm{Po}(\lambda))$ and are minimized at $\lambda = \mu$. If $D(X \| \mathrm{Po}(\mu))$ is finite, this minimizer is unique. ⊛

It is therefore convenient to abbreviate

$$D(X) := D(X \| \mathrm{Po}(\mathbb{E}X)). \tag{7.8}$$

For independent summands with $p_i = \lambda/n$, the bounds below give relative entropy of order $1/n^2$.

7.1.2 Simplest bounds on relative entropy

Unlike Gaussian relative entropy, Poisson relative entropy need not decrease under every convolution. For example,

$$D(\mathrm{Bin}(1,p) \| \mathrm{Po}(p)) = (1-p)\log(1-p) + p, \tag{7.9}$$

$$D(\mathrm{Bin}(2,p) \| \mathrm{Po}(2p)) = 2(1-p)\log(1-p) + 2p - p^2 \log 2. \tag{7.10}$$

Thus adding two IID Bernoulli variables decreases the standardized Poisson divergence only for $p \lesssim 0.7035$.

Lemma 7.5 (Elementary binomial bound).

$$D(\mathrm{Bin}(n,p) \| \mathrm{Po}(np)) \le np^2. \tag{7.11}$$

For $p = \lambda/n$, this is only $O(1/n)$. For dependent variables, the bound includes their total correlation.

Theorem 7.6 (Relative entropy for dependent Bernoulli sums). Let $S_n = \sum_i X_i$ for possibly dependent Bernoulli variables with $\mathbb{E}X_i = p_i$ and $\sum_i p_i = \lambda$. Then

$$D(S_n \| \mathrm{Po}(\lambda)) \le \sum_{i=1}^n p_i^2 + \left(\sum_{i=1}^n H(X_i) - H(X_1, \dots, X_n) \right). \tag{7.12}$$

The second term is the total correlation: it vanishes under independence. For independent summands, a weighted Hölder convolution inequality yields the corresponding Rényi bound.

Theorem 7.7 (Rényi subadditivity for Bernoulli sums). For independent Bernoulli variables and $1 \le \alpha < \infty$,

$$D_\alpha(S_n \| \operatorname{Po}(\lambda)) \le \sum_{i=1}^{n} D_\alpha(X_i \| \operatorname{Po}(p_i)) \le \alpha \sum_{i=1}^{n} p_i^2. \tag{7.13}$$

The proof applies Hölder's inequality to convolution relative to Poisson weights. For $\alpha > 1$, equality in the underlying Hölder inequality forces exponential tilts of Poisson mass functions. For a Bernoulli variable and $\alpha > 1$,

$$D_\alpha(\operatorname{Bern}(p) \| \operatorname{Po}(p)) = p + \frac{1}{\alpha - 1} \log\{(1-p)^\alpha + p\}, \tag{7.14}$$

The expression at $\alpha = 1$ is obtained by taking the limit. The Bernoulli inequality gives the final αp^2 bound.

A sharper expansion gives a bound of order $1/n^2$ when $p = \lambda/n$.

Lemma 7.8 (Binomial entropy bound).

$$D(\operatorname{Bin}(n,p) \| \operatorname{Po}(np)) \le \frac{np^3 + p^2}{2}. \tag{7.15}$$

Hence, for $p = \lambda/n$,

$$D(\operatorname{Bin}(n,\lambda/n) \| \operatorname{Po}(\lambda)) \le \frac{\lambda^3 + \lambda^2}{2n^2}. \tag{7.16}$$

Proof. Expand the exact divergence between the binomial and Poisson laws and use

$$\log \frac{n!}{(n-r)!n^r} = \sum_{s=1}^{r-1} \log\left(1 - \frac{s}{n}\right) \le -\frac{r(r-1)}{2n}.$$

Then substitute the binomial factorial moments and the bound $\log(1-p) \le -p - p^2/2$.

7.2 Fisher information

7.2.1 Standard Fisher information

For a function $f : \mathbb{Z} \to \mathbb{R}$, let

$$\Delta f(x) = f(x+1) - f(x). \tag{7.17}$$

Definition 7.9 (Discrete score and Fisher information). If X has mass function f, define its score on $\{x : f(x) > 0\}$ by

$$\rho_X(x) = \frac{\Delta f(x-1)}{f(x)} = 1 - \frac{f(x-1)}{f(x)}, \tag{7.18}$$

and

$$J(X) = \sum_{x \geq 0} \frac{\{f(x) - f(x-1)\}^2}{f(x)} = \sum_{x \geq 0} \frac{f(x-1)^2}{f(x)} - 1, \tag{7.19}$$

where $f(-1) = 0$ and the sum uses the division conventions above. If $J(X) < \infty$, it also equals $\mathbb{E}\rho_X(X)^2$.

For $X \sim \mathrm{Po}(\lambda)$,

$$\rho_X(x) = 1 - \frac{x}{\lambda}, \qquad J(X) = \frac{1}{\lambda}. \tag{7.20}$$

If a law has bounded support, however, $J(X) = \infty$: at the upper edge the denominator in (7.18) vanishes after the mass function jumps to zero. Adding an independent Poisson variable removes this boundary singularity. The following identities for the backward score are stated for finite J.

The discrete score is chosen to satisfy a Stein identity.

Lemma 7.10 (Discrete Stein identity). If $J(X) < \infty$, then for every test function g for which the expectations are absolutely integrable,

$$\mathbb{E}[\rho_X(X)g(X)] = -\mathbb{E}[\Delta g(X)] = \mathbb{E}g(X) - \mathbb{E}g(X+1). \tag{7.21}$$

If $\mu = \mathbb{E}X$ and $0 < \sigma^2 = \mathrm{Var}\, X < \infty$, minimizing the square of the score after subtracting an affine function gives the standardized information.

Definition 7.11 (Standardized discrete Fisher information).

$$J_{\mathrm{st}}(X) = \sigma^2 J(X) - 1. \tag{7.22}$$

For finite $J(X)$,

$$J_{\mathrm{st}}(X) = \sigma^2 \mathbb{E}\left(\rho_X(X) + \frac{X-\mu}{\sigma^2}\right)^2 \geq 0.$$

If $J(X) = \infty$, set $J_{\mathrm{st}}(X) = \infty$.

For an integer $a \geq 0$ and $Z \sim \mathrm{Po}(\lambda)$, the shifted variable $a + Z$ has $J_{\mathrm{st}}(a+Z) = 0$.

For independent X, Y with finite J, scores of the sum are conditional expectations:

$$\rho_{X+Y}(z) = \mathbb{E}[\rho_X(X) \mid X+Y=z] = \mathbb{E}[\rho_Y(Y) \mid X+Y=z]. \tag{7.23}$$

Conditional Jensen therefore gives, for every $a \in \mathbb{R}$,

$$J(X+Y) \leq a^2 J(X) + (1-a)^2 J(Y), \tag{7.24}$$

If X, Y have finite positive variances, put $\beta_X = \sigma_X^2/(\sigma_X^2 + \sigma_Y^2)$. Then

$$J_{\mathrm{st}}(X+Y) \leq \beta_X J_{\mathrm{st}}(X) + (1-\beta_X) J_{\mathrm{st}}(Y). \tag{7.25}$$

7.2.2 Scaled Fisher information

The scaled score below gives finite information for Bernoulli sums with success probabilities below one.

Definition 7.12 (Scaled Poisson score and information). If X has mass function P and mean $\lambda > 0$, define the scaled score on $\{x : P(x) > 0\}$ by

$$\overline{\rho}_X(x) = \frac{(x+1)P(x+1)}{\lambda P(x)} - 1, \tag{7.26}$$

and

$$K(X) = \frac{1}{\lambda} \sum_{x \geq 0} \frac{\{(x+1)P(x+1) - \lambda P(x)\}^2}{P(x)}. \tag{7.27}$$

The sum uses the division conventions above. When $K(X) < \infty$, it equals $\lambda \mathbb{E}\overline{\rho}_X(X)^2$, and $P(x+1) > 0$ implies $P(x) > 0$.

The scaled score has the following properties. The expectation identities are used when $K(X) < \infty$ and the displayed expectations are absolutely integrable.

1. $K(X) \geq 0$, with equality exactly for $X \sim \mathrm{Po}(\lambda)$.
2. It measures failure of the recursion for factorial moments. For $g(x) = (x)_k$,
$$\mathbb{E}[\overline{\rho}_X(X)g(X)] = \frac{1}{\lambda}\mathbb{E}(X)_{k+1} - \mathbb{E}(X)_k. \tag{7.28}$$
3. If X^* is the size-biased law, $\mathbb{P}(X^* = x) = x\mathbb{P}(X = x)/\lambda$, then
$$\mathbb{E}[\overline{\rho}_X(X)g(X)] = \mathbb{E}g(X^* - 1) - \mathbb{E}g(X). \tag{7.29}$$
4. The Poisson Stein identity
$$\mathbb{E}\{\lambda g(Z+1) - Zg(Z)\} = 0 \tag{7.30}$$
holds for every bounded test function g exactly when $Z \sim \mathrm{Po}(\lambda)$.

The scaled score also satisfies a conditional expectation identity. If independent X, Y have finite K, positive means p, q, and $\alpha_X = p/(p+q)$, $\alpha_Y = q/(p+q)$, then

$$\overline{\rho}_{X+Y}(z) = \mathbb{E}\left[\alpha_X \overline{\rho}_X(X) + \alpha_Y \overline{\rho}_Y(Y) \mid X + Y = z\right]. \tag{7.31}$$

Theorem 7.13 (Subadditivity of scaled Fisher information). For independent nonnegative integer-valued X_i with positive means p_i and $\lambda = \sum_i p_i$,

$$K\left(\sum_{i=1}^n X_i\right) \leq \sum_{i=1}^n \frac{p_i}{\lambda} K(X_i). \tag{7.32}$$

For $X_i \sim \text{Bernoulli}(p_i)$,

$$K(X_i) = \frac{p_i^2}{1-p_i}, \tag{7.33}$$

and therefore

$$\boxed{K(S_n) \le \frac{1}{\lambda}\sum_{i=1}^{n} \frac{p_i^3}{1-p_i}.} \tag{7.34}$$

For $p_i = \lambda/n$, this bound is $O(n^{-2})$.

7.2.3 Dependent variables

Let $P(x,y)$ be a joint mass function with positive means p, q. Assume that its support is closed under decreasing either positive coordinate by one, and that the coordinate scores below have finite second moments. Define coordinate scores, for example

$$\overline{\rho}^{(1)}_{X,Y}(x,y) = \frac{(x+1)P(x+1,y)}{pP(x,y)} - 1. \tag{7.35}$$

The sum score remains a conditional expectation:

$$\overline{\rho}_{X+Y}(z) = \mathbb{E}\left[\alpha_X \overline{\rho}^{(1)}_{X,Y}(X,Y) + \alpha_Y \overline{\rho}^{(2)}_{X,Y}(X,Y) \mid X+Y=z\right]. \tag{7.36}$$

Let

$$M(x,y) = \alpha_X\{\overline{\rho}^{(1)}_{X,Y}(x,y) - \overline{\rho}_X(x)\} + \alpha_Y\{\overline{\rho}^{(2)}_{X,Y}(x,y) - \overline{\rho}_Y(y)\}. \tag{7.37}$$

It vanishes under independence. Expanding a conditional square gives the exact nonnegative identity

$$\begin{aligned} &\frac{\alpha_X^2}{p}K(X) + \frac{\alpha_Y^2}{q}K(Y) - \frac{K(X+Y)}{p+q} + 2\alpha_X\alpha_Y\mathbb{E}[\overline{\rho}_X(X)\overline{\rho}_Y(Y)] \\ &\qquad + 2\mathbb{E}[M(X,Y)\overline{\rho}_{X+Y}(X+Y)] \\ &= \mathbb{E}\left(\alpha_X\overline{\rho}_X(X) + \alpha_Y\overline{\rho}_Y(Y) - \overline{\rho}_{X+Y}(X+Y)\right)^2 \ge 0. \end{aligned} \tag{7.38}$$

Bounds for dependent sums therefore require control of the score correlation and the term involving M.

7.3 Strength of the bounds

A formal Taylor expansion of the discrete de Bruijn integrand introduced below suggests

$$D(P\|\,\text{Po}(\lambda)) \approx \frac{1}{2}\int_0^\infty \frac{K(X+\text{Po}(t))}{\lambda+t}\,\mathrm{d}t. \tag{7.39}$$

Subadditivity gives

$$K(X+\text{Po}(t)) \le \frac{\lambda}{\lambda+t}K(X), \tag{7.40}$$

so the heuristic right side is at most $K(X)/2$. A discrete logarithmic Sobolev inequality gives the following bound.

Theorem 7.14 (Bounds in terms of K). Suppose X has mean λ and interval support, possibly infinite. Then

$$D(X\|\operatorname{Po}(\lambda)) \le K(X), \tag{7.41}$$

$$\sum_{x=0}^{\infty} |P(x) - P_\lambda(x)| \le \sqrt{2K(X)}. \tag{7.42}$$

To see why K is a discrete relative Fisher information, put $f(x) = P(x)/P_\lambda(x)$. Then

$$\Delta f(x) = \frac{(x+1)P(x+1) - \lambda P(x)}{\lambda P_\lambda(x)}, \qquad K(X) = \lambda \sum_x P_\lambda(x) \frac{\{\Delta f(x)\}^2}{f(x)}. \tag{7.43}$$

This is the discrete counterpart of $\int q(f')^2/f$ for continuous relative Fisher information.

Combining (7.34) with (7.41) yields, for independent Bernoulli variables,

$$D(S_n\|\operatorname{Po}(\lambda)) \le \frac{1}{\lambda} \sum_{i=1}^{n} \frac{p_i^3}{1-p_i}, \tag{7.44}$$

which is $O(n^{-2})$ when $p_i = \lambda/n$.

7.4 A Poisson de Bruijn identity

Poisson smoothing is governed by a differential–difference equation. If q_t is the mass function of $Z + \operatorname{Po}(t)$, then

$$\frac{\partial q_t}{\partial t}(x) = q_t(x-1) - q_t(x). \tag{7.45}$$

This follows first for P_t itself and then for every convolution by linearity.

Let $X_t = X + \operatorname{Po}(t)$, let P_t be its law, and define the shifted size-biased law

$$\widetilde{P}_t(r) = \frac{(r+1)\mathbb{P}(X_t = r+1)}{\lambda + t}. \tag{7.46}$$

Theorem 7.15 (Discrete de Bruijn identity). If X is a sum of independent Bernoulli variables with success probabilities strictly below one and $\mathbb{E}X = \lambda > 0$, then

$$\boxed{D(P\|\operatorname{Po}(\lambda)) = \int_0^\infty D(\widetilde{P}_t\|P_t)\,\mathrm{d}t.} \tag{7.47}$$

Proof. For $a > 0$, the fundamental theorem of calculus gives

$$D(P\|\operatorname{Po}(\lambda)) - D(P_a\|\operatorname{Po}(\lambda + a)) = -\int_0^a \frac{\partial}{\partial t} D(P_t\|\operatorname{Po}(\lambda + t))\,\mathrm{d}t.$$

The K bound and (7.40) imply that the terminal divergence tends to zero as $a \to \infty$. Differentiate the entropy and factorial terms using (7.45); summation by parts and the Poisson Stein identity turn the derivative into $-D(\widetilde{P}_t\|P_t)$. Integrating proves (7.47).

The integrand is the relative entropy of the shifted size biased law with respect to P_t. Its quadratic approximation is $K(X_t)/(2(\lambda + t))$.

7.5 L^2 bounds on Poisson distance

7.5.1 L^2 definitions

Definition 7.16 (Poisson–Charlier polynomials).

$$c_k(x;\mu) = \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} \mu^{-j} (x)_j. \tag{7.48}$$

They satisfy

$$\sum_{x} P_\mu(x) c_k(x;\mu) c_\ell(x;\mu) = \mu^{-k} k! \, \mathbf{1}_{\{k=\ell\}}, \tag{7.49}$$

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} c_k(x;\mu) = e^{-t} \left(1 + \frac{t}{\mu}\right)^x. \tag{7.50}$$

If F_X is the mass function of a variable of mean μ, define its relative density $f_X = F_X / P_\mu$. For independent X, Y of means μ, λ, put $\alpha = \mu/(\mu+\lambda)$. Then

$$f_{X+Y} = L_\alpha[f_X, f_Y], \tag{7.51}$$

where

$$L_\alpha[p,q](r) = \sum_{s} \binom{r}{s} \alpha^s (1-\alpha)^{r-s} p(s) q(r-s). \tag{7.52}$$

Proposition 7.17 (Convolution of Charlier coefficients). Suppose $f_X \in L^2(P_\mu)$ and $f_Y \in L^2(P_\lambda)$, and write their orthogonal expansions as

$$f_X = \sum_{k\ge 0} u_X^{(k)} c_k(\,\cdot\,;\mu), \qquad f_Y = \sum_{k\ge 0} u_Y^{(k)} c_k(\,\cdot\,;\lambda). \tag{7.53}$$

Then

$$f_{X+Y} = \sum_{k\ge 0} u_{X+Y}^{(k)} c_k(\,\cdot\,;\mu+\lambda), \qquad u_{X+Y}^{(k)} = \sum_{j=0}^{k} u_X^{(j)} u_Y^{(k-j)}. \tag{7.54}$$

The proof is a calculation using generating functions based on

$$L_\alpha[c_k(\,\cdot\,;\mu), c_\ell(\,\cdot\,;\lambda)] = c_{k+\ell}(\,\cdot\,;\mu+\lambda). \tag{7.55}$$

7.5.2 Sums of Bernoulli variables

For $X \sim \text{Bernoulli}(p)$, the relative density has coefficients

$$u_X^{(k)} = \frac{(-1)^k (1-k) p^k}{k!}. \tag{7.56}$$

In particular,

$$u_X^{(0)} = 1, \quad u_X^{(1)} = 0, \quad u_X^{(2)} = -\frac{p^2}{2}, \quad u_X^{(3)} = \frac{p^3}{3}, \quad u_X^{(4)} = -\frac{p^4}{8}. \tag{7.57}$$

For $S = \sum_i X_i$, write

$$\lambda = \sum_i p_i, \qquad T_r = \sum_i p_i^r, \qquad \bar{p} = \frac{T_2}{\lambda}. \tag{7.58}$$

If Λ_n is the set of partitions of n and m_i counts parts equal to i, the coefficient formula is

$$u_S^{(n)} = (-1)^n \sum_{\substack{\ell \in \Lambda_n \\ m_1 = 0}} \left\{ \prod_s \left(-\frac{T_{\ell_s}}{\ell_s} \right) \right\} \left\{ \prod_i \frac{1}{m_i!} \right\}. \tag{7.59}$$

For example,

$$u_S^{(3)} = \frac{T_3}{3}, \qquad u_S^{(4)} = \frac{T_2^2}{8} - \frac{T_4}{4}. \tag{7.60}$$

Parseval's identity gives

$$\|f_S\|^2_{L^2(P_\lambda)} = \sum_{r=0}^{\infty} \frac{r!}{\lambda^r} \left(u_S^{(r)} \right)^2 \tag{7.61}$$

$$= 1 + \frac{T_2^2}{2\lambda^2} + \frac{2T_3^2}{3\lambda^3} + \frac{24}{\lambda^4} \left(\frac{T_2^2}{8} - \frac{T_4}{4} \right)^2 + \sum_{r=5}^{\infty} \frac{r!}{\lambda^r} \left(u_S^{(r)} \right)^2. \tag{7.62}$$

The first nonconstant coefficient is quadratic in the probabilities. A bound on the full distance also requires control of the remaining terms of the orthogonal expansion.

7.5.3 The normal expansion

The same coefficient algebra appears in the Gaussian case. Define scaled Hermite polynomials $K_r(x;\mu)$ by

$$\sum_{r=0}^{\infty} K_r(x;\mu)\frac{s^r}{r!} = \exp\left(-\frac{s^2}{2\mu} + \frac{sx}{\mu}\right). \tag{7.63}$$

They satisfy

$$\int_{\mathbb{R}} K_r(x;\mu)K_s(x;\mu)\phi_\mu(x)\,\mathrm{d}x = \mu^{-r}r!\,\mathbf{1}_{\{r=s\}}. \tag{7.64}$$

If independent centered variables X, Y of variances μ, λ have relative densities in $L^2(\phi_\mu)$ and $L^2(\phi_\lambda)$, with Hermite coefficients $u_X^{(k)}$ and $u_Y^{(k)}$, then

$$u_{X+Y}^{(k)} = \sum_{j=0}^{k} u_X^{(j)} u_Y^{(k-j)}. \tag{7.65}$$

Thus Charlier and Hermite coefficients obey the same convolution formula.

8 Free Random Variables

8.1 Operators, traces, and spectral distributions

A free random variable is represented by a self-adjoint operator in a von Neumann algebra $\mathcal{A}$ equipped with a normalized trace τ. Expectation is

$$\mathbb{E}[P(X)] = \tau(P(X)).$$

The distribution of X is the probability measure μ_X satisfying

$$\tau(P(X)) = \int P(t)\,\mu_X(\mathrm{d}t)$$

for appropriate functions P. For an $n \times n$ self-adjoint matrix with normalized trace $\tau(A) = \operatorname{tr}(A)/n$, this is exactly the empirical distribution of the eigenvalues.

Definition 8.1 (Wigner semicircular law). The Wigner law with mean μ and variance σ^2 has density

$$w_{\mu,\sigma^2}(x) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - (x-\mu)^2}\,\mathbf{1}_{\{|x-\mu|\le 2\sigma\}}.$$

It plays the role of the Gaussian in free probability.

Its centered moments are zero in odd degree and, in degree $2k$, equal to the Catalan number times σ^{2k}:

$$\mathbb{E}[(W-\mu)^{2k}] = \sigma^{2k}\frac{1}{k+1}\binom{2k}{k}.$$

The corresponding orthogonal polynomials are rescaled Chebyshev polynomials of the second kind, just as Hermite polynomials are natural for the Gaussian and Poisson–Charlier polynomials for the Poisson law.

8.2 Cauchy and R-transforms

For a probability measure μ on $\mathbb{R}$, its Cauchy transform is

$$G_\mu(z) = \int_{\mathbb{R}} \frac{1}{z-t}\,\mu(\mathrm{d}t), \qquad \Im z > 0.$$

The measure can be recovered from the boundary imaginary part of G_μ. If K_μ is the functional inverse of G_μ, the R-transform is

$$R_\mu(z) = K_\mu(z) - \frac{1}{z}.$$

Theorem 8.2 (Linearization of free convolution; Book Theorem 8.1). If X and Y are free, then

$$R_{X+Y}(z) = R_X(z) + R_Y(z).$$

Thus R plays the role that the logarithm of the characteristic function plays for ordinary convolution.

For a centered law of variance σ^2,

$$R_X(z) = \sigma^2 z + O(z^2)$$

near the origin. Therefore the normalized free sum has R-transform converging to $\sigma^2 z$, the transform of the Wigner law.

Theorem 8.3 (Free central limit theorem; Book Theorem 8.2). If $X_1, X_2, \dots$ are free, identically distributed, centered, and have variance σ^2, then

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \Longrightarrow W(0, \sigma^2).$$

8.3 Freeness and conjugate variables

Subalgebras $\mathcal{B}_i \subset \mathcal{A}$ are *free* when every alternating product of centered elements from different subalgebras has trace zero. Random variables are free when the algebras they generate are free. Freeness determines mixed moments from the moments within each subalgebra.

The analogue of differentiation is the derivation ∂_X, defined by $\partial_X X = 1 \otimes 1$, by vanishing on the coefficient algebra, and by the Leibniz rule

$$\partial_X(AB) = (\partial_X A)(1 \otimes B) + (A \otimes 1)(\partial_X B).$$

For a polynomial P in one variable, this is the divided difference

$$(\partial_X P)(s, t) = \frac{P(s) - P(t)}{s - t},$$

with value $P'(s)$ on the diagonal. The first conjugate variable $J_1^{[X]} \in L^2(\mu_X)$ is characterized, for every polynomial P, by

$$\begin{aligned} \tau(J_1^{[X]} P(X)) &= (\tau \otimes \tau)(\partial_X P(X)) \\ &= \iint \frac{P(s) - P(t)}{s - t} \, \mu_X(\mathrm{d}s) \mu_X(\mathrm{d}t). \end{aligned}$$

It is the free version of the negative classical score.

Theorem 8.4 (Score of a free sum; Book Lemma 8.5). If X and Y are free and have conjugate variables in L^2, the conjugate variable of the sum is the conditional expectation of either summand's conjugate variable:

$$J_1^{[X+Y]} = \mathbb{E}\left[J_1^{[X]} \mid X + Y\right] = \mathbb{E}\left[J_1^{[Y]} \mid X + Y\right].$$

This is the exact free counterpart of the classical score identity $\rho_{X+Y} = \mathbb{E}[\rho_X \mid X + Y]$.

For a density $p \in L^3(\mathbb{R})$, the conjugate variable can also be expressed through the Hilbert transform

$$(Hp)(x) = \frac{1}{\pi} \lim_{\varepsilon \downarrow 0} \Re G_p(x + i\varepsilon), \qquad J_1^{[X]} = 2\pi H p.$$

8.4 Free Fisher information and free entropy

Johnson uses the normalization

$$\Phi(X) = \int_{\mathbb{R}} p(x)^3 \, \mathrm{d}x = 3 \int_{\mathbb{R}} p(x)(Hp(x))^2 \, \mathrm{d}x.$$

The Hilbert transform is affine for a Wigner law, exactly as the classical score is affine for a Gaussian.

Theorem 8.5 (Free Cramér–Rao bound; Book Lemma 8.7). If X has mean zero, variance σ^2, and density p, then

$$\Phi(X) \geq \frac{3}{4\pi^2\sigma^2}.$$

Equality holds precisely when

$$(Hp)(x) = \frac{x}{2\pi\sigma^2}$$

on the support of p, which characterizes the matching Wigner law.

The free entropy functional used in the book is the logarithmic energy

$$\Sigma(X) = \iint_{\mathbb{R}^2} p(x)p(y) \log|x-y| \, \mathrm{d}x \, \mathrm{d}y.$$

Among laws with fixed variance, the Wigner law maximizes Σ. Consequently, the entropy deficit from the matching Wigner law is the free analogue of Gaussian relative entropy.

Let X be compactly supported, and let W_t be a free Wigner variable of variance t, free from X. With the normalization above, the free de Bruijn identity is

$$\Sigma(X + W_T) - \Sigma(X) = \frac{2\pi^2}{3} \int_0^T \Phi(X + W_t) \, \mathrm{d}t.$$

For a Wigner law of variance $1+t$, $\Phi = 3/(4\pi^2(1+t))$, so the right side equals $\frac{1}{2}\log(1+T)$.

8.5 Projection inequality and information-theoretic free CLT

For a polynomial g, define the averaged free derivative by

$$(\nabla_Y g)(y) = \int \frac{g(y) - g(z)}{y - z} \, \mu_Y(\mathrm{d}z).$$

Equivalently, $\nabla_Y g(Y) = (1 \otimes \tau_Z)(\partial g(Y, Z))$, where Z has the same law as Y in the second tensor factor. A variable Y has free Poincaré constant R if

$$\mathrm{Var}(g(Y)) \leq R\,\mathbb{E}\Big[(\nabla_Y g(Y))^2\Big]$$

for every polynomial g, and for limits obtained by convergence of both g and $\nabla_Y g$ in $L^2(\mu_Y)$. The Wigner law of variance σ^2 has free Poincaré constant σ^2.

The proof then repeats the classical geometry of Chapter 2. A function of $X + Y$ is projected onto functions of X and of Y; Cauchy–Schwarz bounds the projection error; the Poincaré inequality converts derivative control into variance control; and equality forces the conjugate variable to be linear.

Theorem 8.6 (Free Fisher contraction; Book Theorem 8.3). Let X and Y be free, identically distributed, centered variables with variance $\sigma^2 > 0$, finite free Fisher information Φ, and finite free Poincaré constant R. With the standardized deficit

$$\Phi_{\mathrm{st}}(X) = 4\sigma^2\pi^2\Phi(X) - 3,$$

the projection inequality gives

$$\Phi_{\mathrm{st}}\left(\frac{X+Y}{\sqrt{2}}\right) \le \Phi_{\mathrm{st}}(X)\left(1 - \frac{1}{1 + 8\pi^2 R\Phi(X)/3}\right).$$

Thus the deficit contracts unless it is already zero.

The constant follows from the normalization of the conjugate variable:

$$\mathbb{E}[(J_1^{[X]})^2] = 4\pi^2 \int p(x)(Hp(x))^2 \,\mathrm{d}x = \frac{4\pi^2}{3}\Phi(X).$$

This is the factor that enters Cauchy–Schwarz in the projection proof.

For centered free IID variables with variance $\sigma^2 > 0$ and finite Φ and R, the chapter concludes (Book equations (8.103)–(8.104)) that

$$\Phi_{\mathrm{st}}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) \longrightarrow 0.$$

The corresponding entropy deficit satisfies

$$\Sigma(W(0,\sigma^2)) - \Sigma\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) \longrightarrow 0.$$

9 Appendices A–E

9.1 Appendix A: calculating entropies

9.1.1 Gamma laws

For a Gamma law $U_m \sim \Gamma(m, \theta)$ with shape $m > 0$ and rate $\theta > 0$, the density is

$$f_m(x) = \frac{\theta^m}{\Gamma(m)} x^{m-1} e^{-\theta x}, \qquad x > 0.$$

Its entropy is

$$H(U_m) = m - \log\theta - (m-1)\psi(m) + \log\Gamma(m),$$

where $\psi = \Gamma'/\Gamma$ is the digamma function. The calculation uses

$$\mathbb{E}U_m = \frac{m}{\theta}, \qquad \mathbb{E}\log U_m = \psi(m) - \log\theta.$$

The second identity follows by differentiating the normalizing integral with respect to m.

Stirling and digamma expansions give, for the Gaussian with the same mean and variance,

$$D\left(U_m \middle\| \mathcal{N}\left(\frac{m}{\theta}, \frac{m}{\theta^2}\right)\right) = \frac{1}{3m} + O(m^{-2}).$$

9.1.2 Stable scaling

If q_c is a stable family of index α, then scaling gives

$$H(q_c) = H(q_1) + \frac{1}{\alpha}\log c.$$

The Cauchy and Lévy entropies are given in (5.12).

For densities f, g, the cross-entropy is

$$A_g(f) = -\int f(x)\log g(x)\ \mathrm{d}x.$$

If F is the distribution function of f, $\rho_g = g'/g$, and $0 < g(t) < \infty$, integration by parts gives

$$A_g(f) = -\log g(t) + \int_{-\infty}^{t} \rho_g(y)F(y)\,\mathrm{d}y - \int_{t}^{\infty} \rho_g(y)(1-F(y))\,\mathrm{d}y,$$

provided the integrals are absolutely convergent. This formula is used in Chapter 5 to compare stable laws.

9.2 Appendix B: Poincaré inequalities

Definition 9.1 (Poincaré constant). For a real random variable Y with finite positive variance, its Poincaré constant is

$$R_Y = \sup_g \frac{\mathrm{Var}(g(Y))}{\mathbb{E}[g'(Y)^2]},$$

where the supremum is over absolutely continuous functions with $\mathrm{Var}(g(Y)) > 0$ and finite denominator. A zero denominator with positive variance gives an infinite ratio.

The Poincaré constant satisfies

1. $R_{aX+b} = a^2 R_X$.
2. If X and Y are independent, then $R_{X+Y} \le R_X + R_Y$.
3. $R_X \ge \mathrm{Var}(X)$, with equality exactly for the Gaussian.
4. If $R_X < \infty$, then
$$\mathbb{E}\exp\left(\frac{|X - \mathbb{E}X|}{12\sqrt{R_X}}\right) \le 2,$$
so X has moments of every order.

For a sequence of centered variables Y_n with variance one, $R_{Y_n} \to 1$ implies $Y_n \Rightarrow \mathcal{N}(0,1)$ (Book Lemma B.1). Expectations also converge for continuous functions whose growth is bounded by $C\exp(c|x|)$, with c sufficiently small.

Theorem 9.2 (Convergence of Poincaré constants; Book Theorem B.1). Let $X_1, X_2, \dots$ be IID with mean zero, variance one, finite Fisher information $J(X_1)$, and finite Poincaré constant R_{X_1}. Put

$$U_n = \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

There is a constant C, depending on $J(X_1)$ and R_{X_1}, such that

$$0 \le R_{U_n} - 1 \le \frac{C}{n}.$$

The inequality is also a statement about a spectral gap. If $Dg = g'$ and D^* is its adjoint in $L^2(\mu)$, then the smallest nonzero spectral value of D^*D controls the best Poincaré constant. For a density p with score $\rho = p'/p$,

$$D^* f = -f' - \rho f.$$

Weighted derivatives $Dg = mg'$ are useful for other stable laws. For smooth functions satisfying the integration by parts boundary conditions, their adjoint is

$$D^* f = -mf' - (m' + m\rho)f,$$

where ρ is the score of the underlying density. Weights can be chosen from differentiation identities for orthogonal polynomials.

9.3 Appendix C: de Bruijn's identity

Let U have mean zero, variance one, and density f. Let $Z_t \sim \mathcal{N}(0,t)$ be independent, and let ϕ be the standard normal density. The integral form of de Bruijn's identity is

$$\mathrm{D}(f\|\phi) = \frac{1}{2}\int_0^\infty \left[J(U+Z_t) - \frac{1}{1+t}\right]\,\mathrm{d}t = \frac{1}{2}\int_0^\infty \frac{J_{\mathrm{st}}(U+Z_t)}{1+t}\,\mathrm{d}t. \tag{9.1}$$

The differential identity behind it is

$$\frac{\mathrm{d}}{\mathrm{d}t}H(U+Z_t) = \frac{1}{2}J(U+Z_t).$$

Its proof uses the heat equation. Gaussian convolution produces a density f_t satisfying

$$\partial_t f_t = \frac{1}{2}\partial_{xx} f_t,$$

and integration by parts converts the entropy derivative into Fisher information.

For random vectors, Fisher information is a matrix. The smoothed density satisfies the scalar heat equation

$$\partial_t f_t = \frac{1}{2}\operatorname{tr}(C\nabla^2 f_t).$$

If X is centered with a density and positive definite covariance B, and $Z_{Ct} \sim \mathcal{N}(0,tC)$ is independent with C positive definite, the integral identity for the matching Gaussian is

$$D(X\|\mathcal{N}(0,B)) = \frac{1}{2}\int_0^\infty \operatorname{tr}\Big(C\left[J(X+Z_{Ct}) - (B+tC)^{-1}\right]\Big)\,\mathrm{d}t.$$

This is the analytic input for the multivariate result in Chapter 3.

9.4 Appendix D: entropy power inequality

For an n-dimensional random vector with differential entropy $H(X)$, define entropy power by

$$N(X) = \frac{1}{2\pi e}\exp\left(\frac{2}{n}H(X)\right).$$

Theorem 9.3 (Entropy power inequality; Book Theorem D.1). For independent random vectors X and Y with densities and finite differential entropies,

$$N(X+Y) \geq N(X) + N(Y),$$

with equality exactly for Gaussian vectors with proportional covariance matrices.

The one-dimensional proof combines two earlier ingredients:

- de Bruijn's identity, which differentiates entropy power after Gaussian smoothing;
- Stam's Fisher information inequality,

$$\frac{1}{J(X+Y)} \geq \frac{1}{J(X)} + \frac{1}{J(Y)}.$$

Equality in Stam's inequality requires affine scores and hence Gaussian laws.

Costa's strengthening says that $t \mapsto N(X+Z_t)$ is concave. Differentiating twice gives the corresponding differential inequality for Fisher information.

9.5 Appendix E: comparing modes of convergence

For centered laws with the same variance as their Gaussian target,

$$\begin{aligned}\text{convergence of Fisher information} &\Longrightarrow \text{convergence in relative entropy}\\ &\Longrightarrow \text{convergence in } L^1 \text{ and total variation}\\ &\Longrightarrow \text{weak convergence.}\end{aligned}$$

The first implication follows from the entropy power estimate

$$\mathrm{D}(X\|Z) \le \frac{1}{2}\log(1 + J_{\mathrm{st}}(X)) \le \frac{1}{2}J_{\mathrm{st}}(X)$$

in natural logarithms. The second is Pinsker's inequality.

Barron's examples distinguish convergence in relative entropy from weak and uniform convergence (Book Examples E.1–E.2). There is a density with finite variance for which every normalized IID sum has infinite relative entropy to the Gaussian. The density is sufficiently concentrated to have differential entropy $-\infty$. Another example has relative entropy tending to zero while the convolution densities remain unbounded, so uniform convergence fails.

Theorem 9.4 (Uniform convergence of convolution densities; Book Theorem E.1). Let $X_1, X_2, \dots$ be IID with a density, mean μ, and finite variance $\sigma^2 > 0$. Let g_n be the density of

$$U_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}},$$

and let ϕ be the standard Gaussian density. If

$$\sup_x |g_n(x) - \phi(x)| \longrightarrow 0,$$

then $D(g_n\|\phi) \to 0$.

Proof. Write $\delta_n = \sup_x |g_n(x) - \phi(x)|$ and choose $b_n \to \infty$ slowly enough that $\delta_n \exp(b_n^2/2) \to 0$. On $[-b_n, b_n]$, the ratio g_n/ϕ tends uniformly to one, which controls the entropy integral there. For large n, g_n is uniformly bounded, and

$$\log \frac{g_n(x)}{\phi(x)} \le C + \frac{x^2}{2}.$$

The central limit theorem and $\mathbb{E}U_n^2 = 1$ imply uniform integrability of U_n^2. This controls the positive part of the entropy integral outside $[-b_n, b_n]$; its negative part is bounded by ϕ/e. Both tail contributions therefore tend to zero. ⊛

A local limit theorem (Book Theorem E.2) gives uniform convergence of g_n to ϕ if $g_m \in L^r(\mathbb{R})$ for some m and $1 < r \le 2$. For arbitrary sequences of densities, uniform convergence to the Gaussian can coexist with a nonzero limiting relative entropy when their second moments are not uniformly integrable (Book Example E.3).

Theorem 9.5 (Entropy convergence on the integer lattice; Book Theorem E.3). Let $X_1, X_2, \dots$ be IID integer-valued variables with mean μ, finite variance $\sigma^2 > 0$, and lattice span one. Put $S_n = \sum_{i=1}^n X_i$ and $p_n(k) = \mathbb{P}(S_n = k)$. Define

$$a_n(k) = \frac{1}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{(k - n\mu)^2}{2n\sigma^2}\right), \qquad Z_n = \sum_{k\in\mathbb{Z}} a_n(k),$$

and the probability mass function $q_n(k) = a_n(k)/Z_n$. Then

$$Z_n \longrightarrow 1, \qquad D(p_n\|q_n) \longrightarrow 0.$$

Lattice span one means that the support is not contained in a single congruence class modulo an integer $h > 1$. The lattice local limit theorem controls the central probabilities, and finite variance controls the entropy tails as in the preceding proof. The book first proves $\sum_k p_n(k) \log(p_n(k)/a_n(k)) \to 0$. Normalization adds $\log Z_n$, which also tends to zero.

Lemma 9.6 (Fisher information and Hellinger distance; Book Lemma E.2). Let h be a density with finite Poincaré constant R, and let f be a density with $f \ll h$. Assume that $g = \sqrt{f/h}$ is absolutely continuous and $\int h(g')^2 < \infty$. Define

$$d_H(f,h)^2 = \int (\sqrt{f} - \sqrt{h})^2 \,\mathrm{d}x, \qquad J(f\|h) = \int f(\rho_f - \rho_h)^2 \,\mathrm{d}x.$$

Then

$$\|f - h\|_1 \le 2d_H(f,h) \le \sqrt{2RJ(f\|h)}.$$

Proof. Set $g = \sqrt{f/h}$ and $a = \int \sqrt{fh}$. Since $\int hg^2 = 1$, the Poincaré inequality gives

$$1 - a^2 \le R \int h(g')^2 = \frac{R}{4} J(f\|h).$$

Now $d_H(f,h)^2 = 2(1-a) \le 2(1-a^2)$. Cauchy–Schwarz applied to $(\sqrt{f} - \sqrt{h})(\sqrt{f} + \sqrt{h})$ gives the L^1 bound. ⊛