--	--

Shannon Entropy and the Central Limit Theorem

A thesis presented

by

Miller Smith Puckette

to

The Department of Mathematics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Mathematics

Harvard University

Cambridge, Massachusetts

January, 1986

The fact that a Gaussian random variable has maximum entropy leads to two proofs of the Central Limit Theorem. First the theorem is proved as a result of the maximum entropy property, the convolution inequality for entropy powers, and the weak compactness of a closed tight family of measures. An effective form of the CLT follows from a more analytic approach.

To Pamela.

Contents.

Note to the reader.

As an experiment in mathematical exposition, this paper is written in a block structure like that of a structured computer language. Whenever a statement requires a proof of more than a sentence, the proof is indented and put between two squares (□). As in any mathematical exposition, it is usually clearest to state the result before the proof, but occasionally as the last statement in its own proof. This structured indenting allows the writer to imbed lemmas right where they are needed, without obscuring the flow of logic of the discussion containing them. Also, a reader uninterested in the most tedious details can skip the blocks which contain them without losing the main argument.

1. Introduction.

The Central Limit Theorem derives a remarkably specific conclusion from what appear to be very weak conditions. All we require is an infinite sequence of independent finite-variance random variables satisfying a negligibility condition on their tail content; we then know that the distributions of their partial sums (rescaled to have the same mean and variance) converge uniformly to a normal distribution. Such a rich return for such a scanty assumption suggests that we look for an irresistible force which carries distributions of larger and larger sums of independent random variables toward the normal distribution.

If the normal distribution is the equilibrium at which the partial sum distributions eventually come to rest, no matter what the early distributions may have been, we are drawn to look for an analogy from thermodynamics to explain it. Such an analogy would need a notion of entropy and a second law of thermodynamics to push entropy upward in further and further partial sums. The Gaussian distribution would have the highest entropy possible, and the CLT would amount to the eventual arrival at equilibrium at the state of maximal entropy.

A suitable notion of entropy was defined in [Shannon]. If a random variable X has density $f(x)$, its entropy is given by

$$\mathbf{H}(X) = \int_{-\infty}^{\infty} -f(x)\log f(x)\,dx\ ,$$

taking the integrand to be zero whenever $f(x)=0$. Shannon showed that for a given mean and standard deviation, normal random variables have the greatest entropy.

Entropy is related to the quantity,

$$J(X) = \int_{-\infty}^{\infty} \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\,dx\ .$$

This quantity achieves its minimum for normal random variables; in this case $J(X) = \dfrac{1}{\mathbf{var}(X)}$.

The relationship between J and $\mathbf{H}$ is as follows. Let A_t be an independent normal random variable with variance t. Then,

$$J(X+A_t) = 2\frac{d}{dt}\,\mathbf{H}(X+A_t)\ .$$

For small values of t this gives $J(X) \approx \mathbf{H}(X+A_t) - \mathbf{H}(X)$.

We can now describe the information-theoretic proofs of the CLT which appear in the literature. Throughout this description, we will tacitly assume that all random variables mentioned have smooth densities. This is achieved in fact through Lyapunov's method of adding small independent normal random variables to the random variables in the hypothesis of the CLT. We prove the CLT for the new, smoothed variables and later let the variances of the normal additions approach zero to prove the result for our original random variables.

The first proof is given in [Linnick], who requires only the Lindeberg condition. Let $X_1, X_2, \cdots$ be independent random variables with mean 0 and finite standard deviations, and let $S_n = \sum_1^n X_n$ be their partial sums. Linnick shows that

$$\mathbf{H}\left(\frac{S_{n+1}}{\sigma(S_{n+1})}\right) - \mathbf{H}\left(\frac{S_n}{\sigma(S_n)}\right) = \frac{\sigma^2(X_{n+1})}{2\sigma^2(S_n)} \left[J\left(\frac{S_n}{\sigma(S_n)}\right) - 1 \right] + \varepsilon ,$$

where $|\varepsilon|$ is related to the tail content,

$$\int_{\{|t| > \tau\sigma(S_n)\}} t^2 d\mu ,$$

where μ is the probability measure of X_{n+1} and τ is small. The Lindeberg condition on $\{X_n\}$ allows ε to be controlled; and we know that the term $J(S_n) - \frac{1}{\mathbf{var}(S_n)}$ is nonnegative and zero only when S_n is normal. But the $\mathbf{H}$ terms are bounded above by the maximum entropy property of normal random variables. Hence the J term must approach zero as $n \to \infty$. This implies that the distribution of S_n approaches a normal distribution.

In the case that the X_n are identically distributed, [Brown] provides a proof which relies on a convolution inequality for J. If X_1, X_2 are i.i.d., then we have,

$$J\left(\frac{X_1+X_2}{\sigma(X_1+X_2)}\right) \le J\left(\frac{X_1}{\sigma(X_1)}\right) .$$

This is a special case of equation (3.9). Letting $Z_n = S_{2^n}$, and

$$d_n = J\left(\frac{Z_{n+1}}{\sigma(Z_{n+1})}\right) - J\left(\frac{Z_n}{\sigma(Z_n)}\right) ,$$

we have $d_n \le 0$, hence $d_n \to 0$ since J is bounded below for random variables with variance 1. Brown shows that $d_n \to 0$ implies that the distribution of Z_n approaches a normal distribution. In [Baron], this argument is made more direct, in that it replaces the quantity J with $\mathbf{H}$. He does this by integrating Brown's relation in J above to obtain a similar one for $\mathbf{H}$.

In this work we will be concerned with arguments which prove directly that the Shannon entropy of S_n approaches its extremal value, that of a normal random variable. The idea of using entropy inequalities and a steady-state maximal entropy property to prove limit theorems for stochastic processes is not new; it has been carried out in the case of the ergodic theorem for finite-state Markov chains in [Renyi] and for random walks on a compact topological group in [Csiszár].

We sketch Renyi's proof as follows. Let random variables $X_1, X_2, \cdots$ be the outcomes of a finite-state Markov chain in its steady state; i.e. let the X_n all have the same probability distribution $p_i = \text{Prob}\{X_n = i\}$, and let $p_{ij}^{(n)}$ denote $\text{Prob}\{X_n = j \mid X_1 = i\}$. Let $I(X;Y)$ denote the mutual information between two discrete random variables X and Y defined in [Shannon]. An application of Jensen's inequality for convex functions, in this case the function $f(x) = x \log x$, gives that $I(X_{n+1};X_1) \le I(X_n;X_1)$. Since Shannon information is never negative, $I(X_n;X_1)$ must approach a limit. By the compactness of a finite-deminsional unit cube, there must be a subsequence of $p_{ij}^{(n)}$ (considered as a sequence in n) which converges, say to q_{ij}. Let q_{ij}' be q_{ij} propagated by one step of the Markov chain, and let Q, Q' be random variables obtained by applying the transition matrices q and q' to the random variable X_1; then we have that $I(X_1;Q) = I(X_1;Q')$ since $I(X_1;X_n)$ approaches a limit. But then by the equality condition of the same application of Jensen's theorem, Renyi finds that q_{ij} is independent of i; i.e. that outcomes of the X_n are eventually independent of the first state, and approach the steady state probabilities regardless of its outcome.

[Csiszár] uses the Shannon entropy $\mathbf{H}$ of a random variable on a compact topological group G to prove that the probability measure of the outcome of the nth step of a nondegenerate random walk on G approaches the Haar measure as $n \to \infty$. He starts with a probability density f, and considers the densities $f^{(n)}$ given by convoluting f by itself n times. Jensen's theorem gives that $\mathbf{H}(f^{(n)})$ is an increasing sequence; but since the set of probability densities on G is weak compact, some subsequence of $\{f^{(n)}\}$ must converge, say to g. It follows that $\mathbf{H}(f * g) = \mathbf{H}(g)$, which by the equality condition of Jensen's inequality implies that g is the uniform probability density, and that $\mathbf{H}(f^{(n)}) \to \mathbf{H}(g)$ then implies that $f^{(n)} \to g$. Both Renyi and Csiszár mention that many measures of entropy exist which would work in their proofs.

We will present two proofs of the Central Limit Theorem assuming the Lindeberg condition on $\{X_n\}$. The first relies mainly on the maximum entropy property of **H** and its convolution inequality. If we define the entropy power of X as

$$\Pi(f) = \frac{e^{2\mathbf{H}(f)}}{2\pi e} ,$$

then the maximum entropy property for normal random variables becomes, $\Pi(X) \le \mathbf{var}(X)$, with equality only when X is normal. Now we need to produce an upward force on the entropy of convolutions of random variables. This is given by the convolution inequality for entropy power, which states that

$$\Pi(X+Y) \ge \Pi(X) + \Pi(Y) ,$$

whenever X, Y are independent continuous random variables. Equality holds only when X and Y are both normal. This inequality is stated in [Shannon], but first proved correctly in [Stam].

In chapter 4 we will define the entropy defect of a random variable X with positive variance as $1-\dfrac{\Pi(Y)}{\mathbf{var}(Y)}$, where Y is the sum of X and an independent normal random variable whose variance is that of X. Entropy defect is positive unless X is normal, and the convolution inequality implies that $D(X+Y) \le \max(D(X), D(Y))$ for independent random variables X and Y, with equality only when $D(X)=D(Y)=0$, i.e. X and Y are both normal.

Let $X_1, X_2, \cdots, X_n$ be independent random variables with zero mean. Suppose that their sum S has variance 1. We are interested in showing that if the X_i have little tail content (in a sense we will develop in chapter 2), then S is nearly normal, in the sense that it has a small entropy defect. Suppose we split the X_i into two sequences $X_1, \cdots, X_m$, and $X_{m+1}, \cdots, X_n$, with sums S_1 and S_2. Then since $S = S_1 + S_2$, we have $D(S) \le \max(D(S_1), D(S_2))$. Hence S is at least as close to normal as one of S_1, S_2, and strictly closer if it is not already normal.

To get a CLT out of this, we will need to define a measure of the tail content of the X_i such that the two subsequences, when rescaled to have total variances of one each, will have tail measures which are possibly higher, but controllably so. Then as the total tail measure goes to zero the sequence becomes more and more deeply decomposable in the above way, with each recomposition pushing the entropy defect down. That it limits toward zero will follow from the following theorem.

1.1. Theorem. Let G be a metric space with a continuous binary operation $*$, let $D,V\colon G\to[0,\infty)$, not necessarily continuous, with D bounded, and let $G\supseteq\Sigma\supseteq\Sigma_1\supseteq\Sigma_2\supseteq\cdots$, where Σ is compact. Suppose that $V(g)=1$ for all $g\in\Sigma_1$, and that there is a unique $z\in G$ such that $V(z)=1$, $D(z)=0$. Suppose (a) that for all $\{s_i\in\Sigma_i\}$ with $s_i\to s$ we have $V(s)=1$, and that for any $\{t_i\in\Sigma\}$ with $t_i\to t$ we have $V(t)\le\limsup V(t_i)$, and if $V(t_i)\to V(t)\neq 0$ we have $D(t_i)\to D(t)$; (b) that $V(s*t)=V(s)+V(t)$ and $D(s*t)\le\max(D(s),D(t))$, with equality only if at least one of $D(s*t)$, $V(s)$, $V(t)$ is zero; and (c) that if $i>1$ and $s\in\Sigma_i$, there exist $t,u\in\Sigma$ and $t',u'\in\Sigma_{i-1}$ such that $V(t),V(u)\ge\frac{1}{3}$, $s=t*u$, $D(t)=D(t')$, and $D(u)=D(u')$. Then all sequences $\{s_i\in\Sigma_i\}$ converge to z.

To apply this to the CLT, we will let G be the set of probability measures on the real line; Σ, the set of those having mean 0 and variance ≤ 1, and V, D, and z be variance, entropy defect, and the standard normal measure, respectively. Since V, D are not continuous we introduce the sets Σ_i and the partial continuity condition (a) of the theorem. The choice of Σ_i will amount to the Lindeberg condition. Roughly speaking, the Σ_i will be defined to consist of measures of variance 1 obtained by the convolution of many (about 27^i) measures whose total tail content is small. The proof uses the fact that D is decreased by $*$ to push its supremum over Σ_i toward its minimum value of zero, thereby pushing the elements of Σ_i toward z.

This theorem assumes surprisingly little about V, D, and z; it requires no knowledge about how nearly the composition inequality for D may approach equality for $s\in\Sigma$ other than z, nor about how near D can be to zero for $s\neq z$ and $V(s)=1$. The proof invokes the compactness of Σ to solve these problems.

□ Proof of (1.1). Let

– $$\lambda_i=\sup\{D(g):g\in\Sigma_i\},$$

which is defined because D is bounded. Since the λ_i are nonincreasing and nonnegative, they approach a limit λ. We can then pick $\{s_i\in\Sigma_i\}$ so that $D(s_i)\to\lambda$. We choose $\{t_i,u_i\in\Sigma,\ t'_i,u'_i\in\Sigma_i\}$ as in (c) such that $t_i*u_i=s_{i+1}$. Since Σ is compact so is Σ^3, and so $\{(s_{i+1},t_i,u_i)\}$ has a convergent subsequence. We may replace $\{(s_{i+1},t_i,u_i)\}$ by a further subsequence which converges to some (s,t,u) and so that $\{V(t_i)\}$ has a limit α, and replace t'_i,u'_i accordingly. Note that since $\Sigma_1\supseteq\Sigma_2\supseteq\cdots$ we still have $s_i,t_i,u_i\in\Sigma_i$ after taking subsequences. Since $*$ is continuous we have $s=t*u$. Also, $V(t)\le\limsup V(t_i)=\alpha$; similarly $V(u)\le 1-\alpha$, and by additivity of V, $V(t)+V(u)=1$, so $V(t)=\alpha$. By assumption (a), $D(s_i)\to D(s)$, $D(t_i)\to D(t)$, $D(u_i)\to D(u)$. Since $V(t_i),V(u_i)\geq\frac{1}{3}$, we have $V(t),V(u)\geq\frac{1}{3}$, and by assumption (b),

$$\lambda = D(s) \leq \max(D(t),D(u)) ,$$

but since $D(t_i)=D(t'_i)\leq\lambda_i$ and $D(t_i)\to D(t)$, we have $D(t)\leq\lambda$ and similarly $D(u)\leq\lambda$. It follows that $D(s)=D(t)=D(u)=\lambda$. By hypothesis (b) $D(s)=0$ and hence $\lambda=0$.

Now we choose any sequence $\{s_i\in\Sigma_i\}$. Suppose s_i does not limit to z. Then we may replace s_i by a subsequence which avoids z. Since Σ is compact $\{s_i\}$ has a convergent subsequence; suppose it converges to η. Then $D(\eta)=0$, $V(\eta)=1$, contradicting the uniqueness of z. We conclude that $s_i\to z$, proving the theorem. □

In the next chapter we introduce the Lindeberg form of the CLT, putting the tail condition into a single-parameter form. Chapter 3 proves the properties of entropy hypothesized in (b) and chapter 4 introduces appropriate sets Σ_i for conditions (a) and (c) to hold, proving the Lindeberg form of the CLT.

Our second proof of the CLT, which starts in chapter 5, establishes a rate of convergence result. It is similar to Linnick's proof in that we consider the effect of each X_i on the difference $\mathbf{H}(S_{i+1}+A)-\mathbf{H}(S_i+A)$, where A is a normal random variable independent of any of the X_i. Roughly, we show that

$$\mathbf{H}(S_{i+1}+A) \approx \mathbf{H}(S_i+N_{i+1}+A) ,$$

where N_{i+1} is a normal random variable with the same mean and variance as X_{i+1}. By telescoping, we will arrive at theorem 5.28, which states that if S_n has variance one, then

$$\mathbf{H}(N+A) - \mathbf{H}(S_n+A) \leq \sum_{i=1}^{n} \frac{K}{(\sigma(A))^5}\ \mathbf{E}(\tau(X_i)) ,$$

where N is normal with variance 1 and $\tau(x)=\min(x^2,|x|^3)$. Note that $\tau(X_i)$ always has a finite expectation since X_i^2 does. This theorem functions as an effective form of the CLT since the left hand side of the inequality is nonnegative and zero only if S_n is normal. In chapter 6 we will relate the error estimate $e_1=\mathbf{H}(N+A)-\mathbf{H}(S_n+A)$ to the usual error estimate, $e_2=\sup|\mathrm{Prob}\{S_n\leq t\}-\mathrm{Prob}\{N\leq t\}|$. Since the two measures are quite different, we can only prove weak bounds on the latter one through bounds on the former; the best we will find is equation (6.9), which establishes that e_2 is roughly on the order of $[\sum_{i=1}^{n}\mathbf{E}(\tau(X_i))]^{\frac{1}{8}}$.

To give an estimate of the strength of these bounds, we compare them to the Berry-Esséin bounds in the case that the X_i are identically distributed. If X is a random variable with variance 1 with $\beta = E(|X|^3) < \infty$, and if the X_i above are distributed as $n^{-1/2}X$, then the Berry-Esséin theorem gives $e_2 \le \frac{K_2\beta}{\sqrt{n}}$, where K_2 may be taken to be 2. In comparison, we have $\tau(x) \le |x|^3$, $\mathbf{E}(\tau(X_i)) \le \beta n^{-\frac{3}{2}}$, so theorem 5.28 gives $e_1 \le \frac{K(\sigma(A))\beta}{n}$, but translates to a much weaker bound on e_2. On the other hand, the Berry-Esséin bound does not appear to imply the bound we give on entropy, although stronger results in the literature certainly would.

2. Lindeberg's form of the CLT.

Let $\{X_{ij},\ i \ge 1,\ 1 \le j \le k_i\}$ be a doubly indexed array of RVs, with row sums $S_i = \sum_{j=1}^{k_i} X_{ij}$ and probability measures $\mu_{ij} = \mu_{X_{ij}}$, $\mu_i = \mu_{S_i}$, satisfying the following conditions:

(2.1a) The rows, $\{X_{ij}, j=1,...,k_i\}$, are independent;

(2.1b) Each X_{ij} has mean zero and finite variance;

(2.1c) The row sums S_i have variance one.

(2.1d) $\max_{1 \le j \le k_i} \mathbf{var}(X_{ij})$ approaches zero as $i \to \infty$;

Henceforward we will let **n**, **N**, **ν** denote the standard normal density, probability distribution, and measure, respectively. Lindeberg's form of the CLT is,

2.2. Proposition. Let $(\{X_{ij}\},\{k_i\},\{S_i\},\{\mu_i\},\{\mu_{ij}\})$ satisfy (2.1a-d). If

$$\lim_{i\to\infty} \sum_{j=1}^{k_i} \int_{[-\beta,\beta]^c} x^2 d\mu_{ij}(x) = 0, \quad \text{for all } \beta > 0\ , \tag{2.3}$$

then

$$\lim_{i\to\infty} \sup\{\,|\mathrm{Prob}\{S_i \le t\} - \mathbf{N}(t)|\ :\ t \in \mathbf{R}\} = 0\ . \tag{2.4}$$

Feller has shown a converse, which is restated in [Araujo] to give that (2.1a-c) and (2.4) imply (2.1d) and (2.3). We now show that the following condition is equivalent to (2.3):

(2.5) $$\lim_{i\to\infty}\sum_{j=1}^{k_i}\mathbf{E}(\tau(X_{ij}))=0\ ,$$

where τ is an even continuous function satisfying,

– $\tau(0)=0$ *iff* $x=0$,

– $\tau(x)=x^2$ if $x\ge 1$,

– $\dfrac{\tau(x)}{x^2}$ is nondecreasing for $0\le x\le 1$,

– $\lim\limits_{x\to 0}\dfrac{\tau(x)}{x^2}=0$.

Examples of functions which satisfy these conditions are $\tau(x)=\min(x^2,|x|^3)$ and $\tau(x)=\min(x^2,x^4)$.

□ We may consider only $\beta<1$ in (2.3). Then for any RV X with finite variance,

– $$\mathbf{E}(\tau(X))=\int\tau(x)d\mu_X(x)$$

– $$\ge\int_{[-\beta,\beta]^c}\tau(x)d\mu_X(x)\ge\frac{\tau(\beta)}{\beta^2}\int_{[-\beta,\beta]^c}x^2d\mu_X(x)\ .$$

Assume (2.5) is true and pick $\varepsilon>0$, $\beta>0$. Choose i large enough that

– $$\sum_{j=1}^{k_i}\mathbf{E}(\tau(X_{ij}))\le\frac{\tau(\beta)\varepsilon}{\beta^2}\ ,$$

using the fact that $\tau(\beta)>0$, giving

– $$\sum_{j=1}^{k_i}\int_{[-\beta,\beta]^c}x^2d\mu_{X_{ij}}(x)\le\varepsilon\ ,$$

and so (2.3) is met. On the other hand,

– $$\int\tau(x)d\mu_X(x)\le\frac{\tau(\beta)}{\beta^2}\int_{[-\beta,\beta]}x^2d\mu_X(x)+\int_{[-\beta,\beta]^c}x^2d\mu_X(x)$$

– $$\le\frac{\tau(\beta)}{\beta^2}\int x^2d\mu_X(x)+\int_{[-\beta,\beta]^c}x^2d\mu_X(x)\ .$$

Now assume (2.3) is true and pick $\varepsilon>0$. Choose $\beta>0$ small enough that $\dfrac{\tau(\beta)}{\beta^2}\le\dfrac{\varepsilon}{2}$ and i large enough that

– $$\sum_{j=1}^{k_i}\int_{[-\beta,\beta]^c}x^2d\mu_{X_{ij}}(x)\le\frac{\varepsilon}{2}\ .$$

Then

– $$\sum_{j=1}^{k_i}\int\tau(x)d\mu_{X_{ij}}(x)\le\sum_{j=1}^{k_i}\frac{\tau(\beta)}{\beta^2}\int x^2d\mu_{X_{ij}}(x)+\sum_{j=1}^{k_i}\int_{[-\beta,\beta]^c}x^2d\mu_{X_{ij}}(x)$$

– $$\le\varepsilon\ ,$$

proving that (2.3) implies (2.5). □

Condition (2.5) also clearly implies (2.1d); thus an equivalent statement of the Lindeberg CLT is,

2.6. Proposition. If ($\{X_{ij}\}$,$\{k_i\}$,$\{S_i\}$,$\{\mu_i\}$,$\{\mu_{ij}\}$) satisfy (2.1a-c) and (2.5), then they satisfy (2.4).

Having a condition equivalent to (2.3) without the parameter β allows us to delete the index i from (2.2); an equivalent proposition is:

2.7. Proposition. Let $\varepsilon>0$ and let τ be as above. Then there exists a $\delta>0$ such that for any sequence of independent zero-mean finite-variance RVs $\{X_j,\ 1\le j\le k\}$, if $S=\sum X_j$ has variance 1 and

(2.8) $$\sum_{j=1}^{k}\mathbf{E}(\tau(X_j)) \le \delta\ ,$$

then

(2.9) $$\sup\{\,|\,\mathrm{Prob}\{S\le t\} - \mathbf{N}(t)\,|\,\} \le \varepsilon\ .$$

The converse of this statement is also equivalent to the converse of the Lindeberg form of the CLT. This form of the CLT suggests that we seek to prove it with an explicit formula for $\delta(\varepsilon)$. We will do this in chapter 6.

3. Shannon Entropy.

In this chapter we define the entropy of a random variable with a continuous density function, and develop the properties of entropy which we will need to apply (1.1) to the CLT. This section consists mostly of restatements of (and occasional variations on) standard results about Shannon entropy. The definition of entropy and its fundamental properties are in [Shannon], and relations (3.1), (3.2), (3.9), (3.10) and (3.16) are from [Blachman], whose discussion this section follows. The distance-related inequality (3.6) is closely related to ones found, for example, in [Kemperman].

We will need to use normal distributions, densities, and measures which do not necessarily have mean 0 and standard deviation 1. We will let $\mathbf{n}_a$ denote the normal density with mean 0 and standard deviation a, and similarly for the distribution function $\mathbf{N}_a$ and the measure $\boldsymbol{\nu}_a$. If X is a RV with finite variance we let $\mathbf{n}_X$ denote the density of a *normal mate* for X: a normal *RV* whose mean and variance match those of X. We define $\mathbf{N}_X$, $\boldsymbol{\nu}_X$ correspondingly.

It will often be convenient to deal with the probability density or measure of a RV rather than with the RV itself. Accordingly, we will often use the functions **E**, **var**, and σ on densities and probability measures, as in $\mathbf{E}(f)$ or $\mathbf{E}(\mu)$, where f is a density and μ a probability measure. We extend the notation for normal mates in the same way; for instance $\mathbf{n}_f$ will denote the normal density which matches a density f in mean and variance.

If X is a continuous random variable with density function f we define the *Shannon Entropy* of X by

$$\mathbf{H}(X) = \mathbf{H}(f) = \int \phi(f(x))dx \ ,$$

where

$$\phi(x) = \begin{cases} x\log(\frac{1}{x}), & x>0 \ , \\ 0, & x=0 \ , \end{cases}$$

whenever the integral converges absolutely, and ∞ or –∞ if the integral diverges in one direction. Note that φ is continuous on $[0,\infty)$, C^∞ on $(0,\infty)$, and concave. We immediately see that **H** is translation-invariant and that if $a>0$ and $g(x)=\frac{f(ax)}{a}$, $\mathbf{H}(g) = \mathbf{H}(f)-\log(a)$. If f and g are densities with finite entropy, we have

$$\mathbf{H}(f*g) = \int \phi(\int f(x-t)g(t)dt)dx \tag{3.1}$$

$$\geq \int \phi(f(x))dx$$

by concavity of φ,

$$= \mathbf{H}(f) \ .$$

We now show that the normal density has the greatest entropy possible for its variance, in fact that

$$\mathbf{H}(f) \leq \mathbf{H}(\mathbf{n}_f) = \log(\sqrt{2\pi e}\,\sigma(f)) \ , \tag{3.2}$$

with equality only when f is normal.

□ Proof. Since entropy is translation invariant, we may assume that $\mathbf{E}(f)=0$. Let $\sigma=\sigma(f)$. Then

$$\mathbf{n}_f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{x^2}{2\sigma^2}} \ ,$$

$$\mathbf{H}(\mathbf{n}_f) = \int \mathbf{n}_f(x)[\frac{x^2}{2\sigma^2} + \log(\sqrt{2\pi}\,\sigma)]dx$$

$$= \log(\sqrt{2\pi e}\,\sigma) \ .$$

We have,

$$\mathbf{H}(f)-\mathbf{H}(\mathbf{n}_f)=\int\left[\phi(f(x))+f(x)[\log\sqrt{2\pi}\sigma+\frac{x^2}{2\sigma^2}]\right]dx\ .$$

When $f(x)>0$, the integrand becomes,

$$f(x)[-\log f(x)+\log\sqrt{2\pi}\sigma+\frac{x^2}{2\sigma^2}]$$

$$=f(x)[\log\mathbf{n}_f(x)-\log f(x)]$$

$$=f(x)\log(\frac{\mathbf{n}_f(x)}{f(x)})$$

$$\leq f(x)[\frac{\mathbf{n}_f(x)}{f(x)}-1]$$

$$=\mathbf{n}_f(x)-f(x)\ .$$

If $f(x)=0$ the relation still holds. Then,

(3.3) $$\mathbf{H}(f)-\mathbf{H}(\mathbf{n}_f)\leq\int_S[\mathbf{n}_f(x)-f(x)]dx=0\ ,$$

with equality only if $f=\mathbf{n}_f$ a.e., i.e. f is normal. □

We will need the following two bounds for functions which need not be densities:

3.4. Proposition. If $f\geq 0$ is measurable, $a=\int f(x)dx>0$ and $b=\int x^2 f(x)dx$, then

$$\int\phi(f(x))dx\leq\frac{a}{2}\log(\frac{2\pi eb}{a^3})\ .$$

□ Proof. Let $f(x)=ag(x)$ so that $\int g(x)dx=1$. Then $\phi(f(x))=a\phi(g(x))+\phi(a)g(x)$, so

$$\int\phi(f(x))dx=-a\log a+a\int\phi(g(x))dx\ .$$

We have,

$$\int x^2 g(x)dx=\frac{b}{a}\ ,$$

so by (3.2),

$$\int\phi(g(x))dx\leq\frac{1}{2}\log(\frac{2\pi eb}{a})\ ,$$

$$\int\phi(f(x))dx\leq\frac{a}{2}\log(\frac{2\pi eb}{a^3})\ .\ \square$$

3.5. Proposition. If $f\geq 0$ is measurable, $1\geq a\geq\int f(x)dx>0$ and $b\geq\int x^2 f(x)dx$, then

$$\int\phi(f(x))dx\leq\frac{a}{2}(b+3-3\log a)\ .$$

☐ Proof. The quantity $\frac{a}{2}(b+3-3\log a)$ is an increasing function of a and b for $1\ge a>0$ and $b>0$, and

$$\frac{a}{2}\log\left(\frac{2\pi e b}{a^3}\right)\le\frac{a}{2}\left(b-1+\log\left(\frac{2\pi e}{a^3}\right)\right)$$

$$\le\frac{a}{2}(b+3-3\log a)\ .$$

This and (3.4) imply (3.5). ☐

We will also want to compare the entropies of two functions which are close together; this is handled by the next proposition.

3.6. Proposition. If f and g are probability densities with $f,g\le 1$, then

$$|\mathbf{H}(f)-\mathbf{H}(g)|\le\int\phi(|f(x)-g(x)|)dx+\int|f(x)-g(x)|dx\ .$$

☐ Proof. If $0\le a\le b\le 1$, then

$$\phi(b)\le\phi(a)+\phi(b-a)$$

immediately. Since $\phi'(x)\ge -1$ for $0<x<1$, we also have

$$\phi(b)\ge\phi(a)-(b-a)\ ,$$

and hence

$$|\phi(b)-\phi(a)|\le\phi(|b-a|)+|b-a|\ .$$

This still holds when $a>b$ by switching a and b. Setting $a=f(x)$, $b=g(x)$ and integrating proves (3.6). ☐

In addition to the maximum entropy property (3.2), we need a lower bound for the entropy of the convolution of two densities. This lower bound is given by the convolution inequality for entropy, which we now develop. This discussion differs from [Blachman] by first proving a linear form of the convolution inequality, which might be original here.

We define the quantity

$$J(f)=\int\frac{f'(x)^2}{f(x)}\,dx\ ,$$

when f is a C^1 probability density. We will prove a convolution inequality for J and then relate J to $\mathbf{H}$ to prove the convolution inequality for $\mathbf{H}$. We will need to differentiate under an integral twice, so we introduce a criterion for doing so here.

3.7. Proposition. If $f(x,t)$ is C^1, $a<b$, and the function

$$\sup\{\,|\frac{\partial f(x,t)}{\partial x}|: a<x<b\,\}$$

is integrable, then

$$\frac{d}{dx}\int f(x,t)dt = \int \frac{\partial}{\partial x} f(x,t)dt\ ,\quad a<x<b\ .$$

□ Proof. We have,

$$\frac{d}{dx}\int f(x,t)dt = \lim_{h\to 0}\int \frac{f(x+h,t)-f(x,t)}{h}dt$$

if the limit exists. Let $g(x,t)=\frac{\partial f(x,t)}{\partial x}$. By the mean value theorem,

$$\frac{f(x+h,t)-f(x,t)}{h} = g(x+z(x,t,h),t)\ ,$$

where $|z(x,t,h)|<|h|$. For sufficiently small h,

$$|g(x+z(x,t,h),t)| \le \sup\{\,|g(x,t)|: a<x<b\,\}\ ,$$

which we assumed to be integrable. But

$$\lim_{h\to 0} g(x+z(x,t,h),t) = g(x,t)\ ,$$

and the bounded convergence theorem implies that

$$\lim_{h\to 0}\int \frac{f(x+h,t)-f(x,t)}{h}dt = \int g(x,t)dt\ .\ \square$$

Suppose that f and g are C^1 and $f'(x), g'(x)$ are bounded and in L^1. Then

$$(f*g)'(x) = \frac{d}{dx}\int f(t)g(x-t)dt\ ,$$

and for any $a<b$,

$$\sup\{\,|f(t)g'(x-t)|: a<x<b\,\}$$

is integrable because f is integrable and g' is bounded, so

$$(f*g)'(x) = \int f(t)g'(x-t)dt = (f*(g'))(t).$$

Similarly, $(f*g)(t)=(f'*g)(t)$. We have,

$$J(f*g) = \int\int \left[\frac{(f*g)'(u)}{(f*g)(u)}\right]^2 (f*g)(u)du$$

$$= \int\int \left[\frac{(f*g)'(x+y)}{(f*g)(x+y)}\right]^2 f(x)dx\ g(y)dy\ ,$$

but also,

$$J(f*g) = \int \left[\frac{(f*g)'(u)}{(f*g)(u)} \right] (f*g)'(u)du$$

$$= \int \left[\frac{(f*g)'(u)}{(f*g)(u)} \right] (g'*f)(u)du$$

$$= \int\int \left[\frac{(f*g)'(x+y)}{(f*g)(x+y)} \right] f(x)dx\, g'(y)dy$$

$$= \int\int \left[\frac{(f*g)'(x+y)}{(f*g)(x+y)} \right] \left[\frac{g'(y)}{g(y)} \right] f(x)dx\, g(y)dy\ ,$$

and similarly,

$$J(f*g) = \int\int \left[\frac{(f*g)'(x+y)}{(f*g)(x+y)} \right] \left[\frac{f'(x)}{f(x)} \right] f(x)dx\, g(y)dy\ .$$

We can now derive the convolution inequalities for J. For any real a, b, we have,

$$\int\int \left[a\frac{f'(x)}{f(x)} + b\frac{g'(y)}{g(y)} - (a+b)\frac{(f*g)'(x+y)}{(f*g)(x+y)} \right]^2 f(x)dx\, g(y)dy$$

$$= \int\int \Bigg\{ a^2 \left[\frac{f'(x)}{f(x)} \right]^2 + b^2 \left[\frac{g'(y)}{g(y)} \right]^2 + (a+b)^2 \left[\frac{(f*g)'(x+y)}{(f*g)(x+y)} \right]^2$$

$$- 2a(a+b) \left[\frac{(f*g)'(x+y)}{(f*g)(x+y)} \right] \left[\frac{f'(x)}{f(x)} \right]$$

$$- 2b(a+b) \left[\frac{(f*g)'(x+y)}{(f*g)(x+y)} \right] \left[\frac{g'(y)}{g(y)} \right]$$

$$- 2ab \left[\frac{f'(x)}{f(x)} \right] \left[\frac{g'(y)}{g(y)} \right] \Bigg\} f(x)dx\, g(y)dy$$

$$= a^2 J(f) + b^2 J(g) - (a+b)^2 J(f*g)\ .$$

(The last term of the expansion is zero by the assumption that $f', g' \in L^1$.) Since we started with the integral of a perfect square, we have proved the linear convolution inequality for J:

(3.8) $\quad a^2 J(f) + b^2 J(g) \geq (a+b)^2 J(f*g)\ .$

If $J(f), J(g) < \infty$, substituting $a = J(f)^{-1}$, $b = J(g)^{-1}$ gives,

$$\frac{1}{J(f)} + \frac{1}{J(g)} \geq J(f*g) \left[\frac{1}{J(f)} + \frac{1}{J(g)} \right]^2 ,$$

or

(3.9) $\quad \dfrac{1}{J(f*g)} \geq \dfrac{1}{J(f)} + \dfrac{1}{J(g)}\ ,$

the nonparametric convolution inequality for J.

We wish to relate J to $\mathbf{H}$ in order to get a convolution inequality for $\mathbf{H}$. We will show that

(3.10) $$\frac{d}{dt}\mathbf{H}(f*\mathbf{n}_t) = \frac{J(f*\mathbf{n}_t)}{2} \quad ,$$

whenever f is a continuous density with bounded support.

□ For $t>0$ we have,

$$\frac{d^2}{dx^2}\,\mathbf{n}_t(x) = 2\frac{d}{dt}\,\mathbf{n}_t(x)\ ,$$

$$\frac{d^2}{dx^2}\,(f*\mathbf{n}_t)(x) = 2\frac{d}{dt}\,(f*\mathbf{n}_t)(x)\ .$$

The function $\phi((f*\mathbf{n}_t)(x))$ is C^∞ and its derivative with respect to t is also C^∞. We have,

$$\frac{d}{dt}\phi((f*\mathbf{n}_t)(x)) = (-\log((f*\mathbf{n}_t)(x)) - 1)\frac{d}{dt}(f*\mathbf{n}_t)(x)$$

$$= ½(-\log((f*\mathbf{n}_t)(x)) - 1)\frac{d^2}{dx^2}(f*\mathbf{n}_t)(x)\ .$$

Since f has bounded support, $f*\mathbf{n}_t''$ goes to zero fast enough to make the whole expression satisfy the condition of proposition 3.7 and we may differentiate under the integral sign as follows:

$$\frac{d}{dt}\,H(f*\mathbf{n}_t) = \int\frac{d}{dt}\phi((f*\mathbf{n}_t)(x))\,dx$$

$$= \int[-\log((f*\mathbf{n}_t)(x)) - 1]\,[\frac{d}{dt}\,(f*\mathbf{n}_t)(x)]\,dx$$

$$= \frac{1}{2}\int[-\log((f*\mathbf{n}_t)(x)) - 1]\,[\frac{d^2}{dx^2}\,(f*\mathbf{n}_t)(x)]\,dx$$

$$= -\frac{1}{2}\,[\log((f*\mathbf{n}_t)(x) - 1]\frac{d}{dx}\,(f*\mathbf{n}_t)(x)\,\Big|_{-\infty}^{\infty} + \frac{1}{2}\int\frac{[\frac{d}{dx}\,(f*\mathbf{n}_t)(x)]^2}{(f*\mathbf{n}_t)(x)}\ dx\ ,$$

by integration by parts. Again, since f has bounded support $f*\mathbf{n}'$ goes to zero fast enough that we may ignore the first term above. The second term is $½J(f*\mathbf{n}_t)$, so we have

$$\frac{d}{dt}\,H(f*\mathbf{n}_t) = \frac{J(f*\mathbf{n}_t)}{2}\ .\ \square$$

By the fundamental theorem of calculus,

$$\mathbf{H}(f*\mathbf{n}_t) - \lim_{s\to 0+}\mathbf{H}(f*\mathbf{n}_s) = \frac{1}{2}\int_0^t J(f*\mathbf{n}_u)du\ .$$

But since f is continuous and has bounded support, $f*\mathbf{n}_s\to f$ uniformly as $s\to 0+$, and hence $\phi(f*\mathbf{n}_s)\to\phi(f)$ uniformly. Since for sufficiently small s the contribution to $\mathbf{H}(f*\mathbf{n}_s)$ is uniformly small (in s) outside the support of f, we have $\mathbf{H}(f*\mathbf{n}_s)\to\mathbf{H}(f)$. We now have, (3.11) $$\mathbf{H}(f*\mathbf{n}_t) - \mathbf{H}(f) = \frac{1}{2}\int_0^t J(f*\mathbf{n}_u)du\ .$$

We now prove the linear convolution law for entropy. Suppose for the moment that f, g are continuous and have bounded support. By (3.11), for any $t>0$, $a>0$,

$$\mathbf{H}(f*\mathbf{n}_{at}) - \mathbf{H}(f) = \frac{a}{2}\int_0^t J(f*\mathbf{n}_{au})du\ .$$

Letting $b>0$ we apply this to f, g and $f*g$ and add, giving,

(3.12) $$(a+b)\mathbf{H}(f*g) - a\mathbf{H}(f) - b\mathbf{H}(g)$$
$$-\ [(a+b)\mathbf{H}(f*g*\mathbf{n}_{(a+b)t}) - a\mathbf{H}(f*\mathbf{n}_{at}) - b\mathbf{H}(g*\mathbf{n}_{bt})]$$
$$= \tfrac{1}{2}\int_0^t \left[a^2 J(f*\mathbf{n}_{au}) + b^2 J(g*\mathbf{n}_{bu}) - (a+b)^2 J(f*g*\mathbf{n}_{(a+b)u})\right] du \geq 0\ ,$$

by (3.8). But since $\mathbf{H}(f*\mathbf{n}_t) \geq \mathbf{H}(\mathbf{n}_t) = \log\sqrt{2\pi e t}$ by convexity of $\mathbf{H}$ and $\mathbf{H}(f*\mathbf{n}_t) \leq \mathbf{H}(\mathbf{n}_f*\mathbf{n}_t) = \log\sqrt{2\pi e(t+\mathbf{var}(f))}$ by the maximum entropy property (3.2), we have,

$$\lim_{t\to\infty}\,[\mathbf{H}(f*\mathbf{n}_t) - \log\sqrt{2\pi e t}\,] = 0\ ,$$

and so letting $t\to\infty$ in (3.12) gives,

(3.13) $$(a+b)[\mathbf{H}(f*g) - \log\sqrt{a+b}\,] \geq a[\mathbf{H}(f) - \log\sqrt{a}\,] + b[\mathbf{H}(g) - \log\sqrt{b}\,]\ ,$$

whenever f and g are continuous and have bounded support. Now we remove the assumption of bounded support for f and g. Let $f_R(x)$ be $f(x)$ restricted to a large interval $[-R,R]$, but chopped off smoothly to preserve the C^0 assumption, and then renormalized to be a probability density. Let g_R be defined correspondingly. We have shown (3.13) for f_R and g_R. Proposition 3.6 immediately implies (after rescaling so that f, f_R are bounded by one) that $\lim_{R\to\infty}\mathbf{H}(f_R) = \mathbf{H}(f)$, proving (3.13) for f and g. (3.13) is the linear form of the convolution inequality for entropy; it might be original here. If we let

$$\overline{\mathbf{H}}(f) = \mathbf{H}(f) - \mathbf{H}(\mathbf{n}_f)\ ,$$

then the maximum entropy property (3.2) becomes,

$$\overline{\mathbf{H}}(f) \leq 0\ ,$$

and setting $a=\mathbf{var}(f)$, $b=\mathbf{var}(g)$ in (3.13) gives,

(3.14) $$\mathbf{var}(f*g)\overline{\mathbf{H}}(f*g) \geq \mathbf{var}(f)\overline{\mathbf{H}}(f) + \mathbf{var}(g)\overline{\mathbf{H}}(g)\ ,$$

whenever f and g are continuous.

To state the classical form of the convolution inequality for entropy, we define the entropy power of f as

$$\Pi(f) = \frac{e^{2\mathbf{H}(f)}}{2\pi e} .$$

Entropy power scales as variance: $\Pi(\frac{1}{a}\, f(\frac{x}{a})) = a^2\Pi(f)$. The maximum entropy property (3.2) becomes,

(3.15) $\Pi(f) \le \mathbf{var}(f)$,

with equality *iff* f is normal. If f and g are densities we let $a=e^{2\mathbf{H}(f)}$, $b=e^{2\mathbf{H}(g)}$ in (3.13), giving

$$(e^{2\mathbf{H}(f)}+e^{2\mathbf{H}(g)})[\mathbf{H}(f*g)-\frac{1}{2}\log(e^{2\mathbf{H}(f)}+e^{2\mathbf{H}(g)})] \ge 0,$$

which implies,

3.16. Proposition. If f and g are continuous probability densities, then

$$\Pi(f*g) \ge \Pi(f) + \Pi(g) ,$$

with equality *iff* both f and g are Gaussian. This is the usual form of the convolution inequality for entropy, as stated in [Shannon] and [Blachman].

4. Topological proof of the CLT.

It is our goal to apply theorem 1.1 to prove the Lindeberg CLT. Let G be the set of probability measures, taken with the weak topology. Let $*$ and V be convolution and variance, respectively, and let D be entropy defect, which we define as $D(\mu)=1-\frac{\Pi(\mu*\mathbf{v}_\mu)}{\mathbf{var}(\mu*\mathbf{v}_\mu)}$, or ∞ if $\mathbf{var}(\mu)=0$. This quantity is unchanged by rescaling μ. By the maximum entropy property (3.15), $D(\mu)\ge 0$ with equality *iff* $\mu*\mathbf{v}_\mu$ is normal, which implies that μ is normal since convolution by a normal is one to one. It is interesting that this fact is the only remaining dependence on Fourier analysis in this proof; it results from the fact that the Fourier transform of ν has no zeros, so multiplication by it is invertible. Result (3.16) gives,

$$D(\mu*\eta) \le \frac{\mathbf{var}(\mu)D(\mu)+\mathbf{var}(\eta)D(\eta)}{\mathbf{var}(\mu)+\mathbf{var}(\eta)} \le \max(D(\mu),D(\eta)) ,$$

and the additivity of variance gives the rest of hypothesis (b) of theorem 1.1. We still need to define Σ and Σ_i satisfying (a) and (c), and show that convergence in G implies the desired conclusion (2.4) of the CLT. We will need the following standard fact:

4.1. Proposition. If f is a continuous nonnegative function, $\{\mu_i\}$, μ are measures with $\mu_i \underset{\text{weak}}{\to} \mu$ and $\int f(x)d\mu_i(x) \le \varepsilon$ for all i, then $\int f(x)d\mu(x) \le \varepsilon$.

Let Σ denote the topological space of probability measures μ such that $\int x^2 d\mu(x) \le 1$, taken with the weak topology. By (4.1) Σ is closed, and by Prohorov's Theorem [Billingsley, p. 35], Σ is compact.

For $i \ge 1$ we define Σ_i to be the set of all measures $\mu = \mu_1 * \cdots * \mu_k$ such that $\mathbf{E}(\mu_i)=0$, $\mathbf{var}(\mu)=1$, and

$$\sum_{j=1}^{k} \int \tau(x) d\mu_j(x) \le 27^{-i} ,$$

where $\tau(x)=\min(x^2, |x|^3)$. We now show that this choice of Σ_i satisfies hypothesis (c).

□ Choose $\mu \in \Sigma_i$ for $i > 1$, and let $\mu_1, \cdots, \mu_k$ be as in the definition of Σ_i. We will show that $\mathbf{var}(\mu_j) < \frac{1}{3}$ for $1 \le j \le k$.

□ Certainly,

$$\int \tau(x) d\mu_j(x) \le \frac{1}{27} .$$

Since $\tau(x) = x^2$ outside $[-1,1]$,

$$\int_{[-1,1]^c} x^2 d\mu_j(x) \le \frac{1}{27} .$$

Let $a = \int_{[-1,1]} d\mu$. By the power mean inequality,

$$\int_{[-1,1]} x^2 d\left(\frac{\mu_j}{a}\right)(x) \le \left[\int_{[-1,1]} |x|^3 d\left(\frac{\mu_j}{a}\right)(x) \right]^{\frac{2}{3}} ,$$

$$\int_{[-1,1]} x^2 d\mu_j(x) \le a^{\frac{1}{3}} \left[\int_{[-1,1]} |x|^3 d\mu_j(x) \right]^{\frac{2}{3}}$$

$$\le \frac{1}{9} ,$$

since $a \le 1$, and

$$\int x^2 d\mu_j(x) \le \frac{1}{27} + \frac{1}{9} < \frac{1}{3} . \ \square$$

Hence we can pick m and $\eta = \mu_1 * \cdots * \mu_m$ such that $\frac{1}{3} \le \mathbf{var}(\eta) \le \frac{2}{3}$. Let η' be η rescaled to have variance 1 and let μ_j' be μ_j rescaled accordingly. Then

– $$\sum_1^m \int \tau(x) d\mu_j{}'(x) \le \sum_1^k \int \tau(x) d\mu_j{}'(x)$$

– $$\le 27 \sum_1^k \int \tau(x) d\mu_j(x)$$

(since we have rescaled by at most a factor of 3 and $\tau(3x) \le 27\tau(x)$),

– $$\le 27^{-(i-1)} ,$$

so $\eta' \in \Sigma_{i-1}$ and $D(\eta') = D(\eta)$. Similarly $\mu_{m+1} * \cdots * \mu_k$ has the same entropy defect as a member of Σ_{i-1}, proving (c). □

Let L be the set of limit points of $\{s_i \in \Sigma_i\}$ as in the theorem. Now we work toward condition (a) of theorem 1.1. That $*$ is continuous is problem 1 in [Billingsley, P. 28]; it follows immediately from his theorem 3.2 on page 21, which states that the weak topology on the product of two separable metric spaces is the product of the weak topologies on the two spaces.

We now verify the convergence of the functions V and D as needed for condition (a). In fact, we will see that $\mathbf{H}(\mu * \mathbf{n})$ is a continuous function of μ on Σ, but $\mathbf{var}(\mu)$ is easily seen not to be. Since D depends on variance it is not continuous either. However, we will be able to show that variance converges for convergent sequences $\{\mu_i \in \Sigma_i\}$, which we then use to show that D does the same. In the next three propositions we will treat continuity of $\mathbf{H}(\mu * \mathbf{n})$ and the first moment, and then the convergence of variance.

4.2. Proposition. For any $\gamma > 0$, the function $F : \Sigma \to R$ defined by $F(\mu) = \mathbf{H}(\mu * \mathbf{n}_\gamma)$ is defined and continuous.

□ Proof. By rescaling by $\gamma^{-1/2}$, it is sufficient to show that $\mathbf{H}(\mu * \mathbf{n})$ is defined for all μ with finite variance, and that if $\{\mu_i\}$, μ have variance at most $\gamma^{-1/2}$ and $\mu_i \underset{\text{weak}}{\to} \mu$, then $\mathbf{H}(\mu_i * \mathbf{n}) \to \mathbf{H}(\mu * \mathbf{n})$.

Choose $\varepsilon > 0$. Let $d(\mu * \mathbf{n})(x) = g(x) dx$. We then have $g(x) = \int \mathbf{n}(x - t) d\mu(t)$. Since $\mathbf{n}(x) < 1$, $\mathbf{n}'(x) < 1$, we have $g(x) < 1$, $g'(x) < 1$. By the definition of entropy,

– $$F(\mu) = \int \phi(g(x)) dx ,$$

if the integral exists. The integrand is nonnegative since $g(x) < 1$ and ϕ is nonnegative on $[0,1]$. But (3.3) bounds the integral above, so it converges absolutely and $F(\mu)$ is defined.

We now show that there is a bounded interval M = [-m,m] large enough that

(4.3) $$\int_{M^c} \phi(g(x))dx < \frac{\varepsilon}{3},$$

regardless of μ.

□ By Chebychev's inequality, for any $\eta>0$ we can pick m large enough that

– $$\int_{M^c} g(x)dx < \eta .$$

Also,

– $$\int_{M^c} x^2 g(x)dx < 1 ,$$

so by (3.5),

– $$\int_{M^c} \phi(g(x))dx \le \frac{\eta}{2}(4-3\log\eta) .$$

By choosing η sufficiently small we can ensure (4.3). □

Since ϕ is continuous on $[0,1]$ it is uniformly continuous there. Choose $\delta>0$ such that $|\phi(a)-\phi(b)| \le \frac{\varepsilon}{6m}$ whenever $|a-b| \le \delta$. Pick an increasing sequence $-m=x_1, \cdots, x_k=m$ such that $x_{j+1}-x_j < \frac{\delta}{3}$. Now let $\mu_i \underset{\text{weak}}{\rightarrow} \mu$ and let $d(\mu_i * \mathbf{n})(x) = g_i(x)dx$. Then (4.3) is also true with g replaced by g_i. Pick i large enough that $|g(x_j)-g_i(x_j)| < \frac{\delta}{3}$ for all $1 \le j \le k$ by weak convergence. Then for any $-m<x<m$ we pick an x_j with $|x_j-x| < \frac{\delta}{3}$. Since $g'<1$, $g_i'<1$, we have $|g(x_j)-g(x)| < \frac{\delta}{3}$, $|g_i(x_j)-g_i(x)| < \frac{\delta}{3}$,

– $$|g_i(x)-g(x)| \le |g_i(x_j)-g(x_j)| + |g(x_j)-g(x)| + |g_i(x_j)-g_i(x)|$$

– $$\le \delta ,$$

and by our choice of δ,

– $$|\phi(g_i(x)) - \phi(g(x))| \le \frac{\varepsilon}{6M} ,$$

– $$\int_M |\phi(g_i(x)) - \phi(g(x))| dx \le \frac{\varepsilon}{3} ,$$

– $$|F(\mu_i)-F(\mu)|$$

– $$= \left| \int \phi(g_i(x))dx - \int \phi(g(x))dx \right| \le \int_M |\phi(g_i(x)) - \phi(g(x))| dx$$

$$+ \int_{M^c} |\phi(g_i(x))| dx + \int_{M^c} |\phi(g(x))| dx \le \varepsilon ,$$

proving that F is continuous. □

We now prove continuity of first moments in Σ.

4.4. Proposition. If $\{\mu_i\}$ are probability measures satisfying $\int x^2 d\mu_i(x) \le 1$ for all $i \ge 1$, and if $\mu_i \underset{\text{weak}}{\to} \mu$, then $\int x d\mu_i(x) \to \int x d\mu(x)$.

□ Proof. Choose $\varepsilon>0$. Let $M=[-\frac{3}{\varepsilon}, \frac{3}{\varepsilon}]$. Then

$$\int_{M^c} |x|\, d\mu_i(x) \le \frac{\varepsilon}{3} \int_{M^c} x^2 d\mu_i(x)$$

$$\le \frac{\varepsilon}{3}\ .$$

By (4.1),

$$\int_{M^c} |x|\, d\mu(x) \le \frac{\varepsilon}{3}\ .$$

Since $\mu_i \underset{\text{weak}}{\to} \mu$ we may choose i sufficiently large that

$$\left| \int_M x d\mu(x) - \int_M x d\mu_i(x) \right| \le \frac{\varepsilon}{3}\ .$$

Then

$$\left| \int x d\mu(x) - \int x d\mu_i(x) \right| \le$$

$$\le \left| \int_M x d\mu(x) - \int_M x d\mu_i(x) \right| + \left| \int_{M^c} x d\mu(x) \right| + \left| \int_{M^c} x d\mu_i(x) \right| \le \varepsilon\ ,$$

proving 4.4. □

Now suppose $\mu_i \in \Sigma$, $\mu_i \to \mu$. By the above we have $\int x d\mu_i(x) \to \int x d\mu(x)$, and by 4.1 we have $\int x^2 d\mu_i(x) \le \lim\ \sup \int x^2 d\mu(x)$. The second part of condition (a) of theorem 1.1 now follows: that $\mathrm{var}(\mu) \le \lim\sup \mathrm{var}(\mu_i)$. Now we prove the first (and hardest) part of the condition.

4.5. Proposition. Let $(\{X_{ij}\},\{k_i\},\{S_i\},\{\mu_i\},\{\mu_{ij}\})$ satisfy (2.1a-d) and

$$\lim_{i\to\infty} \sum_{j=1}^{k_i} \int_{[-1,1]^c} x^2 d\mu_{X_{ij}}(x) = 0\ . \tag{4.6}$$

(This is implied, e.g., by (2.5) but is weaker, being only the $\beta=1$ case of (2.3).) Suppose $\mu_i \underset{\text{weak}}{\to} \mu$. Then

– $\int x^2 d\mu_i(x) \to \int x^2 d\mu(x)$.

□ Proof. From weak convergence we have $\int d\mu = 1$ and from (4.4) we have $\int x d\mu(x) = 0$. Let $\mu_{ij} = \mu_{X_{ij}}$. We use (4.6) to get tail control on zeroth and first moments:

(4.7) $$\int_{[-1,1]^c} d\mu_{ij} \le \int_{[-1,1]^c} |x|\, d\mu_{ij}(x) \le \int_{[-1,1]^c} x^2 d\mu_{ij}(x) ,$$

so

(4.8) $$\lim_{i\to\infty} \sum_j \int_{[-1,1]^c} d\mu_{ij} = \lim_{i\to\infty} \sum_j \int_{[-1,1]^c} |x|\, d\mu_{ij}(x) = 0 .$$

We define the measures η_{ij} to be μ_{ij} restricted to $[-1,1]$, and $\eta_i = \eta_{i1} * \cdots * \eta_{ik_i}$. We now claim that

(4.9) $$\eta_i \underset{\text{weak}}{\to} \mu .$$

□ To show this, note from the definition of η_{ij} that $\eta_i \le \mu_i$, and yet from (4.6) we have that $\int d(\mu_i - \eta_i) \to 0$. These and $\mu_i \underset{\text{weak}}{\to} \mu$ imply that $\eta_i \underset{\text{weak}}{\to} \mu$. □

It follows immediately that $\int d\eta_i \to 1$, and from (4.7) that $\int x d\eta_i(x) \to 0$. We now show that

(4.10) $$\lim_{i\to\infty} \left[\int x^2 d\eta_i(x) - \int x^2 d\mu_i(x) \right] = 0 .$$

□ We have,

– $$\int x^2 d\eta_i(x) = \sum_j \pi_{ij} \int x^2 d\eta_{ij}(x) + 2 \sum_{j\le k} \pi_{ijk} \int x d\eta_{ij}(x) \int x d\eta_{ik}(x) ,$$

where $\pi_{ij} = \prod_{n\ne j} \int d\eta_{in}$, $\pi_{ijk} = \prod_{n\ne j,k} \int d\eta_{in}$. We have $\pi_{ij} \le 1$, $\pi_{ijk} \le 1$, and since $\pi_{ij} \ge \sum_j (1 - \int d\eta_{ij})$, (4.7) implies that $\lim_{i\to\infty} \pi_{ij} = 1$. From (4.6) we have

– $$\lim_{i\to\infty} \left[\int x^2 d\mu_i(x) - \sum_j \int x^2 d\eta_{ij}(x) \right]$$

– $$= \lim_{i\to\infty} \sum_j \left[\int x^2 d\mu_{ij}(x) - \int x^2 d\eta_{ij}(x) \right]$$

– $$= \lim_{i\to\infty} \sum_j \int_{[-1,1]^c} x^2 d\mu_{ij}(x) = 0 .$$

Also,

– $$\sum_{j\le k} \left| \int x d\eta_{ij}(x) \int x d\eta_{ik}(x) \right| \le \Big(\sum_j \left| \int x d\eta_{ij}(x) \right| \Big)^2 .$$

By (4.7) and since $\pi_{ijk} \le 1$,

$$\lim_{i\to\infty} \left| \sum_{j\ge k} \pi_{ijk} \int x d\eta_{ij}(x) \int x d\eta_{ik}(x) \right| = 0 .$$

(4.10) follows immediately. □

It now suffices to prove that

(4.11) $$\lim_{i\to\infty} \int x^2 d\eta_i(x) = \int x^2 d\mu(x) .$$

Since η_{ij} is supported on $[-1,1]$, we have

$$\left| \int x^3 d\eta_{ij}(x) \right| \le \int x^2 d\eta_{ij}(x) ,$$

and hence

(4.12) $$\sum_j \left| \int x^3 d\eta_{ij}(x) \right| \le 1 .$$

Similarly,

(4.13) $$\sum_j \int x^4 d\eta_{ij}(x) \le 1 ,$$

and by (4.7) we have,

(4.14) $$\sum_j \left| \int x d\eta_{ij}(x) \right| \le 1 .$$

This allows us to control the fourth moment of η_i:

$$\begin{aligned} \int x^4 d\eta_i(x) &= \sum_j \pi_{ij} \int x^4 d\eta_{ij}(x) \\ &+ 4\sum_{j,k} \pi_{ijk} \left[\int x^3 d\eta_{ij}(x) \right] \left[\int x d\eta_{ik}(x) \right] \\ &+ 3\sum_{j,k} \pi_{ijk} \left[\int x^2 d\eta_{ij}(x) \right] \left[\int x^2 d\eta_{ik}(x) \right] \\ &+ 6\sum_{j,k,l} \pi_{ijkl} \left[\int x^2 d\eta_{ij}(x) \right] \left[\int x d\eta_{ik}(x) \right] \left[\int x d\eta_{il}(x) \right] \\ &+ \sum_{j,k,l,m} \pi_{ijklm} \left[\int x d\eta_{ij}(x) \right] \left[\int x d\eta_{ik}(x) \right] \left[\int x d\eta_{il}(x) \right] \left[\int x d\eta_{im}(x) \right] , \end{aligned}$$

where in the above sums none of j,k,l,m are equal and $\pi_{ij} = \prod_{n\ne j} \int d\eta_{in}$, $\pi_{ijk} = \prod_{n\ne j,k} \int d\eta_{in}$, $\pi_{ijkl} = \prod_{n\ne j,k,l} \int d\eta_{in}$, $\pi_{ijklm} = \prod_{n\ne j,k,l,m} \int d\eta_{in}$, so

$$\begin{aligned} \int x^4 d\eta_i(x) &\le \sum_j \int x^4 d\eta_{ij}(x) + 4 \left| \sum_j \int x^3 d\eta_{ij}(x) \right| \left| \sum_k \int x d\eta_{ik}(x) \right| \\ &+ 3 \left| \sum_j \int x^2 d\eta_{ij}(x) \right| \left| \sum_k \int x^2 d\eta_{ik}(x) \right| \end{aligned}$$

$$+\,6\left|\sum_j \int x^2 d\eta_{ij}(x)\right|\left|\sum_k \int x d\eta_{ik}(x)\right|\left|\sum_l \int x d\eta_{il}(x)\right|$$

$$+\left|\sum_j \int x d\eta_{ij}(x)\right|\left|\sum_k \int x d\eta_{ik}(x)\right|\left|\sum_l \int x d\eta_{il}(x)\right|\left|\sum_m \int x d\eta_{im}(x)\right|$$

$$\le 15\,.$$

By (4.1) and (4.9) we also have $\int x^4 d\mu(x) \le 15$. Pick $\varepsilon>0$ and let $M=[-(\frac{45}{\varepsilon})^{½},(\frac{45}{\varepsilon})^{½}]$. Then

$$\int_{M^c} x^2 d\eta_i(x) \le \frac{\varepsilon}{45}\int_{M^c} x^4 d\eta_i(x) \le \frac{\varepsilon}{3}\;,$$

and similarly,

$$\int_{M^c} x^2 d\mu(x) \le \frac{\varepsilon}{3}\;.$$

By (4.9) we can choose i large enough that

$$\left|\int_M x^2 d\mu(x) - \int_M x^2 d\eta_i(x)\right| \le \frac{\varepsilon}{3}\;,$$

and hence

$$\left|\int x^2 d\mu(x) - \int x^2 d\eta_i(x)\right| \le \varepsilon\,. \quad \square$$

It immediately follows that if $\mu_i \in \Sigma_i$, $\mu_i \to \mu$, then var(μ)=1, the first part of condition (a) in theorem 1.1. The remainder of the condition is the statement about D which we now verify.

□ Suppose $\mu_i \in \Sigma$, $\mu_i \to \mu$, and var$(\mu_i) \to$ var(μ). We wish to show that $D(\mu_i) \to D(\mu)$, which is equivalent to showing that $\mathbf{H}(\mu_i * \nu_{\mu_i}) \to \mathbf{H}(\mu * \nu_\mu)$. If $a<b$ and η is any measure with finite variance, (3.1) gives,

$$\mathbf{H}(\eta * \nu_b) \ge \mathbf{H}(\eta * \nu_a)\,.$$

On the other hand, by (3.10) we have,

$$\mathbf{H}(\eta * \nu_b) = \mathbf{H}(\eta * \nu_a) + \int_a^b J(\eta * \nu_t)dt$$

$$\le \mathbf{H}(\eta * \nu_a) + \int_a^b J(\nu_t)dt$$

by (3.9),

$$= \mathbf{H}(\eta * \nu_a) + \int_a^b \frac{1}{t}\,dt$$

$$= \mathbf{H}(\eta * \nu_a) + \log(\frac{b}{a})\,.$$

Thus, since $\mathbf{var}(\mu_i) \to \mathbf{var}(\mu)$, we have

$$\lim_{i\to\infty} \quad |\mathbf{H}(\mu_i * \nu_{\mu_i}) - \mathbf{H}(\mu_i * \nu_\mu)| \to 0 .$$

But by (4.2),

$$\lim_{i\to\infty} \quad \mathbf{H}(\mu_i * \nu_\mu) \to \mathbf{H}(\mu * \nu_\mu) ,$$

so

$$\lim_{i\to\infty} \quad \mathbf{H}(\mu_i * \nu_{\mu_i}) \to \mathbf{H}(\mu * \nu_\mu) . \quad \square$$

We have now verified that G, Σ, $\{\Sigma_i\}$, V, D, and $*$ meet all the conditions of (1.1), proving the following result.

4.15. Proposition. If $(\{X_{ij}\},\{k_i\},\{S_i\},\{\mu_i\},\{\mu_{ij}\})$ satisfy (2.1a-d) and if

$$(4.16) \qquad \sum_{j=1}^{n} \int \tau(x) d\mu_i(x) \le 27^{-i} ,$$

then $\mu_i \underset{\text{weak}}{\to} \nu$.

We now show that (4.15) is equivalent to (2.2).

$\square$ First we note that the condition (4.16) may be replaced by the weaker condition (2.5); we still must have $\mu_i \underset{\text{weak}}{\to} \nu$, because otherwise we could pick a subsequence which contradicted (4.15).

We need only show that weak convergence of measures is equivalent to uniform convergence of the probability distribution as in (2.4). If $\{\mu_i\}$ and μ are probability measures, then by the discussion on [Billingsley, p.18], the following are equivalent:

1. $\mu_i \underset{\text{weak}}{\to} \mu$;
2. $\mu_i((-\infty,t\,]) \to \mu((-\infty,t\,])$, for all t with $\mu(\{t\})=0$.

If we let μ be the standard normal measure $\boldsymbol{\nu}$, then (2) implies that $\mathrm{Prob}(S_i \le t) = \mu_i((-\infty,t\,]) \to \mathbf{N}(t)$ pointwise, but bounded nondecreasing functions which converge pointwise to a continuous function must converge uniformly. Hence, (1) is equivalent to (2.4), proving the CLT in the form (2.2). $\square$

5. The rate of convergence of entropy.

In this chapter we work toward an effective form of the CLT which provides an upper bound on the quantity $\mathbf{H}(\mathbf{n}_\mu * \mathbf{n}_\gamma) - \mathbf{H}(\mu * \mathbf{n}_\gamma)$ where $\gamma > 0$ and $\mu = \mu_1 * \cdots * \mu_m$. The approach is to show that convolving μ_i to the density $\mu_1 * \cdots * \mu_{i-1} * \mathbf{n}_\gamma$ adds almost as much entropy to the density as convolving its normal mate would. But when all the μ_i are replaced by their normal mates, the convolution is normal, supplying a comparison between $\mathbf{H}(\mu * \mathbf{n}_\gamma)$ and $\mathbf{H}(\mathbf{n}_\mu * \mathbf{n}_\gamma)$.

First we show that if two probability densities f and $\tilde{f}$ are supported on a bounded interval, and if f, f' have equal zeroth, first, and second moments, then the difference $H(f * g) - H(\tilde{f} * g)$ is on the order of the third moments of f and f', if g is a density of a random variable with a normal random variable added into it. We will need the bounded support assumption to control $\left| \frac{\mathbf{n}'''}{\mathbf{n}} \right|$; later we will work around this restriction.

5.1. Proposition. Let f, $\tilde{f}$, g be continuous probability densities such that

– $f(x) = \tilde{f}(x) = 0$ for $|x| > 2$,

– $\int x f(x) dx = \int x \tilde{f}(x) dx = \int x g(x) dx = 0$,

and

– $\int x^2 f(x) dx = \int x^2 \tilde{f}(x) dx$.

Let

– $\varepsilon = \int |x|^3 (f(x) + \tilde{f}(x)) dx$,

– $\beta = \int (1 + x^2) g(x) dx = 1 + \mathbf{var}(g)$.

Then for some universal constant $K_{(5.1)}$ we have,

– $\left| \mathbf{H}(f * g * \mathbf{n}) - \mathbf{H}(\tilde{f} * g * \mathbf{n}) \right| \le K_{(5.1)} \varepsilon \beta$.

□ Proof. Let

– $a = f * \mathbf{n}$, $\tilde{a} = \tilde{f} * \mathbf{n}$,

– $b = f * g * \mathbf{n}$, $\tilde{b} = \tilde{f} * g * \mathbf{n}$,

– $T = \{ t : |a(t) - \tilde{a}(t)| \le \frac{a(t)}{4} \}$,

– $U = \{ u : |b(u) - \tilde{b}(u)| \le \frac{b(u)}{2} \}$.

We now obtain a bound on $|a - \tilde{a}|$:

(5.2) $\quad |(a - \tilde{a})(x)| \le \frac{\varepsilon}{6} \sup\{ |\mathbf{n}'''(x+y)| : |y| < 2 \}$.

□ Let $e(s) = f(s) - \tilde{f}(s)$. If $|s| < 2$ then by Taylor's theorem,

$$\mathbf{n}(x-s) = \mathbf{n}(x) - s\mathbf{n}'(x) + \frac{1}{2} s^2 \mathbf{n}''(x) - \frac{1}{6} s^3 \mathbf{n}'''(x+y(x,s)) ,$$

where $|y(x,s)| < 2$. Hence,

$$(a-\tilde{a})(x) = (e * \mathbf{n})(x) = \int_{-\infty}^{\infty} e(s)\mathbf{n}(x-s)ds$$

$$= \int_{-\infty}^{\infty} e(s)\left[\mathbf{n}(x) - s\mathbf{n}'(x) + \frac{1}{2} s^2 \mathbf{n}''(x) - \frac{1}{6} s^3 \mathbf{n}'''(x+y(x,s))\right] ds$$

$$= \mathbf{n}(x)\int e(s)ds \; - \mathbf{n}'(x)\int se(s)ds \; + \frac{1}{2}\, \mathbf{n}''(x)\int s^2 e(s)ds$$

$$- \frac{1}{6}\int \mathbf{n}'''(x+y(x,s))s^3 e(s)ds$$

$$= -\frac{1}{6}\int \mathbf{n}'''(x+y(x,s))s^3 e(s)ds ,$$

$$|(a-\tilde{a})(x)| \le \frac{1}{6} \sup\{ |\mathbf{n}'''(x+y)| : |y| < 2\} \int |s^3 e(s)| ds$$

$$\le \frac{\varepsilon}{6} \sup\{ |\mathbf{n}'''(x+y)| : |y| < 2\} . \quad \square$$

By the mean value theorem,

$$a(x) = \int_{-2}^{2} f(t)\mathbf{n}(x-t)dt = \mathbf{n}(x+z) ,$$

for some $|z| \le 2$. By (5.2), for some $|y| < 2$,

$$\frac{|(a-\tilde{a})(x)|}{a(x)} \le \frac{|-(x+y)^3 + 3(x+y)|\,\varepsilon}{6} \; \frac{\mathbf{n}(x+y)}{\mathbf{n}(x+z)} \tag{5.3}$$

$$= \frac{|-(x+y)^3 + 3(x+y)|\,\varepsilon}{6} \; e^{(z-y)x + \frac{1}{2}(z^2-y^2)}$$

$$\le \frac{(|x|^3 + 6|x|^2 + 15|x| + 14)\,\varepsilon}{6} \; e^{4|x|+2}$$

$$\le 6e^2 e^{7|x|}\varepsilon .$$

We now show that for some $K_{(5.4)}$,

$$\int_{T^c} |a(t) - \tilde{a}(t)| dt \le K_{(5.4)}\varepsilon^2 . \tag{5.4}$$

□ By (5.3),

$$|a(x) - \tilde{a}(x)| \le \frac{a(x)}{4} \quad \text{for} \; |x| < M ,$$

where

$$M = -\frac{1}{7} \log(24e^2\varepsilon) .$$

Suppose $M \geq 30$. The definition of T gives,

– $\mathrm{T} \supseteq (-M, M)$,

(5.5) $$\int_{\mathrm{T}^c} |a(t) - \tilde{a}(t)| \, dt \leq 2 \int_M^{\infty} 2\mathbf{n}(\min(0, |x| - 2)) dx ,$$

since $a = f * \mathbf{n}$ and f is supported on $[-2, 2]$,

– $$\leq 4 \int_{M-2}^{\infty} \mathbf{n}(x) dx$$

– $$\leq 8\mathbf{n}(M-2) ,$$

since $\int_u^{\infty} \mathbf{n}(x) dx \leq 2\mathbf{n}(u)$ for $u > 0$,

– $$\leq \frac{8}{\sqrt{2\pi}} e^{-14M} ,$$

since for $M \geq 30$, $\dfrac{(M-2)^2}{2} \geq 14M$,

– $$= \frac{8}{\sqrt{2\pi}} e^{2\log(24e^2\varepsilon)} ,$$

– $$= K_{(5.5)} \varepsilon^2 ,$$

where $K_{(5.5)} = 2304e^4 (\frac{2}{\pi})^{½}$. On the other hand, suppose $M < 30$. Then by the definition of M,

– $$\frac{1}{7} \log(24e^2\varepsilon) \geq -30 ,$$

– $$\varepsilon^2 \geq \frac{e^{-424}}{576} ,$$

– $$\int_{\mathrm{T}^c} |a(t) - \tilde{a}(t)| \, dt \leq \frac{e^{424}}{288} \varepsilon ,$$

proving (5.4) with $K_{(5.4)} = \dfrac{e^{424}}{288}$. □

From this we now show that

(5.6) $$\int_{\mathrm{U}^c} |b(u) - \tilde{b}(u)| \, du \leq 2K_{(5.4)} \varepsilon^2 .$$

□ If $u \in \mathrm{U}^c$ we have,

(5.7) $$\int_{\mathrm{T}} |a(t) - \tilde{a}(t)| \, g(u-t) dt \leq \frac{1}{4} \int_{\mathrm{T}} a(t) g(u-t) dt$$

by the definition of T,

– $$\leq \frac{1}{4} \int a(t) g(u-t) dt = \frac{1}{4} b(u)$$

$$\leq \frac{1}{2}\ |b(u)-\tilde{b}(u)|\ ,$$

by the definition of U. Also,

(5.8) $$\int |a(t)-\tilde{a}(t)|\ g(u-t)dt \geq \left|\int (a(t)-\tilde{a}(t))\ g(u-t)dt\right|$$

$$= |b(u)-\tilde{b}(u)|\ .$$

Subtracting (5.7) from (5.8),

$$\int_{T^c} |a(t)-\tilde{a}(t)|\ g(u-t)dt \geq \frac{1}{2}\ |b(u)-\tilde{b}(u)|\ ,\ \text{ all } u \in U^c\ .$$

This means that at least half of the contribution of $|a-\tilde{a}|$ to large values of $|b-\tilde{b}|$ comes from large values of $|a-\tilde{a}|$. We have,

$$\int_{T^c} |a(t)-\tilde{a}(t)|\,dt = \int_{T^c} |a(t)-\tilde{a}(t)|\ \left[\int g(u-t)du\right]\ dt$$

$$= \int \left[\int_{T^c} |a(t)-\tilde{a}(t)|\ g(u-t)dt\right] du$$

$$\geq \int_{U^c} \left[\int_{T^c} |a(t)-\tilde{a}(t)|\ g(u-t)dt\right] du$$

$$\geq \int_{U^c} \left[\frac{1}{2}\ |b(u)-\tilde{b}(u)|\right] du\ .$$

This and (5.4) imply (5.6). □

We will now get a bound on $\int |\phi(b(x))-\phi(\tilde{b}(x))|\,dx$ by bounding it separately both inside and outside U. First we show from (5.6) that both b and $\tilde{b}$ contribute negligibly outside U.

□ If $u \in U^c$ then $b(u) \leq 2|b(u)-\tilde{b}(u)|$ and

$$\tilde{b}(u) \leq b(u)+|b(u)-\tilde{b}(u)| \leq 3|b(u)-\tilde{b}(u)|\,.$$

By (5.6),

$$\int_{U^c} b(u)du \leq 6K_{(5.4)}\varepsilon^2\ ,$$

$$\int_{U^c} \tilde{b}(u)du \leq 6K_{(5.4)}\varepsilon^2\ .$$

Also,

$$\int_{U^c} u^2 b(u)du \leq \int u^2 b(u)du \leq \beta+4 \leq 5\beta\ ,$$

$$\int_{U^c} u^2 \tilde{b}(u)du \leq 5\beta\ ,$$

and hence by (3.5),

$$\int_{U^c} |\phi(b(u)) - \phi(\tilde b(u))|\,du \le 6K_{(5.4)}\varepsilon^2\,(5\beta+3-3\log(6K_{(5.4)}\varepsilon^2))\ .$$

Since $\varepsilon\le 16$ and $\varepsilon^2 \log\varepsilon$ approaches zero faster than ε,

$$\int_{U^c} |\phi(b(u)) - \phi(\tilde b(u))|\,du \le K_{(5.9)}\varepsilon\beta\ , \tag{5.9}$$

for some $K_{(5.9)}$. □

On the other hand, inside U we can relate $\phi(b)-\phi(\tilde b)$ to $b-\tilde b$ smoothly. If $u\in U$, then $\tilde b(u) \ge \frac{b(u)}{2}$ and since $b=f*g*\mathbf{n}$ and $\mathbf{n}(u)<1$, $b(u)<1$ and similarly $\tilde b(u)<1$. We have,

$$|\phi(b(u))-\phi(\tilde b(u))| \le |b(u)-\tilde b(u)|\ \sup\{|\phi'(t)| : \frac{b(u)}{2} \le t \le 1\} \tag{5.10}$$

$$= |b(u)-\tilde b(u)|\ \sup\{(-1-\log t) : \frac{b(u)}{2} \le t\le 1\}$$

$$= |b(u)-\tilde b(u)|\ (\log 2 - 1 - \log b(u))\ .$$

We now need a bound on the quantity $-\log b(u)$. We have,

$$\sigma(f*g) = \sqrt{\mathbf{var}(f)+\mathbf{var}(g)} \le \sqrt{4+\mathbf{var}(g)} \le 2+\sigma\ ,$$

where $\sigma=\sigma(g)$. By Chebychev's inequality,

$$\int_{-2\sigma-4}^{2\sigma+4} (f*g)(x)dx \ge \frac{3}{4}\ ,$$

$$b(u) = \int (f*g)(x)\mathbf{n}(u-x)dx \ge \int_{-2\sigma-4}^{2\sigma+4} (f*g)(x)\mathbf{n}(u-x)dx$$

$$\ge \frac{3}{4} \inf_{\{|x|<2\sigma+4\}} \mathbf{n}(u-x)$$

$$= \frac{3}{4}\,\mathbf{n}(|u|+2\sigma+4)\ ,$$

so

$$|\log b(u)| \le |\log(\frac{3}{4\sqrt{2\pi}}) + \frac{(|u|+2\sigma+4)^2}{2}|$$

$$\le \frac{(|u|+2\sigma+4)^2}{2}$$

$$\le \frac{3}{2}(u^2+4\sigma^2+16)$$

$$\le 24(u^2+\beta)\ .$$

Using this in (5.10),

$$|\phi(b(u))-\phi(\tilde b(u))| \le |b(u)-\tilde b(u)|(23+\log 2)(u^2+\beta)\ ,$$

(5.11) $$\int_U |\phi(b(u))-\phi(\tilde{b}(u))|\,du \le K_{(5.11)}\int |b(u)-\tilde{b}(u)|\,(u^2+\beta)du ,$$

where $K_{(5.11)}=23+\log 2$. We now need to bound the zeroth and second moments of $|b-\tilde{b}|$. By (5.2),

$$\int |a(t)-\tilde{a}(t)|\,dt < K_1\varepsilon ,$$

$$\int t^2 |a(t)-\tilde{a}(t)|\,dt < K_2\varepsilon ,$$

where

$$K_1=\frac{1}{6}\int \sup\{|\mathbf{n}'''(z)| : |z-x|<2\}dx ,\ \ K_2=\frac{1}{6}\int x^2 \sup\{|\mathbf{n}'''(z)| : |z-x|<2\}dx .$$

Convolving by g,

$$\int |b(u)-\tilde{b}(u)|\,du = \int |a(u)-\tilde{a}(u)|\,du < K_1\varepsilon ,$$

$$\int u^2 |b(u)-\tilde{b}(u)|\,du = \int u^2 |a(u)-\tilde{a}(u)|\,du \int g(u)du$$

$$+\int |a(u)-\tilde{a}(u)|\,du \int u^2 g(u)du$$

$$\le (K_2+K_1\beta)\varepsilon .$$

Using these estimates in (5.11),

$$\int_U |\phi(b(u))-\phi(\tilde{b}(u))|\,du \le K_{(5.11)}(2K_1+K_2)\varepsilon\beta .$$

This and (5.9) give,

$$\left|\mathbf{H}(b) - \mathbf{H}(\tilde{b})\right| \le K_{(5.1)}\varepsilon\beta ,$$

where $K_{(5.1)} = K_{(5.11)}(2K_1+K_2) + K_{(5.9)}$, proving (5.1). □

The next two propositions replace the two densities in (5.1) with an arbitrary density f and its normal mate $\mathbf{n}_f$. First we handle the easier problem of replacing $\tilde{f}$ in (5.1) with its normal mate (and hence that of f.) The strategy is to let $\tilde{f}$ be a density of bounded support which is essentially the same as $\mathbf{n}_f$.

5.12. Proposition. Let f, g be probability densities such that $f(x)=0$ for $|x|>1$ and

$$\int xf(x)dx = \int xg(x)dx = 0.$$

Let

$$\varepsilon=\int |x|^3 f(x)dx,$$

$$\beta=\int (1+|x|^2)g(x)dx = \mathbf{var}(g)+1 ,$$

as in (5.1). Then for some universal constant $K_{(5.12)}$ we have,

– $$|\mathbf{H}(f * g * \mathbf{n}) - \mathbf{H}(\mathbf{n}_f * g * \mathbf{n})| \le K_{(5.12)}\varepsilon\beta\ .$$

Proof.

□ We construct a probability density $\tilde{f}(x)$ with the same mean and standard deviation as f, but which satisfies

(5.13a) $\tilde{f}(x)=0$, for all $|x|>2$,

(5.13b) $\int|\tilde{f}(x)-\mathbf{n}_f(x)|\,dx < K_{(5.13b)}\varepsilon^2$,

(5.13c) $\int|x|^3\tilde{f}(x)dx < K_{(5.13c)}\varepsilon$,

for universal constants $K_{(5.13b)}, K_{(5.13c)}$.

□ We will choose $\tilde{f}$ to be $\mathbf{n}_f$ truncated to [-1,1], with a small adjustment h to give it the proper zeroth and second moments. Let

– $$m_0 = \int_{[-1,1]^c} \mathbf{n}_f(x)dx\ ,\ \ m_2 = \int_{[-1,1]^c} x^2\mathbf{n}_f(x)dx\ .$$

We now show that $m_2 < 4m_0$.

Let $\sigma=\sigma(f)$, which cannot exceed 1 because f is supported in [−1,1]. The statement is true in effect because $\frac{m_2}{m_0}$ is an increasing function of $\sigma(f)$, so we need only check the statement in the special case that σ=1. Let

– $$a(t)=\frac{\mathbf{n}_\sigma(t)}{\int_1^\infty \mathbf{n}_\sigma(t)dt}\ ,$$

– $$b(t)=\frac{\mathbf{n}_1(t)}{\int_1^\infty \mathbf{n}_1(t)dt}\ .$$

By inspection, $\frac{b(t)}{a(t)}$ is an increasing function of t. Hence, for $x>1$ we have

– $$\int_x^\infty b(t)dt \le \frac{b(x)}{a(x)}\int_x^\infty a(t)dt\ ,$$

– $$\int_1^x b(t)dt \ge \frac{b(x)}{a(x)}\int_1^x a(t)dt\ ,$$

or dividing,

$$\frac{\int_x^\infty b(t)dt}{\int_1^x b(t)dt} \le \frac{\int_x^\infty a(t)dt}{\int_1^x a(t)dt} ,$$

or adding one and reciprocating,

$$\int_1^x b(t)dt \ge \int_1^x a(t)dt .$$

By integration by parts,

$$\int_1^\infty t^2 b(t)dt - \int_1^\infty t^2 a(t)dt = \int_1^\infty \left[\int_1^t a(u)du - \int_1^t b(u)du \right] 2tdt \ge 0 ,$$

or

$$\frac{\int_1^\infty t^2 \mathbf{n}_1(t)dt}{\int_1^\infty \mathbf{n}_1(t)dt} \ge \frac{\int_1^\infty t^2 \mathbf{n}_\sigma(t)dt}{\int_1^\infty \mathbf{n}_\sigma(t)dt} .$$

The right hand side of the equation is equal to $\frac{m_2}{m_0}$ and the left is approximately 2.52, from which we conclude that $m_2 < 4m_0$. □

We can thus construct a nonnegative integrable function h which is zero outside $(-2,2)$ and whose zeroth, first, and second moments are m_0, 0, and m_2, respectively. Let

$$\tilde{f}(x) = h(x) + \begin{cases} \mathbf{n}_f(x) , & -1 \le x \le 1 , \\ 0 , & \text{otherwise} , \end{cases}$$

so that $\tilde{f}$ has the same zeroth, first, and second moments as $\mathbf{n}_f$, and hence the same as f.

Let $\sigma^2 = v = \mathbf{var}(f)$. The tail of a normal density falls faster than the reciprocal of any polynomial, so for some universal constant $K_{(5.14)}$ we have,

$$m_0 = \int_{[-\sigma^{-1},\sigma^{-1}]^c} \mathbf{n}(x)dx \le K_{(5.14)} v^3 .$$

By the power mean inequality $v^3 \le \varepsilon^2$, so

$$\int | \tilde{f}(x) - \mathbf{n}_f(x) | dx \le 2m_0 \le 2K_{(5.14)} \varepsilon^2 ,$$

verifying (5.13b). Now let

$$K_{(5.15)} = \int |x|^3 \mathbf{n}_f(x)dx \left[\int x^2 \mathbf{n}_f(x)dx \right]^{-3/2} , \tag{5.15}$$

which is independent of f. Then we have

$$\int |x|^3 \mathbf{n}_f(x)dx = K_{(5.15)} \left[\int x^2 \mathbf{n}_f(x)dx \right]^{3/2}$$

– $$= K_{(5.15)} \left[\int x^2 f(x) dx \right]^{3/2}$$

– $$\le K_{(5.15)} \int |x|^3 f(x) dx$$

by the power mean inequality,

– $$= K_{(5.15)} \varepsilon ,$$

and

– $$\int |x|^3 \tilde{f}(x) dx \le \int |x|^3 \mathbf{n}_f(x) dx + 8 \int h(x) dx$$

– $$\le K_{(5.15)} \varepsilon + 8 K_{(5.14)} \varepsilon^2$$

– $$\le (K_{(5.15)} + 8 K_{(5.14)}) \varepsilon ,$$

verifying (5.13c). □

Let $k = |\tilde{f} * g * \mathbf{n} - \mathbf{n}_f * g * \mathbf{n}|$. By (3.6),

(5.16) $$|\mathbf{H}(\tilde{f} * g * \mathbf{n}) - \mathbf{H}(\mathbf{n}_f * g * \mathbf{n})| \le \int k(x) dx + \int \phi(k(x)) dx .$$

We wish to estimate $\mathbf{H}(k)$ using (3.5), so we now estimate its zeroth and second moments. We have,

– $$k = |g * \mathbf{n} * (f - \mathbf{n}_f)| \le g * \mathbf{n} * |f - \mathbf{n}_f| ,$$

– $$\int k(x) dx \le \int |\tilde{f}(x) - \mathbf{n}_f(x)| dx$$

– $$\le K_{(5.13b)} \varepsilon^2 ;$$

and

– $$\int x^2 k(x) dx \le \int x^2 (g * \mathbf{n} * |\tilde{f} - \mathbf{n}_f|)(x) dx$$

– $$= \int x^2 |\tilde{f}(x) - \mathbf{n}_f(x)| dx + \beta \int |\tilde{f}(x) - \mathbf{n}_f(x)| dx$$

– $$\le 2(1+\beta) \le 4\beta ,$$

so 3.5 gives,

– $$\mathbf{H}(k) \le \frac{K_{(5.13b)} \varepsilon^2}{2} \left(4\beta + 3 - 3 \log (K_{(5.13b)} \varepsilon^2)\right) .$$

Since $\varepsilon \le 1$ and $\beta \ge 1$, for a sufficiently large $K_{(5.17)}$ we have,

(5.17) $$|\mathbf{H}(\tilde{f} * g * \mathbf{n}) - \mathbf{H}(\mathbf{n}_f * g * \mathbf{n})| \le K_{(5.17)} \varepsilon \beta .$$

By (5.1),

– $$\left| \mathbf{H}(f * g * \mathbf{n}) - \mathbf{H}(\tilde{f} * g * \mathbf{n}) \right| \le K_{(5.1)} \varepsilon \beta .$$

The proposition follows from the triangle inequality. □

Now we work toward an inequality which does not require that f have bounded support. The idea is to write f as a weighted average of densities f_1 supported on $[-1,1]$ and f_2 supported on $(-1,1)^c$. We use (5.12) on f_1 and allow f_2 to be as bad as possible. The concavity of **H**, considered as a function on the set of probability densities, allows us to combine entropy estimates for f_1, f_2 into one for f.

5.18. Proposition. Let f, g be probability densities such that

$$\int xf(x)dx = \int xg(x)dx = 0 .$$

Then for some universal constant $K_{(5.18)}$ we have,

$$\frac{1}{2}\log\left(\frac{\mathbf{var}(f)+\beta}{\beta}\right) + \mathbf{H}(g*\mathbf{n}) - \mathbf{H}(f*g*\mathbf{n})$$

$$\leq K_{(5.18)}\int \min(|x|^2,\beta|x|^3)f(x)dx ,$$

where

$$\beta = \mathbf{var}(g)+1 .$$

□ Proof. Since the statement is unchanged by substituting $f(-x)$ for $f(x)$ and $g(-x)$ for $g(x)$, we may assume without loss of generality that,

$$\int_{-\beta^{-1}}^{\beta^{-1}} xf(x)dx \geq 0 .$$

Then since the function

$$h(x) = \int_{-\beta^{-1}}^{x\beta^{-1}} tf(t)dt$$

is continuous and $h(0) \leq 0 \leq h(1)$, for some $0 \leq \alpha \leq 1$ we have

$$\int_{-\beta^{-1}}^{\alpha\beta^{-1}} tf(t)dt = 0 . \tag{5.19}$$

Let $a_1 = \int_{-\beta^{-1}}^{\alpha\beta^{-1}} f(t)dt$ and $a_2 = 1-a_1$. Then we may write f as a weighted average $f = a_1\tilde{f}_1 + a_2\tilde{f}_2$ where $\tilde{f}_1$, $\tilde{f}_2$ are probability densities zero outside $[-\beta^{-1},\alpha\beta^{-1}]$ and $[-\beta^{-1},\alpha\beta^{-1}]^c$, respectively. Let

$$f_1 = a_1\tilde{f}_1 ,\ v_1 = \mathbf{var}(\tilde{f}_1) ,$$

$$f_2 = a_2\tilde{f}_2 ,\ v_2 = \mathbf{var}(\tilde{f}_2) ,$$

$$v = \mathbf{var}(f) = a_1v_1 + a_2v_2 .$$

Since $\beta \geq 1$, f_1 satisfies the conditions of proposition (5.12), so

$$\mathbf{H}(\mathbf{n}_{\tilde{f}_1} * g * \mathbf{n}) - \mathbf{H}(\tilde{f}_1 * g * \mathbf{n}) \leq K_{(5.12)}\beta\int |x|^3\tilde{f}_1(x)dx .$$

Applying (3.14) to $\mathbf{n}_{\tilde{f}_1}$ and $g*\mathbf{n}$ gives,

$$\mathbf{var}(\mathbf{n}_{\tilde{f}_1}*g*\mathbf{n})\overline{\mathbf{H}}(\mathbf{n}_{\tilde{f}_1}*g*\mathbf{n}) \geq \mathbf{var}(g*\mathbf{n})\overline{\mathbf{H}}(g*\mathbf{n})$$

$$\geq \mathbf{var}(\mathbf{n}_{\tilde{f}_1}*g*\mathbf{n})\overline{\mathbf{H}}(g*\mathbf{n}) \ ,$$

since $\overline{\mathbf{H}}(g*\mathbf{n}) \leq 0$, and hence by the definition of $\overline{\mathbf{H}}$,

$$\mathbf{H}(\mathbf{n}_{\tilde{f}_1}*g*\mathbf{n}) - \mathbf{H}(g*\mathbf{n}) \geq \frac{1}{2}\log(\frac{\mathbf{var}(\mathbf{n}_{\tilde{f}_1}*g*\mathbf{n})}{\mathbf{var}(g*\mathbf{n})}) = \frac{1}{2}\log(\frac{v_1+\beta}{\beta}) \ .$$

We thus have,

$$\mathbf{H}(\tilde{f}_1*g*\mathbf{n}) \geq \mathbf{H}(g*\mathbf{n}) + \frac{1}{2}\log(\frac{v_1+\beta}{\beta}) - K_{(5.12)}\beta\int|x|^3\tilde{f}_1(x)dx \ .$$

By (3.1),

$$\mathbf{H}(\tilde{f}_2*g*\mathbf{n}) - \mathbf{H}(g*\mathbf{n}) \geq 0 \ .$$

and by concavity of $\mathbf{H}$,

$$\mathbf{H}(f*g*\mathbf{n}) \geq a_1\mathbf{H}(\tilde{f}_1*g*\mathbf{n}) + a_2\mathbf{H}(\tilde{f}_2*g*\mathbf{n}) \ .$$

Combining the three above inequalities gives,

$$\mathbf{H}(f*g*\mathbf{n}) \geq \mathbf{H}(g*\mathbf{n}) + \frac{a_1}{2}\log(\frac{v_1+\beta}{\beta}) - a_1K_{(5.12)}\beta\int|x|^3\tilde{f}_1(x)dx \tag{5.21}$$

$$= \mathbf{H}(g*\mathbf{n}) + \frac{a_1}{2}\log(\frac{v_1+\beta}{\beta}) - K_{(5.12)}\beta\int|x|^3 f_1(x)dx \ .$$

We need to relate the quantities, $a_1\log(\frac{v_1+\beta}{\beta})$, a_2v_2 and $\log(\frac{v+\beta}{\beta})$. First we prove the plausible inequality,

$$v_1 \leq v \leq v_2 \ . \tag{5.22}$$

□ In fact, if $b<0<c$ and r, s are densities with mean zero supported in $[b,c]$ and $(b,c)^c$, respectively, we have,

$$(\frac{c-b}{2})^2 = \int(\frac{c-b}{2})^2 r(t)dt$$

$$\geq \int(x - \frac{c+b}{2})^2 r(t)dt$$

since $(\frac{c-b}{2})^2 \geq (x-\frac{c+b}{2})^2$ for $x\in[b,c]$,

$$= \mathbf{var}(r) + (\frac{c+b}{2})^2 \ ,$$

so

$$\mathbf{var}(r) \leq -bc \ ;$$

and in the same way, $\mathbf{var}(s) \geq -bc$ since $(\frac{c-b}{2})^2 \leq (x - \frac{c+b}{2})^2$ for $x \in [b,c]^c$. We then have $v_1 \leq v_2$, and since $v = a_1 v_1 + a_2 v_2$, we have $v_1 \leq v \leq v_2$. □

We now show that for any $1 \leq x \leq y$, $u, v > 0$, $u+v=1$,

(5.23) $\log(ux+vy) \leq u \log x + v(y-1)$.

□ Since $\log'(t) < 1$ for $t \geq 1$, we have,

$$\log(ux+vy) = \log(x+v(y-x)) \leq \log x + v(y-x).$$

Since $\log x \leq x-1$,

$$v \log x \leq v(x-1) .$$

Adding the two gives (5.23). □

Letting u,v,x,y be a_1, a_2, $\frac{v_1+\beta}{\beta}$, $\frac{v_2+\beta}{\beta}$ respectively,

$$\log(\frac{v+\beta}{\beta}) \leq a_1 \log(\frac{v_1+\beta}{\beta}) + a_2 \frac{v_2}{\beta}$$

$$\leq a_1 \log(\frac{v_1+\beta}{\beta}) + \int x^2 f_2(x) dx ,$$

and (5.21) now gives,

(5.24) $$\mathbf{H}(f * g * \mathbf{n}) \geq \mathbf{H}(g * \mathbf{n}) + \frac{1}{2} \log(\frac{v+\beta}{\beta}) - \frac{1}{2} \int x^2 f_2(x) dx$$

$$- K_{(5.12)} \beta \int |x|^3 f_1(x) dx .$$

We now relate the two integral terms in (5.24) to the one in the proposition we are proving.

□ Let $A = [-\beta^{-1}, \beta^{-1}]$, $B = [-\beta^{-1}, \alpha\beta^{-1}]$, $C = [\alpha\beta^{-1}, \beta^{-1}]$. We have,

$$\int_{-\infty}^{-\beta^{-1}} x^2 f(x) dx \geq \beta^{-1} \int_{-\infty}^{-\beta^{-1}} |x| f(x) dx$$

$$= \beta^{-1} \int_{\alpha\beta^{-1}}^{\infty} |x| f(x) dx$$

by (5.19),

$$\geq \beta^{-1} \int_C |x| f(x) dx$$

$$\geq \int_C x^2 f(x) dx ,$$

so *a fortiori*,

– $$\int_{A^c} x^2 f(x)dx \ge \int_C x^2 f(x)dx\ ,$$

– $$\beta\int_A |x|^3 f(x)dx + \int_{A^c} x^2 f(x)dx \ge \int_C (x^2 - \beta|x|^3) f(x)dx\ .$$

Adding the left hand side to both sides,

– $$2\left[\beta\int_A |x|^3 f(x)dx + \int_{A^c} x^2 f(x)dx\right]$$

– $$\ge \int_C (x^2 - \beta|x|^3) f(x)dx + \beta\int_A |x|^3 f(x)dx + \int_{A^c} x^2 f(x)dx$$

– $$= \beta\int_B |x|^3 f(x)dx + \int_{B^c} x^2 f(x)dx\ ,$$

– $$2\int \min(x^2, \beta|x|^3) f(x)dx \ge \int x^2 f_2(x)dx + \beta\int |x|^3 f_1(x)dx\ .\ \square$$

Using this in (5.24),

– $$\mathbf{H}(f*g*\mathbf{n}) \ge \mathbf{H}(g*\mathbf{n}) + \frac{1}{2}\log(\frac{v+\beta}{\beta})$$

$$-\,2\max(half, K_{(5.12)})\int \min(x^2, \beta^2|x|^3) f(x)dx\ ,$$

proving (5.18). $\square$

We now remove the restriction that the normal density in (5.18) have variance one by renormalization. This will be of great importance in the next section, in which we will have to use normal densities of small variances to get tight bounds on the distribution functions of sums of random variables. Let f, g be continuous probability densities such that

– $$\int xf(x)dx = \int xg(x)dx = 0\ .$$

Let $0 \le a,b,c \le 1$ and let $a = \mathbf{var}(f)$, $b = \mathbf{var}(g)$. Let $\gamma = \sqrt{c}$. We want an upper bound for the quantity $\frac{1}{2}\log(\frac{a+b+c}{b+c}) + \mathbf{H}(g*\mathbf{n}_c) - \mathbf{H}(f*g*\mathbf{n}_c)$.

Let $\tilde{f}(x) = \gamma f(\gamma x)$, $\tilde{g}(x) = \gamma g(\gamma x)$. By (5.18),

– $$\frac{1}{2}\log(\frac{\mathbf{var}(\tilde{f}*\tilde{g}*\mathbf{n})}{\mathbf{var}(\tilde{g}*\mathbf{n})}) + \mathbf{H}(\tilde{g}*\mathbf{n}) - \mathbf{H}(\tilde{f}*\tilde{g}*\mathbf{n})$$

$$\le K_{(5.18)}\int \min(x^2, (1+\mathbf{var}(\tilde{g}))|x|^3)\tilde{f}(x)dx$$

– $$= K_{(5.18)}\int \min(x^2, (1+\frac{b}{\gamma^2})|x|^3)\tilde{f}(x)dx\ .$$

We have,

$$-\qquad \log\left(\frac{\mathbf{var}(\tilde{f} * \tilde{g} * \mathbf{n})}{\mathbf{var}(\tilde{g} * \mathbf{n})}\right) = \log\left(\frac{\mathbf{var}(f * g * \mathbf{n}_c)}{\mathbf{var}(g * \mathbf{n}_c)}\right),$$

$$-\qquad \mathbf{H}(\tilde{g} * \mathbf{n}) - \mathbf{H}(\tilde{f} * \tilde{g} * \mathbf{n}) = \mathbf{H}(g * \mathbf{n}_c) - \mathbf{H}(f * g * \mathbf{n}_c),$$

$$-\qquad \int \min(x^2, (1+\frac{b}{\gamma^2})|x|^3)\tilde{f}(x)dx = \frac{1}{\gamma^2}\int \min(x^2, \frac{\gamma^2+b}{\gamma^3}|x|^3)f(x)dx,$$

and substituting,

$$(5.25)\qquad \frac{1}{2}\log\left(\frac{a+b+c}{b+c}\right) + \mathbf{H}(g * \mathbf{n}_c) - \mathbf{H}(f * g * \mathbf{n}_c)$$

$$\leq K_{(5.18)}\frac{1}{\gamma^2}\int \min(x^2, \frac{\gamma^2+b}{\gamma^3}|x|^3)f(x)dx.$$

We now show that (5.25) holds with the probability densities f, g replaced by arbitrary probability measures μ, η with finite variance.

□ Proof. Let $\varepsilon>0$, $a=\mathbf{var}(\mu)$, $b=\mathbf{var}(\eta)$ and $\gamma=\sqrt{c}$ satisfy $0<a,b,c<1$. Let $f=\mu * \mathbf{n}_\varepsilon$, $g=\eta * \mathbf{n}_\varepsilon$. By (5.25),

$$-\qquad \frac{1}{2}\log\left(\frac{a+b+c+2\varepsilon}{b+c+\varepsilon}\right) + \mathbf{H}(\eta * \mathbf{n}_\varepsilon * \mathbf{n}_c) - \mathbf{H}(\mu * \eta * \mathbf{n}_{2\varepsilon} * \mathbf{n}_c)$$

$$-\qquad \leq K_{(5.18)}\frac{1}{\gamma^2}\int \min(x^2, \frac{\gamma^2+b+\varepsilon}{\gamma^3}|x|^3)d(\mu * \mathbf{n}_\varepsilon)(x).$$

All the quantities above are continuous for $\varepsilon\geq 0$ and so the inequality also holds for $\varepsilon=0$:

$$(5.26)\qquad \frac{1}{2}\log\left(\frac{a+b+c}{b+c}\right) + \mathbf{H}(\eta * \mathbf{n}_c) - \mathbf{H}(\mu * \eta * \mathbf{n}_c)$$

$$\leq K_{(5.18)}\frac{1}{\gamma^2}\int \min(x^2, \frac{\gamma^2+b}{\gamma^3}|x|^3)d\mu(x). \quad \square$$

We are now ready to prove the main result of this chapter.

5.28. Theorem. Let $\{X_i,\ 1\leq i\leq k\}$ be independent random variables and suppose that $S=\sum_1^k X_j$ has variance 1. Let A be a normal random variable, independent of all the X_i, with variance $\gamma^2<1$. Then for a universal constant $K_{(5.28)}$, we have

$$-\qquad \mathbf{H}(\mathbf{n}_{A+S}) - \mathbf{H}(A+S) \leq \frac{K_{(5.28)}}{\gamma^5}\sum_1^k \mathbf{E}(\tau(X_i)),$$

where

$$-\qquad \tau(x) = \min(x^2, |x|^3).$$

□ Proof. We let S_0 be the zero RV; for $1\leq i\leq k$ we let $S_i=\sum_{j=1}^i X_i$ and $\mu=\mu_{X_i}$, $\eta=\mu_{S_{i-1}}$. By (5.26),

$$\frac{1}{2}\log\left(\frac{\mathbf{var}(A+S_i)}{\mathbf{var}(A+S_{i-1})}\right)+\mathbf{H}(A+S_{i-1})-\mathbf{H}(A+S_i)$$

$$\le \frac{K_{(5.18)}}{\gamma^2}\int \min\left(|x|^2, \frac{\gamma^2+\mathbf{var}(S_{i-1})}{\gamma^3}|x|^3\right)d\mu(x)\ ,$$

and since $\gamma<1$, $\mathbf{var}(S_{i-1})<1$, we have

$$\frac{1}{2}\log\left(\frac{\mathbf{var}(A+S_i)}{\mathbf{var}(A+S_{i-1})}\right)+\mathbf{H}(A+S_{i-1})-\mathbf{H}(A+S_i)$$

$$\le \frac{K_{(5.18)}}{\gamma^2}\int \min\left(|x|^2, \frac{2}{\gamma^3}|x|^3\right)d\mu(x)$$

$$\le \frac{2K_{(5.18)}}{\gamma^5}\int \min(x^2, |x|^3)d\mu(x)$$

$$= \frac{2K_{(5.18)}}{\gamma^5}\mathbf{E}(\tau(X_i))\ .$$

Summing over i,

$$\frac{1}{2}\log\left(\frac{\mathbf{var}(A+S)}{\mathbf{var}(A)}\right)+\mathbf{H}(A)-\mathbf{H}(A+S)\le \frac{2K_{(5.18)}}{\gamma^5}\sum_{i=1}^{k}\mathbf{E}(\tau(X_i))\ ,$$

$$\mathbf{H}(\mathbf{n}_{A+S})-\mathbf{H}(A+S)\le \frac{2K_{(5.18)}}{\gamma^5}\sum_{i=1}^{k}\mathbf{E}(\tau(X_i))\ ,$$

which proves (5.28) with $K_{(5.28)}=2K_{(5.18)}$. □

6. The rate of convergence of the probability distribution.

We show how the bound on entropy given in theorem (5.28) implies a bound on the cumulative distribution function of the convolution of many random variables.

Let f be a continuous probability density, $\mathbf{E}(f)=0$, $1\le\mathbf{var}(f)\le 2$, and $f(x)<M$ for $M>1$. Later we will let f be the density of $A+S$ from (5.28). By (3.3),

$$\mathbf{H}(\mathbf{n}_f)-\mathbf{H}(f)=\int f(x)\log\left(\frac{f(x)}{\mathbf{n}_f(x)}\right)dx$$

$$=\int \mathbf{n}_f(x)\frac{f(x)}{\mathbf{n}_f(x)}\log\left(\frac{f(x)}{\mathbf{n}_f(x)}\right)dx$$

$$=\int \mathbf{n}_f(x)\left[\frac{f(x)}{\mathbf{n}_f(x)}\log\left(\frac{f(x)}{\mathbf{n}_f(x)}\right)-\frac{f(x)}{\mathbf{n}_f(x)}+1\right]dx,$$

so letting

$$\theta(t)=t\log t-t+1\ ,$$

we have

(6.1) $\qquad \mathbf{H}(\mathbf{n}_f) - \mathbf{H}(f) = \int \mathbf{n}_f(x)\theta(\frac{f(x)}{\mathbf{n}_f(x)})\, dx$.

We will show that for any $R > 1$,

(6.2) $\qquad \theta(t) \geq \frac{1}{2e^2} \frac{\log R}{R} (t-1)^2, \quad 0 \leq t \leq R$.

☐ Suppose $t \leq e^2$. We have

– $\qquad \theta'(t) = \log(t)\,, \quad \theta''(t) = \frac{1}{t} \geq \frac{1}{e^2}$,

and $\theta(1) = \theta'(1) = 0$. By Taylor's theorem,

– $\qquad \theta(t) = \frac{1}{2}\, \theta''(y)(t-1)^2$

for some $y = y(t) \leq e^2$,

– $\qquad \geq \frac{(t-1)^2}{2e^2}$.

Since $R > 1$ we have $\frac{\log R}{R} < 1$ and

– $\qquad \theta(t) \geq \frac{1}{2e^2} \frac{\log R}{R} (t-1)^2$.

On the other hand, suppose $t > e^2$. Then $\frac{t \log t}{2} \geq t$ and

– $\qquad \theta(t) \geq \frac{t \log t}{2} \geq \frac{\log t}{2t} (t-1)^2$

– $\qquad \geq \frac{\log R}{2R} (t-1)^2$,

since $\frac{\log t}{t}$ is decreasing for $t > e$. ☐

Letting $t = \frac{f(x)}{\mathbf{n}_f(x)}$, $R = \frac{M}{\mathbf{n}_f(x)}$ in (6.2),

– $\qquad \theta(\frac{f(x)}{\mathbf{n}_f(x)}) \geq \frac{1}{2e^2} \frac{\log(\frac{M}{\mathbf{n}_f(x)})}{\frac{M}{\mathbf{n}_f(x)}} (\frac{f(x)}{\mathbf{n}_f(x)} - 1)^2$

– $\qquad = \frac{1}{2e^2 M \mathbf{n}_f(x)} (\log(M) + \log(\frac{1}{\mathbf{n}_f(x)}))(f(x) - \mathbf{n}_f(x))^2$.

Since $1 \leq \mathbf{var}(f) \leq 2$,

– $\qquad \log(\mathbf{n}_f(x)) \leq \log(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{4}}) \leq -\frac{x^2+1}{4}$,

and $M > 1$. Thus,

$$\theta\left(\frac{f(x)}{\mathbf{n}_f(x)}\right) \geq \frac{1}{8e^2 M \mathbf{n}_f(x)} \, (1+x^2)\,(f(x) - \mathbf{n}_f(x))^2 \, .$$

Equation (6.1) now becomes,

(6.3) $$\mathbf{H}(\mathbf{n}_f) - \mathbf{H}(f) \geq \frac{K_{(6.3)}}{M} \int (1+x^2)(f(x) - \mathbf{n}_f(x))^2 dx \, ,$$

where

$$K_{(6.3)} = \frac{1}{8e^2} \, .$$

We now relate the RHS of (6.3) to the quantity $\int |f(x) - \mathbf{n}_f(x)| dx$.

□ By the Cauchy inequality,

$$\int (1+x^2) |g(x)| \frac{dx}{1+x^2} \leq \left[\int \frac{dx}{1+x^2} \right]^{1/2} \left[\int (1+x^2)^2 |g(x)|^2 \frac{dx}{1+x^2} \right]^{1/2} ,$$

$$\int |g(x)| dx \leq \sqrt{\pi} \left[\int (1+x^2)(g(x))^2 dx \right]^{1/2} . \ \square$$

Now we let

$$F(x) = \int_{-\infty}^{x} f(t) dt$$

denote the cumulative distribution associated with the density f. Then,

(6.4) $$|F(t) - \mathbf{N}_f(t)| \leq \int |f(x) - \mathbf{n}_f(x)| dx$$

$$\leq \sqrt{\pi} \left[\int (1+x^2)(g(x))^2 dx \right]^{1/2}$$

$$\leq K_{(6.4)} \sqrt{M} (\mathbf{H}(\mathbf{n}_f) - \mathbf{H}(f))^{1/2} ,$$

where $K_{(6.4)} = \left(\frac{\pi}{K_{(6.3)}}\right)^{1/2}$.

Now suppose that $f = g * \mathbf{n}_{\gamma^2}$, where g is a probability density with mean 0 and variance 1, and $\gamma \leq 1$. Let $G(x) = \int_{-\infty}^{x} g(t) dt$. Then since $\mathbf{n}_{\gamma^2}(x) < \frac{1}{\gamma}$, $|f(x)| < \frac{1}{\gamma}$ and by (6.4),

$$|F(x) - \mathbf{N}_f(x)| \leq \eta \, ,$$

where

(6.5) $$\eta = \frac{K_{(6.4)}}{\sqrt{\gamma}} (\mathbf{H}(\mathbf{n}_f) - \mathbf{H}(f))^{1/2} .$$

A bound on the difference $F - \mathbf{N}_f$ allows us to bound $G - \mathbf{N}_g$ as follows.

□ Let $\beta > 0$, $S = [-\beta, \beta]$ and let

(6.6) $$\varepsilon = \int_{S^c} \mathbf{n}_{\gamma^2}(x)dx$$

$$= 2\int_{\frac{\beta}{\gamma}}^{\infty} \mathbf{n}(x)dx$$

$$\le \sqrt{\frac{2}{\pi}}\ \frac{\gamma}{\beta}\ e^{-\frac{1}{2}(\frac{\beta}{\gamma})^2} .$$

Let $y(x)$ be defined so that $|y(x)| \le \beta$ and

$$\int_S G(x-t)\mathbf{n}_{\gamma^2}(t)dt = G(x+y(x))\int_S \mathbf{n}_{\gamma^2}(t)dt ,$$

and hence

$$G*\mathbf{n}_{\gamma^2}(x)-\varepsilon \le \int_S G(x-t)\mathbf{n}_{\gamma^2}(t)dt$$

$$=G(x+y(x))\int_S \mathbf{n}_{\gamma^2}(t)dt$$

$$= (1-\varepsilon)G(x+y(x)) \le G(x+y(x)) .$$

Similarly choose $z(x)$ so that $|z(x)| \le \beta$ and

$$\mathbf{N}_g * \mathbf{n}_{\gamma^2}(x) \ge \mathbf{N}_g(x+z(x))-\varepsilon .$$

We now have,

$$G(x) \ge G(x-\beta+y(x-\beta))$$

by monotonicity of G,

$$\ge (G*\mathbf{n}_{\gamma^2})(x-\beta)-\varepsilon$$

by definition of $y(x)$,

$$\ge (\mathbf{N}_g * \mathbf{n}_{\gamma^2})(x-\beta)-\varepsilon-\eta$$

by choice of η,

$$\ge (\mathbf{N}_g * \mathbf{n}_{\gamma^2})(x+\beta)-\varepsilon-\eta-2\beta$$

since $\mathbf{N}_g{}'(x)<1$,

$$\ge (\mathbf{N}_g)(x+\beta+z(x+\beta))-2\varepsilon-\eta-2\beta$$

by definition of $z(x)$,

$$\ge (\mathbf{N}_g)(x)-2\varepsilon-\eta-2\beta$$

by monotonicity of $\mathbf{N}_g$. Similarly we have,

$$G(x) \le (\mathbf{N}_g)(x)+2\varepsilon+\eta+2\beta ,$$

and combining,

– $|G(x) - (\mathbf{N}_g)(x)| \le 2\varepsilon+\eta+2\beta$.

Now we choose

– $\beta=\gamma\sqrt{2-2\log\gamma}$,

where the quantity inside the radical is positive since $\gamma<1$. By (6.6),

– $$\varepsilon \le \sqrt{\frac{2}{\pi}}\ \frac{1}{\sqrt{2-2\log\gamma}}\ e^{-(1-\log\gamma)} \le \frac{\gamma}{e\sqrt{\pi}} \ .$$

Plugging in β, ε and (6.5), and using $\gamma\le 1$ and $\beta\le\sqrt{2}\,\gamma(1-\log\gamma)$,

(6.7) $$|G(x) - (\mathbf{N}_g)(x)| \le \frac{K_{(6.4)}}{\sqrt{\gamma}}\ (\mathbf{H}(\mathbf{n}_f)-\mathbf{H}(f))^{\frac{1}{2}} + K_{(6.7)}\gamma(1-\log(\gamma)) \ ,$$

where

– $$K_{(6.7)} = (2\sqrt{2}+\frac{2}{e\sqrt{\pi}}) \ . \ \square$$

Now take $\{X_i\}$, S, and γ as in theorem 5.28, and let G be the probability distribution of S. Let

(6.8) $$\delta = \sum_1^k \mathbf{E}(\tau(X_i) \ ,$$

where

– $\tau(x) = \min(x^2, |x|^3)$.

By theorem 5.28,

– $\mathbf{H}(\mathbf{n}_{A+S}) - \mathbf{H}(A+S) \le \gamma^{-5}\delta$,

and by (6.7),

– $|G(x) - (\mathbf{N}_g)(x)| \le K_{(6.4)}\gamma^{-3}\delta^{\frac{1}{2}} + K_{(6.7)}\gamma(1-\log(\gamma))$.

Setting $\gamma=\delta^{\frac{1}{8}}$ gives,

(6.9) $$|G(x) - (\mathbf{N}_g)(x)| \le K_{(6.9)}\delta^{\frac{1}{8}}(1-\log(\delta)) \ ,$$

where

– $K_{(6.9)} = 8(K_{(6.7)}+K_{(6.4)})$.

This proves 2.7 for $\tau(x) = \min(x^2, |x|^3)$, with the explicit formula $\delta=h^{-1}(\varepsilon)$ where $h(t)=K_{(6.9)}t^{1/8}(1-\log t)$ for $t<1$.

Appendix. Comparison with the methods of Linnick and Brown.

In the introduction we mentioned information-theoretic proofs of the CLT due to Linnick and Brown. The purpose of this appendix is to sketch them in greater detail, showing the similarities and contrasts with the proofs given here. No effort has been made to make these proofs either similar or dissimilar to the previous ones; in fact this author was unaware of them until late in this work. But this subject still has potential for further work, and this appendix might be helpful to future investigators who wish to make the most effective use of what has been done before them.

We start with Linnick, who was the first to make a connection between information theory and limit theorems for stochastic processes. Throughout this discussion, we will assume that the $\{X_i\}$ are independent RVs with mean 0 and finite variance, but we will drop the assumption that $\mathrm{var}(S_n)=1$. We recall the definition $\overline{\mathbf{H}}(X)=\mathbf{H}(X)-\mathbf{H}(\mathbf{n}_X)$, and now set $\bar{J}(X)=J(\frac{X}{\sigma(X)})-1$, which Linnick shows to be nonnegative. In discussing Linnick we assume as he has that the X_n are independent sums of bounded RVs with normal ones. As in the second proof given here, Linnick worked incrementally, considering the effect of each X_{n+1} on S_n, proving a result relating $\mathbf{H}(S_{n+1})$ to $\mathbf{H}(S_n)$. Given any $\delta,\tau>0$ we say that X_i satisfies the Lindeberg condition "individually" if

$$ - \qquad \int_{\{|x|>\tau\sigma(S_i)\}} x^2\, d\mu_{X_i}(x) < \delta . $$

Late in the paper it is mentioned that, if the real Lindeberg condition is met then for any $\delta,\tau>0$ we can show that for n sufficiently large the X_i for $i<n$ which fail to meet the Lindeberg condition individually contribute only a small portion of the variance of S_n, and hence may be dropped without changing the outcome of the theorem; in this way we can assume that all the X_i indeed meet the Lindeberg condition individually. In this case, Linnick's main lemma becomes (in the language of this work) that for any $\eta>0$ we can choose $\delta,\tau>0$ such that if all the X_i meet the Lindeberg condition individually for δ,τ, we have

$$ - \qquad \overline{\mathbf{H}}(S_{n+1}) - \overline{\mathbf{H}}(S_n) = \frac{\sigma^2(X_{n+1})}{2\sigma^2(S_n)}\,[\bar{J}(S_n)+\varepsilon] , $$

for some ε with $|\varepsilon|<\eta$. This plays a role comparable to that played in this work by equation (5.26), which telescopes into theorem 5.27.

Linnick uses this lemma to prove the CLT as follows. First he shows that if Y is a continuous RV with mean 0 and variance 1, if $\bar{J}(Y)$ is small, then so are $\sup\{|\mathrm{Prob}\{Y<x\}-\mathbf{N}(x)|\}$, and hence $|\bar{\mathbf{H}}(Y)|$. Since $\bar{\mathbf{H}}(S_n)$ is bounded above, $\bar{J}(S_n)$ must eventually reach a small value $\eta_2(\eta)$ (but not necessarily maintain it.) Hence $|\bar{\mathbf{H}}(S_n)|$ reaches another small value $\eta_3(\eta)$. Linnick then proves by induction on n that, possibly reducing η more, $|\bar{\mathbf{H}}(S_n)|$ can be made to remain below η_3; for if $\bar{J}(X_n)$ is smaller than η than $0\geq\bar{\mathbf{H}}(S_{n+1})\geq\bar{\mathbf{H}}(S_n)$, but otherwise X_n is nearly normal and hence so is X_{n+1}, limiting $|\mathbf{H}(S_{n+1})|$. Linnick could then immediately conclude that the rescaled probability distributions of the $\{X_n\}$ approach the normal distribution, but instead he goes back through his main lemma to show that $\bar{J}(S_n)$ is then small sufficiently often to prove the conclusion of the CLT. He appears only to have learned late in his own work that a rigorous proof existed that $\bar{\mathbf{H}}(X)\leq 0$.

The proof in [Brown] is more direct. He works in the i.i.d. case, using the quantity J. If f is a differentiable probability density with finite variance such that $J(f)<\infty$ (which we can guarantee by the usual method of convolution by an independent normal density,) and if $f_1=f$, $f_{n+1}=f_n * f_n$ for $n\geq 1$, then by the same calculation as in the proof of (3.8), we have

$$2J(f_n) - 4J(f_{n+1}) = \iint\left[\frac{f_n'(x)}{f_n(x)} + \frac{f_n'(y)}{f_n(y)} - 2\frac{f_{n+1}'(x+y)}{f_{n+1}(x+y)}\right]^2 f_n(x)dx\, f_n(y)dy\ .$$

Setting d_n to this quantity, we have $d_n^2\geq 0$ with equality *iff* f_n is Gaussian. Moreover, since $J(f_n)\geq\dfrac{1}{\mathbf{var}(f_n)}$, and since

$$0\leq d_n = 2\left(J(f_n)-\frac{1}{\mathbf{var}(f_n)}\right) - 4\left(J(f_{n+1})-\frac{1}{\mathbf{var}(f_{n+1})}\right),$$

we conclude that $d_n\to 0$. Brown uses the integral expression for d_n and an argument using Hermite polynomials to show that the f_n (appropriately rescaled) approach the normal density.

The two proofs given in this work are similar to these earlier ones in the use of information-theoretic quantities and inequalities which have extremal properties for normal random variables. The convolution inequalities used by Brown are also used here; in the first proof the use of a convolution inequality is especially evident in the assumption that $D(x*y)\leq\min(D(x),D(y))$. The first CLT proof here is more closely related to that of Brown than that of Linnick. The most significant difference appears to be in the use of a compactness argument, like those of Renyi and Csizar. Thus we arrive at a conclusion such as $(X_n,Y_n,Z_n)\to(X,Y,Z)$ where X,Y,Z satisfy the equality condition of a convolution inequality and hence are normal, instead of calculating from the fact that $\bar{J}(X_{n+1})-\bar{J}(X_n)\rightarrow 0$ that X_n must approach a normal RV in distribution. The avoidance of the need for analysis of such error terms in the convolution inequalities has come at the cost of proposition 4.5. Essentially, if we want to make a compactness argument, we must insure that the limiting distributions are not point mass distributions. It is nessary to prove that if the probability measures of the $\dfrac{S_n}{\mathrm{var}(S_n)}$ approach a limit, that that limit has variance 1. It is interesting that in a proof of the CLT, the central difficulty should be to avoid such tail behavior that the limiting distribution might have the wrong variance.

References.

Araujo, A. and Giné, E., *The Central Limit Theorem for Baanach-valued random variables.* Wiley 1980 (Wiley series in probability and mathematical statistics.)

Billingsley, P., *Convergence of Probability Measures.* Wiley 1968 (Wiley series in probability and mathematical statistics.)

Baron, A., to appear in *Annals of Probability Jan. 1986.*

Blachman, N. M., The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*. Apr 1965, pp 267-271.

Brown, L., A proof of the Central Limit Theorem motivated by the Cramér-Rao inequality. In Kalianpur, Krishnaiah, Ghosh, eds., *Statistics and Probability: essays in honor of C. R. Rao,* North-Holland 1982.

Csiszár, I., A note on limiting distributions on topological groups. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* 1964, pp 595-599

Kemperman, J. H. B., On the optimum rate of transmitting information. In *Lecture Notes in Mathematics* 89, pp. 126-169.

Linnick. An information-theoretic proof of the Central Limit Theorem with the Lindeberg condition. *Theory of Probability and its Applications.* 4(1959) pp 288-299.

Renyi, A. On measures of entropy and information. 4th Berkeley symposium on mathematical statistics and probability, 1961, v 2 pp 251-280.

Shannon, C. E., A mathematical theory of communication. *Bell System Technical Journal,* Vol. 27, Oct. 1948, pp 623-656.

Stam, A. J., Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control,* Vol. 2, June 1959, pp 101-112.