MONOTONIC CENTRAL LIMIT THEOREM FOR DENSITIES

by

Andrew R. Barron

TECHNICAL REPORT NO. 50

March 5, 1984

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION

GRANT ECS82-11568

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

MONOTONIC CENTRAL LIMIT THEOREM FOR DENSITIES[1]

Andrew R. Barron

Stanford University

ABSTRACT

The probability density function f_n for the normalized sum (of independent and identically distributed random variables with finite variance) converges to the normal density ϕ in a natural sense. It is shown that the Kullback-Leibler divergence (relative entropy) $\int f_n \log f_n/\phi$ converges to zero as $n \to \infty$, provided the divergence is finite for some n. Furthermore, the convergence is monotone along the powers of two subsequences $n_k = 2^k n_0$. This result extends classical limit theorems for densities. The proof does not involve the usual Fourier transform technique, but follows instead from fundamental properties of Shannon entropy and Fisher information.

AMS 1980 subject classifications. Primary 60F05; secondary 94A17, 62B10.

[1] This work was partially supported by NSF Contract ECS 82-11568.

Key words and phrases. Central limit theorem, local limit theorem, Kullback-Leibler divergence, Shannon entropy, Fisher information, convolution inequalities.

1. Introduction

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean μ, variance σ^2, and probability density function $f(x)$. Define the normalized sums

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu). \tag{1.1}$$

The probability density function for S_n is denoted by $f_n(x)$ (the normalized n-fold convolution of f). Let Z be a normal random variable with mean zero and variance σ^2. The normal distribution has the probability density function

$$\phi(x) = (2\pi\sigma^2)^{-1/2} e^{-x^2/2\sigma^2}. \tag{1.2}$$

A central limit theorem for densities (also referred to as a local limit theorem) is an assertion that f_n converges to ϕ in some sense as $n \to \infty$. Prohorov (1952) established L^1 convergence: $\int |f_n - \phi| \to 0$. Ranga Rao and Varadarajan (1960) established pointwise convergence: $f_n(x) \to \phi(x)$ for (Lebesgue) almost every x. Gnedenko (1954) (see also Kolmogorov and Gnedenko 1954, p.224) established uniform convergence: $(\text{ess})\sup_x |f_n(x) - \phi(x)| \to 0$, provided f_n is (essentially) bounded for some n. The classical proofs of these results rely on properties of the Fourier transform of the density.

This paper extends the central limit theorem by establishing convergence of the Kullback-Leibler divergence (relative entropy): $\int f_n \log f_n/\phi \to 0$, provided the divergence is finite for some n. Furthermore, the convergence is monotone along the powers of two subsequences $n_k = 2^k n_0$. Our proof follows directly from fundamental properties of Shannon (differential) entropy and Fisher information. The classical results of L^1 convergence of the density and convergence in distribution are established as corollaries, via the inequality $(\int |f_n - \phi|)^2 \leq 2 \int f_n \log f_n/\phi$.

The monotonicity suggests that the Kullback-Leibler divergence is a natural measure of discrepancy between two densities. Indeed, the divergence arises naturally in hypothesis testing (as the asymptotic rate of probability of error -- see Chernoff 1956) and in information theory (as the least upper bound to the asymptotic redundancy of Shannon codes). For general properties of the Kullback-Leibler divergence (also called relative entropy, informational divergence, discrimination information, etc.) see Kullback and Leibler (1951), Kullback (1959,1967), Pinsker (1964), and Csiszár (1967,1975).

The Kullback-Leibler divergence can be expressed as the difference of the Shannon entropy from the maximum entropy for the given variance (attained by the normal density). Thus the monotone decrease of the divergence (to zero) is equivalent to the monotone increase of the entropy (to the entropy of the normal). This characterization of the central limit theorem resembles the second law of thermodynamics.

The outline of our convergence proof is as follows. We show that the Kullback-Leibler divergence of f_n from ϕ is equal to an integral of the difference between the Fisher information and the reciprocal of the variance. (The Fisher information is for f_n convolved with a normal density. The integration is with respect to the variance of the added normal.) The monotone convergence of the Fisher information (to the reciprocal of the variance) is established using the methods of Brown (1982). If the divergence is finite for some n, then by the monotone convergence theorem, the divergence has limit zero.

Linnik (1959) used the information measures of Shannon and Fisher in a proof of convergence in distribution. Rényi (1970, p.601) states that Linnik (1959) established convergence of $\int f_n \log f_n/\phi$ to zero. A reading of Linnik (1959) reveals that convergence was established only for densities of truncated random variables smoothed by addition of independent normal random variables. We show that the divergence $\int f_n \log f_n/\phi$ converges to zero (provided that the divergence is finite for some n). No smoothness conditions are required of f_n for this convergence to hold.

Recently Brown (1982) gave an elegant proof of convergence in distribution. Brown provided a mean squared error interpretation of the monotone decrease in Fisher information. He examined the Fisher information of f_n convolved with a normal density of arbitrary variance. Using properties of the mean squared error, he showed that these smoothed densities converge pointwise (and uniformly on compact subsets) to the normal density. We extend Brown's technique to show that the Fisher informations converge to the reciprocal of the variance.

Motivation for this effort came from a convolution inequality for entropy power proposed by Shannon (1948). Stam (1959) and Blachman (1965) give convolution inequalities for Fisher information; relate a derivative of entropy to Fisher information; and prove the entropy power inequality. The entropy power inequality shows that the divergence is monotone decreasing (along the powers of two) and hence has a limit. To

identify that the limit is zero we use the integral relationship between divergence and Fisher information. This relationship also establishes the monotonicity of divergence, so it is not necessary to invoke the entropy power inequality.

Section 2 relates Shannon entropy and Fisher information. Convolution inequalities for entropy and information are presented in section 3. Section 4 treats the convergence of Fisher information. The convergence of entropy and divergence is established in section 5. Section 6 proves a Lemma involving the derivative of entropy. Finally, section 7 contains examples of densities for which the entropy either remains infinite of becomes finite.

2. Entropy and information

Let X be a random variable with probability measure F (for Borel subsets on the real line). Let Z be a normal random variable with probability measure Φ and probability density function $\phi(x)$. If F has a density function $f(x)$, the Kullback-Leibler divergence of F from Φ is

$$D(F\,||\,\Phi) = D(f\,||\,\phi) = \int f(x) \log f(x)/\phi(x)\, dx; \tag{2.1}$$

otherwise, if F has a singular component, the divergence is defined to be $D(F\,||\,\Phi) = +\infty$. (Here log denotes natural logarithm. We adopt the convention $0 \log 0 = 0$. The divergence always exists, since $D(f\,||\,\phi) = \int (f/\phi) \log (f/\phi)\, d\Phi$ and $u \log u \geq e^{-1}$ for $u \geq 0$.) By Jensen's inequality, the divergence satisfies $D(F\,||\,\Phi) \geq 0$ with equality if and only if $F \equiv \Phi$.

If π is a finite partition of the real line, the corresponding discrete divergence is

$$D_\pi(F\,||\,\Phi) = \sum_{A \in \pi} F(A) \log F(A)/\Phi(A). \tag{2.2}$$

Dobrushin (see Pinsker 1964) established that

$$D(F\,||\,\Phi) = \sup_\pi D_\pi(F\,||\,\Phi), \tag{2.3}$$

where we may restrict the partitions to consist of sets in a generating field.

The divergence is lower semicontinuous with respect to convergence in distribution. Specifically, if F_t is a family of probability measures with weak limit F_{t_0} as $t \to t_0$, then

$$\liminf_{t \to t_0} D(F_t\,||\,\Phi) \geq D(F_{t_0}\,||\,\Phi). \tag{2.4}$$

Lower semicontinuity follows immediately from (2.3) and the continuity of $D_\pi(F_t\,||\,\Phi)$ when π consists of sets with boundary measure zero.

Suppose X has finite variance σ^2. Let $D(X) = D(F\,||\,\Phi)$ denote the divergence of $X \sim F$ from the normal $Z \sim \Phi$ with the same mean and variance as X. Then

$$\begin{aligned} D(X) &= H(Z) - H(X) \\ &= \frac{1}{2} \log 2\pi e\sigma^2 - H(X). \end{aligned} \tag{2.5}$$

Here H is the Shannon differential entropy defined by

$$H(X) = -\int f(x) \log f(x)\, dx \tag{2.6}$$

if X has a density $f(x)$ and $H(X) = -\infty$ otherwise. Equation (2.5) trivially holds if the distribution of X has a singular component. Suppose X has a density $f(x)$. Then (2.5) follows from (2.1), since $\int f \log \phi = \int \phi \log \phi$ where ϕ is the normal density with the same mean and variance as f.

By the positivity of the divergence, the normal has maximum entropy for a given variance, i.e.,

$$H(X) \leq H(Z) = \frac{1}{2} \log 2\pi e\sigma^2. \tag{2.7}$$

Note that the differential entropy is translation invariant

$$H(X+c) = H(X), \tag{2.8}$$

but not scale invariant

$$H(cX) = H(X) + \log |c|; \tag{2.9}$$

whereas the divergence $D(X)$ is both translation and scale invariant.

If X_1 and X_2 are independent random variables, then

$$H(X_1 + X_2) \geq H(X_1) \tag{2.10}$$

with equality if and only if X_2 is almost surely constant. This inequality trivially holds if $H(X_1) = -\infty$. Suppose X_1 has a density f with $H(X_1) > -\infty$. The density of $Y = X_1 + X_2$ is $g(y) = E\, f(y-X_2)$. (Here y is fixed; the expectation is taken with respect to the capitalized variable.) By Jensen's inequality we have $-g(y) \log g(y) \geq -E\, f(y-X_2) \log f(y-X_2)$ (with equality if and only if X_2 is almost surely constant). Integrating yields (2.10).

We shall see that the Shannon entropy of X is related to the Fisher informations of random variables of the form $X + Z$ where Z is an independent normal random variable. Before we introduce properties of the Fisher information note the following properties of convolution with a normal distribution. The random variable $Y = X+Z$ has density $g(y) = E\, \phi(y-X)$ (even if X does not have a density). Furthermore, the normal density $\phi(z)$ has a bounded derivative $\phi'(z)$, so the density for Y has a bounded derivative $g'(y) = E\, \phi'(y-X)$. (The implicit exchange of limit and expectation is valid by application of mean value and bounded convergence theorems.) Note

that the given version of the density $g(y)$ has a derivative $g'(y)$ for *every* y (not only almost everywhere).

Let Y be a random variable with everywhere differentiable density $g(y)$. The "score" function for Y is $\rho(y) = g'(y)/g(y) = (d/dy) \log g(y)$. The Fisher information (for the location family) is defined as the expected square of the score,

$$I(Y) = E\,(g'(Y)/g(Y))^2. \tag{2.11}$$

Note that the Fisher information is also translation invariant

$$I(Y+c) = I(Y), \tag{2.12}$$

but not scale invariant

$$I(cY) = I(Y)/c^2. \tag{2.13}$$

If Y has variance σ^2, define the normalized Fisher information by

$$J(Y) = \sigma^2\, I(Y) - 1. \tag{2.14}$$

This normalized Fisher information is both translation and scale invariant.

The Fisher information (for an everywhere differentiable density $g(y)$ with variance σ^2) satisfies the Cramér-Rao inequality

$$\sigma^2\, I(Y) \geq 1 \tag{2.15}$$

or equivalently

$$J(Y) \geq 0 \tag{2.16}$$

with equality if and only if Y is normal. This inequality is well known under stronger conditions (see, e.g., Cramér 1946, p.475 or Pitman 1979, p.32 and p.37). A proof using everywhere differentiability only is as follows. The inequality trivially holds if $\sigma^2\, I(Y)$ is infinite. Suppose $\sigma^2\, I(Y)$ is finite. Without loss of generality assume that Y has zero mean. Note that $g(y) = 0$ implies $g'(y) = 0$ (since there g attains its minimum). Thus $\int |yg'(y)|\,dy = E\,|Yg'(Y)/g(Y)|$. By the Cauchy-Schwartz inequality this integral is less than or equal to $(\sigma^2\, I(Y))^{1/2}$ (with equality if and only if $g'(Y)/g(Y)$ is proportional to Y, i.e. g normal). Hence $yg'(y)$ is integrable and $|\int yg'(y)dy| \leq (\sigma^2 I(Y))^{1/2}$. It remains to show $\int yg'(y)dy = -1$. The derivative $(d/dy)\, yg(y) = yg'(y) + g(y)$ exists everywhere and is integrable. Consequently $yg(y)$ is absolutely continuous (see Rudin 1974, p.179). Now $yg(y) \to 0$ as $|y| \to \infty$ (since $E\,|Y|$ is

finite), so integrating the derivative yields $0 = \int (d/dy)\, yg(y)\, dy$ $= \int yg'(y)dy + \int g(y)dy$. Consequently $\int yg'(y)dy = -1$ and this completes the proof. (A similar argument shows that $\int g'(y)dy = E\, g'(Y)/g(Y) = 0$ whenever $I(Y)$ is finite.)

The normalized Fisher information (for an everywhere differentiable density $g(y)$) can be expressed as an L^2 distance between the scores $g'(y)/g(y)$ and $\phi'(y)/\phi(y)$. Here ϕ is the normal density with the same mean and variance as g . The normal is the unique density with linear score. Expanding the square we find

$$J(Y) = \sigma^2\, E\, (\, \frac{g'(Y)}{g(Y)} - \frac{\phi'(Y)}{\phi(Y)}\,)^2 \tag{2.17}$$

Let Y_1 and Y_2 be independent random variables and suppose Y_1 has a density g with bounded derivative, then

$$I(Y_1 + Y_2) \leq I(Y_1) \tag{2.18}$$

with equality if and only if Y_2 is almost surely constant. This inequality trivially holds if $I(Y_1)$ is infinite. Suppose $I(Y_1)$ is finite. The sum $S = Y_1 + Y_2$ has a density $g_*(s) = E\, g(s-Y_2)$ with bounded derivative $g_*'(s) = E\, g'(s-Y_2)$. The score for the sum S is the conditional expectation of the score for the summand Y_1 . Specifically, $g_*'(S)/g_*(S) = E[g'(Y_1)/g(Y_1)\,|\,S]$. Hence $(g_*'(S)/g_*(S))^2 \leq E[(g'(Y_1)/g(Y_1))^2\,|\,S]$ (with equality if and only if Y_2 is almost surely constant). Taking the expectation yields (2.18).

The following lemma relates Shannon entropy and Fisher information.

Lemma 2.1

Let X be a random variable with finite variance σ^2. Let Z be an independent normal random variable with the same variance as X. The divergence $D(X) = H(Z) - H(X)$ is equal to an integral of normalized Fisher informations,

$$D(X) = \frac{1}{2} \int_0^1 J(\, \sqrt{1-t}X + \sqrt{t}Z\,) \frac{dt}{1-t}$$

$$= \frac{1}{2} \int_0^\infty J(\, X + \sqrt{\tau}Z\,) \frac{d\tau}{1+\tau}. \tag{2.19}$$

Proof

Let $Y_t = \sqrt{1-t}X + \sqrt{t}Z$. For each $t \in (0,1]$, the density $g_t(y)$ for Y_t is everywhere differentiable with respect to y, so by the Cramér-Rao inequality $J(Y_t)$ is nonnegative. Thus the integral $\int_0^1 J(Y_t)\, dt/1-t$ exists, although it is possibly infinite. The normal density has bounded derivative, so by (2.18) the Fisher information satisfies $I(Y_t) \leq I(\sqrt{t}Z) = 1/t\sigma^2$. Thus the integrand $J(Y_t)/1-t = (\sigma^2 I(Y_t) - 1)/1-t$ is bounded above by $1/t$. The entropy of Y_t satisfies $H(\sqrt{t}Z) \leq H(Y_t) \leq H(Z)$. Thus $H(Y_t)$ is finite for $t \in (0,1]$. In section 6 we verify that $H(Y_t)$ is differentiable with respect to t for $t \in (0,1)$ and

$$\frac{d}{dt} H(Y_t) = \frac{1}{2} J(Y_t)/1-t. \tag{2.20}$$

This derivative exists and is bounded for all $t \in [a,b]$ where $0<a<b<1$, so we may integrate the derivative to obtain

$$H(Y_b) - H(Y_a) = \frac{1}{2} \int_a^b J(Y_t) \frac{dt}{1-t}. \tag{2.21}$$

Now $0 \leq H(Z) - H(Y_b) \leq (1/2) \log b$, so $H(Y_b) \to H(Z)$ as $b \to 1$. Also $H(Y_a) \geq H(\sqrt{1-a}X) = H(X) + (1/2) \log (1-a)$, so $\liminf H(Y_a) \geq H(X)$ as $a \to 0$. Note that $Y_a \to X$ in probability (and hence in distribution) as $a \to 0$. By lower semicontinuity of the divergence, $\liminf D(Y_a) \geq D(X)$. Consequently, $\limsup H(Y_a) \leq H(X)$. Thus $H(Y_a) \to H(X)$ as $a \to 0$ (even if $H(X) = -\infty$). If the integral on (0,1) is finite, then letting $a \to 0$ and $b \to 1$ in (2.21) we obtain

$$H(Z) - H(X) = \frac{1}{2} \int_0^1 J(Y_t) \frac{dt}{1-t}. \tag{2.22}$$

If the integral on (0,1) is infinite, then by Fatou's lemma the left side is also infinite (i.e., $H(X) = -\infty$). Thus (2.22) holds regardless. The second integral in (2.19) follows by changing variables to $\tau = t/1-t$. This completes the proof of Lemma 2.1.

Note that Lemma 2.1 can be interpreted as expressing the Shannon entropy $H(X)$ in terms of the Fisher informations $I(X + \sqrt{\tau}Z)$,

$$H(X) = \frac{1}{2} \log 2\pi e\sigma^2 - \frac{\sigma^2}{2} \int_0^\infty \left(I(X + \sqrt{\tau}Z) - \frac{1}{(1+\tau)\sigma^2}\right) d\tau. \tag{2.23}$$

3. Convolution inequalities

Convolution inequalities are easily established for Fisher information. Because of Lemma 2.1, analogous inequalities also hold for Shannon entropy. Using these convolution inequalities, we find that the entropy $H(S_n)$ and the Fisher information $I(S_n + \sqrt{\tau}Z)$ are monotone along the powers of two subsequences $n_k = 2^k n_0$ and "nearly" monotone for the entire sequence.

Recall from section 2 that the score function for a differentiable density $g(y)$ is defined by $\rho(y) = g'(y)/g(y)$. Also recall that the score for the sum of independent random variables is the conditional expectation for the score of either summand (provided the density of each summand has a bounded derivative).

Suppose that independent random variables Y_1 and Y_2 have densities (g_1 and g_2 respectively) with bounded derivatives. Let $\alpha_1, \alpha_2 \geq 0$, $\alpha_1 + \alpha_2 = 1$. Then the Fisher informations satisfy

$$I(Y_1 + Y_2) \leq \alpha_1^2 I(Y_1) + \alpha_2^2 I(Y_2) \tag{3.1}$$

with equality only if Y_1 and Y_2 are independent normal random variables (with variances proportional to α_i). This convolution inequality has been presented in various equivalent forms by Stam (1959), Blachman (1964), and Brown (1982). If either $I(Y_1)$ or $I(Y_2)$ is infinite, (3.1) trivially holds. Suppose $I(Y_1)$ and $I(Y_2)$ are finite. Inequality (3.1) is a simple consequence of the following result. Let ρ_1, ρ_2, and ρ_* denote the score functions for Y_1, Y_2, and $S = Y_1 + Y_2$ respectively. The difference between the right and left sides of (3.1) is the mean squared error for the estimation of the average score $\alpha_1\rho_1(Y_1) + \alpha_2\rho_2(Y_2)$ by the score $\rho_*(S)$ of the sum. Specifically,

$$\alpha_1^2 I(Y_1) + \alpha_2^2 I(Y_2) - I(Y_1 + Y_2) = E\,(\alpha_1\rho_1(Y_1) + \alpha_2\rho_2(Y_2) - \rho_*(Y_1 + Y_2))^2. \tag{3.2}$$

To see why (3.2) holds, note that $\rho_*(S)$ is the conditional expectation $\rho_*(S) = E[\alpha_1\rho_1(Y_1) + \alpha_2\rho_2(Y_2) \mid S]$ (since $\rho_*(S) = E[\rho_i(Y_i) \mid S]$ for $i = 1,2$). Therefore, $\rho_*(S)$ has the least mean squared error among estimates depending only on the sum. Equation (3.2) is the corresponding Pythagorean relation. (From (3.2), equality holds in (3.1) if and only if $\alpha_1\rho_1(Y_1) + \alpha_2\rho_2(Y_2) = \rho_*(S)$ almost surely. Blachman (1965) verified that this equality condition requires normal g_1 and g_2.)

By replacing Y_i with $\sqrt{\alpha_i}\,Y_i$ we obtain the following result from (3.1) and (2.13). If Y_1 and Y_2 are independent random variables possessing densities with

bounded derivatives and if $\alpha_1,\alpha_2 \geq 0$, $\alpha_1+\alpha_2 = 1$, then the Fisher informations satisfy

$$I(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) \leq \alpha_1 I(Y_1) + \alpha_2 I(Y_2) \tag{3.3}$$

with equality if and only if Y_1 and Y_2 are independent normal random variables with the same variance. In particular, if Y_1 and Y_2 are independent and identically distributed, then

$$I(\frac{Y_1 + Y_2}{\sqrt{2}}) \leq I(Y_1) \tag{3.4}$$

with equality if and only if the distribution is normal.

Let $Y_1, Y_2, \ldots, Y_m$ be independent random variables. Suppose each Y_i has a density with a bounded derivative. Let $\alpha_i \geq 0$, $\sum_{i=1}^m \alpha_i = 1$. By induction from the case $m=2$ we have

$$I(\sum_{i=1}^m \sqrt{\alpha_i}\, Y_i) \leq \sum_{i=1}^m \alpha_i\, I(Y_i) \tag{3.5}$$

with equality if and only if the Y_i are independent normal random variables with the same variance. When the Y_i have a common variance, (3.5) also holds for normalized Fisher information,

$$J(\sum_{i=1}^m \sqrt{\alpha_i}\, Y_i) \leq \sum_{i=1}^m \alpha_i\, J(Y_i). \tag{3.6}$$

Suppose $Y_1, Y_2, \ldots, Y_m$ are independent and identically distributed. A useful special case of inequality (3.5) is

$$I(\frac{Y_1 + Y_2 + \ldots + Y_m}{\sqrt{m}}) \leq I(Y_1). \tag{3.7}$$

Because of Lemma 2.1, convolution inequalities for Fisher information immediately yield inequalities for Shannon entropy.

Lemma 3.1

Let $X_1, X_2, \ldots, X_m$ be independent random variables with the same variance σ^2. Let $\alpha_i \geq 0$, $\sum_{i=1}^m \alpha_i = 1$. The entropy satisfies

$$H(\sum_{i=1}^m \sqrt{\alpha_i}X_i) \geq \sum_{i=1}^m \alpha_i\, H(X_i) \tag{3.8}$$

or equivalently, the divergence satisfies

$$D(\sum_{i=1}^{m} \sqrt{\alpha_i} X_i) \leq \sum_{i=1}^{m} \alpha_i D(X_i) \tag{3.9}$$

with equality if and only if the X_i are independent normal random variables with the same variance. In particular, for X_1 and X_2 independent and identically distributed,

$$H(\frac{X_1 + X_2}{\sqrt{2}}) \geq H(X_1) \tag{3.10}$$

with equality if and only if the distribution is normal. Also, if $X_1, X_2, \ldots, X_m$ are independent and identically distributed then

$$H(\frac{X_1 + X_2 + \ldots + X_m}{\sqrt{m}}) \geq H(X_1). \tag{3.11}$$

Proof

Set $Y_i = X_i + \sqrt{\tau} Z_i$ where the Z_i are independent normals with the same variance σ^2. For $\tau > 0$, the density of Y_i has a bounded derivative. Thus we may apply the convolution inequality (3.6). Direct substitution into (2.19) proves Lemma 3.1.

Remarks

For symmetric random variables there is a simple information theoretic proof of (3.10). (Let $S = (X_1 + X_2)/\sqrt{2}$ and $T = (X_1 - X_2)/\sqrt{2}$. Then $2H(X_1) = H(X_1, X_2) = H(S,T) \leq H(S) + H(T) = 2H(S)$.) Lemma 3.1 also follows from Shannon's entropy power inequality (Shannon 1948, Stam 1959, Blachman 1965), which is equivalent to

$$e^{2H(\sum_{i=1}^{m} \sqrt{\alpha_i} X_i)} \geq \sum_{i=1}^{m} \alpha_i e^{2H(X_i)}. \tag{3.12}$$

Here the X_i are independent random variables, but the X_i need not have identical variance. Equality holds if and only if the X_i are independent normal random variables. By the convexity of the exponential, (3.12) implies (3.8). The weaker and more easily established inequality (3.8) is adequate for this paper.

Now let $X_1, X_2, \ldots$ be independent and identically distributed random variables with finite variance. Let $S_n = (X_1 + X_2 + \ldots + X_n)/\sqrt{n}$ be the normalized sum. For $n = mp$ a product of any two positive integers, the normalized sum may be grouped as $S_{mp} = \sum_{i=1}^{m} S_p^{(i)}/\sqrt{m}$ where the $S_p^{(i)}$ are independent copies of S_p. From inequality

(3.11) we have

$$H(S_{mp}) \geq H(S_p). \tag{3.13}$$

Similarly, from inequality (3.8) we obtain

$$H(S_{p+q}) \geq \frac{p}{p+q} H(S_p) + \frac{q}{p+q} H(S_q) \tag{3.14}$$

for any positive integers p and q. These inequalities (3.13) and (3.14) suggest that the entropy is nearly monotone increasing. Consider doubling the sample size. We obtain

$$H(S_{2n}) \geq H(S_n) \tag{3.15}$$

or equivalently

$$D(S_{2n}) \leq D(S_n). \tag{3.16}$$

Thus the entropy is monotone increasing and the divergence is monotone decreasing along powers of two subsequences $n_k = 2^k n_0$ (for any initial n_0).

The following Lemma shows that the entropy $H(S_n)$ has a limit equal to the supremum (as one would expect for a sequence with nearly monotone behavior).

Lemma 3.2

The entropy $H(S_n)$ of the normalized sum satisfies

$$\lim_n H(S_n) = \sup_n H(S_n) \tag{3.17}$$

or equivalently, the divergence satisfies

$$\lim_n D(S_n) = \inf_n D(S_n). \tag{3.18}$$

Proof

Let $\epsilon > 0$ and let p be such that $H(S_p) \geq \sup_n H(S_n) - \epsilon$. Let $n = mp + r$ where the remainder r is less than p. Using inequalities (2.10) and (3.13) we have

$$\begin{aligned} H(S_n) &\geq H((\frac{mp}{n})^{1/2} S_{mp}) \\ &= H(S_{mp}) + \frac{1}{2} \log \frac{n-r}{n} \\ &\geq H(S_p) + \frac{1}{2} \log (1 - \frac{p}{n}) \end{aligned} \tag{3.19}$$

which converges to $H(S_p)$ as $n \to \infty$. Thus $\liminf H(S_n) \geq \sup_n H(S_n) - \epsilon$. Hence

(3.17) or equivalently (3.18) follows by letting $\epsilon \to 0$. This completes the proof of Lemma 3.2.

The analogous results hold for Fisher information. Let $Y_1, Y_2, \ldots$ be independent and identically distributed random variables. Suppose the distribution has a density with bounded derivative. Let $S_n' = (Y_1 + Y_2 + \ldots + Y_n)/\sqrt{n}$ be the normalized sum. Then the Fisher information decreases with doubling sample size

$$I(S_{2n}') \leq I(S_n'). \tag{3.20}$$

Furthermore,

$$I(S_{mp}') \leq I(S_p') \tag{3.21}$$

and

$$I(S_{p+q}') \leq \frac{p}{p+q} I(S_p') + \frac{q}{p+q} I(S_q'). \tag{3.22}$$

From inequalities (2.18) and (3.21) we deduce that, for $n = mp + r$,

$$I(S_n') \leq I((\frac{mp}{n})^{1/2} S_{mp}') = \frac{n}{n-r} I(S_{mp}') \leq \frac{n}{n-p} I(S_p'). \tag{3.23}$$

Using (3.23) with p such that $I(S_p')$ is near the infimum, we obtain the following.

Lemma 3.3 $\lim_n I(S_n') = \inf_n I(S_n')$.

Lemmas 3.2 and 3.3 show that limits must exist for the entropy $H(S_n)$ and for the Fisher information $I(S_n + \sqrt{\tau} Z)$. The remaining challenge is to identify the limits (as the corresponding entropy and information of the normal random variable).

4. Convergence of the Fisher information

In this section we prove that the Fisher information $I(S_n + \sqrt{\tau}Z)$ (for the normalized sum smoothed by addition of independent normal) converges to the limit suggested by the Cramér-Rao inequality. Much of this section is a summary of the results of Brown (1982).

From section 3, the difference in Fisher information $I(Y_1) - I((Y_1 + Y_2)/\sqrt{2})$ (where Y_1 and Y_2 are independent and identically distributed with density with bounded derivative) is twice the mean squared error for the estimation of the average score $(\rho(Y_1) + \rho(Y_2))/2$ by the score $\rho_*(S)$ of the sum $S = Y_1 + Y_2$. Now the score $\rho_*(S)$ is linear in S only for the normal distribution. This motivates establishing that the mean squared error cannot be much less than the mean squared error for the best linear estimate.

Let Z, Z_1, and Z_2 be independent and identically distributed normal random variables. Let v be an arbitrary measurable function with finite $E\, v^2(Z)$. Let $v_{NL}(S) = E[v(Z_1) + v(Z_2) \mid S]$ be the best (minimum mean squared error) non-linear estimate of $v(Z_1) + v(Z_2)$ based on $S = (Z_1 + Z_2)/\sqrt{2}$. The associated mean squared error is

$$\text{mse}_{NL} = E\,(v(Z_1) + v(Z_2) - v_{NL}(\frac{Z_1 + Z_2}{\sqrt{2}}))^2. \tag{4.1}$$

Let $\alpha_0 + \alpha_1 Z$ be the best linear estimate of $v(Z)$. Then $v_L(S) = 2\alpha_0 + \sqrt{2}\alpha_1 S$ is the best linear estimate of $v(Z_1) + v(Z_2)$ based on S. The mean squared error is

$$\text{mse}_L = E\,(v(Z_1) + v(Z_2) - v_L(\frac{Z_1 + Z_2}{\sqrt{2}}))^2$$

$$= 2\,E\,(v(Z) - \alpha_0 - \alpha_1 Z)^2. \tag{4.2}$$

Brown (1982) established the following remarkable result.

Lemma 4.1 $\text{mse}_{NL} \leq \text{mse}_L \leq 2\,\text{mse}_{NL}$.

Note that $\text{mse}_{NL} \leq \text{mse}_L$ is trivial and that equality holds if and only if v is linear. By inspection of the proof, equality holds in $\text{mse}_L \leq 2\,\text{mse}_{NL}$ if and only if v is quadratic.

The essence of Brown's proof is as follows. Without loss of generality we may assume that the normal has zero mean and unit variance. Let $v(z) = \sum_{k=0}^{\infty} \alpha_k H_k(z)$ be the expansion of v in terms of orthogonal Hermite polynomials. Then $\text{mse}_L = 2\sum_{k=2}^{\infty} \alpha_k^2 k!$ (since $EH_k(Z)H_m(Z) = k!$ for $k=m$ and $=0$ for $k \neq m$). Similarly $\text{mse}_{NL} = \sum_{k=2}^{\infty} (2\alpha_k^2 - \beta_k^2)k!$ where the β_k are the coefficients in the expansion $v_{NL}(s) = \sum_{k=0}^{\infty} \beta_k H_k(s)$. Also $v_{NL}(s) = 2\int v((s+t)/\sqrt{2})\phi(t)dt$. Suppose $v(z) = H_k(z)$. Then $\beta_m = \int v_{NL}(s)H_m(s)\phi(s)ds/m! = 2\int\int H_k((s+t)/\sqrt{2})H_m(s)\phi(s)\phi(t)dsdt/m!$. Now the Hermite polynomials satisfy $H_m(s)\phi(s) = (-1)^m(d^m/dt^m)\phi(s)$, so we may integrate by parts (m times) to obtain $\beta_m = 2(1/\sqrt{2})^m \int\int H_k^{(m)}((s+t)/\sqrt{2})\phi(s)\phi(t)dsdt/m!$ $= 2(1/\sqrt{2})^m \int H_k^{(m)}(z)\phi(z)dz/m! = 2(1/\sqrt{2})^m \int H_k(z)H_m(z)\phi(z)dz/m! = 2(1/\sqrt{2})^k$ for $k=m$ and $=0$ for $k \neq m$. Thus for general $v(z) = \sum_{k=0}^{\infty} \alpha_k H_k(z)$ we have $\beta_k = 2(1/\sqrt{2})^k \alpha_k$ and $\text{mse}_{NL} = \sum_{k=2}^{\infty}(2 - 4/2^k)\alpha_k^2 k! \geq \sum_{k=2}^{\infty}\alpha_k^2 k! = \text{mse}_L/2$ which is the desired result.

We are now prepared to examine the convergence of Fisher information. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean μ and variance σ^2. These random variables need not have a density. Let S_n be the normalized sum $S_n = \sum_{i=1}^{n}(X_i - \mu)/\sqrt{n}$. Let Z be a normal random variable (independent of the X_i) with mean zero and variance σ^2. We are interested in the sequence of Fisher informations $I(S_n + \sqrt{\tau}Z)$.

As in Brown (1982), it is convenient to consider the sequence of random variables $Y_i = X_i + \sqrt{\tau}Z_i$, $i=1,2,\ldots$ where the Z_i are independent and identically distributed copies of the normal Z. The normalized sums of this sequence are $S_n' = \sum_{i=1}^{n}(Y_i - \mu)/\sqrt{n}$. Note that S_n' has the same distribution as $S_n + \sqrt{\tau}Z$.

Whenever the sample size is doubled we have from inequality (3.21) that $I(S_{2n}') \leq I(S_n')$. Consider a powers of two subsequence $n_k = 2^k n_0$, $k=0,1,\ldots$ with arbitrary initial n_0. The Fisher information $I(S_{n_k}')$ is monotone decreasing and hence has a limit. Furthermore $((1+\tau)\sigma^2)^{-1} \leq I(S_n') \leq (\tau\sigma^2)^{-1}$, so the limit is finite. The challenging part is to show the limit equals the lower bound $((1+\tau)\sigma^2)^{-1}$.

Brown considered the difference sequence $I(S_{n_k}') - I(S_{n_{k+1}}')$. Since $I(S_{n_k}')$ converges to a finite limit, the difference sequence must converge to zero. From section 3,

$$I(S'_{n_k}) - I(S'_{n_{k+1}}) = \frac{1}{2}\int\int(\rho_k(y)+\rho_k(y')-\sqrt{2}\rho_{k+1}(\frac{y+y'}{\sqrt{2}}))^2 g_k(y)g_k(y')dydy' \quad (4.3)$$

where $g_k(y)$ is the density for $S_{n_k} + \sqrt{\tau}Z$ and $\rho_k(y) = g_k'(y)/g_k(y)$. If the density is lower bounded by a normal density, Lemma 4.1 can be applied. Let ϕ_τ denote the normal density with mean zero and variance $\tau\sigma^2$. Brown established that

$$g_k(y) \geq c_\tau \phi_{\tau/2}(y) \quad (4.4)$$

where $c_\tau = (3/4\sqrt{2})e^{-4/\tau}$. (This inequality follows from $g_k(y) = E\phi_\tau(y-S_{n_k}) \geq P\{|S_{n_k}| \leq 2\sigma\}\min_{|x|\leq 2\sigma}\phi_\tau(y-x) \geq (3/4)\phi_\tau(|y|+2\sigma) \geq c_\tau\phi_{\tau/2}(y)$ where the lower bound on $P\{|S_{n_k}| \leq 2\sigma\}$ is from Chebyshev's inequality.) Thus from (4.3), (4.4) and Lemma 4.1

$$I(S'_{n_k}) - I(S'_{n_{k+1}}) \geq \frac{c_\tau^2}{2}\int\int(\rho_k(y)+\rho_k(y')-\sqrt{2}\rho_{k+1}(\frac{y+y'}{\sqrt{2}}))^2\phi_{\tau/2}(y)\phi_{\tau/2}(y')dydy'$$

$$\geq \frac{c_\tau^2}{2}\int(\rho_k(y) - b_k + y/a_k\sigma^2)^2\,\phi_{\tau/2}(y)dy \quad (4.5)$$

for some constants a_k and b_k. The difference $I(S'_{n_k}) - I(S'_{n_{k+1}})$ converges to zero, so as in Brown (1982),

$$\lim_{k\to\infty}\int(\rho_k(y) - b_k + y/a_k\sigma^2)^2\,\phi_{\tau/2}(y)dy = 0. \quad (4.6)$$

Now $\rho_k(y) - b_k + y/a_k\sigma^2 = (d/dy)(\log g_k(y) - \log\phi_{a_k}(y-b_k))$. Thus Brown readily established that $g_k(y) - c_k\phi_{a_k}(y-b_k) \to 0$ uniformly on compact subsets, for some sequence of constants c_k. But $g_k(y)$ is a probability density with mean zero and variance $(1+\tau)\sigma^2$. So we expect that $c_k \to 1$, $b_k \to 0$, and $a_k \to (1+\tau)$, which Brown verified using Chebyshev's inequality and a truncation argument. Thus

$$\lim_{k\to\infty}\int(\frac{g_k'(y)}{g_k(y)} - \frac{\phi_{1+\tau}'(y)}{\phi_{1+\tau}(y)})^2\,\phi_{\tau/2}(y)dy = 0. \quad (4.7)$$

Also

$$\lim_{k\to\infty} g_k(y) = \phi_{1+\tau}(y) \quad (4.8)$$

uniformly on compact subsets.

Brown used (4.8) and let $\tau \to 0$ to show convergence in distribution. We use (4.7), (4.8) and a uniform integrability argument to show convergence of $I(S_n + \sqrt{\tau}Z)$ to the

reciprocal of the variance.

Lemma 4.2

For each $\tau > 0$, the Fisher information converges to the lower bound,

$$\lim_{n\to\infty} I(S_n + \sqrt{\tau}Z) = \frac{1}{(1+\tau)\sigma^2}. \tag{4.9}$$

Equivalently, the normalized Fisher information converges to zero,

$$\lim_{n\to\infty} J(S_n + \sqrt{\tau}Z) = 0. \tag{4.10}$$

Furthermore the convergence is monotone along the powers of two subsequences, $n_k = 2^k n_0$.

Proof

Fix $\tau \geq 0$. Let $g_k(y) = g_{k,\tau}(y)$ be the density for $S_{n_k} + \sqrt{\tau}Z$. From (4.7), the score $g_k'(y)/g_k(y)$ converges to $\phi_{1+\tau}'(y)/\phi_{1+\tau}(y)$ in normal $(0,\tau\sigma^2/2)$ measure and hence in Lebesgue measure. From (4.8), the density $g_k(y)$ converges to $\phi_{1+\tau}(y)$ in Lebesgue measure. Thus $(g_k'(y)/g_k(y))^2 g_k(y)$ converges to $(\phi_{1+\tau}'(y)/\phi_{1+\tau}(y))^2\phi_{1+\tau}(y)$ in measure. Consequently,

$$\lim_{k\to\infty} \int \left(\frac{g_k'(y)}{g_k(y)}\right)^2 g_k(y)\, dy = \frac{1}{(1+\tau)\sigma^2} \tag{4.11}$$

provided the integrand is uniformly integrable. Lemma 4.3 shows that if the sequence of densities g_k has bounded divergence from the normal, then g_k is uniformly integrable. Now

$$\begin{aligned} D(g_k \,||\, \phi_{1+\tau}) &= H(\sqrt{1+\tau}Z) - H(S_{n_k} + \sqrt{\tau}Z) \\ &\leq H(\sqrt{1+\tau}Z) - H(\sqrt{\tau}Z) \\ &= \frac{1}{2}\log\frac{1+\tau}{\tau} \end{aligned} \tag{4.12}$$

uniformly in k . Thus $g_{k,\tau}(y)$ is uniformly integrable for each $\tau > 0$. Lemma 4.4 uses the Cauchy-Schwartz inequality to show that

$$\left(\frac{g_{k,\tau}'(y)}{g_{k,\tau}(y)}\right)^2 g_{k,\tau}(y) \leq \frac{c}{\tau\sigma^2}\, g_{k,2\tau}(y) \tag{4.13}$$

where $c = 4\sqrt{2}e^{-1}$ is a constant. The right side of (4.13) is uniformly integrable, so the left side is also uniformly integrable. Thus (4.11) is valid, which means $\lim I(S_{n_k} + \sqrt{\tau}Z) = ((1+\tau)\sigma^2)^{-1}$ or equivalently $\lim J(S_{n_k} + \sqrt{\tau}Z) = 0$. By Lemma 3.3 the entire sequence must have the same limit as the subsequence. This completes the proof of Lemma 4.2.

In the above proof we used the following simple results.

Lemma 4.3

If a sequence of probability densities g_k has bounded divergence from a normal density ϕ, then g_k is uniformly integrable.

Proof

Since normal measure is equivalent to Lebesgue measure, it suffices to show that g_k/ϕ is uniformly Φ–integrable. Let $\gamma = \sup_k D(g_k \,||\, \phi)$ which is finite by assumption. Now

$$\int_{\{\frac{g_k}{\phi} > r\}} g_k \leq \frac{1}{\log r} \int g_k (\log \frac{g_k}{\phi})^+ \leq \frac{\gamma + e^{-1}}{\log r} \tag{4.14}$$

which converges to zero uniformly in k as $r \to \infty$. Thus g_k is uniformly integrable. This completes the proof of Lemma 4.3.

Lemma 4.4

Let $g_\tau(y)$ be the density for $Y = X + \sqrt{\tau}Z$ where X is an arbitrary random variable and Z is an independent standard normal random variable. Then

$$\left(\frac{g_\tau'(y)}{g_\tau(y)}\right)^2 g_\tau(y) \leq \frac{c}{\tau}\, g_{2\tau}(y) \tag{4.15}$$

where $c = 4\sqrt{2}e^{-1}$ is a constant.

Proof

Let ϕ_τ be the normal density for $\sqrt{\tau}Z$. By the Cauchy-Schwartz inequality

$$(g_\tau'(y))^2 = (E\phi_\tau'(y-X))^2 = (E-\frac{y-X}{\tau}\phi_\tau^{1/2}(y-X)\ \phi_\tau^{1/2}(y-X))^2$$

$$\leq E\ (\frac{y-X}{\tau})^2\phi_\tau(y-X)\ \ g_\tau(y)$$

$$\leq E\ \frac{c}{\tau}\ \phi_{2\tau}(y-X)\ \ g_\tau(y)$$

$$= \frac{c}{\tau}\ g_{2\tau}(y)\ g_\tau(y) \qquad (4.16)$$

which is the desired result.

5. Strengthened central limit theorem

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean μ and variance σ^2. Let F_n be the probability measure for the normalized sum $S_n = \sum_{i=1}^{n} (X_i - \mu)/\sqrt{n}$. Let Φ be the probability measure and ϕ the density for the normal Z with mean zero and variance σ^2.

We say that S_n converges to Z in *information* if for some n (and hence for all larger n) S_n has a density f_n and

$$\lim_{n\to\infty} D(f_n \,||\, \phi) = 0. \tag{5.1}$$

Now S_n has the same mean and variance as Z, so convergence in information is equivalent to convergence of the entropy,

$$\lim_{n\to\infty} H(S_n) = H(Z). \tag{5.2}$$

Recall that $H(Z)$ is the maximum entropy for the given variance. If $H(S_n)$ is finite ($>-\infty$) for some n, say $n = m$, then $H(S_n)$ is finite for all $n \geq m$ (from inequality (2.10)). In particular, if X_1 has a density $f(x)$ with finite entropy $H(X_1)$, then $H(S_n)$ is finite for all n.

The main result of this paper is the following.

Theorem

If the entropy $H(S_n)$ is finite for some n, then S_n converges to Z in information. Furthermore the convergence in (5.1) and (5.2) is monotone along the powers of two subsequences $n_k = n_0\, 2^k$ (for any initial n_0).

Thus $D(F_n \,||\, \Phi)$ converges to either zero or infinity and

$$\lim_{n\to\infty} D(F_n \,||\, \Phi) = \begin{cases} 0 & \text{if } D(F_n \,||\, \Phi) < \infty \text{ for some } n \\ \infty & \text{if } D(F_n \,||\, \Phi) = \infty \text{ for all } n. \end{cases} \tag{5.3}$$

Proof

Monotonicity $H(S_{n_{k+1}}) \geq H(S_{n_k})$ follows from inequality (3.15). Also $D(F_{n_k} \,||\, \Phi) = D(S_{n_k}) = H(Z) - H(S_{n_k})$ is monotone decreasing and hence has a limit as $k \to \infty$. From Lemma 2.1 we have

$$D(F_{n_k} \,||\, \Phi) = \frac{1}{2} \int_0^\infty J(S_{n_k} + \sqrt{\tau} Z) \, \frac{d\tau}{1+\tau}. \tag{5.4}$$

From Lemma 4.2 the integrand is monotone decreasing with limit zero. If $D(F_n \,||\, \Phi)$ is eventually finite, the monotone convergence theorem applies and

$$\lim_{k \to \infty} D(F_{n_k} \,||\, \Phi) = 0. \tag{5.5}$$

By Lemma 3.2, convergence of the subsequence implies convergence of the entire sequence to the same limit. This completes the proof.

Corollary 1

If the entropy is eventually finite, then the density f_n converges to ϕ in L^1,

$$\lim_{n \to \infty} \int |f_n(x) - \phi(x)| \, dx = 0. \tag{5.6}$$

Furthermore, the probability measure F_n converges to Φ in the uniform setwise sense,

$$\lim_{n \to \infty} \sup_A |F_n(A) - \Phi(A)| = 0 \tag{5.7}$$

where the supremum is over all Borel sets A.

Proof

The corollary is an immediate consequence of the following chain of inequalities due to Csiszár (1967), Kullback (1967), and Kemperman (1967)

$$2 \sup_A |F_n(A) - \Phi(A)|^2 = 2(F_n(A_n) - \Phi(A_n))^2 = \frac{1}{2}(\int |f_n - \phi|)^2$$

$$\leq D_{\pi_n}(F_n \,||\, \Phi) \leq D(f_n \,||\, \phi) \tag{5.8}$$

where $A_n = \{x : \phi(x) > f_n(x)\}$ and π_n is the partition consisting of A_n and its complement.

Remarks

The classical central limit theorem states convergence in distribution. Specifically,

$$\lim_{n \to \infty} |F_n(A) - \Phi(A)| = 0 \tag{5.9}$$

for each set A with boundary measure zero. Uniform setwise convergence is clearly

stronger because convergence holds for every Borel set and because $\lim |F_n(A_n) - \Phi(A_n)| = 0$ for arbitrarily varying sets A_n .

Another characterization of convergence in distribution is that

$$\lim_{n\to\infty} E\ h(S_n) = E\ h(Z) \tag{5.10}$$

for all bounded uniformly continuous functions h . A consequence of convergence in information is that (5.10) holds for the considerably larger class of measurable functions h for which $E\ e^{\alpha h(Z)}$ is finite for all α in some neighborhood of zero (see Csiszár 1975). In particular, (5.10) holds for functions $h(x)$ bounded by some multiple of $x^2 + 1$.

Is the quadratic $\log \phi(S_n)$ a consistent approximation of the log-likelihood $\log f_n(S_n)$? The L^1 approximation error is

$$E\ |\ \log f_n(S_n) - \log \phi(S_n)| = \int f_n\ |\ \log \frac{f_n}{\phi}\ |. \tag{5.11}$$

This divergence-like quantity is upper bounded by

$$\int f_n\ |\ \log \frac{f_n}{\phi}\ | \leq D(f_n\,||\,\phi) + \int\ |f_n - \phi\,|$$

$$\leq D(f_n\,||\,\phi) + (2D(f_n\,||\,\phi))^{1/2}. \tag{5.12}$$

(This bound follows from $\int f_n(\log f_n/\phi)^- = \int_{A_n} f_n \log \phi/f_n \leq F_n(A_n) \log \Phi(A_n)/F_n(A_n)$ $\leq \Phi(A_n) - F_n(A_n) = (1/2)\int |f_n - \phi|$ where $A_n = \{x{:}\phi(x) > f_n(x)\}$.) A similar bound, but with a constant larger than 2, was given by Pinsker (1964). From (5.12) we immediately have the following result.

Corollary 2

If $H(S_n)$ is finite for some n, then $\log f_n(S_n) - \log \phi\,(S_n)$ converges to zero in L^1 (and hence in probability) as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} E\ |\ \log f_n(S_n) - \log \phi(S_n)| = 0. \tag{5.13}$$

6. The entropy derivative

The derivative of entropy with respect to the variance of added normal is one-half the Fisher information. Stam (1959) credits de Bruijn with the discovery of this fact. Stam (1959, p.108) suggests that the random variable X must have a density with some conditions. A schetch of the proof was given in Blachman (1965), but without justification of the exchanges of differentiation and integration. We offer a proof that makes no assumptions on the distribution of X other than finite variance.

Lemma 6.1

Let X be a random variable with finite variance. Let Z be an independent standard normal random variable. For all $\tau > 0$,

$$\frac{\partial}{\partial \tau} H(X + \sqrt{\tau}Z) = \frac{1}{2} I(X + \sqrt{\tau}Z). \tag{6.1}$$

By change of variables we obtain the following.

Corollary

Let X be a random variable with finite variance. Let Z be an independent normal random variable with the same variance as X. For all $0 < t < 1$,

$$\frac{\partial}{\partial t} H(\sqrt{1-t}X + \sqrt{t}Z) = \frac{1}{2} J(\sqrt{1-t}X + \sqrt{t}Z)/1\text{-}t. \tag{6.2}$$

Proof of Lemma 6.1

For $\tau > 0$, the entropy $H(X + \sqrt{\tau}Z)$ and the Fisher information $I(X + \sqrt{\tau}Z)$ are finite. Let ϕ_τ be the normal density with mean zero and variance τ. Then $g_\tau(y) = E\ \phi_\tau(y-X)$ is the density for $Y = X + \sqrt{\tau}Z$. The proof of Lemma 6.1 is immediate from the following three lemmas.

Lemma 6.2 $\quad \frac{\partial}{\partial \tau} g_\tau(y) = \frac{1}{2} \frac{\partial^2}{\partial y^2} g_\tau(y)$.

Lemma 6.3 $\quad \frac{\partial}{\partial \tau} H(X + \sqrt{\tau}Z) = -\int (\frac{\partial}{\partial \tau} g_\tau(y)) \log g_\tau(y)\ dy$.

Lemma 6.4 $\quad I(X + \sqrt{\tau}Z) = -\int (\frac{\partial^2}{\partial y^2} g_\tau(y)) \log g_\tau(y)\ dy$.

To prove these lemmas we shall make frequent use of the following bounds on the derivatives of the normal density $\phi_\tau(z)$. The first derivative is $(\partial/\partial z)\,\phi_\tau(z) = -(z/\tau)\,\phi_\tau(z)$. The absolute value of this derivative is bounded by the constant $(2\pi e\tau^2)^{-1/2}$ and dominated by the function $2(e\tau)^{-1/2}\,\phi_{2\tau}(z)$. The second derivative with respect to z and first derivative with respect to the variance τ satisfy

$$\frac{\partial}{\partial\tau}\,\phi_\tau(z) = \frac{1}{2}\,\frac{\partial^2}{\partial z^2}\,\phi_\tau(z) = \frac{1}{2\tau}\left(\frac{z^2}{\tau} - 1\right)\phi_\tau(z). \tag{6.3}$$

The absolute value of this derivative is bounded by the constant $(8\pi\tau^3)^{-1/2}$ and dominated by the function $2\sqrt{2}(e\tau)^{-1}\,\phi_{2\tau}(z)$.

Proof of Lemma 6.2

We wish to show that $g_\tau(y)$ satisfies the "diffusion" equation

$$\frac{\partial}{\partial\tau}\,g_\tau(y) = \frac{1}{2}\,\frac{\partial^2}{\partial y^2}\,g_\tau(y). \tag{6.4}$$

By equation (6.3), the normal $\phi_\tau(z)$ satisfies the diffusion equation. Now $g_\tau = E\,\phi_\tau(y-Z)$. Thus (6.4) follows provided the partial derivatives exist and satisfy

$$\frac{\partial}{\partial\tau}\,E\,\phi_\tau(y-X) = E\,\frac{\partial}{\partial\tau}\,\phi_\tau(y-X) \tag{6.5}$$

and

$$\frac{\partial^2}{\partial y^2}\,E\,\phi_\tau(y-X) = E\,\frac{\partial^2}{\partial y^2}\,\phi_\tau(y-X). \tag{6.6}$$

To verify (6.5) let $0 < a < \tau$. Then $|(\partial/\partial\tau)\phi_\tau(y-X)|$ is bounded by $(8\pi a^3)^{-1/2}$, uniformly in a neighborhood of τ. The mean value theorem and the bounded convergence theorem apply to verify (6.5). To verify (6.6), note that $|(\partial/\partial y)\phi_\tau(y-X)| \le (2\pi e\tau^2)^{-1/2}$ and $|(\partial^2/\partial y^2)\,\phi_\tau(y-X)| < (2\pi\tau^3)^{-1/2}$ uniformly in y. Twice applying the mean value and bounded convergence theorems verifies (6.6). This completes the proof of Lemma 6.2.

Proof of Lemma 6.3

We wish to show that the entropy $H(X + \sqrt{\tau}Z)$ is differentiable with respect to τ and

$$\frac{\partial}{\partial \tau} H(X+\sqrt{Z}) = -\int (\frac{\partial}{\partial \tau} g_\tau(y)) \log g_\tau(y)\, dy. \tag{6.7}$$

By calculus

$$\frac{\partial}{\partial \tau} (g_\tau(y) \log g_\tau(y)) = \frac{\partial}{\partial \tau} g_\tau(y) + (\frac{\partial}{\partial \tau} g_\tau(y)) \log g_\tau(y). \tag{6.8}$$

Thus the desired result follows from

$$\frac{\partial}{\partial \tau} \int g_\tau(y) \log g_\tau(y)\, dy = \int \frac{\partial}{\partial \tau} (g_\tau(y) \log g_\tau(y))\, dy \tag{6.9}$$

and

$$0 = \frac{\partial}{\partial \tau} \int g_\tau(y)\, dy = \int (\frac{\partial}{\partial \tau} g_\tau(y))\, dy. \tag{6.10}$$

We first verify (6.10). Let $a,b,$ be such that $0 < a < \tau < b$. Then the derivative of the normal with respect to the variance τ is dominated, uniformly in a neighborhood of τ, by

$$|\frac{\partial}{\partial \tau} \phi_\tau(y-X)| \leq c\ \phi_{2b}(y-X) \tag{6.11}$$

where $c = 2(ea)^{-1}(2b/a)^{1/2}$. Thus

$$|\frac{\partial}{\partial \tau} g_\tau(y)| \leq E\ |\frac{\partial}{\partial \tau} \phi_\tau(y-X)| \leq c\ g_{2b}(y). \tag{6.12}$$

Hence $(\partial/\partial\tau)\, g_\tau(y)$ is dominated, uniformly in a neighborhood of τ, by an integrable function. The mean value and dominated convergence theorems apply to verify (6.10).

The verification of (6.9) is similar, except that a bound for $|\log g_\tau(y)|$ is necessary. From inequality 3.4 (due to Brown 1982),

$$c_\tau\, \phi_{\tau/2}(y) \leq g_\tau(y) \leq 1/\sqrt{2\pi\tau} \tag{6.13}$$

where $c_\tau = (3/4\sqrt{2})e^{-4/\tau}$. Let $0 < a < \tau < b$ and for convenience make $a \leq 1/2\pi \leq b$. Then

$$|\log g_\tau(y)| \leq \log \frac{1}{\sqrt{2\pi a}} - \log (\frac{c_a}{\sqrt{\pi b}} e^{-y^2/a})$$

$$= \frac{1}{a}(y^2 + c') \tag{6.14}$$

where $c' = 4 + a \log (4/3)(b/a)^{1/2}$. Using (6.8), (6.12), and (6.14) we obtain

$$|\frac{\partial}{\partial\tau}(g_\tau(y)\log g_\tau(y))| \le \frac{c}{a}(y^2 + c' + a)\, g_{2b}(y) \tag{6.15}$$

uniformly in a neighborhood of τ. Now X has finite variance, so $Y = X + \sqrt{2b}Z$ has finite variance. Hence the bound in (6.15) is integrable. (6.9) follows by application of the mean value and dominated convergence theorems. This completes the proof of Lemma 6.3.

Proof of Lemma 6.4

We wish to show that the Fisher information satisfies

$$I(X + \sqrt{\tau}Z) = -\int(\frac{\partial^2}{\partial y^2}g_\tau(y))\log g_\tau(y)\, dy. \tag{6.16}$$

Together Lemmas 6.2 and 6.3 show that $-((\partial^2/\partial y^2)\, g_\tau(y))\log g_\tau(y)$ is integrable. The Fisher information $I(X + \sqrt{\tau}Z)$ is finite, which means that $((\partial/\partial y)\, g_\tau(y))^2/g_\tau(y)$ is integrable. Thus the following derivative exists everywhere and is integrable

$$\frac{\partial}{\partial y}[(\frac{\partial}{\partial y}g_\tau(y))\log g_\tau(y)] = (\frac{\partial^2}{\partial y^2}g_\tau(y))\log g_\tau(y) + (\frac{\partial}{\partial y}g_\tau(y))^2/g_\tau(y). \tag{6.17}$$

Hence $((\partial/\partial y)\, g_\tau(y))\log g_\tau(y)$ is absolutely continuous (see Rudin 1974, p.179). Now

$$|(\frac{\partial}{\partial y}g_\tau(y))\log g_\tau(y)| = |\frac{\frac{\partial}{\partial y}g_\tau(y)}{\sqrt{g_\tau(y)}}|\ \ |2\sqrt{g_\tau(y)}\log\sqrt{g_\tau(y)}| \tag{6.18}$$

which tends to zero as $|y| \to \infty$, because the square of the first factor on the right side is integrable and because $u\log u \to 0$ as $u \to 0$. Thus integrating (6.17) on $[-B,B]$ and letting $B \to \infty$ we obtain

$$0 = \int(\frac{\partial^2}{\partial y^2}g_\tau(y))\log g_\tau(y)\, dy + \int(\frac{\partial}{\partial y}g_\tau(y))^2/g_\tau(y)\, dy \tag{6.19}$$

which is the desired result. This completes the proof of Lemma 6.4.

7. Examples

Can the normalized sum S_n have infinite divergence from the normal for every n? If the divergence is initially infinite, can it become finite for larger n (and hence converge to zero) ? Both possibilities are illustrated by modifying an example in Kolmogorov and Gnedenko (1954, p.223).

Consider the following parametric family of probability densities, for $r > 0$

$$f_r(x) = \begin{cases} \dfrac{r\, x^{-1}}{\log^{1+r} x^{-1}} & \text{for } 0 < x \leq e^{-1} \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

The support set $[0,e^{-1}]$ is compact, so all moments are finite. Note that this density is unbounded in neighborhoods of zero. Let $H(f_r)$ denote the differential entropy. Straightforward calculation yields

$$H(f_r) = \begin{cases} -\infty & \text{for } 0 < r \leq 1 \\ -(r^2(r-1))^{-1} - \log r & \text{for } r > 1. \end{cases} \tag{7.2}$$

Let $f_r^{(n)}$ denote the n-fold convolution of f_r . The entropy of $f_r^{(n)}$ (and the divergence from normality) is finite if and only if $n > 1/r$, i.e.,

$$H(f_r^{(n)}) \begin{cases} = -\infty & \text{for } n \leq 1/r \\ > -\infty & \text{for } n > 1/r. \end{cases} \tag{7.3}$$

To justify (7.3) it is shown that for s in a neighborhood of zero

$$\frac{1}{n!}\, f_{nr}(s) \leq f_r^{(n)}(s) \leq \frac{2^{n-1}}{n}\, f_{nr}(\frac{s}{n}); \tag{7.4}$$

whereas outside neighborhoods of zero, $f_r^{(n)}(s)$ is bounded. Thus the entropy of $f_r^{(n)}$ is finite if and only if the entropy of f_{nr} is finite. So (7.3) follows from (7.2).

We verify inequality (7.4) for $n=2$. The density f_r is monotone decreasing in the interval $(0,e^{-(r+1)})$. Hence for s in this interval, the density $f_r^{(2)}$ satisfies

$$f_r^{(2)}(s) = \int_0^s f_r(x)\, f_r(s-x)dx \geq f_r(s) \int_0^s f_r(x)dx = \frac{f_r(s)}{\log^r s^{-1}} = \frac{1}{2}\, f_{2r}(s). \tag{7.5}$$

and

$$f_r^{(2)}(s) = \int_0^{s/2} f_r(x)\, f_r(s-x)dx + \int_{s/2}^s f_r(x)\, f_r(s-x)dx$$

$$\leq f_r(\tfrac{s}{2}) \int_0^{s/2} f_r(x)dx + f_r(\tfrac{s}{2}) \int_{s/2}^{s} f_r(s-x)dx$$

$$= f_{2r}(\tfrac{s}{2}). \tag{7.5}$$

For $n>2$, similar arguments establish (7.4) by induction.

The central limit theorem requires that the density $f_r^{(n)}$ be normalized to have zero mean and constant variance. Since the entropy of this normalized density is finite for all $n > 1/r$, the convergence theorem (of section 5) applies. We conclude that this density converges in information to the normal density.

In this example the normalization pushes the support of the unbounded portion of the density toward $-\infty$. To trap the singularity at zero, we start with the symmetrized density $(f_r(x) + f_r(-x))/2$. In this case the normalized n-fold convolution still converges in information, even though pointwise convergence fails at the point $x = 0$.

For an example where the density does not converge in information to the normal, let

$$f_r(x) = \begin{cases} \dfrac{r\ x^{-1}}{\log x^{-1}\ (\log\log x^{-1})^{1+r}} & \text{for } 0 < x < e^{-e} \\ 0 & \text{otherwise.} \end{cases} \tag{7.8}$$

This density has a sharper peak at zero than the previous example. The associated entropy is

$$H(f_r) = -\infty \quad \text{for all } r > 0. \tag{7.9}$$

As in the previous example, the n-fold convolution $f_r^{(n)}$ satisfies

$$f_r^{(n)}(s) \geq \frac{1}{n!} f_{nr}(s) \tag{7.10}$$

for s in a neighborhood of zero. Therefore

$$H(f_r^{(n)}) = -\infty \quad \text{for all } n. \tag{7.11}$$

After normalization to zero mean and constant variance, the entropy is still $-\infty$. Consequently the divergence from the normal is infinite for all n.

Acknowledgments

The results of this paper were conjectured by Professor Tom Cover. He showed that Shannon's entropy power inequality implies the monotonicity of entropy and divergence and he posed the problem of identifying the limit. Professors Persi Diaconis and Imre Csiszár are also acknowledged for their helpful suggestions.

References

Blachman, N.M. (1965). The convolution inequality for entropy powers. *IEEE Transactions on Information Theory.* **IT-11** 267-271.

Brown, L.D. (1982). A proof of the central limit theorem motivated by the Cramér-Rao inequality. *Statistics and Probability: Essays in Honor of C.R. Rao.* ed. by G. Kallianpur, P.R. Krishnaiah, J.K. Ghosh. North-Holland, Amsterdam.

Chernoff, H. (1956). Large sample theory -- parametric case. *Ann. Math. Statist.* **27** 1-22.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299-318.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146-158.

Gnedenko, B.V. (1954). Local limit theorem for densities. *Doklady Akad. Nauk. SSSR.* **95** 5-7.

Kolmogorov, A.N. and Gnedenko, B.V. (1954). *Limit Distributions for Sums of Independent Random Variables.* Translated by K.L. Chung, Addison-Wesley, Reading, Mass.

Kullback, S. (1967). A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory.* **IT-13** 126-127.

Kullback, S. (1959). *Information Theory and Statistics.* Wiley, New York.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79-86.

Linnik, Yu. V. (1959). An information-theoretic proof of the central limit theorem with the Lindeberg condition. *Theory Probab. Appl.* **4** 288-299.

Pinsker, M.S. (1964). *Information and Information Stability of Random Variables.* Translated by A. Feinstein, Holden-Day, San Francisco.

Pitman, E.J.G. (1979). *Some Basic Theory for Statistical Inference.* Chapman and Hall, London.

Prohorov, Yu. V. (1952). On a local limit theorem for densities. *Doklady Akad. Nauk. SSSR.* **83** 797-800.

Ranga Rao, R. and Varadarajan, V.S. (1960). A limit theorem for densities. *Sankhya.* **22** 261-266.

Rényi, A. (1970). *Probability Theory.* North-Holland, Amsterdam.

Rudin, W. (1974). *Real and Complex Analysis.* McGraw-Hill, New York.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 623-656.

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

Information Systems Lab
Stanford, University
Stanford, CA 94305