CONVERGENCE TO FISHER INFORMATION AND THE CENTRAL LIMIT THEOREM

Bogdan Gheorghe MUNTEANU*, **Doru LUCULESCU***

*„Henri Coandă" Air Forces Academy, Braşov

***Abstract**: In this paper, we give conditions for a $O(1/n)$ rate of convergence of Fisher information J and relative entropy D in the Central Limit Theorem. We use the Poincaré inequality and the theory of projections in L^2 to provide a better understanding of the decrease in Fisher information implied by results of Barron and Brown.*

***Keywords**: Normal convergence, entropy, Fisher information, Poincaré inequalities, rates of convergence.*

1. INTRODUCTION

Bounds on Shannon entropy and Fisher information have long been used in proofs of central limit theorems, based on quantification of the change in information as a result of convolution, as in the papers of Linnik (1959), [15], Shimizu (1975, [17]), Brown (1982, [5]), Barron (1986, [2]) and Jhonson (2000, [11]). Each of these papers have a final step involving completeness or uniform integrability in which a limit is taken without explicitly bounding the information distance from the normal distribution.

The purpose of the present paper is to provide an explicit rate of convergence of information distances, under certain natural conditions on the random variables. Let $X_1, X_2, ..., X_n$ be independent identically distributed random variables with mean zero, variance σ^2 and density function $p(x)$ satisfying Poincaré conditions (relating L^2 norms of mean zero functions to L^2 norms of their derivative), and let $\Phi_{\sigma^2}(x)$ be the corresponding $N(0,\sigma^2)$ density. The relative entropy distance is:

$$D(X) = \int p(x) \ln \frac{p(X)}{\Phi_{\sigma^2}(X)} dx \qquad (01)$$

In the case of random variables with differentiable densities, *the Fisher information distance* is

$$J(X) = \sigma^2 E\left[\frac{d}{dx}\ln p(X) - \frac{d}{dx}\ln \Phi_{\sigma^2}(X)\right]^2 \quad (02)$$

which is related to *the Fisher information* $I(X) = E\left[\frac{d}{dx}\ln p(X)\right]^2$. This is an L^2 norm between derivatives of log-densities, and gives a natural measure of convergence, stronger than those in existing theorems, as described in Lemma 1.6 in [13], where it is shown that if X is a random variable with density f and φ is a standard normal, then:

$$\sup_x |f(x) - \varphi(x)| \le \left(1 + \sqrt{\frac{6}{\pi}}\right)\sqrt{J(X)} \qquad (03)$$

and

$$\int |f(x) - \varphi(x)| dx \le 2 d_H(f,\varphi) \le \sqrt{2}\sqrt{J(X)} \quad (04)$$

Where $d_H(f,\varphi)$ is the Hellinger distance:

$$\left(\int \left|\sqrt{f(x)} - \sqrt{\varphi(x)}\right|^2 dx\right)^{1/2} \qquad (05)$$

At the same time, this lemma shows the relationship between convergence in Fisher information and several weaker forms of convergence. Recent work by Ball et al (2002,

[1]) has also considered the rate of convergence of these quantities. Their paper provides similar results, but by a very different method, involving transportation costs and a variation characterization of Fisher information.

The very important point of the present paper is the proof of the relationships (9) and (10) respectively, describing the Fisher information.

In examination the Fisher information a central role is played by *the score function:*

$$\rho(y) = \frac{d}{dy} \ln p(y) = \frac{p'(y)}{p(y)} \tag{06}$$

The score function of the sum of independent random variables, via a conditional expectation, has been used in proving the convolution inequalities for Fisher information and Shannon entropy (in the work of Stam [16], Blachman [3] and others).

In particular, Y_1 and Y_2 are independent and identically distributed with score function ρ, then the score of the sum $Y_1 + Y_2$ is the projection of $(\rho(Y_1) + \rho(Y_2))/2$ onto the linear space of functions of $Y_1 + Y_2$, so by the Pythagorean identity and rescaling, it follows:

$$\frac{I(Y_1) + I(Y_2)}{2} - I\left(\frac{Y_1 + Y_2}{\sqrt{2}}\right) = 2E\left(\overline{\rho}(Y_1 + Y_2) - \frac{\rho(Y_1) + \rho(Y_2)}{2}\right)^2 \tag{07}$$

Papers by Shimizu [17], Brown [5] and Barron [2] quantify the change in Fisher information with each doubling of the sample size, deducing convergence to the normal distribution along the powers of two subsequence and convergence of the entire information sequence $I(U_n)$, by subadditivity of $nI(U_n)$. However, these papers only rarely consider the behavior of the Fisher information for $X \sim Y + Z_\tau$ (for Z_τ a small normal perturbation).

In general, we can conclude that if the Fisher information $I(S_k)$ is always finite, since it is decreasing and bounded below, this difference sequence $I(S_k) - I(S_{k+1})$ tends to zero. The expression (07) measures the squared L^2 difference between a "ridge function" (a function of the sum $Y_1 + Y_2$) and an additive function (a function of the form $g_1(Y_1) + g_2(Y_2)$. From calculus it is known that, in general, the only functions:

$$g(y_1, y_2) = g_1(y_1) + g_2(y_2) \tag{08}$$

that are both ridge and additive are the affine functions $g_1(y_1) = ay_1 + b_1$ and $g_2(y_2) = ay_2 + b_2$ with a, b_1, b_2 constants, that is, the functions for which the derivatives $g_i'(y_i)$, $i = 1,2$ are constant and to one other.

By Lemma 3.1 of [5] (see also Barron, [2]) we have:

Lemma 1.1 [5]. For any two functions f and g there exist some constants a,b such that:

$$E(g(Y_1) - a_1 Y_1 - b) \le E(f(Y_1 + Y_2) - g(Y_1) - g(Y_2))^2 \tag{09}$$

where Y_1 and Y_2 are independent identically distributed normally.

The main technique used in the present paper will enable us to generalize Lemma 1.1 to a wider class of random variables Y_1, Y_2. For example, consider any Y_1 and Y_2 independent identically distributed with finite Fisher information I. Proposition 2.1 suggests us to take a differentiable ridge function $f(y_1 + y_2)$ with closest additive function g. Then for a certain constant μ:

$$E(g'(Y_1) - \mu)^2 \le I\, E(f(Y_1 + Y_2) - g(Y_1) - g(Y_2))^2 \tag{10}$$

Our proof starts with $f(Y_1 + Y_2)$, finds its additive part with $g(y_1) = E_{Y_2} f(y_1 + Y_2)$ and recognizes that:

$$g'(y_1) = -E_{Y_2} f(y_1 + Y_2) \cdot \rho(Y_2)$$

In this paper, the use of the Cauchy-Schwarz inequality completes the proof as detailed in Section 2.

Poincaré constant is the notion which provides a relationship between L^2 norms on functions and the L^2 norms on their derivatives:

Definition 1.1. Given a random variable Y, define the Poincaré constant R_Y:

$$R_Y = \sup_{g \in H_1(Y)} \frac{Eg^2(Y)}{Eg'(Y)^2} \quad (11)$$

where $H_1(Y)$ is the space of absolutely continuous functions g such that $\mathrm{Var}\, g(Y) > 0,\ Eg(Y) = 0$ and $Eg^2(Y) < \infty$ and the restricted Poincaré constant R_Y^*:

$$R_Y^* = \sup_{g \in H_1^*(Y)} \frac{Eg^2(Y)}{Eg'(Y)^2} \quad (12)$$

where $H_1^*(Y) = H_1(Y) \cap \{g : Eg'(Y) = 0\}$.

For certain Y, R_Y is infinite. However, R_Y is finite for the normal and other log-concave distributions (see, for example, [12], [9], [8], [7], [4]). Because $H_1^*(Y) \subseteq H_1(Y)$, then $R_Y^* \le R_Y$, that is we maximize over a smaller set of functions.

Further, for $Z \sim N(0, \sigma^2)$, $R_Z^* = \sigma^2/2$, with $g(x) = x^2 - \sigma^2$.

The other important definition that we need is that of weak differentiability, introduced in [10]. Brown and Gajek [6] and Lehmann and Casella [14] discuss this condition and provide easier check conditions under which it will hold.

Definition 1.2. A random variable Y has weakly differentiable density p if there exists a function $\rho \in L^2(p)$ (that is $E\rho^2(Y) < \infty$) such that for all f with $Ef(Y+u)^2 < \infty$, the function $g(u) = Ef(Y+u)$ has a derivative g'(u) equal to $-E[f(Y+u)\rho(Y)]$.

Poincaré constants are not finite for all distributions Y. Indeed, by Borovkov and Utev [4] if $R_Y < \infty$, then by considering $g_n(x) = |x|^n$, we inductively deduce that all the moments of Y are finite

$$R_Y = \sup\left[E\,Y^{2n} / \left(n^2 E\,Y^{2(n-1)}\right)\right] < \infty \Leftrightarrow$$
$$E\,Y^{\alpha} < \infty$$

The Berry-Essen theorem asserts that only $2+\delta$th moment conditions suffice to ensure an explicit $O\left(1/n^{\delta/2}\right)$ rate of weak convergence, for $0 < \delta \le 1$. The paper by Johnson and Barron describe a proof of Fisher information convergence under only second moment conditions, though without an explicit rate:

Theorem 1.1. [13]. Let $X_1, X_2, \ldots$ be weakly differentiable, independent identically distributed with finite variance σ^2 functions and define the normalized sum:

$$U_n = \left(\sum_{i=1}^{n} X_i\right) / \sqrt{n\sigma^2}\,.$$

If $J(U_m)$ is finite for some m, then $\lim_{n\to\infty} J(U_n) = 0$. This theorem extends Lemma 2 of Barron (1986, [2]), which holds only when *X* is of the form $Y + Z_\tau$ where Z_τ is a normal perturbation.

2. PROJECTION OF FUNCTIONS IN L^2

Although the main application of the following proposition will concern score functions, we present it as an abstract result concerning projection of functions in $L^2(Y_1, Y_2)$.

Proposition 2.1 [13]. Consider independent random variables Y_1, Y_2 with weakly differentiable densities and functions f, h_1, h_2 such that $E[f(Y_1+Y_2)]^2 < \infty$ and $E[f(Y_1+Y_2)] = 0$.

We find functions g_1, g_2 and a constant μ such that for any $\beta \in [0,1]$ we have:

$$E[f(Y_1+Y_2) - h_1(Y_1) - h_2(Y_2)]^2 \ge$$
$$E(g_1(Y_1) - h_1(Y_1))^2 + E(g_2(Y_2) - h_2(Y_2))^2 +$$
$$\left(\bar{I}\right)^{-1}\left(\beta\, E(g_1'(Y_1) - \mu)^2 + (1-\beta) E(g_2'(Y_2) - \mu)^2\right)$$

where $\bar{I} = (1-\beta) I(Y_1) + \beta I(Y_2)$.

Remark 2.1. We note that μ has the same value in both cases (for r_1 and r_2), because:

$$\mu = -E[g_2(Y_2)\rho_2(Y_2)] = E(g_2(Y_2))' \\ = -E[g_1(Y_1)\rho_1(Y_1)] = E(g_1(Y_1))' \quad (13)$$

Remark 2.2. In general, this inequality holds for any weakly differentiable Y_1, Y_2 with finite Fisher information (that is the score function is in L^2), whereas previous such expression have only held in the case of $Y_i \sim U_i + Z_\tau$ for some U_i.

Remark 2.3. This inequality holds for independent random variables that are not identically distributed. One may prove Central Limit Theorems giving information convergence to the normal for random variables satisfying a uniform Lindeberg type condition [11]. In certain cases we can provide a rate of convergence.

Remark 2.4. We can produce a similar expression using a similar method for finite dimensional random vectors Y_1 and Y_2. Weak differentiability can be defined in this case and $\rho_i = \frac{\partial}{\partial x_i} \ln p(x)$ will usually be the ith component of the score vector function ρ. In this case a similar analysis can lead to an alternative proof of the theorems in [13].

3. RATE OF CONVERGENCE OF FISHER INFORMATION

Let us extend Lemma 1.1 from the case of normal Y_1 and Y_2 to more general distributions, providing an explicit exponential rate of convergence of Fisher information, have finite restricted Poincaré constants R_1^* and R_2^*.The following lemma holds

Lemma 3.1. Let $S = Y_1 + Y_2$, where Y_1, Y_2 are independent and Y_2 is weakly differentiable with respect to the score function ρ_2. Then S is weakly differentiable with respect to the score function $\bar{\rho}(s) = E[\rho_2(Y_2)|S = s]$. Hence for independent weakly differentiable random variables Y_1 and Y_2 with respect to the score functions ρ_1 and ρ_2, $E(\rho_i(Y_i)) = 0$, $i = 1,2$ and writing $\bar{\rho}$ for the score function of S, we have:

$$\frac{I(Y_1) + I(Y_2)}{2} - I\left(\frac{Y_1 + Y_2}{\sqrt{2}}\right) = 2E\left(\bar{\rho}(S) - \frac{\rho_1(Y_1) + \rho_2(Y_2)}{2}\right) \quad (14)$$

Proposition 3.1. Let Y_1 and Y_2 be two independent identically distributed random variables which are weakly differentiable and have the variance σ^2 and restricted Poincaré constant R^*. Then

$$J\left(\frac{Y_1 + Y_2}{\sqrt{2}}\right) \le J(Y_1)\left(\frac{2R^*}{\sigma^2 + 2R^*}\right) \quad (15)$$

By performing succesive projections onto smaller additive spaces a more careful analysis generalises Proposition 3.1 to obtain the very important result of this section (see Theorem 3.1), for a given function f, define a series of functions by $f_n = f$ and, for m<n,

$$f_m\left(\frac{X_1 + \ldots + X_m}{\sqrt{n}}\right) = E_{X_{m+1}} f_{m+1}\left(\frac{X_1 + \ldots + X_m + X_{m+1}}{\sqrt{n}}\right) \quad (16)$$

Further, define:

$g(u) = \sqrt{n} E f\left(\frac{X_1 + \ldots + X_{n-1} + u}{\sqrt{n}}\right)$. At step i we approximate the function f by $f_i\left(\frac{X_1 + \ldots + X_i}{\sqrt{n}}\right)$ plus a sum of $g(X_j)$ for $j > i$, which is the best approximation onto the linear space of such partially additive functions.

Lemma 3.2. Define the squared distance between successive projections by:

$$t_i = E\left[f_i\left(\frac{X_1 + \ldots + X_i}{\sqrt{n}}\right) - f_{i-1}\left(\frac{X_1 + \ldots + X_{i-1}}{\sqrt{n}}\right) - \frac{1}{\sqrt{n}} g(X_i)\right]^2$$

Then, for independent identically distributed and weakly differentiable X_i, we have

$$t_i \ge \frac{i-1}{nI(X)} E[g'(X) - \mu]^2 \quad (17)$$

Lemma 3.3. For independente identically distributed X_i the sum of these squared

distances t_i is $s_n = \sum_{i=1}^{n} t_i$, where

$$s_m = E\left[f_m\left(\frac{X_1 + ... + X_m}{\sqrt{n}} \right) - \sum_{i=1}^{m} \frac{g(X_i)}{\sqrt{n}} \right]^2 \quad (18)$$

Combining Lemmas 3.2 and 3.3, we deduce that:

$$t_i \geq \frac{i-1}{nI(X)} E[g'(X) - \mu]^2 \Bigg| \sum_{i=1}^{n} \Rightarrow$$

$$s_n \geq \frac{1}{nI(X)} E[g'(X) - \mu]^2 \sum_{i=1}^{n} (i-1) \quad (19)$$

$$= \frac{n-1}{2I(X)} E[g'(X) - \mu]^2$$

Using this, one can extend the Brown inequality Lemma 1.1 (with a constant depending on $I(Y_1)$ and R_{Y_1} to a class of random variables wider than just normals. Since linear score functions correspond to the family of normal distributions, equations (7) and (10) provide a mean to prove the following Central Limit Theorems, which involves the Fisher information:

Theorem 3.1. Let $X_1, X_2, ..., X_n$ be a sequence of independent identically distributed random variables and with finite variance σ^2 and define the normalized sum $U_n = \left(\sum_{i=1}^{n} X_i \right) / \sqrt{n\sigma^2}$. If X_i are weakly differentiable with finite restricted Poincaré constant R^* then:

$$J(U_n) \leq \frac{2R^*}{(n-1)\sigma^2 + 2R^*} J(X), \quad \forall n \quad (20)$$

If X_i have finite Poincaré constant R, then

$$D(U_n) \leq \frac{2R}{(n-1)\sigma^2 + 2R} D(X), \quad \forall n \quad (21)$$

REFERENCES

1. Ball, K., Barthe, F., Naor, A., *Entropy jumps in the presence of a spectral gap*, preprint, 2002;
2. Barron, A., *Entropy and the Central Limit Theorem*, Ann. Probab., 141(1986);
3. Blachman, N.M., *The convolution inequality for entropy powers*, IEEE Trans. Inform. Theory, 11(1965), p.267-271;
4. Borovkov, A.A., Utev, S.A., *On an inequality and a related characterization of the normal distribution*, Theory Probab. Appl., 28(1984), p.219-228;
5. Brown, L.D., *A proof of the Central Limit Theorem motivated by the Cramer-Rao inequality*, Statistics and Probability: Essays in Honour of C.R. Rao, North-Holland, New-York, 1982, p.141-148;
6. Brown, L.D., Gajek, L., *Information inequalities for the Bayes risk*}, Ann. Statist., 18(1990), p.1578-1594;
7. Cacoullos, TH., *On upper and lower bounds for the variance of a function of a random variable*, Ann. Probab., 10(1982), p. 799-809;
8. Chen, L.H.Y., *An inequality for the normal distribution*, J. Multivariate Anal., 12(1982), p.306-315;
9. Chernoff, H., *A note on an inequality involving the normal distribution*, Ann. Probab., 9(1981), p.533-535;
10. Fabian, V., Hannan, J., *On the Cramer-Rao inequality*, Ann. Statist., 5(1977), p.197-205;
11. Johnson, O.T., *Entropy inequalities and the Central Limit Theorem*, Stochastic Process Appl., 88(2000), p.291-304;
12. Klaasen, C.A.J., *On a inequality of Chernoff*, Ann. Probab., 13(1985), p.966-974;
13. Johnson O.T., Suhov, Y.M., *Entropy and random vectors*, J. Statist Phys., 104(2001), p.147-167;
14. Lehmann, E., Casella, G., *Theory of point estimation*, 2nd ed., Springer Texts in Statistics, Springer, New-York, 1998;
15. Linnik, Y.V., *An information-theoretic proof of the Central Limit Theorem with the Lindeberg Condition*, Theory Probab. Appl., 4(1959), p.288-299;
16. Stam, A.J., *Some inequalities satisfied by the quantities of information of Fisher and Shannon*, Inform. and Control, 2(1959), p.101-112;
17. Shimizu, R., *On Fisher's amount of information for location family*, Statistical Distributions, 3(1975), p.305-312.