The Annals of Probability
1990, Vol. 18, No. 2, 840–850

THE CRAMÉR–RAO FUNCTIONAL AND LIMITING LAWS

By Eduardo Mayer-Wolf

University of Massachusetts and University of North Carolina

Some versions of the Fisher information matrix and the Cramér–Rao inequality are considered. We study properties of the Fisher matrix such as continuity and convexity and use the Cramér–Rao functional as a variational tool to prove convergence to Gaussian laws. These concepts are generalized to non-Gaussian limiting laws.

1. Introduction and notation. In the context of statistical estimation, the Cramér–Rao inequality is well known. Namely, the Fisher information matrix J_θ of order m,

$$J_\theta = E_\theta \frac{(\nabla_\theta\, p_\theta(X))(\nabla_\theta\, p_\theta(X))^T}{p_\theta(X)^2}$$

(∇_θ is the gradient operator with respect to θ, E_θ the expectation with density p_θ) is associated with any member of a family $\{p_\theta\}_{\theta \in H}$ of probability densities on $\mathbb{R}^n$, sufficiently smooth in $\theta \in H$ (an open set in $\mathbb{R}^m$). The Cramér–Rao inequality provides, in terms of J_θ, a lower bound for the error matrix of any unbiased estimator $\hat{\theta} = h(X)$:

$$E_\theta(\theta - \hat{\theta})(\theta - \hat{\theta})^T \ge J_\theta^{-1}, \qquad \forall\, \theta \in H. \tag{1.1}$$

(Throughout, the matrix inequality $A \ge B$ means that $A - B$ is a symmetric nonnegative matrix.) Consider the following particular case [cf. Pitman (1979), pages 36–39] in which $n = m$: If $p(x)$ is a "reasonable" density on $\mathbb{R}^n$ with $\mu = \int_{\mathbb{R}^n} x p(x)\, dx$ and if $p_\theta(x) = p(x - \theta)$ for each θ in some neighborhood of the origin, then clearly $h(x) = x - \mu$ is an unbiased estimator of the location parameter θ. For $\theta = 0$, (1.1) becomes

$$\operatorname{cov}(X) \ge J^{-1}, \tag{1.2}$$

where X is a random vector distributed with density p and J $(= J_0)$ is given by $E[(\nabla p(X))(\nabla p(X))^T/p(X)^2]$. Note that (1.2) is an inequality on the density p, not necessarily connected with any estimation setup, and has a simple direct proof.

Here, for (1.2), we shall assume fewer regularity conditions on p [$p \in H(\mathbb{R}^n)$ as in (2.2)] than the usual proofs require [cf. Rao (1973)].

This work concentrates on two issues connected with inequality (1.2). In Section 2 we shall consider properties of the Fisher matrix $J = J(\mu)$ as a function of probability measures (such as continuity, convexity, etc.). The main result in Section 2 is Theorem 2.1, which implies in particular that conver-

Received November 1987; revised November 1988.
AMS 1980 *subject classifications*. Primary 60F05; secondary 49B50.
Key words and phrases. Fisher matrix, Cramér–Rao functional, convergence in variation.

gence in distribution is equivalent to convergence in total variation for sequences of distributions with bounded trace of the Fisher information. Theorem 2.2 is known with additional assumptions [Barron (1986); Stam (1959)]. It is included here as an immediate consequence of Theorem 2.1. In Section 3 we shall exploit the simple fact that equality is achieved in (1.2) iff μ is Gaussian to establish that probability measures which "nearly" achieve equality in (1.2) are "close" to being normal, thus providing a tool for proving convergence to Gaussian laws.

In particular, we discuss in Section 3 the use of these methods by Brown (1982) and Barron (1986) in the classical CLT. Here we just mention that asymptotic equality in (1.2) for the normalized sums of iid's has not yet been established in full generality. On the other hand, we have succeeded in using the results of Section 3, extended to random measures, to prove a CLT in nonlinear filtering. Namely, if J_t^τ denotes the conditional probability measure of the state of a diffusion process x_t conditioned on the paths y_0^t of a second observation process, with τ denoting the *signal-to-noise ratio* in the observation process, then in probability as $\tau \to \infty$, J_t^τ, suitably normalized, converges to a Gaussian law [Mayer-Wolf (1987)].

Section 4 extends (1.2) to an inequality involving a weighted Fisher matrix. This extension allows one to obtain, along the lines of Section 3, convergence to limiting laws other than normal.

The following notation will be used. For any topological space X, $B(X)$ will denote the Borel σ-algebra, $P(X)$ is the set of probability measures on $(X, B(X))$ and λ_n is the Lebesgue measure on $(\mathbb{R}^n, B(\mathbb{R}^n))$. Weak convergence in $P(X)$ will be denoted as usual by $\Rightarrow$. We use the standard $C^k(X)$, $C_0(X)$ and $L^p(X)$ function spaces $1 \le p \le \infty$ (with respective norms $\|\cdot\|_\infty$, $\|\cdot\|_\infty, \|\cdot\|_p$). Also, for any finite signed measure μ on $(X, B(X))$, its total variations norm is $\|\mu\|_1 = \sup_{f \in C_0(X), \|f\|_\infty = 1} |\int_X f(x)\mu(dx)|$. In $P(X)$ we say that μ_n converges to μ in total variation and write $\mu_n \to_v \mu$ *if* $\|\mu - \mu_n\|_1 \to 0$.

The trace of a matrix $A = (a_{ij})$ is $\operatorname{tr} A = \sum_i a_{ii}$ and ∇ is the formal (weak) gradient operator acting on functions of n real variables $\nabla f = (\partial f/\partial x_i)_{i=1}^n$. Finally, the n-dimensional Gaussian densities are denoted by $\varphi_{a,\Lambda}(x) = ((2\pi)^n \det \Lambda)^{-1/2} \exp(-\frac{1}{2}(x-a)^T \Lambda^{-1}(x-a))$, the corresponding distribution functions by $\phi_{a,\Lambda}$ and the measure they induce on $(\mathbb{R}^n, B(\mathbb{R}^n))$ by $\mu_{a,\Lambda}$.

2. The Fisher matrix and its properties. For a region $D \subset \mathbb{R}^n$, let $W^{1,2}(D)$ be the Sobolev space of functions belonging to $L^2(D)$ together with its first (weak) derivatives [cf. Adams (1975)] equipped with the norm $\|f\|_{1,2} = (\|f\|_2^2 + \||\nabla f|\|_2^2)^{1/2}$, and define

$$\tilde{H}(D) = \left\{\mu \in P(D) | p = \frac{d\mu}{d\lambda_n} \text{ exists and } p^{1/2} \in W^{1,2}(D)\right\} \tag{2.1}$$

and

$$H(D) = \left\{\mu \in \tilde{H}(D) | \int_D |x|^2 \mu(dx) < \infty\right\}. \tag{2.2}$$

For $\mu \in \tilde{H}(D)$, we define the Fisher matrix

$$J(\mu) = 4\int_D (\nabla p^{1/2}(x))(\nabla p^{1/2}(x))^T\,dx, \tag{2.3}$$

while if $\mu \in H(D)$,

$$\bar{\mu} = \int_D x\mu(dx) \tag{2.4}$$

and

$$\Sigma(\mu) = \int_D (x-\bar{\mu})(x-\bar{\mu})^T\mu(dx). \tag{2.5}$$

REMARKS. (a) Setting $A = \{x \in D | p(x) = 0\}$, $\mu(A) = \int_D 1_A(x)p(x)\,dx = 0$. Then the matrix $J(\mu)$ can be rewritten as any of the following equivalent expressions:

$$\begin{aligned} &\int_D \frac{(\nabla p(x))(\nabla p(x))^T}{p(x)}\,dx \\ &\quad = \int_D \left(\frac{\nabla p(x)}{p(x)}\right)\left(\frac{\nabla p(x)}{p(x)}\right)^T \mu(dx) \\ &\quad = \int_D (\nabla \ln p(x))(\nabla \ln p(x))^T\mu(dx) \\ &\quad = -\int_D \nabla\nabla^T \ln p(x)\mu(dx). \end{aligned} \tag{2.6}$$

(The last term requires further regularity for p.)

(b) For $n = 1$, it follows from Sobolev's imbedding theorem [Adams (1975)] that if $\mu \in \tilde{H}(D)$, then p is necessarily bounded and continuous. For $n \geq 2$, this is no longer true.

(c) We shall not distinguish between a random vector X and the measure μ_X it induces. Thus, terminology such as $X \in \tilde{H}(D)$, $J(X)$ or "the mean of μ" will be adopted freely.

THEOREM 2.1. *Let D be a region in $\mathbb{R}^n$ and $(\mu_k)_{k\geq 1}$ a sequence in $\tilde{H}(D)$ with densities p_k and $\mu_k \Rightarrow \mu \in P(D)$ as $k \to \infty$. Furthermore, assume that*

$$\operatorname{tr} J(\mu_k) \leq M < \infty, \qquad \forall\, k \geq 1. \tag{2.7}$$

Then:

(i) $\mu \in \tilde{H}(D)$ *(with density p).*
(ii) $\operatorname{tr} J(\mu) \leq \liminf \operatorname{tr} J(\mu_k)$.
(iii) $p_k \to p$ *in L^1 as $k \to \infty$ (equivalently $\mu_k \to_v \mu$).*

PROOF. First we note that if $\nu \in \tilde{H}(D)$ with density π, then

$$\|\pi^{1/2}\|_{1,2}^2 = 1 + \tfrac{1}{4}\operatorname{tr} J(\nu). \tag{2.8}$$

Assume now that D is a bounded region. By (2.7), $p_k^{1/2}$ is a bounded sequence in $W^{1,2}(D)$ and thus by the Rellich–Kondrachov compact imbedding theorem [Adams (1975), Theorem 6.2] a compact sequence in $L^2(D)$. It follows that there exists a subsequence of $p_k^{1/2}$ (which, without loss of generality we shall assume to be $p_k^{1/2}$ itself) such that $p_k^{1/2} \to g_0$ in the weak topology of $W^{1/2}(D)$ and $p_k^{1/2} \to \tilde{g}_0$ in $L^2(D)$. It is straightforward to verify that $g_0 = \tilde{g}_0$ a.e. We have

$$\|p_k - g_0^2\|_1 \le \|p_k^{1/2} + g_0\|_2 \|p_k^{1/2} - g_0\|_2 \le 2\|p_k^{1/2} - g_0\|_2 \to 0,$$

which proves that g_0^2 is indeed μ's density p and that conclusion (iii) holds. Since $g_0 \in W^{1,2}(D)$, (i) also holds. Furthermore, since in any Banach space the norm is a lower semicontinuous function in the weak topology and in view of (2.8),

$$\operatorname{tr} J(\mu) = 4\|g_0\|_{1,2}^2 - 4 \le 4 \liminf_{k\to\infty} \|p_k^{1/2}\|_{1,2}^2 - 4 = \liminf_{k\to\infty} \operatorname{tr}(J(\mu_k)),$$

which proves (ii).

Now let D be an arbitrary region, not necessarily bounded. Choose $\psi \in C^2(\mathbb{R})$ such that $0 \le \psi(\rho) \le 1$ $\forall$ $\rho \in \mathbb{R}$, $\psi|_{(-\infty,0]} \equiv 1$, $\psi|_{[1,\infty)} \equiv 0$ and $(\psi'^2(\rho))/(\psi(\rho)) \le C_1 < \infty$ $\forall$ $\rho \in (0,1)$. [An example of such a function is $\psi(\rho) = \exp\{-\rho^3/(1-\rho^2)\}$, $0 < \rho < 1$.] For a sufficiently large positive N, we truncate any probability measure η with density π to a probability measure η^N by defining its density $\pi^N(x) = \alpha^N(\pi)\psi(|x| - N)\pi(x)$, where $\alpha^N(\pi)$ is a positive normalizing factor. Note that as $N \to \infty$, $\alpha^N(\eta) \to 1$ and $\pi^N \to \pi$ in $L_1(D)$, uniformly on any tight family of probability measures. Furthermore, if $\eta \in \tilde{H}(D)$, one can verify directly that $\eta^N \in \tilde{H}(D)$ and

$$\operatorname{tr} J(\eta^N) \le \frac{1}{\alpha_N(\eta)}\left[\operatorname{tr}(J(\eta)) + C\eta(\{|x| > N\})\right], \tag{2.9}$$

where C is a constant independent of N and η.

Returning to the given sequence μ_k, for any sufficiently large N it is obvious that $\mu_k^N \Rightarrow \mu^N$ and from (2.9) it follows that (2.7) is satisfied for $\{\mu_k^N\}$, hence the theorem's conclusions (i), (ii) and (iii) hold for the truncated measures, i.e., $\mu^N \in \tilde{H}(D \cap \{|x| \le N+1\})$ (with a density we shall call p^N by an abuse of notation) and $p_k^N \to p^N$ in L^1.

Now, being weakly convergent, $\{\mu_k\}$ is a tight sequence so that, by a previous remark, $\{p_k^N\}_{N\ge 1}$ are L^1-Cauchy sequences uniformly in $k = 1, 2, \dots$, by which $\{p^N\}_{N\ge1}$ is also Cauchy, thus converging in L^1 to a density p which is obviously μ's density. From (2.9) and the tightness of $\{\mu_k\}$, it follows that for each $\epsilon > 0$,

$$\begin{aligned}\int_{\{|x|\le N\}} \frac{|\nabla p|^2}{p}(x)\,dx &\le \alpha_N(\mu)\operatorname{tr}\big(J(\mu^N)\big) \le \alpha_N(\mu) \liminf_{k\to\infty} \operatorname{tr}\big(J(\mu_k^N)\big)\\ &\le \alpha_N(\mu) \liminf_{k\to\infty} \frac{1}{\alpha_N(\mu_k)}\left[\operatorname{tr}(J(\mu_k)) + C\mu_k(\{|x| > N\})\right]\\ &\le (1+\epsilon) \liminf_{k\to\infty} \operatorname{tr} J(\mu_k) + \epsilon\end{aligned}$$

for all N sufficiently large, from which we obtain the theorem's first two conclusions. Finally, conclusion (iii) follows by using the equivalent result for the truncated measures and a standard $\epsilon/3$ argument. □

REMARKS. (a) Among other things, what Theorem 2.1 says is that in $P(D)$, equipped with the topology of weak convergence:

(1) For each $M > 0$, $\tilde{H}_M(D) = [\mu \in \tilde{H}(D) \mid \mathrm{tr}(J(\mu)) \le M]$ is closed.
(2) $\mathrm{tr}\, J(\mu)$ is lower semicontinuous on $\tilde{H}(D)$.

(b) Let $w \in \mathbb{R}^n$ have positive components and consider the weighted trace of a matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$, namely,

$$\mathrm{tr}_w\, A = \sum_{i=1}^{n} w_i a_{ii}. \tag{2.10}$$

Then Theorem 2.1 obviously remains true if tr is replaced everywhere by tr_w. [The Sobolev space $W^{1,2}(D)$ should be equipped in this case with the equivalent norm $\|f\|_{1,2,w} = (\|f\|_2^2 + \sum_{i=1}^n w_i \|\partial f/\partial x_i\|_2)^{1/2}$.]

LEMMA 2.1. *Let X be a random vector in $\tilde{H}(\mathbb{R}^n)$, $b \in \mathbb{R}^n$ and A a nonsingular $n \times n$ matrix. Then $AX + b \in \tilde{H}(\mathbb{R}^n)$ and $J(AX+b) = A^{-1T}J(X)A^{-1}$.*

PROOF. Follows immediately from the definitions. □

COROLLARY 2.1. *Conclusion* (ii) *of Theorem* 2.1 *can be replaced by the stronger statement*:

(ii′) *For each $\alpha \in \mathbb{R}^n$, $\alpha^T J(\mu)\alpha \le \liminf_{k\to\infty} \alpha^T J(\mu_k)\alpha$.*

In particular the function $\alpha^T J(\cdot)\alpha$, extended to $+\infty$ on $P(D)\setminus \tilde{H}(D)$, is lower semicontinuous on $P(D)$.

PROOF. First let $\epsilon > 0$ and conclude (ii) of Theorem 2.1 with tr_{w_ϵ} replacing tr (as in the remark following the theorem) where $w_\epsilon = (1, \epsilon, \epsilon, \dots, \epsilon)^T \in \mathbb{R}^n$. Letting $\epsilon \to 0$ we obtain $J(\mu)_{11} \le \liminf_{k\to\infty} J(\mu_k)_{11}$.

Next, given $0 \ne \alpha \in \mathbb{R}^n$, consider a nonsingular matrix A with α as its first column. Letting $\tilde{\mu} = \mu \circ A$, $\tilde{\mu}_k = \mu_k \circ A$ and using Lemma 2.1,

$$\begin{aligned}\alpha^T J(\mu)\alpha &= \big(A^T J(\mu)A\big)_{11} = J(\tilde{\mu})_{11} \le \liminf_{k\to\infty} J(\tilde{\mu}_k)_{11}\\ &= \liminf_{k\to\infty}\big(A^T J(\mu_k)A\big)_{11} = \liminf_{k\to\infty} \alpha^T J(\mu_k)\alpha.\end{aligned}$$ □

LEMMA 2.2. *For any region $D \subset \mathbb{R}^n$, $\tilde{H}(D)$ is a convex subset of $P(D)$ and $J(\cdot)$ is a convex (matrix-valued) function on $\tilde{H}(D)$.*

PROOF. If $\mu_i \in \tilde{H}(D)$ with respective densities p_i, $i = 1, 2$, and if $\alpha, \beta \ge 0$ and $\alpha + \beta = 1$, set $\mu = \alpha\mu_1 + \beta\mu_2$ (whose density is $p = \alpha p_1 + \beta p_2$). The lemma follows then from the equality

$$\left(\frac{(\nabla p)(\nabla p)^T}{p}\right)(x) = \alpha\left(\frac{(\nabla p_1)(\nabla p_1)^T}{p_1}\right)(x) + \beta\left(\frac{(\nabla p_2)(\nabla p_2)^T}{p_2}\right)(x)$$
$$- \alpha\beta\left(\frac{(p_1 \nabla p_2 - p_2 \nabla p_1)(p_1 \nabla p_2 - p_2 \nabla p_1)^T}{p p_1 p_2}\right)(x),$$

which holds for any x for which $p(x) > 0$. □

THEOREM 2.2. *If $\mu \in \tilde{H}(D_1)$ and $\eta \in P(D_2)$, then $\mu * \eta \in \tilde{H}(D_1 + D_2)$ and*

$$J(\mu * \eta) \leq J(\mu). \tag{2.11}$$

PROOF. First, if μ_{x_0} is the shifted measure $\mu_{x_0}(B) = \mu(B - x_0)$, obviously $J(\mu_{x_0}) = J(\mu)$ (see also Lemma 2.1). So by extending Lemma 2.2 to all finite convex combinations, inequality (2.11) is true for η of the form

$$\eta = \sum_{i=1}^{m} \alpha_i \, \delta_{x_i}. \tag{2.12}$$

Now, an arbitrary η can be approximated weakly by a sequence (η_k) of measures of the form (2.12) which thus satisfy $J(\mu * \eta_k) \leq J(\mu)$. Since obviously $\mu * \eta_k \Rightarrow \mu * \eta$, (2.11) follows from Corollary 2.1. □

COROLLARY 2.2. *For every $\mu \in \tilde{H}(\mathbb{R}^n)$, there exists a sequence (μ_k) in $\tilde{H}(\mathbb{R}^n)$ with μ_k having positive C^∞ densities for all k such that $\mu_k \Rightarrow \mu$ and $J(\mu_k) \to J(\mu)$ as $k \to \infty$.*

PROOF. Simply define $\mu_k = \mu * \mu_{0, k^{-1}I}$ [recall $\mu_{\alpha, \Lambda} \sim N(\alpha, \Lambda)$]. Obviously $\mu_k \Rightarrow \mu$ and from Theorem 2.2, $J(\mu_k) \leq J(\mu)$ $\forall$ k. These two facts and Theorem 2.1 imply that $\operatorname{tr} J(\mu) \leq \liminf_{k \to \infty} \operatorname{tr}(J(\mu_k))$, from which the corollary follows. □

For completeness' sake we shall include the following result on the Fisher matrix of marginal laws.

LEMMA 2.3. *Let X be a random vector in $\tilde{H}(\mathbb{R}^n)$ and consider the partition $X = (X^{(1)}, X^{(2)})$, $X^{(1)} \in \mathbb{R}^k$, $X^{(2)} \in \mathbb{R}^{n-k}$ with $1 \leq k < n$. Then $X^{(1)} \in \tilde{H}(\mathbb{R}^k)$ and*

$$J(X^{(1)}) \leq (J(X)_{i,j})_{i,j=1}^{k}.$$

PROOF. See Bobrovsky, Mayer-Wolf and Zakai (1987), Proposition 1. □

Finally, we state the Cramér–Rao inequality (1.2).

THEOREM 2.3. *Let X be a random vector in $H(\mathbb{R}^n)$. Then*

$$\Sigma(X) \geq J(X)^{-1} \tag{2.13}$$

with equality if and only if X is Gaussian.

REMARKS. (a) The theorem implicitly states that $J(X)$ is nonsingular.

(b) Equivalent formulations of (2.13) are

$$J(X) \geq \Sigma(X)^{-1}, \qquad \Sigma(X)^{1/2} J(X) \Sigma(X)^{1/2} \geq I$$

and so on.

(c) The assumption $X \in H(\mathbb{R}^n)$ allows us to verify that

$$\int (x - EX)(\nabla p)^T(x)\, dx = -I$$

(by integration by parts) and the proof may be completed as in Rao (1973), page 327. A detailed proof of a more general result appears in Bobrovsky, Mayer-Wolf and Zakai (1987).

Motivated by (2.13), we define the Cramér–Rao "functional"

$$\Gamma(\mu) = \Sigma(\mu)^{1/2} J(\mu) \Sigma(\mu)^{1/2} - I. \tag{2.14}$$

3. The Cramér–Rao functional and divergence from normality. A natural question which arises from Theorem 2.3 is whether the matrix valued Cramér–Rao functional $\Gamma(\mu)$ defined in (2.14) can serve as an effective measure of divergence of μ from normality. In this section we shall provide a positive answer in the sense that (loosely) $\Gamma(\mu) \to 0$ implies that μ_n tends to a Gaussian probability measure.

LEMMA 3.1. *Let $\mu \in H(\mathbb{R})$ have density p, mean a and variance σ^2. Then*

$$\|p - \varphi_{a_0, \sigma_0^2}\|_1 \leq \sqrt{\frac{\pi}{2}\Gamma(\mu)} + \frac{1}{\sigma_0}\left(\sqrt{\frac{2}{\pi}}|a - a_0| + 2|\sigma - \sigma_0|\right). \tag{3.1}$$

PROOF. First assume $a = a_0$ and $\sigma = \sigma_0$ and without loss of generality we may assume they are, respectively, 0 and 1 since both the LHS of (3.1) and $\Gamma(\mu)$ are invariant under an affine transformation of the underlying random variables.

Denote

$$g(x) = p'(x) + xp(x). \tag{3.2}$$

Note that

$$\begin{aligned} \|g\|_1 &= \int_{-\infty}^{\infty} \left|\frac{p'(x)}{p(x)} + x\right| \mu(dx) \leq \left(\int_{-\infty}^{\infty} \left|\frac{p'(x)}{p(x)} + x\right|^2 \mu(dx)\right)^{1/2} \\ &= \left(J(\mu) + 2\int_{-\infty}^{\infty} xp'(x)\, dx + 1\right)^{1/2} = \Gamma(\mu)^{1/2}. \end{aligned} \tag{3.3}$$

Now solving the ODE (3.2), we may write

$$p(x) = e^{-x^2/2}\left[\frac{1}{\sqrt{2\pi}} + \int_{-\infty}^{\infty} e^{s^2/2} g(x)\left(\phi_{0,1}(s) - 1_{[0,\infty)}(s-x)\right) ds\right],$$

so that

$$\begin{aligned}\|p - \varphi_{0,1}\|_1 &\le \int_{-\infty}^{\infty} e^{-x^2/2}\int_{-\infty}^{\infty} e^{s^2/2}|g(s)||\phi_{0,1}(s) - 1_{[0,\infty)}(s-x)|\,ds\,dx\\ &= \int_{-\infty}^{\infty} e^{s^2/2}|g(s)|\,ds\left[\int_{-\infty}^{s} e^{-x^2/2}(1-\phi_{0,1}(s))\,dx\right.\\ &\qquad\left. + \int_{s}^{\infty} e^{-x^2/2}\phi(s)\,dx\right]\\ &= 2\sqrt{2\pi}\int_{-\infty}^{\infty} e^{s^2/2}|g(s)|\phi_{0,1}(s)(1-\phi_{0,1}(s))\,ds.\end{aligned}$$

Now the continuous function $h(s) = e^{s^2/2}\phi_{0,1}(s)(1-\phi_{0,1}(s))$ tends to 0 at $\pm\infty$ so that it is bounded. Actually it is not hard to prove that $\|h\|_\infty = h(0) = \frac{1}{4}$, from which we conclude $\|p - \varphi_{0,1}\|_1 \le \sqrt{\pi/2}\,\|g\|_1$, which together with (3.3) proves the lemma in this particular case.

In the general case just note that

$$\|p - \varphi_{a_0,\sigma_0^2}\|_1 \le \|p - \varphi_{a,\sigma^2}\|_1 + \|\varphi_{a,\sigma^2} - \varphi_{a_0,\sigma_0^2}\|_1$$

and it is a straightforward exercise to verify that

$$\|\varphi_{a,\sigma^2} - \varphi_{a_0,\sigma_0^2}\|_1 \le \frac{1}{\sigma_0}\left(\sqrt{\frac{2}{\pi}}\,|a-a_0| + 2|\sigma-\sigma_0|\right). \qquad \square$$

REMARKS. (a) An inequality

$$\|p - \varphi_{a_0,\sigma_0^2}\|_q \le C\left(\Gamma(\mu)^{1/2} + |a-a_0| + |\sigma-\sigma_0|\right) \tag{3.1'}$$

can be obtained along the same lines as Lemma 3.1 for $q = \infty$ and by interpolation for $1 < q < \infty$ as well.

(b) A sharper inequality than (3.1) follows from the chain (take $a_0 = 0$, $\sigma_0 = 1$) $\|p - \varphi_{0,1}\|_1^2 \le 2\int p \log(p/\varphi_{0,1}) \le \log(1+\Gamma(p))$ [Csiszár (1967), Theorem 4.1 and Stam (1959), Equation (2.3)]. However, this proof does not extend to bounds on the uniform norm as in the previous remark.

THEOREM 3.1. *Let Z_j, $j \ge 1$, be a sequence of random vectors in $H(\mathbb{R}^n)$. If* (i) $EZ_j \to a$, $\text{cov}(Z_j) \to \Lambda$ *(positive definite) and* (ii) $\Gamma(Z_j) \to 0$, *then* $Z_j \to_v Z \sim N(a, \Lambda)$.

PROOF. The case $n = 1$ follows directly from Lemma 3.1. For the general case choose $\alpha \in \mathbb{R}^n$ with $|\alpha| = 1$ and an orthonormal matrix A with α^T as its first row. For any random vector X, it follows from Lemmas 2.1 and 2.3 that

$$J(\alpha^T X) = J((AX)_1) \le (J(AX))_{11} = \alpha^T J(X)\alpha, \tag{3.4}$$

so that denoting $\tilde{Z}_j = (\operatorname{cov}(Z_j))^{-1/2} Z_j$, we have $E\alpha^T \tilde{Z}_j \to \alpha^T \Lambda^{-1/2} a$, $\operatorname{cov}(\alpha^T \tilde{Z}_j) = 1$ and, by (3.4),

$$\Gamma(\alpha^T \tilde{Z}_j) \le \alpha^T J(\tilde{Z}_j)\alpha^T - 1 = \alpha^T(\Gamma(Z_j) + I)\alpha - 1 \to 0,$$

so that from the scalar case we conclude that for all $\alpha \in \mathbb{R}^n$, $\alpha^T \tilde{Z}_j \to_v \alpha^T \tilde{Z}$, where $\tilde{Z} \sim N(\Lambda^{-1/2} a, I)$. By the Cramér–Wold technique [Billingsley (1968), Theorem 7.7] it follows that $\tilde{Z}_j \Rightarrow \tilde{Z}$ or $Z_j \Rightarrow Z \sim N(a, \Lambda)$.

Note that only weak convergence was obtained because the Cramér–Wold method does not extend to L^1 convergence. However, denoting $\Lambda_j = \operatorname{cov}(Z_j)$,

$$\begin{aligned} \operatorname{tr} J(Z_j) &= \operatorname{tr}\{(I + \Gamma(Z_j))\Lambda_j^{-1}\} \le \operatorname{tr}(I + \Gamma(Z_j))\operatorname{tr}(\Lambda_j^{-1}) \\ &\le \left\{n + \sup_j (\operatorname{tr}\Gamma(Z_j))\right\} \sup_j \operatorname{tr}(\Lambda_j^{-1}) < \infty, \end{aligned}$$

so by Theorem 2.1, $Z_j \Rightarrow Z$ implies $Z_j \to_v Z$. □

As an illustration, we shall conclude this section by relating its contents to the classical central limit theorem where X_n are i.i.d. mean 0, variance 1 random variables in $H(\mathbb{R})$ and $S_n = n^{-1/2}(\sum_{i=1}^n X_i)$. Direct applications of the results in Section 2 yield simple proofs to known convolution inequalities [cf. Stam (1959), Equation (2.9)]; in particular $(n+m)J(S_{n+m}) \le nJ(S_n) + mJ(S_m)$, which shows by subadditivity that $L = \lim_{n\to\infty} J(S_n)$ exists. This was enough for Brown (1982) to provide a new proof of the classical CLT. Barron (1986) then showed that $L = 1$ if we assume that $X_1 = U + \delta Z$ [$\delta > 0$ and $Z \sim N(0,1)$ independent of U], thus proving a total variation CLT in this case. Moreover he obtained a CLT for "convergence in entropy" which still implies convergence in L^1 but is weaker than $L = 1$. The question whether $L = 1$ in general remains open. [We note that L^1 convergence has already been established in Prohorov (1952); for $X_1 \in H$ it also follows directly from our Theorem 2.1.]

4. Non-Gaussian limit laws. A natural question which arises is whether the techniques of Section 3 can be modified so as to yield limiting laws other than Gaussian. In this section we shall show a way of doing this; we restrict ourselves to the scalar case $n = 1$.

Let $\sigma(x)$ be a measurable function of a real variable satisfying

$$0 < c_1 \le \sigma(x) \le c_2 < \infty, \qquad \forall\, x \in \mathbb{R}, \tag{4.1}$$

and consider the positive measure λ_σ on $\mathbb{R}$ defined by $d\lambda_\sigma/d\lambda = \sigma$ (λ being the Lebesgue measure). If $D = (a, b)$, define just as in (2.1),

$$\tilde{H}_\sigma(D) = \left\{\mu \in P(D) \mid p \doteq \frac{d\mu}{d\lambda} \text{ exists and } (\sigma p)^{1/2} \in W^{1,2}(D,\sigma)\right\}, \tag{4.2}$$

where $W^{1,2}(D,\sigma)$ is the weighted Sobolev space $\{f\colon f, f' \in L^2(D, \lambda_\sigma)\}$. In other words, $\mu \in \tilde{H}_\sigma(D)$ if its density p satisfies

$$J_\sigma(\mu) := \int_D \frac{((\sigma p)'(x))^2}{p(x)} dx < \infty. \tag{4.3}$$

[As before, for a random variable X, $J_\sigma(X)$ stands for $J_\sigma(\mu_X)$.] Also $H_\sigma(D) = \{\mu \in \tilde{H}_\sigma(D) | \Sigma(\mu) < \infty\}$.

THEOREM 4.1. *If $\mu \in H_\sigma(\mathbb{R})$ with density p, then*

$$J_\sigma(\mu)\Sigma(\mu) \geq \left(\int_{-\infty}^{\infty} \sigma(x)\mu(dx)\right)^2. \tag{4.4}$$

Equality is achieved in (4.4) *if and only if there exist constants $\alpha \in \mathbb{R}$, $\beta > 0$ and $C(\alpha, \beta) > 0$ such that*

$$p(x) = \varphi^{[\sigma]}_{\alpha,\beta^2} := \frac{C(\alpha,\beta)}{\sigma(x)} \exp\left(-\frac{1}{\beta^2}\int_\alpha^x \frac{(t-\alpha)}{\sigma(t)}\,dt\right). \tag{4.5}$$

PROOF. Inequality (4.4) is simply the Cauchy–Schwarz inequality $(f, g)^2 \leq \|f\|^2\|g\|^2$ in $L^2(\mathbb{R}, \mu)$ when $f(x) = (x - \bar{\mu})$ and $g = (\sigma p)'/p$. The fact that $|(f,g)| = |\int_{-\infty}^{\infty}\sigma(x)\mu(dx)|$ follows in the same way as for the particular case $\sigma \equiv 1$ (Theorem 2.3). Similarly, the functions $\varphi^{[\sigma]}_{\alpha,\beta^2}$ defined in (4.5) are the only solutions to the equation $(\sigma p)'(x) = \gamma p(x)(x - \alpha)$, arising from the condition for equality in the Cauchy–Schwarz inequality, which are probability densities. □

We shall denote the probability measure whose density is $\varphi^{[\sigma]}_{\alpha,\beta^2}$ by $\mu^{[\sigma]}_{\alpha,\beta^2}$ and motivated by Theorem 4.1 define the nonnegative functional $\Gamma_\sigma(\mu) = J_\sigma(\mu)\Sigma(\mu) - (\int_{-\infty}^{\infty}\sigma(x)\mu(dx))^2$. Let $F(x) = \int_{x_0}^x (1/\sigma(t))\,dt$ with $x_0 = 0$ and for any probability measure μ, $\Sigma_\sigma(\mu) = \int_{-\infty}^{\infty}(x - \bar{\mu})F(x)\mu(dx)$ [finite or $+\infty$; note that the value of x_0 in the definition of $F(x)$ is immaterial].

We also observe that $\bar{\mu}^{[\sigma]}_{\alpha,\beta^2} = \alpha$ and $\Sigma_\sigma(\mu^{[\sigma]}_{\alpha,\beta^2}) = \beta^2$, so that the natural parameters of the family of measures achieving equality in (4.4) are the mean $\bar{\mu}$ and the "distorted variance" $\Sigma_\sigma(\mu)$. Note as well that $J_\sigma(X) = J(F(X))$.

THEOREM 4.2. *Let σ be a function which satisfies* (4.1) *and (μ_j) a sequence of measures in $H_\sigma(D)$ such that when $j \to \infty$:*

(i) $\bar{\mu}_j \to \alpha \in \mathbb{R}$ *and* $\Sigma_\sigma(\mu_j) \to \beta^2 > 0$.

(ii) $\Gamma_\sigma(\mu_j) \to 0$.

Then $\mu_j \to_v \mu^{[\sigma]}_{\alpha,\beta^2}$.

PROOF. Denote $\alpha_j = \bar{\mu}_j$ and $\beta_j^2 = \Sigma_\sigma(\mu_j)$. Since

$$\|\mu_j - \mu^{[\sigma]}_{\alpha,\beta^2}\|_1 \leq \|\mu_j - \mu^{[\sigma]}_{\alpha_j,\beta_j^2}\|_1 + \|\mu^{[\sigma]}_{\alpha_j,\beta_j^2} - \mu^{[\sigma]}_{\alpha,\beta^2}\|_1$$

and the second term in the RHS obviously tends to 0, it suffices to show that

$\|\mu_j - \mu^{[\sigma]}_{\alpha_j, \beta_j^2}\|_1 \to 0$. For that we proceed just as in Lemma 3.1: Denote

$$K_j = \left(\int_{-\infty}^{\infty} \sigma(x)\mu_j(dx)\right) \Big/ \left(\int_{-\infty}^{\infty} (x - \alpha_j)^2 \mu_j(dx)\right), \tag{4.6}$$
$$g_j(x) = (\sigma p_j)'(x) + K_j(x - \alpha_j)p_j(x), \qquad p_j(x) = d\mu_j/d\lambda(x).$$

Solving (4.6), we obtain

$$\begin{aligned}(4.7)\qquad &\left(p_j(x) - \varphi^{[\sigma]}_{\alpha_j, \beta_j^2}(x)\right)\\ &\quad = \frac{e^{-K_j G_j(x)}}{\sigma(x)} \int_{-\infty}^{\infty} g_j(t) e^{K_j G_j(t)} \left[\phi^{[\sigma]}_{\alpha_j, \beta_j^2}(t) - \mathbf{1}_{[0,\infty)}(t - x)\right] dt,\end{aligned}$$

where $G_j(u) = \int_{\alpha_j}^{u} (v - \alpha_j)/(\sigma(v))\, dv$ and $\phi^{[\sigma]}_{\alpha, \beta^2}(u) = \int_{-\infty}^{u} \varphi^{[\sigma]}_{\alpha, \beta^2}(v)\, dv$. From (4.7), straightforward computations (which are direct generalizations of those in Lemma 3.1) show that

$$\|p_j - \varphi^{[\sigma]}_{\alpha_j, \beta_j^2}\|_1 \le C\|g_j\|_1 \le C'\Gamma_\sigma(\mu_j)^{1/2},$$

where the positive constants C and C' are independent of j (because $\inf_j \beta_j > 0$), thus completing the proof. □

Acknowledgments. This paper is based on parts of my Ph.D. thesis. I would like to thank my advisors, Professor M. Zakai and Professor P. Feigin, for their important ideas and suggestions. I would also like to express my appreciation to the referee for his keen interest and numerous suggestions.

REFERENCES

ADAMS, R. A. (1975). *Sobolev Spaces*. Academic, New York.

BARRON, A. (1986). Entropy and the central limit theorem. *Ann. Probab.* **14** 336–342.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BOBROVSKY, B. Z., MAYER-WOLF, E. and ZAKAI, M. (1987). Some classes of global Cramér–Rao inequalities. *Ann. Statist.* **15** 1421–1438.

BROWN, L. D. (1982). A proof of the central limit theorem motivated by the Cramér–Rao inequality. In *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh, eds.) 141–148. North-Holland, Amsterdam.

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

MAYER-WOLF, E. (1987). Asymptotic laws for the nonlinear filtering problem. Ph.D. dissertation, Technion. (In Hebrew.)

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. *Dokl. Akad. Nauk SSSR* **83** 797–800.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

STAM, A. J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inform. and Control* **2** 101–112.

VAN TREES, H. (1968). *Estimation and Modulation Theory* **1** Wiley, New York.

DEPARTMENT OF MATHEMATICS
TECHNION, I.I.T.
HAIFA 32000
ISRAEL