19

Another Look at Differentiability in Quadratic Mean

David Pollard[1]

ABSTRACT This note revisits the delightfully subtle interconnections between three ideas: differentiability, in an $\mathcal{L}^2$ sense, of the square-root of a probability density; local asymptotic normality; and contiguity.

19.1 A mystery

The traditional regularity conditions for maximum likelihood theory involve existence of two or three derivatives of the density functions, together with domination assumptions to justify differentiation under integral signs. Le Cam (1970) noted that such conditions are unnecessarily stringent. He commented:

> Even if one is not interested in the maximum economy of assumptions one cannot escape practical statistical problems in which apparently "slight" violations of the assumptions occur. For instance the derivatives fail to exist at one point x which may depend on θ, or the distributions may not be mutually absolutely continuous or a variety of other difficulties may occur. The existing literature is rather unclear about what may happen in these circumstances. Note also that since the conditions are imposed upon probability densities they may be satisfied for one choice of such densities but not for certain other choices.

Probably Le Cam had in mind examples such as the double exponential density, $\frac{1}{2}\exp(-|x-\theta|)$, for which differentiability fails at the point $\theta = x$. He showed that the traditional conditions can be replaced by a simpler assumption of differentiability in quadratic mean (DQM): differentiability in norm of the square root of the density as an element of an $\mathcal{L}^2$ space. Much asymptotic theory can be made to work under DQM. In particular, as Le Cam showed, it implies a quadratic approximation property for the log-likelihoods known as local asymptotic normality (LAN).

Le Cam's idea is simple but subtle. When I first encountered the LAN property I wrongly dismissed it as nothing more than a Taylor expansion to quadratic terms of the log-likelihood. Le Cam's DQM result showed otherwise:

[1]Yale University

one appears to get the benefit of the quadratic expansion without paying the twice-differentiability price usually demanded by such a Taylor expansion. How can that happen?

My initial puzzlement was not completely allayed by a study of several careful accounts of LAN, such as those of Le Cam (1970; 1986, Section 17.3), Ibragimov & Has'minskii (1981, page 114), Millar (1983, page 105), Le Cam & Yang (1990, page 101), or Strasser (1985, Chapter 12). None of the proofs left me with the feeling that I really understood why second derivatives are not needed. (No criticism of those authors intended, of course.)

Eventually it dawned on me that I had overlooked a vital ingredient in the proofs: the square root of a density is not just an element of an $\mathcal{L}^2$ space: *it is an element with norm* 1. By rearranging some of the standard arguments I hope to convince the gentle reader of this note that the fixed norm is the real reason for why an assumption of one-times differentiability (in quadratic mean) can convey the benefits usually associated with two-times differentiability. I claim that the Lemma in the next Section is the key to understanding the role of DQM.

19.2 A lemma

The concept of differentiability makes sense for maps into an arbitrary normed space $(\mathcal{L}, \|\cdot\|)$. For the purposes of my exposition, it suffices to consider the case where the norm is generated by an inner product, $\langle\cdot,\cdot\rangle$. In fact, $\mathcal{L}$ will be $\mathcal{L}^2(\lambda)$, the space of functions square-integrable with respect to some measure λ, but that simplification will play no role for the moment.

A map ξ from $\mathbb{R}^k$ into $\mathcal{L}$ is said to be differentiable at a point θ_0 with derivative Δ, if $\xi(\theta) = \xi(\theta_0) + \Delta(\theta - \theta_0) + r(\theta)$ near θ_0, where $\|r(\theta)\| = o(|\theta - \theta_0|)$ as θ tends to θ_0. The derivative Δ is linear; it may be identified with a k-vector of elements from $\mathcal{L}$.

For a differentiable map, the Cauchy-Schwarz inequality implies that $\langle \xi(\theta_0), r(\theta)\rangle = o(|\theta - \theta_0|)$. It would usually be a blunder to assume naively that the bound must therefore be of order $O(|\theta - \theta_0|^2)$; typically, higher-order differentiability assumptions are needed to derive approximations with smaller errors. However, if $\|\xi(\theta)\|$ is constant—that is, if the function is constrained to take values lying on the surface of a sphere—then the naive assumption turns out to be no blunder. Indeed, in that case, $\langle \xi(\theta_0), r(\theta)\rangle$ can be written as a quadratic in $\theta - \theta_0$ plus an error of order $o(|\theta - \theta_0|^2)$. The sequential form of the assertion is more convenient for my purposes.

(1) **Lemma** *Let $\{\delta_n\}$ be a sequence of constants tending to zero. Let ξ_0, ξ_1, ... be elements of norm one for which $\xi_n = \xi_0 + \delta_n W + r_n$, with W a fixed element of $\mathcal{L}$ and $\|r_n\| = o(\delta_n)$. Then $\langle \xi_0, W\rangle = 0$ and $\langle \xi_0, r_n\rangle = -\frac{1}{2}\delta_n^2 \|W\|^2 + o(\delta_n^2)$.*

Proof. Because both ξ_n and ξ_0 have unit length,

$$\begin{aligned} 0 = \|\xi_n\|^2 - \|\xi_0\|^2 &= 2\delta_n\langle\xi_0, W\rangle && \text{order } O(\delta_n) \\ &\quad + 2\langle\xi_0, r_n\rangle && \text{order } o(\delta_n) \\ &\quad + \delta_n^2\|W\|^2 && \text{order } O(\delta_n^2) \\ &\quad + 2\delta_n\langle W, r_n\rangle + \|r_n\|^2 && \text{order } o(\delta_n^2). \end{aligned}$$

On the right-hand side I have indicated the order at which the various contributions tend to zero. (The Cauchy-Schwarz inequality delivers the $o(\delta_n)$ and $o(\delta_n^2)$ terms.) The exact zero on the left-hand side leaves the leading $2\delta_n\langle\xi_0, W\rangle$ unhappily exposed as the only $O(\delta_n)$ term. It must be of smaller order, which can happen only if $\langle\xi_0, W\rangle = 0$, leaving

$$0 = 2\langle\xi_0, r_n\rangle + \delta_n^2\|W\|^2 + \mathrm{o}(\delta_n^2),$$

as asserted. □

Without the fixed length property, the inner product $\langle\xi_0, r_n\rangle$, which inherits $\mathrm{o}(\delta_n)$ behaviour from $\|r_n\|$, might not decrease at the $\mathrm{O}(\delta_n^2)$ rate.

19.3 A theorem

Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures on a space $(\mathcal{X}, \mathcal{A})$, indexed by a subset Θ of $\mathbb{R}^k$. Suppose P_θ has density $f(x, \theta)$ with respect to a sigma-finite measure λ.

Under the classical regularity conditions—twice continuous differentiability of $\log f(x, \theta)$ with respect to θ, with a dominated second derivative—the likelihood ratio

$$\prod_{i\le n} \frac{f(x_i, \theta)}{f(x_i, \theta_0)}$$

enjoys the LAN property. Write $L_n(t)$ for the likelihood ratio evaluated at θ equal to $\theta_0 + t/\sqrt{n}$. The property asserts that, if the $\{x_i\}$ are sampled independently from P_{θ_0}, then

$$L_n(t) = \exp\left(t'S_n - \tfrac{1}{2}t'\Gamma t + o_p(1)\right) \qquad \text{for each } t, \tag{2}$$

where Γ is a fixed matrix (depending on θ_0) and S_n has a centered asymptotic normal distribution with variance matrix Γ.

Formally, the LAN approximation results from the usual pointwise Taylor expansion of the log density $g(x, \theta) = \log f(x, \theta)$, following a style of argument familiar to most graduate students. For example, in one dimension,

$$\begin{aligned} \log L_n(\theta_0 + t/\sqrt{n}) &= \sum_{i\le n}\left(g(x_i, \theta_0 + t/\sqrt{n}) - g(x_i, \theta_0)\right) \\ &= \frac{t}{\sqrt{n}}\sum_{i\le n} g'(x_i, \theta_0) + \frac{t^2}{2n}\sum_{i\le n} g''(x_i, \theta_0) + \dots, \end{aligned}$$

which suggests that S_n be the standardized score function,

$$\frac{1}{\sqrt{n}} \sum_{i \le n} g'(x_i, \theta_0) \rightsquigarrow N\big(0, \operatorname{var}_{\theta_0} g'(x, \theta_0)\big),$$

and Γ should be the information function,

$$-P_{\theta_0} g''(x, \theta_0) = \operatorname{var}_{\theta_0} g'(x, \theta_0).$$

The dual representation for Γ allows one to eliminate all mention of second derivatives from the statement of the LAN approximation, which hints that two derivatives might not really be needed, as Le Cam (1970) showed.

In general, the family of densities is said to be differentiable in quadratic mean at θ_0 if the square root $\xi(x, \theta) = \sqrt{f(x, \theta)}$ is differentiable in the $\mathcal{L}^2(\lambda)$ sense: for some k-vector $\Delta(x)$ of functions in $\mathcal{L}^2(\lambda)$,

$$\xi(x, \theta) = \xi(x, \theta_0) + (\theta - \theta_0)' \Delta(x) + r(x, \theta), \tag{3}$$

where

$$\lambda |r(x, \theta)|^2 = o(|\theta - \theta_0|^2) \qquad \text{as } \theta \to \theta_0.$$

Let us abbreviate $\xi(x, \theta_0)$ to $\xi_0(x)$ and $\Delta(x)/\xi_0(x)$ to $D(x)$. From (3) one almost gets the LAN property.

(4) **Theorem** *Assume the DQM property (3). For each fixed t the likelihood ratio has the approximation, under $\{\mathbb{P}_{n,\theta_0}\}$,*

$$L_n(t) = \exp\left(t' S_n - \tfrac{1}{2} t' \Gamma t + o_p(1)\right),$$

where

$$S_n = \frac{2}{\sqrt{n}} \sum_{i \le n} D(x_i) \rightsquigarrow \mathrm{N}(0, \mathbb{I}_0) \qquad \text{and} \qquad \Gamma = \tfrac{1}{2}\mathbb{I}_0 + \tfrac{1}{2}\mathbb{I},$$

with $\mathbb{I}_0 = 4\lambda(\Delta\Delta' \{\xi_0 > 0\})$ and $\mathbb{I} = 4\lambda(\Delta\Delta')$.

Notice the slight difference between Γ and the limiting variance matrix for S_n. At least formally, $2D(x)$ equals the derivative of $\log f(x, \theta)$: ignoring problems related to division by zero and distinctions between pointwise and $\mathcal{L}^2(\lambda)$ differentiability, we have

$$2D(x) = \frac{2}{\sqrt{f(x, \theta_0)}} \frac{\partial}{\partial \theta} \sqrt{f(x, \theta_0)} = \frac{\partial}{\partial \theta} \log f(x, \theta_0).$$

Also, Γ again corresponds to the information matrix, expressed in its variance form, except for the intrusion of the indicator function $\{\xi_0 > 0\}$. The extra indicator is necessary if we wish to be careful about 0/0. Its presence is related to the property called contiguity—another of Le Cam's great ideas—as is explained in Section 5.

At first sight the derivation of Theorem 4 from assumption (3) again appears to be a simple matter of a Taylor expansion to quadratic terms of the log likelihood ratio. Writing $R_n(x) = r(x, \theta_0 + t/\sqrt{n})/\xi_0(x)$, we have

$$\begin{aligned}\log L_n(t) &= \sum_{i\le n} 2\log \frac{\xi(x_i, \theta_0 + t/\sqrt{n})}{\xi(x_i, \theta_0)} \\ &= \sum_{i\le n} 2\log\left(1 + \frac{t'}{\sqrt{n}} D(x_i) + R_n(x_i)\right).\end{aligned}$$

From the Taylor expansion of $\log(\cdot)$ about 1, the sum of logarithms can be written as a formal series,

$$\begin{aligned}&2\sum_{i\le n}\left(\frac{t}{\sqrt{n}} D(x_i) + R_n(x_i)\right) - \sum_{i\le n}\left(\frac{t'}{\sqrt{n}} D(x_i) + R_n(x_i)\right)^2 + \ldots \\ &\quad = \frac{2t'}{\sqrt{n}}\sum_{i\le n} D(x_i) + 2\sum_{i\le n} R_n(x_i) - \frac{1}{n}\sum_{i\le n}\left(t'D(x_i)\right)^2 + \ldots \qquad (5)\end{aligned}$$

The first sum on the right-hand side gives the $t'S_n$ in Theorem 4. The law of large numbers gives convergence of the third term to $t'P_{\theta_0}DD't$. Mere one-times differentiability might not seem enough to dispose of the second sum. Each summand has standard deviation of order $o(1/\sqrt{n})$, by DQM. A sum of n such terms could crudely be bounded via a triangle inequality, leaving a quantity of order $o(\sqrt{n})$, which clearly would not suffice. In fact the sum of the $R_n(x_i)$ does not go away in the limit; as a consequence of Lemma 1, it contributes a fixed quadratic in t. That contribution is the surprise behind DQM.

19.4 A proof

Let me write $\mathbb{P}_n$ to denote calculations under the assumption that the observations $x_1, \ldots, x_n$ are sampled independently from P_{θ_0}. The ratio $f(x_i, \theta_0 + t/\sqrt{n})/f(x_i, \theta_0)$ is not well defined when $f(x_i, \theta_0) = 0$, but under $\mathbb{P}_n$ the problem can be neglected because

$$\mathbb{P}_n\{f(x_i, \theta_0) = 0 \text{ for at least one } i\} = 0.$$

For other probability measures that are not absolutely continuous with respect to $\mathbb{P}_n$, one should be more careful. It pays to be quite explicit about behaviour when $f(x_i, \theta_0) = 0$ for some i, by including an explicit indicator function $\{\xi_0 > 0\}$ as a factor in any expressions with a ξ_0 in the denominator.

Define D_i to be the random vector $\Delta(x_i)\{\xi_0(x_i) > 0\}/\xi_0(x_i)$, and, for a fixed t, define

$$R_{i,n} = r(\xi_i, \theta_0 + t/\sqrt{n})\{\xi_0(x_i) > 0\}/\xi_0(x_i).$$

Then

$$\frac{\xi(x_i, \theta_0 + t/\sqrt{n})}{\xi_0(x_i)}\{\xi_0(_i) > 0\} = 1 + t'D_i + R_{i,n}.$$

The random vector D_i has expected value $\lambda(\xi_0\Delta)$, which, by Lemma 1, is zero, even without the traditional regularity assumptions that justify differentiation under an integral sign. It has variance $\frac{1}{4}\mathbb{I}_0$. It follows by a central limit theorem that

$$S_n = \frac{2}{\sqrt{n}} \sum_{i\le n} D_i \rightsquigarrow \mathrm{N}(0, \mathbb{I}_0).$$

Also, by a (weak) law of large numbers,

$$\frac{1}{n} \sum_{i\le n} D_i D_i' \to \mathbb{P}_n(D_1 D_1') = \tfrac{1}{4}\mathbb{I}_0 \qquad \text{in probability.} \tag{6}$$

To establish rigorously the near-LAN assertion of Theorem 4, it is merely a matter of bounding the error terms in (5) and then justifying the treatment of the sum of the $R_n(x_i)$. Three facts are needed.

(7) **Lemma** *Under $\{\mathbb{P}_n\}$, assuming DQM,*

(a) $\max_{i\le n} |D_i| = o_p(\sqrt{n})$,

(b) $\max_{i\le n} |R_{i,n}| = o_p(1)$,

(c) $\sum_{i\le n} 2R_{i,n} \to -\frac{1}{4}t'\mathbb{I}t$ *in probability.*

Let me first explain how Theorem 4 follows from Lemma 7. Together the two facts (a) and (b) ensure that with high probability $\log L_n(t)$ does not involve infinite values. For $(t'D_i/\sqrt{n}) + R_{i,n} > -1$ we may then an appeal to the Taylor expansion

$$\log(1+y) = y - \tfrac{1}{2}y^2 + \tfrac{1}{2}\beta(y),$$

where $\beta(y) = \mathrm{o}(y^2)$ as y tends to zero, to deduce that $\log L_n(t)$ equals

$$\frac{2}{\sqrt{n}} \sum_{i\le n} t'D_i + 2\sum_{i\le n} R_{i,n} - \sum_{i\le n} \left(\frac{t'D_i}{\sqrt{n}} + R_{i,n}\right)^2 + \sum_{i\le n} \beta\left(\frac{t'D_i}{\sqrt{n}} + R_{i,n}\right),$$

which expands to

$$t'S_n + 2\sum_{i\le n} R_{i,n} - \frac{1}{n}\sum_{i\le n}(t'D_i)^2 - \frac{2}{\sqrt{n}}\sum_{i\le n} t'D_i R_{i,n} - \sum_{i\le n} R_{i,n}^2 + \mathrm{o}_p(1)\sum_{i\le n}\left(\frac{|D_i|^2}{n} + R_{i,n}^2\right).$$

Each of the last three sums is of order $\mathrm{o}_p(1)$ because $\sum_{i\le n} |D_i|^2/n = O_p(1)$ and

$$\begin{aligned} \mathbb{P}_n \textstyle\sum_{i\le n} R_{i,n}^2 &= n\lambda\big(\xi_0^2 r(x_1, \theta_0 + t/\sqrt{n})\{\xi_0 > 0\}/\xi_0^2\big) \\ &\le n\lambda |r(\cdot, \theta_0 + t/\sqrt{n})|^2 \\ &= o(1). \end{aligned} \tag{8}$$

By virtue of (6) and (c), the expansion simplifies to

$$t'S_n - \tfrac{1}{4}t'\mathbb{I}t - \tfrac{1}{4}t'\mathbb{I}_0 t + \mathrm{o}_p(1),$$

as asserted by Theorem 4.

Proof of Lemma 7. Assertion (a) follows from the identical distributions:

$$\begin{aligned}\mathbb{P}_n\{\max_{i\le n}|D_i| > \epsilon\sqrt{n}\} &\le \sum_{i\le n}\mathbb{P}_n\{|D_i| > \epsilon\sqrt{n}\}\\ &= n\mathbb{P}_n\{|\Delta_1| > \epsilon\sqrt{n}\}\\ &\le \epsilon^{-2}\lambda\Delta_1^2\{|\Delta_1| > \xi_0\epsilon\sqrt{n}\}\\ &\to 0 \qquad \text{by Dominated Convergence.}\end{aligned}$$

Assertion (b) follows from (8):

$$\mathbb{P}_n\{\max_{i\le n}|R_{i,n}| > \epsilon\} \le \epsilon^{-2}\mathbb{P}_n\sum_{i\le n}R_{i,n}^2 \to 0.$$

Only Assertion (c) involves any subtlety. The variance of the sum is bounded by $4\sum_{i\le n}\mathbb{P}_n R_n(x_i)^2$, which tends to zero. The sum of the remainders must lie within $o_p(1)$ of its expected value, which equals

$$2nP_{\theta_0}R_{1,n} = 2n\lambda\big(\xi_0 r(\cdot,\theta_0+t/\sqrt{n})\big),$$

an inner product between two functions in $\mathcal{L}^2(\lambda)$. Notice that the ξ_0 factor makes the indicator $\{\xi_0 > 0\}$ redundant.

It is here that the unit length property becomes important. Specializing Lemma 1 to the case $\delta_n = 1/\sqrt{n}$, with $\xi_n(x) = \xi(x,\theta_0+t/\sqrt{n})$ and $W = t'\Delta$, we get the approximation to the sum of expected values of the $R_{i,n}$, from which Assertion (c) follows. □

A slight generalization of the LAN assertion is possible. It is not necessary that we consider only parameters of the form $\theta_0 + t/\sqrt{n}$ for a fixed t. By arguing almost as above along convergent subsequences of $\{t_n\}$ we could prove an analog of Theorem 4 if t were replaced by a bounded sequence $\{t_n\}$ such that $\theta_0 + t_n/\sqrt{n} \in \Theta$. The extension is significant because (Le Cam 1986, page 584) the slightly stronger result forces a form of differentiability in quadratic mean.

19.5 Contiguity and disappearance of mass

For notational simplicity, consider only the one-dimensional case with the typical value $t = 1$. Let ξ_n^2 be the marginal density, and $\mathbb{Q}_n$ be the joint distribution, for $x_1,\ldots,x_n$ sampled with parameter value $\theta_0 + 1/\sqrt{n}$. As before, ξ_0^2 and $\mathbb{P}_n$ correspond to θ_0. The measure $\mathbb{Q}_n$ is absolutely continuous with respect to $\mathbb{P}_n$ if and only if it puts zero mass in the set

$$A_n = \{\xi_0(x_i) = 0 \text{ for at least one } i \le n\}.$$

Writing α_n for $\lambda\xi_n^2\{\xi_0 = 0\}$, we have

$$\mathbb{Q}_n A_n = 1 - \prod_{i\le n}\big(1 - \mathbb{Q}_n\{\xi_0(x_i) = 0\}\big) = 1 - (1-\alpha_n)^n.$$

By direct calculation,

$$\alpha_n = \lambda\big(r_n + \Delta/\sqrt{n}\big)^2\{\xi_0 = 0\} = \lambda\Delta^2\{\xi_0 = 0\}/n + o(1/n).$$

The quantity $\tau = \lambda\Delta^2\{\xi_0 = 0\}$ has the following significance. Under $\mathbb{Q}_n$, the number of observations landing in A_n has approximately a Poisson(τ) distribution; and $\mathbb{Q}_n A_n \to 1 - e^{-\tau}$.

In some asymptotic sense, the measure $\mathbb{Q}_n$ becomes more nearly absolutely continuous with respect to $\mathbb{P}_n$ if and only if $\tau = 0$. The precise sense is called contiguity: *the sequence of measures $\{\mathbb{Q}_n\}$ is said to be contiguous with respect to $\{\mathbb{P}_n\}$ if $\mathbb{Q}_n B_n \to 0$ for each sequence of sets $\{B_n\}$ such that $\mathbb{P}_n B_n \to 0$.* Because $\mathbb{P}_n A_n = 0$ for every n, the condition $\tau = 0$ is clearly necessary for contiguity. It is also sufficient.

Contiguity follows from the assertion that L, the limit in distribution under $\{\mathbb{P}_n\}$ of the likelihood ratios $\{L_n(1)\}$, have expected value one. ("Le Cam's first lemma"—see the theorem on page 20 of Le Cam and Yang, 1990.) The argument is simple: If $\mathbb{P}L = 1$ then, to each $\epsilon > 0$ there exists a finite constant C such that $\mathbb{P}L\{L < C\} > 1-\epsilon$. From the convergence in distribution, $\mathbb{P}_n L_n\{L_n < C\} > 1 - \epsilon$ eventually. If $\mathbb{P}_n B_n \to 0$ then

$$\begin{aligned}\mathbb{Q}_n B_n &\le \mathbb{P}_n B_n L_n\{L_n < C\} + \mathbb{Q}_n\{L_n \ge C\}\\ &\le C\mathbb{P}_n B_n + 1 - \mathbb{P}_n L_n\{L_n < C\}\\ &< 2\epsilon \qquad \text{eventually.}\end{aligned}$$

For the special case of the limiting $\exp(N(\mu, \sigma^2))$ distribution, where $\mu = -\frac{1}{4}\mathbb{I}_0 - \frac{1}{4}\mathbb{I}$ and $\sigma^2 = \mathbb{I}_0$, the requirement becomes

$$1 = \mathbb{P}\exp\big(N(\mu, \sigma^2)\big) = \exp\big(\mu + \tfrac{1}{2}\sigma^2\big).$$

That is, contiguity obtains when $\mathbb{I}_0 = \mathbb{I}$ (or equivalently, $\lambda(\Delta^2\{\xi_0 = 0\}) = 0$), in which case, the limiting variance of S_n equals Γ. This conclusion plays the same role as the traditional dual representation for the information function. As Le Cam & Yang (1990, page 23) commented, "The equality ... is the classical one. One finds it for instance in the standard treatment of maximum likelihood estimation under Cramér's conditions. There it is derived from conditions of differentiability under the integral sign." The fortuitous equality is nothing more than contiguity in disguise.

From the literature one sometimes gets the impression that $\lambda\Delta^2\{\xi_0 = 0\}$ is always zero. It is not.

(9) **Example** Let λ be Lebesgue measure on the real line. Define

$$f_0(x) = x\{0 \le x \le 1\} + (2 - x)\{1 < x \le 2\}.$$

For $0 \le \theta \le 1$ define densities

$$f(x, \theta) = (1 - \theta^2)\, f_0(x) + \theta^2\, f_0(x - 2).$$

Notice that

$$\lambda\left|\sqrt{f(x,\theta)} - \sqrt{f(x,0)} - \theta\sqrt{f(x,1)}\right|^2 = (\sqrt{1-\theta^2} - 1)^2 = O(\theta^4). \tag{10}$$

The family of densities is differentiable in quadratic mean at $\theta = 0$ with derivative $\Delta(x) = \sqrt{f(x,1)}$. For this family, $\lambda\Delta^2\{\xi_0 = 0\} = 1$.

The near-LAN assertion of Theorem 4 degenerates: $\mathbb{I}_0 = 0$ and $\mathbb{I} = 4$, giving $L_n(t) \to \exp\left(-t^2\right)$ in probability, under $\{\mathbb{P}_{n,\theta_0}\}$. Indeed, as Aad van der Vaart has pointed out to me, the limiting experiment (in Le Cam's sense) for the models $\{\mathbb{P}_{n,t/\sqrt{n}} : 0 \le t \le \sqrt{n}\}$ is not the Gaussian translation model corresponding to the LAN condition. Instead, the limit experiment is $\{\mathbb{Q}_t : t \ge 0\}$, with $\mathbb{Q}_t$ equal to the Poisson(t^2) distribution. That is, for each finite set T and each h, under $\{\mathbb{P}_{n,h/\sqrt{n}}\}$ the random vectors

$$\left(\frac{d\mathbb{P}_{n,t/\sqrt{n}}}{d\mathbb{P}_{n,h/\sqrt{n}}} : t \in T\right)$$

converge in distribution to

$$\left(\frac{d\mathbb{Q}_t}{d\mathbb{Q}_h} : t \in T\right),$$

as a random vector under the $\mathbb{Q}_h$ distribution. □

The counterexample would not work if θ were allowed to take on negative values; one would need $\Delta(x) = -\sqrt{f(x,1)}$ to get the analog of (10) for negative θ. The failure of contiguity is directly related to the fact that $\theta = 0$ lies on boundary of the parameter interval.

In general, $\lambda\Delta\Delta'\{\xi_0 = 0\}$ must be zero at all interior points of the parameter space where DQM holds. On the set $\{\xi_0 = 0\}$ we have $0 \le \sqrt{n}\xi(x, \theta_0 + t/\sqrt{n}) = t'\Delta + \sqrt{n}r_n$, where $\|\sqrt{n}r_n\| \to 0$. Along a subsequence, $\sqrt{n}r_n \to 0$, leaving the conclusion that $t'\Delta \ge 0$ almost everywhere on the set $\{\xi_0 = 0\}$. At an interior point, t can range over all directions, which forces $\Delta = 0$ almost everywhere on $\{\xi = 0\}$; at an interior point, $\Delta\Delta'\{\xi = 0\} = 0$ almost everywhere. More generally, one needs only to be able to approach θ_0 from enough different directions to force $\Delta = 0$ on $\{\xi_0 = 0\}$—as in the concept of a contingent in Le Cam & Yang (1990, Section 6.2).

The assumption that θ_0 lies in the interior of the parameter space is not always easy to spot in the literature.

Some authors, such as Le Cam & Yang (1990, page 101), prefer to dispense with the dominating measure λ, by recasting differentiability in quadratic mean as a property of the densities $d\mathbb{P}_\theta / d\mathbb{P}_{\theta_0}$, whose square roots correspond to the ratios $\xi(x,\theta)\{\xi_0 > 0\}/\xi_0(x)$. With that approach, the behaviour of Δ on the set $\{\xi_0 = 0\}$ must be specified explicitly. The contiguity requirement—that P_θ puts, at worst, mass of order $o(|\theta - \theta_0|^2)$ in the set $\{\xi_0 = 0\}$—is then made part of the definition of differentiability in quadratic mean.

19.6 References

Ibragimov, I. A. & Has'minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

Le Cam, L. (1970), 'On the assumptions used to prove asymptotic normality of maximum likelihood estimators', *Annals of Mathematical Statistics* **41**, 802–828.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.

Millar, P. W. (1983), 'The minimax principle in asymptotic statistical theory', *Springer Lecture Notes in Mathematics* pp. 75–265.

Strasser, H. (1985), *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*, De Gruyter, Berlin.