**IMS Collections**

# Asymptotics of Maximum Likelihood without the LLN or CLT or Sample Size Going to Infinity

**Charles J. Geyer**

*University of Minnesota*

**Abstract:** If the log likelihood is approximately quadratic with constant Hessian, then the maximum likelihood estimator (MLE) is approximately normally distributed. No other assumptions are required. We do not need independent and identically distributed data. We do not need the law of large numbers (LLN) or the central limit theorem (CLT). We do not need sample size going to infinity or anything going to infinity.

Presented here is a combination of Le Cam style theory involving local asymptotic normality (LAN) and local asymptotic mixed normality (LAMN) and Cramér style theory involving derivatives and Fisher information. The main tool is convergence in law of the log likelihood function and its derivatives considered as random elements of a Polish space of continuous functions with the metric of uniform convergence on compact sets. We obtain results for both one-step-Newton estimators and Newton-iterated-to-convergence estimators.

## 1. Introduction

The asymptotics of maximum likelihood is beautiful theory. If you can calculate two derivatives of the log likelihood, you can find the maximum likelihood estimate (MLE) and its asymptotic normal distribution.

But why does this work? And when does it work? The literature contains many different treatments, many theorems with long lists of assumptions, each slightly different from the others, and long messy calculations in their proofs. But they give no insight because neither assumptions nor calculations are sharp. The beauty of the theory is hidden by the mess.

In this article we explore an elegant theory of the asymptotics of likelihood inference inspired by the work of Lucien Le Cam, presented in full generality in Le Cam (1986) and in somewhat simplified form in Le Cam and Yang (2000). Most statisticians find even the simplified version abstract and difficult. Here we attempt to bring this theory down to a level most statisticians can understand. We take from Le Cam two simple ideas.

- If the log likelihood is approximately quadratic with constant Hessian, then the MLE is approximately normal. This is the theory of locally asymptotically normal (LAN) and locally asymptotically mixed normal (LAMN) models.

School of Statistics
University of Minnesota
313 Ford Hall
224 Church St. S. E.
Minneapolis, MN 55455
e-mail: geyer@umn.edu

*AMS 2000 subject classifications:* Primary 62F12, 60F05; secondary 62F40

*Keywords and phrases:* maximum likelihood, locally asymptotically normal, Newton's method, parametric bootstrap, no-$n$ asymptotics, quadraticity

- Asymptotic theory does not need $n$ going to infinity. We can dispense with sequences of models, and instead compare the actual model for the actual data to an LAN or LAMN model.

Although these ideas are not new, being "well known" to a handful of theoreticians, they are not widely understood. When I tell typical statisticians that the asymptotics of maximum likelihood have nothing to do with the law of large numbers (LLN) or the central limit theorem (CLT) but rather with how close the log likelihood is to quadratic, I usually get blank stares. That's not what they learned in their theory class. Worse, many statisticians have come to the conclusion that since the "$n$ goes to infinity" story makes no sense, asymptotics are bunk. It is a shame that so many statisticians are so confused about a crucial aspect of statistics.

### *1.1. Local Asymptotic Normality*

Chapter 6 of Le Cam and Yang (2000) presents the theory of local asymptotically normal (LAN) and locally asymptotically mixed normal (LAMN) sequences of models. This theory has no assumption of independent and identically distributed data (nor even stationarity and weak dependence). The $n$ indexing an LAN or LAMN sequence of models is just an index that need not have anything to do with sample size or anything analogous to sample size. Thus the LLN and the CLT do not apply, and asymptotic normality must arise from some other source.

Surprising to those who haven't seen it before, asymptotic normality arises from the log likelihood being asymptotically quadratic. A statistical model with exactly quadratic log likelihood

$$l(\theta) = U + Z'\theta - \tfrac{1}{2}\theta' K \theta, \qquad \theta \in \mathbb{R}^p \tag{1}$$

where $U$ is a random scalar, $Z$ is a random $p$ vector, and $K$ is a random $p \times p$ symmetric matrix, has observed Fisher information $K$ and maximum likelihood estimator (MLE)

$$\hat{\theta} = K^{-1} Z \tag{2}$$

(if $K$ is positive definite). If $K$ is constant, then the distribution of the MLE is

$$\hat{\theta} \sim \mathcal{N}(\theta, K^{-1}), \tag{3}$$

the right hand side being the multivariate normal distribution with mean vector $\theta$ and variance matrix $K^{-1}$ (Corollary 2.2 below). If $K$ is random, but *invariant in law*, meaning its distribution does not depend on the parameter $\theta$, then (3) holds conditional on $K$ (Theorem 2.1 below).

The essence of likelihood asymptotics consists of the idea that if the log likelihood is only approximately of the form (1) with $K$ invariant in law, then (3) holds approximately too. All likelihood asymptotics that produce conclusions resembling (3) are formalizations of this idea. The idea may be lost in messy proofs, but it's what really makes likelihood inference work the way it does.

### *1.2. "No n" Asymptotics*

By "no $n$ asymptotics" we mean asymptotics done without reference to any sequence of statistical models. There is no $n$ going to infinity. This is not a new idea, as the following quotation from Le Cam (1986, p. xiv) shows.

> From time to time results are stated as limit theorems obtainable as something called $n$ "tends to infinity." This is especially so in Chapter 7 [Some Limit Theorems] where the results are just limit theorems. Otherwise we have made a special effort to state the results in such a way that they could eventually be transformed into approximation results. Indeed, limit theorems "as $n$ tends to infinity" are logically devoid of content about what happens at any particular $n$. All they can do is suggest certain approaches whose performance must then be checked on the case at hand. Unfortunately the approximation bounds we could get were too often too crude and cumbersome to be of any practical use. Thus we have let $n$ tend to infinity, but we urge the reader to think of the material in approximation terms, especially in subjects such as ones described in Chapter 11 [Asymptotic Normality—Global].

Le Cam's point that asymptotics "are logically devoid of content about what happens at any particular $n$" refers to the fact that convergence of a sequence tells us nothing about any initial segment of the sequence. An estimator $\hat{\theta}_n$ that is equal to 42 for all $n < 10^{10}$ and equal to the MLE thereafter is asymptotically equivalent to the MLE. Strictly speaking, asymptotics—at least the usual story about $n$ going to infinity—does not distinguish between these estimators and says you might just as well use one as the other.

The story about $n$ going to infinity is even less plausible in spatial statistics and statistical genetics where every component of the data may be correlated with every other component. Suppose we have data on school districts of Minnesota. How does Minnesota go to infinity? By invasion of surrounding states and provinces of Canada, not to mention Lake Superior, and eventually by rocket ships to outer space? How silly does the $n$ goes to infinity story have to be before it provokes laughter instead of reverence?

Having once seen the absurdity of the $n$ goes to infinity story in any context, it becomes hard to maintain its illusion of appropriateness in any other context. A convert to the "no $n$" view always thinks $n = 1$. You always have one "data set," which comprises the data for an analysis. And it isn't going to infinity or anywhere else.

But how do "no $n$ asymptotics" work? It is not clear (to me) what Le Cam meant by "eventually transformed into approximation results" because he used a convergence structure (Le Cam and Yang, 2000, Chapter 6, Definitions 1, 2, and 3) so weak that, although topological (Beattie and Butzmann, 2002, Proposition 1.7.15), it seems not metrizable. So at this point I part company with Le Cam and present my own take on "no $n$ asymptotics" (although our Section 2 follows Le Cam closely, our Section 3 and Appendix B seem new). It has the following simple logic.

- All assumptions are packed into a single convergence in law statement involving the log likelihood.
- The conclusion is a convergence in law statement about an estimator.
- Hence delta and epsilon arguments using metrics for convergence in law can replace sequential arguments.

One source of this scheme is Geyer (1994), which did not use the whole scheme, but which in hindsight should have. The conclusion of Lemma 4.1 in that article is a "single convergence in law statement about the log likelihood" that incorporates most of the assumptions in that article. It is a forerunner of our treatment here.

At this point we introduce a caveat. By "no $n$" asymptotics we do not mean the letter $n$ is not allowed to appear anywhere, nor do we mean that sequences are not allowed to be used. We mean that we do not use sequences associated with a story about the index $n$ being sample size or anything analogous to sample size.

We do use sequences as a purely technical tool in discussing continuity. To prove a function $f : U \to V$ between metric spaces with metrics $d_U$ and $d_V$ continuous, one can show that for every $x \in U$ and every $\epsilon > 0$ there exists a $\delta$, which may depend on $x$ and $\epsilon$, such that $d_V\big(f(x), f(y)\big) < \epsilon$, whenever $d_U(x, y) < \delta$. Alternatively, one can show that $f(x_n) \to f(x)$, whenever $x_n \to x$. The mathematical content is the same either way.

In particular, it is more convenient to use sequences in discussing convergence of probability measures on Polish spaces. The space of all probability measures on a Polish space is itself a metric space (Billingsley, 1999, p. 72). Hence delta-epsilon arguments can be used to discuss convergence, but nearly the whole probability literature on convergence of probability measures uses sequences rather than delta-epsilon arguments, so it is easier to cite the literature if we use sequences. The mathematical content is the same either way.

The point is that we do not reify $n$ as sample size. So perhaps a longer name such as "$n$ not reified as sample size" asymptotics would be better than the shorter "no $n$" asymptotics, but we will continue with the shorter name.

What is important is "quadraticity." If the log likelihood for the actual model for the actual data is nearly quadratic, then the MLE has the familiar properties discussed in Section 1.1, but no story about sample size going to infinity will make the actual log likelihood more quadratic than it actually is or the actual MLE more nearly normal.

I concede that metrics for convergence in law are unwieldy and might also give "approximation bounds . . . too crude and cumbersome to be of any practical use." But, unlike Le Cam, I am not bothered by this because of my familiarity with computer intensive statistical methods.

### *1.3. Asymptotics is Only a Heuristic*

(This slogan means what Le Cam meant by "all they can do is suggest certain approaches".) We know that asymptotics often works well in practical problems because we can check the asymptotics by computer simulation (perhaps what Le Cam meant by "checked on the case at hand"), but conventional theory doesn't tell us why asymptotics works when it does. It only tells us that asymptotics works for sufficiently large $n$, perhaps astronomically larger than the actual $n$ of the actual data. So that leaves a theoretical puzzle.

- Asymptotics often works.
- But it doesn't work for the reasons given in proofs.
- It works for reasons too complicated for theory to handle.

I am not sure about "too complicated . . . to handle." Perhaps a computer-assisted proof could give "approximation bounds . . . of practical use," what Le Cam wanted but could not devise. But when I think how much computer time a computer-assisted proof might take and consider alternative ways to spend the computer time, I do not see how approximation bounds could be as useful as a parametric bootstrap, much less a double parametric bootstrap (since we are interested in likelihood theory for parametric models we consider only the parametric bootstrap).

A good approximation bound, even if such could be found, would only indicate whether the asymptotics work or don't work, but a bootstrap of approximately pivotal quantities derived from the asymptotics not only diagnoses any failure of the asymptotics but also provides a correction, so the bootstrap may work when asymptotics fails. And the double bootstrap diagnoses any failure of the single bootstrap and provides further correction, so the double bootstrap may work when the single bootstrap fails (Beran, 1987, 1988; Geyer, 1991; Hall, 1992; Newton and Geyer, 1994).

With ubiquitous fast computing, there is no excuse for not using the bootstrap to improve the accuracy of asymptotics in every serious application. Thus we arrive at the following attitude about asymptotics

- Asymptotics is only a heuristic. It provides no guarantees.
- If worried about the asymptotics, bootstrap!
- If worried about the bootstrap, iterate the bootstrap!

However, the only justification of the bootstrap is asymptotic. So this leaves us in a quandary of circularity.

- The bootstrap is only a heuristic. It provides no guarantees.
- All justification for the bootstrap is asymptotic!
- In order for the bootstrap to work well, one must bootstrap approximately asymptotically pivotal quantities!

(the "approximately" recognizes that something less than perfectly pivotal, for example merely variance stabilized, is still worth using).

In practice, this "circularity" does not hamper analysis. In order to devise good estimates one uses the asymptotics heuristic (choosing the MLE perhaps). In order to devise "approximately asymptotically pivotal quantities" one again uses the asymptotics heuristic (choosing log likelihood ratios perhaps). But when one calculates probabilities for tests and confidence intervals by simulation, the calculation can be made arbitrarily accurate for any given $\theta$. Thus the traditional role of asymptotics, approximating $P_\theta$, is not needed when we bootstrap. We only need asymptotics to deal with the dependence of $P_\theta$ on $\theta$. Generally, this dependence never goes entirely away, no matter how many times the bootstrap is iterated, but it does decrease (Beran, 1987, 1988; Hall, 1992).

The parametric bootstrap simulates multiple data sets $y_1^*$, $\ldots$, $y_n^*$ from $P_{\hat{\theta}(y)}$, where $y$ is the real data and $\hat{\theta}$ some estimator. The double parametric bootstrap simulates multiple data sets from each $P_{\hat{\theta}(y_i^*)}$. Its work load can be reduced by using importance sampling (Newton and Geyer, 1994). Assuming $\theta \mapsto P_\theta$ is continuous, these simulations tell everything about the model for $\theta$ in the region filled out by the $\hat{\theta}(y_i^*)$. But nothing prevents one from simulating from $\theta$ in a bigger region if one wants to.

The parametric bootstrap is very different from the nonparametric bootstrap in this respect. The nonparametric bootstrap is inherently an asymptotic (large $n$) methodology because resampling the data, which in effect substitutes the empirical distribution of the data for the true unknown distribution, only "works" when the empirical distribution is close to the true distribution, which is when the sample size is very large. When the parametric bootstrap simulates a distribution $P_\theta$ it does so with error that does not depend on sample size but only on how long we run the computer. When a double parametric bootstrap simulates $P_\theta$ for many different $\theta$ values, we learn about $P_\theta$ for these $\theta$ values and nearby $\theta$ values. If these $\theta$ values are densely spread over a region, we learn about $P_\theta$ for all $\theta$ in the region. So long as the true unknown $\theta$ value lies in that region, we approximate the true unknown distribution with accuracy that does not depend on sample size. Thus if one does enough simulation, the parametric bootstrap can be made to work for small sample sizes in a way the nonparametric bootstrap cannot. Since this is not a paper about the bootstrap, we will say no more on the subject.

Nevertheless, asymptotics often does "work" and permits simpler calculations while providing more insight. In such cases, the bootstrap when used as a diagnostic (Geyer, 1991) proves its own pointlessness. A single bootstrap often shows that it cannot improve the answer provided by asymptotics, and a double bootstrap often shows that it cannot improve the answer provided by the single bootstrap.

Since asymptotics is "only a heuristic," the only interesting question is what form of asymptotics provides the most useful heuristic and does so in the simplest fashion. This article is my attempt at an answer.

### *1.4. An Example from Spatial Statistics*

Geyer and Møller (1994) give a method of simulating spatial point processes and doing maximum likelihood estimation. In their example, a Strauss process (Strauss, 1975), they noted that the asymptotic distribution of the MLE appeared to be very close to normal, although the best asymptotic results they were able to find in the literature (Jensen, 1991, 1993) only applied to Strauss processes with very weak dependence, hence very close to a Poisson process (Geyer and Møller, 1994, Discussion), which unfortunately did not include their example.

From a "no $n$" point of view, this example is trivial. A Strauss process is a two-parameter full exponential family. In the canonical parameterization, which Geyer and Møller (1994) were using, the random part of the log likelihood is linear in the parameter, hence the Hessian is nonrandom. Hence, according to the theory developed here, the MLE will be approximately normal so long as the Hessian is approximately constant over the region of parameter values containing most of the sampling distribution of the MLE.

Geyer and Møller (1994) did not then understand the "no $n$" view and made no attempt at such verification (although it would have been easy). Direct verification of quadraticity is actually unnecessary here, because the curved part of the log likelihood (in the canonical parameterization) of an exponential family is proportional to the cumulant generating function of the canonical statistic, so the family is nearly LAN precisely when the distribution of the canonical statistic is nearly normal, which Geyer and Møller (1994) did investigate (their Figure 2).

Thus there is no need for any discussion of anything going to infinity. The asymptotics here, properly understood, are quite simple. As we used to say back in the sixties, the "$n$ goes to infinity" story is part of the problem not part of the solution.

Although the exponential family aspect makes things especially simple, the same sort of thing is true in general. When the Hessian is random, it is enough that it be nearly invariant in law.

### *1.5. An Example from Statistical Genetics*

Fisher (1918) proposed a model for quantitative genetics that has been widely used in animal breeding and other areas of biology and indirectly led to modern regression and analysis of variance. The data $Y$ are multivariate normal, measurements of some quantitative trait on individuals, decomposed as

$$Y = \mu + B + E$$

where $\mu$ is an unknown scalar parameter, where $B$ and $E$ are independent multivariate normal, $\mathcal{N}(0, \sigma^2 A)$ and $\mathcal{N}(0, \tau^2 I)$, respectively, where $\sigma^2$ and $\tau^2$ are unknown parameters called the *additive genetic* and *environmental* variance, respectively, $A$ is a known matrix called the *numerator relationship matrix* in the animal breeding literature (Henderson, 1976), and $I$ is the identity matrix. The matrix $A$ is determined solely by the relatedness of the individuals (relative to a known pedigree). Every element of $A$ is nonzero if every individual is related to every other, and this implies all components of $Y$ are correlated.

In modern terminology, this is a mixed model with fixed effect vector $\mu$ and variance components $\sigma^2$ and $\tau^2$, but if the pedigree is haphazard so $A$ has no regular structure, there is no scope for telling "$n$ goes to infinity" stories like Miller (1977) does for mixed models for simple designed experiments. Our "no $n$" asymptotics do apply. The log likelihood may be nearly quadratic, in which case we have the "usual" asymptotics.

The supplementary web site `http://purl.umn.edu/92198` at the University of Minnesota Digital Conservancy gives details of two examples with 500 and 2000 individuals. An interesting aspect of our approach is that its intimate connection with Newton's method (Section 3.5 and Appendix B) forced us to find a good starting point for Newton's method (a technology transfer from spatial statistics) and investigate its quality. Thus our theory helped improve methodology.

Readers wanting extensive detail must visit the web site. The short summary is that the example with 500 individuals is not quite in asymptopia, the parametric bootstrap being needed for bias correction in constructing a confidence interval for logit heritability ($\log \sigma^2 - \log \tau^2$). But when this example was redone with 2000 individuals, the bias problem went away and the bootstrap could not improve asymptotics.

### *1.6. Our Regularity Conditions*

Famously, Le Cam, although spending much effort on likelihood, did not like *maximum likelihood.* Le Cam (1990) gives many examples of the failure of maximum likelihood. Some are genuine examples of bad behavior of the MLE. Others can be seen as problems with the "$n$ goes to infinity" story as much as with maximum likelihood. I have always thought that article failed to mention Le Cam's main reason for dislike of maximum likelihood: his enthusiasm for weakest possible regularity conditions. He preferred conditions so weak that nothing can be proved about the MLE and other estimators must be used instead (Le Cam and Yang, 2000, Section 6.3).

His approach does allow the "usual asymptotics of maximum likelihood" to be carried over to quite pathological models (Le Cam and Yang, 2000, Example 7.1) but only by replacing the MLE with a different estimator. The problem with this approach (as I see it) is that the resulting theory no longer describes the MLE, hence is no longer useful to applied statisticians. (Of course, it would be useful if applied statisticians used such pathological models and such estimators. As far as I know, they don't.)

Thus we stick with old-fashioned regularity conditions involving derivatives of the log likelihood that go back to Cramér (1946, Chapters 32 and 33). We shall investigate the consequences of being "nearly" LAMN in the sense that the log likelihood and its first two derivatives are near those of an LAMN model. These conditions are about the weakest that still permit treatment of Newton's method, so our theorems apply to the way maximum likelihood is done in practice. Our approach has the additional benefit of making our theory no stranger than it has to be in the eyes of a typical statistician.

## 2. Models with Quadratic Log Likelihood

### 2.1. Log Likelihood

The log likelihood for a parametric family of probability distributions having densities $f_\theta$, $\theta \in \Theta$ with respect to a measure $\lambda$ is a random function $l$ defined by

$$l(\theta) = u(X) + \log f_\theta(X), \qquad \theta \in \Theta, \tag{4}$$

where $X$ is the random data for the problem and $u$ is any real valued function on the sample space that does not depend on the parameter $\theta$. In this article, we are only interested families of almost surely positive densities so the argument of the logarithm in (4) is never zero and the log likelihood is well defined. This means all the distributions in the family are absolutely continuous with respect to each other.

Then for any bounded random variable $g(X)$ and any parameter values $\theta$ and $\theta + \delta$ we can write

$$\begin{aligned} E_{\theta+\delta}\{g(X)\} &= \int g(x) f_{\theta+\delta}(x) \lambda(dx) \\ &= \int g(x) \frac{f_{\theta+\delta}(x)}{f_\theta(x)} f_\theta(x) \lambda(dx) \\ &= \int g(x) e^{l(\theta+\delta) - l(\theta)} f_\theta(x) \lambda(dx) \\ &= E_\theta\bigl\{g(X) e^{l(\theta+\delta) - l(\theta)}\bigr\} \end{aligned} \tag{5}$$

The assumption of almost surely positive densities is crucial. Without it, the second line might not make sense because of division by zero.

### 2.2. Quadratic Log Likelihood

Suppose the log likelihood is defined by (1). The random variables $U$, $Z$, and $K$ are, of course, functions of the data for the model, although the notation does not indicate this explicitly.

The constant term $U$ in (1) analogous to the term $u(X)$ in (4) is of no importance. We are mainly interested in log likelihood ratios

$$l(\theta + \delta) - l(\theta) = (Z - K\theta)'\delta - \tfrac{1}{2}\delta' K \delta, \tag{6}$$

in which $U$ does not appear.

**Theorem 2.1** (LAMN). *Suppose* (1) *is the log likelihood of a probability model, then*

(a) *$K$ is almost surely positive semi-definite.*

*Also, the following two conditions are equivalent (each implies the other).*

(b) *The conditional distribution of $Z$ given $K$ for parameter value $\theta$ is $\mathcal{N}(K\theta, K)$.*
(c) *The distribution of $K$ does not depend on the parameter $\theta$.*

Any model satisfying the conditions of the theorem is said to be LAMN (locally asymptotically mixed normal). Strictly speaking, “locally asymptotically” refers to sequences converging to such a model, such as those discussed in Section 3, but

"MN" by itself is too short to make a good acronym and LAMN is standard in the literature.

Our theorem is much simpler than traditional LAMN theorems (Le Cam and Yang, 2000, Lemmas 6.1 and 6.3) because ours is not asymptotic and the traditional ones are. But the ideas are the same.

*Proof.* The special case of (5) where $g$ is identically equal to one with (6) plugged in gives

$$1 = E_{\theta+\delta}(1) = E_\theta\bigl\{e^{(Z - K\theta)'\delta - \frac{1}{2}\delta' K \delta}\bigr\} \tag{7}$$

Averaging (7) and (7) with $\delta$ replaced by $-\delta$ gives

$$E_\theta\bigl\{\cosh[(Z - K\theta)'\delta] \exp[-\tfrac{1}{2}\delta' K\delta]\bigr\} = 1. \tag{8}$$

Now plug in $s\delta$ for $\delta$ in (8), where $s$ is scalar, and use the fact that the hyperbolic cosine is always greater than one giving

$$\begin{aligned} 1 &= E_\theta\bigl\{\cosh[s(Z - K\theta)'\delta] \exp[-\tfrac{1}{2}s^2\delta' K\delta]\bigr\} \\ &\ge E_\theta\bigl\{\exp[-\tfrac{1}{2}s^2\delta' K\delta] I_{(-\infty,-\epsilon)}(\delta' K\delta)\bigr\} \\ &\ge \exp(\tfrac{1}{2}s^2\epsilon) P_\theta\bigl\{\delta' K\delta < -\epsilon\bigr\} \end{aligned}$$

For any $\epsilon > 0$, the first term on the right hand side goes to infinity as $s \to \infty$. Hence the second term on the right hand side must be zero. Thus

$$P_\theta\{\delta' K\delta < -\epsilon\} = 0, \qquad \epsilon > 0$$

and continuity of probability implies the equality holds for $\epsilon = 0$ as well. This proves (a).

Replace $g(X)$ in (5) by $h(K)$ where $h$ is any bounded measurable function giving

$$E_{\theta+\delta}\{h(K)\} = E_\theta\bigl\{h(K) e^{(Z - K\theta)'\delta - \frac{1}{2}\delta' K \delta}\bigr\}. \tag{9}$$

Assume (b). Then the moment generating function of $Z$ given $K$ for the parameter $\theta$ is

$$E_\theta\bigl\{e^{Z'\delta} \bigm| K\bigr\} = e^{\theta' K\delta + \frac{1}{2}\delta' K\delta} \tag{10}$$

and this implies

$$E_\theta\bigl\{e^{(Z - K\theta)'\delta - \frac{1}{2}\delta' K \delta} \bigm| K\bigr\} = 1. \tag{11}$$

Plugging (11) into (9) and using the iterated expectation theorem we get

$$\begin{aligned} E_{\theta+\delta}\{h(K)\} &= E_\theta\Bigl\{h(K) E_\theta\bigl\{e^{(Z - K\theta)'\delta - \frac{1}{2}\delta' K \delta} \bigm| K\bigr\}\Bigr\} \\ &= E_\theta\{h(K)\} \end{aligned}$$

which, $h$ being arbitrary, implies (c). This proves (b) implies (c).

Now drop the assumption of (b) and assume (c), which implies the left hand side of (9) does not depend on $\delta$, hence

$$E_\theta\{h(K)\} = E_\theta\bigl\{h(K) e^{(Z - K\theta)'\delta - \frac{1}{2}\delta' K \delta}\bigr\}. \tag{12}$$

By the definition of conditional expectation (12) holding for all bounded measurable functions $h$ implies (11) and hence (10), which implies (b). This proves (c) implies (b). □

**Corollary 2.2** (LAN)**.** *Suppose* (1) *is the log likelihood of an identifiable probability model and $K$ is constant, then*

(a) *$K$ is positive definite.*

*Moreover,*

(b) *The distribution of $Z$ for the parameter value $\theta$ is $\mathcal{N}(K\theta, K)$.*

Any model satisfying the conditions of the corollary is said to be LAN (locally asymptotically normal). Strictly speaking, as we said about LAMN, the "locally asymptotically" refers to sequences converging to such a model, but we also use it for the model itself.

*Proof.* The theorem applied to the case of constant $K$ gives (a) and (b) with "positive definite" in (a) replaced by "positive semi-definite." So we only need to prove that identifiability implies positive definiteness.

If $K$ were not positive definite, there would be a nonzero $\delta$ such that $\delta' K \delta = 0$, but this would imply $K\delta = 0$ and the distribution of $Z$ for the parameter value $\theta + \delta$ would be $\mathcal{N}(K\theta, K)$. Hence the model would not be identifiable (since $Z$ is a sufficient statistic, if the distribution of $Z$ is not identifiable, neither is the model). □

We cannot prove the analogous property, $K$ almost surely positive definite, for LAMN. So we will henceforth assume it. (That this doesn't follow from identifiability is more a defect in the notion of identifiability than in LAMN models.)

### *2.3. Examples and Non-Examples*

The theorem provides many examples of LAMN. Let $K$ have any distribution that is almost surely positive-definite-valued, a Wishart distribution, for example. Then let $Z \mid K$ be $\mathcal{N}(K\theta, K)$.

The corollary provides a more restricted range of examples of LAN. They are the multivariate normal location models with nonsingular variance matrix that does not depend on the parameter.

A non-example is the AR(1) autoregressive model with known innovation variance and unknown autoregressive parameter. Let $X_0$ have any distribution not depending on the parameter $\theta$, and recursively define

$$X_n = \theta X_{n-1} + Z_n \tag{13}$$

where the $Z_i$ are independent and identically $\mathcal{N}(0, 1)$ distributed. The log likelihood is

$$l_n(\theta) = -\tfrac{1}{2} \sum_{i=1}^{n} (X_i - \theta X_{i-1})^2$$

which is clearly quadratic with Hessian $-K_n = -\sum_{i=1}^n X_{i-1}^2$. From (13)

$$E_\theta(X_n^2 | X_0) = \theta^2 E_\theta(X_{n-1}^2 | X_0) + 1$$

and from this we can derive

$$E_\theta(K_n | X_0) = \frac{n-1}{1-\theta^2} + \left[ X_0^2 - \frac{1}{1-\theta^2} \right] \frac{1-\theta^{2(n-1)}}{1-\theta^2}, \qquad \theta \neq 1,$$

which is enough to show that the distribution of $K_n$ depends on $\theta$ so this model is not LAMN (in the "no $n$" sense we are using the term in this article, although it is LAN in the limit as $n \to \infty$ for some values of $\theta$).

## 3. Likelihood Approximation

It is plausible that a model that is “nearly” LAN has an MLE that is “nearly” normally distributed and a model that is “nearly” LAMN has an MLE that is “nearly” conditionally normally distributed. In order to make these vague statements mathematically precise, we need to define what we mean by “nearly” and explore its mathematical consequences.

### *3.1. Convergence in Law in Polish Spaces*

Since log likelihoods are random functions, it makes sense to measure how close one is to another in the sense of convergence in law. In order to do that, we need a theory of convergence in law for function-valued random variables. We use the simplest such theory: convergence in law for random variables taking values in a Polish space (complete separable metric space). Billingsley (1999, first edition 1968) has the theory we will need. Other sources are Fristedt and Gray (1997) and Shorack (2000).

A sequence of random elements $X_n$ of a Polish space $S$ *converges in law* to another random element $X$ of $S$ if $E\{f(X_n)\} \to E\{f(X)\}$, for every bounded continuous function $f : S \to \mathbb{R}$. This convergence is denoted

$$X_n \xrightarrow{\mathcal{L}} X$$

and is also called *weak convergence*, a term from functional analysis, and (when the Polish space is $\mathbb{R}^p$) *convergence in distribution.*

The theory of convergence in law in Polish spaces is often considered an advanced topic, but our use of it here involves only

- the mapping theorem (Billingsley, 1999, Theorem 2.7)
- the Portmanteau theorem (Billingsley, 1999, Theorem 2.1)
- Slutsky’s theorem (Billingsley, 1999, Theorem 3.1)
- Prohorov’s theorem (Billingsley, 1999, Theorems 5.1 and 5.2)
- the subsequence principle (Billingsley, 1999, Theorem 2.6)

which should be in the toolkit of every theoretical statistician. They are no more difficult to use on random elements of Polish spaces than on random vectors. The last three of these are only used in Appendix C where we compare our theory to conventional theory. Only the mapping and Portmanteau theorems are needed to understand the main text.

### *3.2. The Polish Spaces $C(W)$ and $C^2(W)$*

Let $W$ be an open subset of $\mathbb{R}^p$ and let $C(W)$ denote the space of all continuous real-valued functions on $W$. The topology of *uniform convergence on compact sets* for $C(W)$ defines $f_n \to f$ if

$$\sup_{x \in B} |f_n(x) - f(x)| \to 0, \qquad \text{for every compact subset } B \text{ of } W.$$

The topology of *continuous convergence* for $C(W)$ defines $f_n \to f$ if

$$f_n(x_n) \to f(x), \qquad \text{whenever } x_n \to x.$$

Kuratowski (1966, Section 20, VIII, Theorem 2) shows these two topologies for $C(W)$ coincide. Bourbaki (1998, Ch. 10, Sec. 3.3, Corollary, part (b)) shows that $C(W)$ is a Polish space.

Let $C^2(W)$ denote the space of all twice continuously differentiable functions $W \to \mathbb{R}$ with the topology of uniform convergence on compact sets (or continuous convergence) of the functions and their first and second partial derivatives. Then $C^2(W)$ is isomorphic to a subspace of $C(W)^{1+p+p\times p}$, in fact a closed subspace because uniform convergence of derivatives implies sequences can be differentiated term by term. Hence $C^2(W)$ is a Polish space, (Fristedt and Gray, 1997, Propositions 2 and 3 of Section 18.1).

We will consider an almost surely twice continuously differentiable log likelihood whose parameter space is all of $\mathbb{R}^p$ to be a random element of $C^2(\mathbb{R}^p)$ and will restrict ourselves to such log likelihoods (for measurability issues, see Appendix A). We find a way to work around the assumption that the parameter space be all of $\mathbb{R}^p$ in Section 3.6.

### 3.3. Sequences of Statistical Models

In order to discuss convergence we use sequences of models, but merely as technical tools. The $n$ indexing a sequence need have nothing to do with sample size, and models in the sequence need have nothing to do with each other except that they all have the same parameter space, which is $\mathbb{R}^p$. As we said in Section 1.2, we could eliminate sequences from the discussion if we wanted to. The models have log likelihoods $l_n$ and true parameter values $\psi_n$.

Merely for comparison with conventional theory (Appendix C) we also introduce a "rate" $\tau_n$. In conventional asymptotic theory $\tau_n = \sqrt{n}$ plays an important role, so we put it in our treatment too. In the "no $n$" view, however, where the models in the sequence have "nothing to do with each other" there is no role for $\tau_n$ to play, and we set $\tau_n = 1$ for all $n$.

Define random functions $q_n$ by

$$q_n(\delta) = l_n(\psi_n + \tau_n^{-1}\delta) - l_n(\psi_n), \qquad \delta \in \mathbb{R}^p. \tag{14}$$

These are also log likelihoods, but we have changed the parameter from $\theta$ to $\delta$, so the true value of the parameter $\delta$ is zero, and subtracted a term not containing $\delta$, so $q_n(0) = 0$.

Our key assumption is

$$q_n \xrightarrow{\mathcal{L}} q, \qquad \text{in } C^2(\mathbb{R}^p) \tag{15}$$

where $q$ is the log likelihood of an LAMN model

$$q(\delta) = \delta' Z - \tfrac{1}{2}\delta' K \delta, \qquad \delta \in \mathbb{R}^p, \tag{16}$$

there being no constant term in (16) because $q_n(0) = 0$ for all $n$ implies $q(0) = 0$. A consequence of LAMN is that $e^{q(\delta)}$ is for each $\delta$ a probability density with respect to the measure governing the law of $q$, which is the measure in the LAMN model for $\delta = 0$ because $e^{q(0)} = 1$. This implies

$$E\{e^{q(\delta)}\} = 1, \qquad \delta \in \mathbb{R}^p. \tag{17}$$

Please note that, despite our assuming LAMN with "N" standing for normal, we can say we are not actually assuming asymptotic normality. Asymptotic (conditional) normality comes from the equivalence of the two conditions in the LAMN theorem (Theorem 2.1). We can say that we are assuming condition (c) of the theorem and getting condition (b) as a consequence. Normality arises here from the log likelihood being quadratic and its Hessian being invariant in law. Normality is not assumed, and the CLT plays no role.

### 3.4. Contiguity

In Le Cam's theory, property (17) is exceedingly important, what is referred to as *contiguity* of the sequence of probability measures having parameter values $\psi_n + \tau_n^{-1}\delta$ to the sequence having parameter values $\psi_n$ (Le Cam and Yang, 2000, Theorem 1 of Chapter 3).

Contiguity (17) does not follow from the convergence in law (15) by itself. From that we can only conclude by Fatou's lemma for convergence in law (Billingsley, 1999, Theorem 3.4)

$$E\bigl\{e^{q(\delta)}\bigr\} \le 1, \qquad \delta \in \mathbb{R}^p. \tag{18}$$

That we have the equality (17) rather than the inequality (18) is the contiguity property. In our "no $n$" theory (17) arises naturally. We always have it because "improper" asymptotic models (having densities that don't integrate to one) make no sense.

### 3.5. One Step of Newton's Method

Our presentation is closely tied to Newton's method. This is a minor theme in conventional theory (one-step Newton updates of root-$n$-consistent estimators are asymptotically equivalent to the MLE), but we make it a major theme, bringing our theory closer to actual applications. We consider two kinds of estimator: one-step-Newton estimators, like those in conventional theory, and Newton-iterated-to-convergence estimators (Appendix B), what in applications are usually deemed MLE.

To treat these estimators we need to consider not only the log likelihood but also the starting point for Newton's method. The one-step Newton map $G$ is defined for an objective function $q$ and a current iterate $\delta$ by

$$G(q, \delta) = \delta + \bigl(-\nabla^2 q(\delta)\bigr)^{-1} \nabla q(\delta) \tag{19}$$

when $-\nabla^2 q(\delta)$ is positive definite (otherwise Newton's method makes no sense as an attempt at maximization). In order to deal with this possible failure of Newton's method, we allow $\delta$ and $G(q, \delta)$ to have the value NaO (not an object), which, like NaN (not a number) in computer arithmetic, propagates through all operations. Addition of a new isolated point to a Polish space produces another Polish space.

**Lemma 3.1.** *If $q$ is strictly concave quadratic and $\delta \neq \mathrm{NaO}$, then $G(q, \delta)$ is the unique point where $q$ achieves its maximum.*

*Proof.* Suppose

$$q(\delta) = u + z'\delta - \tfrac{1}{2}\delta' K \delta, \tag{20}$$

where $K$ is positive definite, so

$$\begin{aligned} \nabla q(\delta) &= z - K\delta \\ \nabla^2 q(\delta) &= -K \end{aligned}$$

Then

$$G(q, \delta) = \delta + K^{-1}(z - K\delta) = K^{-1} z,$$

which is the solution to $\nabla q(\delta) = 0$, and hence by strict concavity, the unique global maximizer of $q$. □

**Lemma 3.2.** *The Newton map $G$ is continuous at points $(q, \delta)$ such that $q$ is strictly concave quadratic.*

*Proof.* Suppose $\delta_n \to \delta$ and $q_n \to q$ with $q$ given by (20). If $\delta = \text{NaO}$ the conclusion is trivial. Otherwise, because convergence in $C^2(\mathbb{R}^p)$ implies continuous convergence of first and second derivatives

$$-\nabla^2 q_n(\delta_n) \to -\nabla^2 q(\delta) = K,$$

which is positive definite, so $G(q_n, \delta_n) \neq \text{NaO}$ for sufficiently large $n$, and

$$\nabla q_n(\delta_n) \to \nabla q(\delta) = z - K\delta,$$

so

$$G(q_n, \delta_n) \to G(q, \delta) = K^{-1} z.$$

□

**Theorem 3.3.** *Let $\tilde{\delta}_n$ be a sequence of random elements of $\mathbb{R}^d$ and $q_n$ a sequence of log likelihoods having parameter space $\mathbb{R}^d$ and true parameter value zero. Suppose*

$$\big(q_n, \tilde{\delta}_n\big) \xrightarrow{\mathcal{L}} (q, \tilde{\delta}) \tag{21}$$

*where $q$ is the log likelihood of an LAMN model* (16)*, and define $\hat{\delta}_n = G(q_n, \tilde{\delta}_n)$ where $G$ is given by* (19)*. Then*

$$-\nabla^2 q_n(\tilde{\delta}_n) \xrightarrow{\mathcal{L}} K \tag{22a}$$

$$-\nabla^2 q_n(\hat{\delta}_n) \xrightarrow{\mathcal{L}} K \tag{22b}$$

$$\hat{\delta}_n \xrightarrow{\mathcal{L}} K^{-1} Z \tag{22c}$$

$$\big(-\nabla^2 q_n(\hat{\delta}_n)\big)^{1/2} \hat{\delta}_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, I) \tag{22d}$$

*where the matrix square root in* (22d) *is the symmetric square root when the matrix is positive definite and* NaO *otherwise. Moreover,* (21)*,* (22a)*,* (22b)*,* (22c)*, and* (22d) *hold jointly.*

The convergence in law (21) takes place in $C^2(\mathbb{R}^p) \times \mathbb{R}^p$, the product of Polish spaces being a Polish space (Billingsley, 1999, Appendix M6). In (22d) $I$ is the $p \times p$ identity matrix. Since the random variables on the right hand sides of all these limits are never NaO, the left hand sides are NaO with probability converging to zero.

Some would rewrite (22d) as

$$\hat{\theta}_n \approx \mathcal{N}\left(\psi_n, \big(-\nabla^2 l_n(\hat{\theta}_n)\big)^{-1}\right), \tag{23}$$

where the double wiggle means "approximately distributed" or something of the sort and where we have shifted back to the original parameterization

$$\begin{aligned} \tilde{\theta}_n &= \psi_n + \tau_n^{-1} \tilde{\delta}_n \\ \hat{\theta}_n &= \psi_n + \tau_n^{-1} \hat{\delta}_n \\ &= G(l_n, \tilde{\theta}_n) \end{aligned}$$

Strictly speaking, (23) is mathematical nonsense, having no mathematical content except by allusion to (22d), but it is similar to

$$\hat{\theta}_n \approx \mathcal{N}\left(\psi, I_n(\hat{\theta}_n)^{-1}\right) \tag{24}$$

familiar from conventional treatments of the asymptotics of maximum likelihood, where $\psi$ is the true parameter value and $I_n(\theta)$ is expected Fisher information for sample size $n$.

*Proof.* We claim the map

$$\begin{pmatrix} q \\ \delta \end{pmatrix} \mapsto \begin{pmatrix} q \\ \delta \\ \nabla^2 q(\delta) \\ \nabla^2 q\bigl(G(q, \delta)\bigr) \\ G(q, \delta) \\ \Bigl(-\nabla^2 q\bigl(G(q, \delta)\bigr)\Bigr)^{1/2} G(q, \delta) \end{pmatrix} \tag{25}$$

is continuous on $C^2(\mathbb{R}^d) \times \mathbb{R}^2$ at points where $q$ is strictly concave quadratic. By definition of product topology, (25) is continuous if each component is continuous. The first two components are also continuous by definition of product topology, the fifth by Lemmas 3.1 and 3.2, then the third and fourth by continuous convergence, and the sixth by matrix multiplication being a continuous operation, and matrix square root being a continuous operation on the set of positive definite matrices (Horn and Johnson, 1985, Problem 7.2.18) and continuous convergence.

Now all of the assertions of the theorem follow from the mapping theorem (Billingsley, 1999, Theorem 2.7), the only non-obvious limit being (22d), which is clearly $K^{-1/2} Z$. Since $Z$ is $\mathcal{N}(0, K)$ given $K$ by $q$ being LAMN with true parameter value zero, $K^{-1/2} Z$ is $\mathcal{N}(0, I)$ conditional on $K$. Since the conditional distribution does not depend on $K$, it is also the unconditional distribution. □

**Corollary 3.4.** *If $K = -\nabla^2 q(\delta)$ in the theorem is constant, then* (22c) *can be replaced by*

$$\hat{\delta}_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, K^{-1}) \tag{26}$$

The proof is obvious. Some would rewrite (26) as

$$\hat{\theta}_n \approx \mathcal{N}\left(\psi_n, (\tau_n^2 K)^{-1}\right), \tag{27}$$

from which we see that $K$ plays the role of expected Fisher information for sample size one and $\tau_n^2 K$ plays the role of expected Fisher information for sample size $n$, though $K$ isn't the expectation of anything in our setup.

### *3.6. Bounding the Approximation*

Suppose we wish to use (22d) to produce a confidence region. Another application of the mapping theorem gives

$$\hat{\delta}_n' \bigl(-\nabla^2 q_n(\hat{\delta}_n)\bigr) \hat{\delta}_n \xrightarrow{\mathcal{L}} \text{chi}^2(p),$$

where the right hand side denotes a chi-square random variable with $p$ degrees of freedom. If $\kappa$ is the upper $\alpha$ quantile of this distribution, then by the Portmanteau theorem (Billingsley, 1999, Theorem 2.1)

$$\liminf_{n\to\infty} \Pr \left\{ \hat{\delta}_n' \bigl( -\nabla^2 q_n(\hat{\delta}_n) \bigr) \hat{\delta}_n < \kappa \right\} \ge 1 - \alpha$$

and, mapped to the original parameterization, this tells us that

$$\left\{ \theta \in \mathbb{R}^p : (\hat{\theta}_n - \theta)' \bigl( -\nabla^2 l_n(\hat{\theta}_n) \bigr) (\hat{\theta}_n - \theta) < \kappa \right\}$$

is a $1 - \alpha$ asymptotic confidence region for the true unknown parameter value $\psi_n$.

So far so conventional, but now we want to do a finer analysis using the joint convergence of $\tilde{\theta}_n$ and $\hat{\theta}_n$. For any open set $W$ we get

$$\begin{aligned} \liminf_{n\to\infty} \Pr \left\{ \hat{\delta}_n' \bigl( -\nabla^2 q_n(\hat{\delta}_n) \bigr) \hat{\delta}_n < \kappa \text{ and } \tilde{\delta}_n \in W \text{ and } \hat{\delta}_n \in W \right\} & \\ \ge 1 - \alpha - \Pr(\tilde{\delta} \notin W \text{ or } \hat{\delta} \notin W) & \end{aligned} \tag{28}$$

(using Bonferroni on the right hand side), and now we note that we only evaluate $\nabla q_n$ and $\nabla^2 q_n$ at $\tilde{\delta}_n$ and $\hat{\delta}_n$, and hence for this statement to hold we only need the convergence in law (21) relative to $W$, that is, for (28) to hold we only need (21) to hold in $C^2(W) \times W$.

This section is our counterpart of the conventional “regularity condition” that the true parameter value be an interior point of the parameter space. But here we see that if $\tilde{\theta}_n$ is a very bad estimator of $\psi_n$, then $W$ may need to be very large in order to have $\Pr(\tilde{\delta} \notin W \,\text{or}\, \hat{\delta} \notin W)$ very small. The conventional regularity condition vastly oversimplifies what is really needed.

## 4. Discussion

It has not escaped our notice that the “no $n$” view advocated here leaves no place for a lot of established statistical theory: Edgeworth expansions, rates of convergence, the Bayes information criterion (BIC), and the subsampling bootstrap, to mention just a few. Can all of this useful theory be replaced by some “no $n$” analog? Only time will tell. Our goal is merely to explicate this “no $n$” view of maximum likelihood. We are not advocating political correctness in asymptotics.

It does seem obvious that “no $n$” asymptotics based on LAN and LAMN theory says nothing about situations where no likelihood is specified (as in quasilikelihood, estimating equations, and much of nonparametrics and robustness) or where the likelihood is incorrectly specified or where the likelihood is correctly specified but the parameter is infinite-dimensional (as in nonparametric maximum likelihood). Thus it is unclear how this “no $n$” view can be extended to these areas.

Another classical result that does not transfer to the “no $n$” worldview is the asymptotic efficiency of maximum likelihood. In an LAN model, it is true that the MLE is the best equivariant-in-law estimator (van der Vaart, 2000, Proposition 8.4), but James-Stein estimators (James and Stein, 1961) show that the MLE is not the best estimator (where “best” means minimum mean squared error). The “no $n$” view has no room for other interpretations: if one likes James-Stein estimators, then one cannot also consider the MLE asymptotically efficient (because LAN models are already in asymptopia). The argument leading to Le Cam’s almost sure convolution theorem (van der Vaart, 2000, Section 8.6) cannot be transferred to the "no $n$" world (because it would have to prove the MLE as good as James-Stein estimators in LAN models, and it isn't).

The application described in Section 1.4 convinced us of the usefulness of this "no $n$" view in spatial statistics, statistical genetics, and other areas where complicated dependence makes difficult the invention of reasonable "$n$ goes to infinity" stories, much less the proving of anything about them. But having seen the usefulness of the "no $n$" view in any context, one wants to use it in every context. Having understood the power of "quadraticity" as an explanatory tool," many opportunities to use it arise. When a user asks whether $n$ is "large enough" when the log likelihood is nowhere near quadratic, now the answer is "obviously not." When a user asks whether there is a need to reparameterize when the Wald confidence regions go outside the parameter space, now the answer is "obviously."

We imagine some readers wondering whether our ideas are mere "generalized abstract nonsense." Are we not essentially assuming what we are trying to prove? Where are all the deltas and epsilons and inequality bashing that one expects in "real" real analysis? We believe such "inequality bashing" should be kept separate from the main argument, because it needlessly restricts the scope of the theory. Le Cam thought the same, keeping separate the arguments of Chapters 6 and 7 in Le Cam and Yang (2000).

For readers who want to see that kind of argument, we have provided Lemma C.1 in Appendix C that says our "key assumption" (15) is weaker than the "usual regularity conditions" for maximum likelihood (Ferguson, 1996, Chapter 18) in all respects except the requirement that the parameter space be all of $\mathbb{R}^p$, which we worked around in Section 3.6. A similar lemma using a different Polish space with weaker topology is Lemma 4.1 in Geyer (1994), where a "single convergence in law statement about the log likelihood" (the conclusion of the lemma) is shown to follow from complicated analytic regularity conditions (the hypotheses of the lemma), modeled after Pollard (1984, pp. 138 ff.).

Despite all the theory we present, our message is very simple: if the log likelihood is nearly quadratic with Hessian nearly equivariant in law, then the "usual" asymptotics of maximum likelihood hold. The only point of all the theory is to show that conventional theory, which does not support our message, can be replaced by theory that does.

## Appendix A: Measurability in $C(W)$ and $C^2(W)$

Let $W$ be an open subset of $\mathbb{R}^p$, and let $B_n$ be an increasing sequence of compact subsets of $W$ whose union is $W$. Then Rudin (1991, Example 1.44) gives an explicit metric for the space $C(W)$ defined in section 3.2

$$d(f, g) = \max_{n \geq 1} \frac{2^{-n} \|f - g\|_{B_n}}{1 + \|f - g\|_{B_n}} \tag{29}$$

where for any compact set $B$ we define

$$\|f\|_B = \sup_{x \in B} |f(x)|.$$

A map $F : \Omega \to C(W)$ is measurable if and only if its uncurried form $(\omega, \theta) \mapsto F(\omega)(\theta)$ is Carathéodory, meaning

- $\omega \mapsto F(\omega)(\theta)$ is measurable for each fixed $\theta \in W$ and

- $\theta \mapsto F(\omega)(\theta)$ is continuous for each fixed $\omega \in \Omega$

(Rudin, 1991, Example 1.44; Aliprantis and Border, 1999, Corollary 4.23 and Theorem 4.54). The isomorphism between $C^2(W)$ and a closed subspace of $C(W)^{1+p+p\times p}$ means a function from a measurable space to $C^2(W)$ is measurable if and only if it and its first and second partial derivatives have Carathéodory uncurried forms.

## Appendix B: Newton Iterated to Convergence

In this appendix we consider Newton's method iterated to convergence. Of course Newton need not converge, so we have to deal with that issue. Define Newton iterates

$$\begin{aligned} \delta_0 &= \delta \\ \delta_n &= G(q, \delta_{n-1}), \qquad n > 0. \end{aligned}$$

If the sequence $\{\delta_n\}$ converges, then we define

$$G^\infty(q, \delta) = \lim_{n \to \infty} \delta_n \tag{30}$$

and otherwise we define $G^\infty(q, \delta) = \text{NaO}$. This function $G^\infty$ is called the infinite-step Newton map.

**Lemma B.1.** *If $q$ is strictly concave quadratic and $\delta \neq \text{NaO}$, then $G^\infty(q, \delta)$ is the unique point where $q$ achieves its maximum.*

*Proof.* By Lemma 3.1 the Newton sequence converges in one step in this case and has the asserted properties. □

**Lemma B.2.** *The infinite-step Newton map $G^\infty$ is continuous at points $(q, \delta)$ such that $q$ is strictly concave quadratic.*

*Proof.* Suppose $\tilde{\delta}_n \to \tilde{\delta}$ and $q_n \to q$, and suppose $q$ is strictly concave quadratic given by (20). Let $\hat{\delta} = K^{-1} z$ be the unique point at which $q$ achieves its maximum, and let $B$ be a compact convex set sufficiently large so that it contains $\hat{\delta}$ and $\tilde{\delta}$ in its interior and

$$\sup_{\delta \in \partial B} q(\delta) < q(\tilde{\delta}),$$

where $\partial B$ denotes the boundary of $B$. Let $S$ be the unit sphere. Then

$$(\delta, t) \mapsto t' \bigl( -\nabla^2 q_n(\delta) - K \bigr) t$$

converges continuously to zero on $B \times S$. Hence

$$\sup_{\substack{\delta \in B \\ t \in S}} \bigl| t' \bigl( -\nabla^2 q_n(\delta) - K \bigr) t \bigr| \to 0. \tag{31a}$$

Also

$$q_n(\hat{\delta}) - \sup_{\delta \in \partial B} q_n(\delta) \to q(\hat{\delta}) - \sup_{\delta \in \partial B} q(\delta) \tag{31b}$$

and

$$q_n(\tilde{\delta}_n) - \sup_{\delta \in \partial B} q_n(\delta) \to q(\tilde{\delta}) - \sup_{\delta \in \partial B} q(\delta) \tag{31c}$$

and the limits in both (31b) and (31c) are positive.

Choose $N$ such that for all $n \geq N$ the left hand side of (31a) is strictly less than one-third of the smallest eigenvalue of $K$ and the left hand sides of both (31b) and (31c) are positive. For $n \geq N$, $-\nabla^2 q_n(\delta)$ is positive definite for $\delta \in B$, hence $q_n$ is strictly concave on $B$. Let $\hat{\delta}_n$ denote the (unique by strict concavity) point at which $q_n$ achieves its maximum over $B$. Since the left hand side of (31b) is positive, $\hat{\delta}_n$ is in the interior of $B$ and $\nabla q_n(\hat{\delta}_n) = 0$. By Theorem 1 and Example 2 of Veselić (2004) Newton's method applied to $q_n$ starting at $\tilde{\delta}_n$ converges to $\hat{\delta}_n$ for $n \geq N$.

Consider the functions $H_n$ and $H$ defined by $H_n(\delta) = G(q_n, \delta)$ and $H(\delta) = G(q, \delta)$. Then for $n \geq N$, we have $\hat{\delta}_n = H_n(\hat{\delta}_n)$. By Lemma 3.2 $H_n$ converges to $H$ uniformly on compact sets. By Lemma 3.1 $\hat{\delta} = H(\delta)$ whenever $\delta \neq \text{NaO}$. Thus, if $W$ is a neighborhood of $\hat{\delta}$, then $H_n$ maps $B$ into $W$ for sufficiently large $n$, which implies $\hat{\delta}_n \in W$ for sufficiently large $n$, which implies $\hat{\delta}_n \to \hat{\delta}$. □

**Theorem B.3.** *Theorem 3.3 holds with $G$ replaced by $G^\infty$.*

Just change the proof to use Lemmas B.1 and B.2 instead of Lemmas 3.1 and 3.2.

When Newton's method converges, under these conditions (the Hessian is continuous but not necessarily Lipschitz), then its convergence is superlinear (Fletcher, 1987, Theorem 6.2.3). If the Hessian does happen to be Lipschitz, then Newton's method converges quadratically (Fletcher, 1987, Theorem 6.2.1).

It is clear from the proof of Lemma B.2 that the argument of Section 3.6 carries over to this situation. Now we evaluate $\nabla q_n$ and $\nabla^2 q_n$ at an infinite sequence of points, but as the last paragraph of the proof makes clear, all but the first of these points lie in an arbitrarily small neighborhood of $\hat{\delta}_n$ for sufficiently large $n$. There is "almost" no difference between one-step Newton and Newton iterated to convergence.

Please note that this appendix is an endorsement of Newton's method only for objective functions that are "nearly" strictly concave quadratic so that Newton's method "nearly" converges in one step. If Newton's method does not "nearly" converge in one step, then one is crazy to use it and should instead use some form of "safeguarding" such as line search (Fletcher, 1987, pp. 21–40) or trust regions (Fletcher, 1987, Chapter 5).

## Appendix C: Comparison with Classical Theorems

This appendix compares our "regularity conditions" with the "usual regularity conditions" for maximum likelihood. So we leave "no $n$" territory and return to the conventional "$n$ goes to infinity" story. We adopt "usual regularity conditions" similar to those of (Ferguson, 1996, Chapter 18). Suppose we have independent and identically distributed data $X_1$, $X_2$, …, $X_n$, so the log likelihood is the sum of independent and identically distributed terms

$$l_n(\theta) = \sum_{i=1}^{n} h_i(\theta)$$

where

$$h_i(\theta) = \log\left(\frac{f_\theta(X_i)}{f_\psi(X_i)}\right)$$

and where $\psi$ is the true parameter value. We assume

(a) the parameter space is all of $\mathbb{R}^p$,
(b) $h_i$ is twice continuously differentiable,
(c) There exists an integrable random variable $M$ such that

$$\left\| \nabla^2 h_i(\theta) \right\| \le M, \qquad \text{for all } \theta \text{ in some neighborhood of } \psi,$$

(the norm here being the sup norm),
(d) the expected Fisher information matrix

$$K = -E_\psi \{ \nabla^2 h_i(\psi) \}$$

is positive definite, and
(e) the identity $\int f_\theta(x) \lambda(dx) = 1$, can be differentiated under the integral sign twice.

**Lemma C.1.** *Under assumptions (a) through (e) above,* (15) *holds with* $\tau_n = \sqrt{n}$.

*Proof.* Assumption (e) implies

$$\operatorname{var}_\psi \{ \nabla h_i(\psi) \} = -E_\psi \{ \nabla^2 h_i(\psi) \}$$

by differentiation under the integral sign (Ferguson, 1996, p. 120), assumption (d) implies both sides are equal to the positive definite matrix $K$, and the central limit theorem (CLT) and assumption (e) imply

$$\nabla q_n(0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla h_i(\psi) \xrightarrow{\mathcal{L}} \mathcal{N}(0, K). \tag{32}$$

Define

$$r_n(\delta) = \delta' \nabla q_n(0) - \frac{1}{2} \delta' K \delta$$

and its derivatives

$$\begin{aligned} \nabla r_n(\delta) &= \nabla q_n(0) - K\delta \\ \nabla^2 r_n(\delta) &= -K \end{aligned}$$

We now show that

$$r_n \xrightarrow{\mathcal{L}} q, \qquad \text{in } C^2(\mathbb{R}^p) \tag{33a}$$

is a consequence of the mapping theorem (Billingsley, 1999, Theorem 2.7) and (32). Define a function $F : \mathbb{R}^p \to C^2(\mathbb{R}^p)$ by

$$F(z)(\delta) = z'\delta - \tfrac{1}{2} \delta' K \delta$$

We claim that $F$ is continuous, which says no more than that

$$z_n \to z \qquad \text{and} \qquad \delta_n \to \delta$$

imply

$$\begin{gathered} z_n' \delta_n - \tfrac{1}{2} \delta_n' K \delta_n \to z'\delta - \tfrac{1}{2} \delta' K \delta \\ z_n - K\delta_n \to z - K\delta \end{gathered}$$

which are obvious (and that second derivatives converge, which is trivial). Then an application of the mapping theorem in conjunction with (32) says

$$F\bigl(\nabla q_n(0)\bigr) \xrightarrow{\mathcal{L}} F(Y) \tag{33b}$$

where $Y$ is a random vector that has the distribution on the right hand side of (32), and that is the desired conclusion: (33a) in different notation.

Now we use Slutsky's theorem (Billingsley, 1999, Theorem 3.1). This says that if we can show

$$q_n - r_n \xrightarrow{P} 0, \qquad \text{in } C^2(\mathbb{R}^p),$$

then we are done. Using the isomorphism of $C^2(\mathbb{R}^p)$ to a subspace of $C(\mathbb{R}^p)^{1+p+p\times p}$ and another application of Slutsky, it is enough to show the separate convergences

$$q_n - r_n \xrightarrow{P} 0 \tag{34a}$$

$$\nabla q_n - \nabla r_n \xrightarrow{P} 0 \tag{34b}$$

$$\nabla^2 q_n - \nabla^2 r_n \xrightarrow{P} 0 \tag{34c}$$

which take place in $C(\mathbb{R}^p)$, in $C(\mathbb{R}^p)^p$, and in $C(\mathbb{R}^p)^{p\times p}$, respectively. But these are equivalent to

$$\sup_{\delta \in B} |q_n(\delta) - r_n(\delta)| \xrightarrow{P} 0 \tag{35a}$$

$$\sup_{\delta \in B} \|\nabla q_n(\delta) - \nabla r_n(\delta)\| \xrightarrow{P} 0 \tag{35b}$$

$$\sup_{\delta \in B} \bigl\|\nabla^2 q_n(\delta) - \nabla^2 r_n(\delta)\bigr\| \xrightarrow{P} 0 \tag{35c}$$

holding for every compact set $B$, which can be seen using the metric (29).

Let $B(\theta, \epsilon)$ denote the closed ball in $\mathbb{R}^p$ centered at $\theta$ with radius $\epsilon$. Then assumption (c) can be stated more precisely as the existence of an $\epsilon > 0$ and an integrable random variable $M$ such that

$$\bigl\|\nabla^2 h_i(\theta)\bigr\| \le M, \qquad \theta \in B(\psi, \epsilon). \tag{36}$$

Define

$$H_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \nabla^2 h_i(\theta)$$

and

$$H(\theta) = -E_\psi\{\nabla^2 h_i(\theta)\}.$$

Theorem 16(a) in Ferguson (1996) says that

$$\sup_{\theta \in B(\psi,\epsilon)} \|H_n(\theta) - H(\theta)\| \xrightarrow{\text{as}} 0 \tag{37}$$

and that $H$ is continuous on $B(\psi, \epsilon)$, the latter assertion appearing in the proof rather than in the theorem statement. Note that $H(\psi) = K$. Also note that

$$\begin{aligned} \nabla^2 q_n(\delta) &= -H_n(\psi + n^{-1/2}\delta) \\ \nabla^2 r_n(\delta) &= -H(\psi) \end{aligned}$$

Hence

$$
\begin{aligned}
\sup_{\delta \in B(0,\eta)} \left\| \nabla^2 q_n(\delta) - \nabla^2 r_n(\delta) \right\| &= \sup_{\delta \in B(0,\eta)} \left\| H_n(\psi - n^{-1/2}\delta) - H(\psi) \right\| \\
&= \sup_{\theta \in B(\psi, n^{-1/2}\eta)} \| H_n(\theta) - H(\psi) \| \\
&\le \sup_{\theta \in B(\psi, n^{-1/2}\eta)} \| H_n(\theta) - H(\theta) \| \\
&\qquad + \sup_{\theta \in B(\psi, n^{-1/2}\eta)} \| H(\theta) - H(\psi) \|
\end{aligned}
$$

and the first term on the right hand side is dominated by the left hand side of (37) for $n$ such that $n^{-1/2}\eta \le \epsilon$ and hence converges in probability to zero by (37), and the second term on the right hand side goes to zero by the continuity of $H$. Since this argument works for arbitrarily large $\eta$, it proves (35c).

Now using some of the facts established above and the Maclaurin series

$$
\nabla q_n(\delta) = \nabla q_n(0) + \int_0^1 \nabla^2 q_n(s\delta)\delta \, ds
$$

we get

$$
\nabla q_n(\delta) - \nabla r_n(\delta) = \int_0^1 \Bigl[ H_n(\psi + n^{-1/2} s\delta) - H(\psi) \Bigr] \delta \, ds.
$$

So

$$
\begin{aligned}
\sup_{\delta \in B(0,\eta)} \| \nabla q_n(\delta) - \nabla r_n(\delta) \| &\le \sup_{0 \le s \le 1} \sup_{\delta \in B(0,\eta)} \left\| H_n(\psi + n^{-1/2} s\delta) - H(\psi) \right\| \|\delta\| \\
&\le \sup_{\delta \in B(0,\eta)} \left\| H_n(\psi + n^{-1/2}\delta) - H(\psi) \right\| \eta
\end{aligned}
$$

and we have already shown that the right hand side converges in probability to zero for any fixed $\eta$, however large, and that proves (35b).

Similarly using the Maclaurin series

$$
q_n(\delta) = \delta' \nabla q_n(0) + \int_0^1 \delta' \nabla^2 q_n(s\delta) \delta (1 - s) \, ds
$$

we get

$$
q_n(\delta) - r_n(\delta) = \int_0^1 \delta' \left[ H_n(\psi + n^{-1/2} s\delta) - H(\psi) \right] \delta (1 - s) \, ds
$$

and the argument proceeds similarly to the other two cases. □

**Lemma C.2.** *The assumption* (21) *of Theorem 3.3 can be weakened to* (15) *and* $\tilde{\delta}_n$ *being tight if the last assertion of the theorem is weakened by replacing* (21) *with* (15).

*Proof.* The random sequences $\tilde{\delta}_n$ and $q_n$ are marginally tight, either by assumption or by the converse half of Prohorov's theorem, hence jointly tight (Billingsley, 1999, Problem 5.9). Hence by the direct half of Prohorov's theorem there exist jointly convergent subsequences. For every subsequence $\bigl(q_{n_k}, \tilde{\delta}_{n_k}\bigr)$ there is a convergent subsubsequence (21) with $n$ replaced by $n_{k_l}$. And this implies all the conclusions of the theorem with $n$ replaced by $n_{k_l}$. But since the limits are the same for all subsequences $n_k$, it follows from the subsequence principle (Billingsley, 1999, Theorem 2.6) that the conclusions hold as is with $n$ rather than $n_{k_l}$. □

The point of Lemma C.2 is to justify Lemma C.1 dealing only with $q_n$ instead of both $q_n$ and $\tilde{\delta}_n$. In the conventional "$n$ is sample size going to infinity" story, Lemma C.2 seems important, saying that it is enough that $\tilde{\theta}_n$ be a $\tau_n$-consistent sequence of estimators. In the "no $n$" world, however, Lemma C.2 loses significance because $\tau_n$-consistency is meaningless when $n$ is not reified.

## References

Aliprantis, C. D. and Border, K. C. (1999). *Infinite Dimensional Analysis: A Hitchhicker's Guide*, Second ed. Springer-Verlag, Berlin.

Beattie, R. and Butzmann, H. P. (2002). *Convergence Structures and Applications to Functional Analysis*. Kluwer Academic, Dordrecht.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–468.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83** 687–697.

Billingsley, P. (1999). *Convergence of Probability Measures*, Second ed. John Wiley & Sons, New York.

Bourbaki, N. (1998). *Elements of Mathematics: General Topology. Chapters 5–10.* Springer-Verlag, Berlin. Translated from the French. Reprint of the 1989 English translation.

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

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* Chapman & Hall, London.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* **52** 399–433.

Fletcher, R. (1987). *Practical Methods of Optimization*, Second ed. Wiley-Interscience, Chichester.

Fristedt, B. and Gray, L. (1997). *A Modern Approach to Probability Theory.* Birkhäuser Boston, Boston, MA.

Geyer, C. J. (1991). Constrained maximum likelihood exemplified by isotonic convex logistic regression. *J. Amer. Statist. Assoc.* **86** 717–724.

Geyer, C. J. (1994). On the Asymptotics of Constrained $M$-Estimation. *Ann. Statist.* **22** 1993–2010.

Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.* **21** 359–373.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Henderson, C. R. (1976). A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. *Biometrics* **32** 69–83.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis.* Cambridge University Press, Cambridge.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.* **I** 361–379. University of California Press, Berkeley, CA.

Jensen, J. L. (1991). A note on asymptotic normality in the thermodynamic limit at low densities. *Adv. in Appl. Math.* **12** 387–399.

Jensen, J. L. (1993). Asymptotic normality of estimates in spatial point processes. *Scand. J. Statist.* **20** 97–109.

Kuratowski, K. (1966). *Topology.* Academic Press, New York. Translated from French by J. Jaworowski.

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

Le Cam, L. (1990). Maximum Likelihood: An Introduction. *Internat. Statist. Rev.* **58** 153–171.

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, Second ed. Springer-Verlag, New York.

Miller, J. J. (1977). Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance. *Ann. Statist.* **5** 746–762.

Newton, M. A. and Geyer, C. J. (1994). Bootstrap recycling: A Monte Carlo algorithm for the nested bootstrap. *J. Amer. Statist. Assoc.* **89** 905–912.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Rudin, W. (1991). *Functional Analysis*, Second ed. McGraw-Hill, New York.

Shorack, G. R. (2000). *Probability for Statisticians.* Springer-Verlag, New York.

Strauss, D. J. (1975). A model for clustering. *Biometrika* **62** 467–475.

van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Veselić, K. (2004). On the convergence of the Newton iteration. *ZAMM Z. Angew. Math. Mech.* **84** 147–157. MR2038336 (2004m:65074)