Le Cam Made Simple: Asymptotics of Maximum Likelihood without the LLN or CLT or Sample Size Going to Infinity

By
Charles J. Geyer

Technical Report No. 643

School of Statistics
University of Minnesota

April 20, 2005

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.

The theory presented here is a combination of Le Cam style involving local asymptotic normality (LAN) and local asymptotic mixed normality (LAMN) and Cramér style 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 the Polish space 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.

Keywords: Locally asymptotically normal (LAN), Maximum likelihood, Newton's method, No-n asymptotics, Parametric bootstrap, Quadraticity.

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 the proof. 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 due mostly to Lucien Le Cam. This theory is presented in full generality in Le Cam (1986) and in somewhat simplified form in Le Cam and Yang (2000). Most statisticians would 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 this large body of theory 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.

- 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 or 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

$$l(\theta) = U + Z'\theta - \tfrac{1}{2}\theta' K \theta, \qquad \theta \in \mathbb{R}^p \tag{1.1}$$

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)

$$\hat{\theta} = K^{-1} Z \tag{1.2}$$

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

$$\hat{\theta} \sim \mathcal{N}(\theta, K^{-1}), \tag{1.3}$$

the right hand side being the multivariate normal distribution with mean vector θ 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 θ, then (1.3) 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 (1.1) with K invariant in law, then (1.3) holds approximately too. All likelihood asymptotics that produce conclusions resembling (1.3) 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 so weak that it seems not even topological much less metrizable (Le Cam and Yang, 2000, Chapter 6, Definitions 1, 2, and 3). So at this point I part company with Le Cam and present my own take on "no n asymptotics." 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.

When we say "delta and epsilon arguments ... can replace sequential arguments" we mean they can in principle. In practice, one continues to use sequential arguments whenever they seem more intuitive. 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 δ, which may depend on x and ϵ, such that $d_V\bigl(f(x), f(y)\bigr) < \epsilon$, whenever $d_U\bigl(x, y\bigr) < \delta$. Alternatively, one can show that $f(x_n) \to f(x)$, whenever $x_n \to x$. The mathematical content is the same either way. This equivalence of sequential and delta-epsilon arguments does not change when U and V are spaces of probability measures on Polish spaces, which are themselves Polish spaces (Billingsley, 1999, p. 72), and the f in question is the map from log likelihood to MLE.

The point is that we do not need to reify n. Sequences are merely a technical tool. 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 (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 θ. Thus the traditional role of asymptotics, approximating P_θ, is not needed when we bootstrap. We only need asymptotics to deal with the dependence of P_θ on θ. 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 $P_{\hat{\theta}(y)}$, where y is the data and $\hat{\theta}$ some estimator. The double parametric bootstrap simulates y_1^*, ..., y_n^* from $P_{\hat{\theta}(y)}$

and then simulates 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 θ in the region filled out by the $\hat{\theta}(y_i^*)$. But nothing prevents one from simulating from θ in a bigger region if one wants to. Thus if one does enough simulation, there is no need for either asymptotics or the bootstrap to "work" in the traditional sense.

But often asymptotics does "work" and thus permits simpler calculations and provides 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 of the "No n" Method

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 the cumulant generating function, 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 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. This 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_θ, $\theta \in \Theta$ with respect to a measure λ is a random function l defined by

$$l(\theta) = u(X) + \log f_\theta(X), \qquad \theta \in \Theta, \tag{2.1}$$

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 θ. In this article, we are only interested families of almost surely positive densities so

the argument of the logarithm in (2.1) 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 θ and $\theta + \delta$ we can write

$$\begin{aligned} E_{\theta+\delta}\{g(X)\} &= \int g(x) f_{\theta+\delta}(x) \lambda(dx) \\ &= \int g(x) \frac{f_{\theta+\delta}(x)}{f_\theta(x)} f_\theta(x) \lambda(dx) \\ &= \int g(x) e^{l(\theta+\delta)-l(\theta)} f_\theta(x) \lambda(dx) \\ &= E_\theta\big\{g(X) e^{l(\theta+\delta)-l(\theta)}\big\} \end{aligned} \tag{2.2}$$

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 (1.1). 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 (1.1) analogous to the term $u(X)$ in (2.1) is of no importance. We are mainly interested in log likelihood ratios

$$l(\theta + \delta) - l(\theta) = (Z - K\theta)'\delta - \tfrac{1}{2}\delta' K \delta, \tag{2.3}$$

in which U does not appear.

Theorem 2.1 (LAMN). *Suppose* (1.1) *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 the parameter value θ is $\mathcal{N}(K\theta, K)$.*

(c) *The distribution of K does not depend on the parameter θ.*

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 (2.2) where g is identically equal to one with (2.3) plugged in gives

$$1 = E_{\theta+\delta}(1) = E_\theta\bigl\{e^{(Z-K\theta)'\delta - \frac{1}{2}\delta' K\delta}\bigr\} \tag{2.4}$$

Averaging (2.4) and (2.4) with δ replaced by $-\delta$ gives

$$E_\theta\bigl\{\cosh[(Z - K\theta)'\delta]\exp[-\tfrac{1}{2}\delta' K\delta]\bigr\} = 1. \tag{2.5}$$

Now plug in $s\delta$ for δ in (2.5), 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 (2.2) by $h(K)$ where h is any bounded measurable function giving

$$E_{\theta+\delta}\{h(K)\} = E_\theta\bigl\{h(K) e^{(Z-K\theta)'\delta - \frac{1}{2}\delta' K\delta}\bigr\}. \tag{2.6}$$

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

$$E_\theta\bigl\{e^{Z'\delta} \bigm| K\bigr\} = e^{\theta' K\delta + \frac{1}{2}\delta' K\delta} \tag{2.7}$$

and this implies

$$E_\theta\bigl\{e^{(Z-K\theta)'\delta - \frac{1}{2}\delta' K\delta} \bigm| K\bigr\} = 1. \tag{2.8}$$

Plugging (2.8) into (2.6) 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 (2.6) does not depend on δ, hence

$$E_\theta\{h(K)\} = E_\theta\bigl\{h(K)e^{(Z-K\theta)'\delta-\frac{1}{2}\delta'K\delta}\bigr\}. \tag{2.9}$$

By the definition of conditional expectation (2.9) holding for all bounded measurable functions h implies (2.8) and hence (2.7), which implies (b). This proves (c) implies (b). □

Corollary 2.2 (LAN). *Suppose* (1.1) *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 θ 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 δ 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 θ, and recursively define

$$X_n = \theta X_{n-1} + Z_n \tag{2.10}$$

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

$$l_n(\theta) = U_n + Z_n \theta - \tfrac{1}{2} K_n \theta^2$$

where $Z_n = \sum_{i=1}^{n} X_i X_{i-1}$ and $K_n = \sum_{i=1}^{n} X_{i-1}^2$ and U_n is irrelevant. From (2.10)

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

we can derive

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

for $\theta \neq 1$, which is enough to show that the distribution of K_n does depend on θ and hence 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 θ).

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, also called the theory of convergence in law for random variables taking values in a Polish space (complete separable metric space). Billingsley (1999) 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

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

or

$$\mathcal{L}(X_n) \to \mathcal{L}(X),$$

the notation $\mathcal{L}(X)$ meaning the law of X. This form of convergence is also called *weak convergence*, a term from functional analysis, and (in more elementary contexts) *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 following tools

- Prohorov's theorem (Billingsley, 1999, Theorems 5.1 and 5.2)
- Skorohod's representation theorem (Billingsley, 1999, Theorem 6.7)
- the mapping theorem (Billingsley, 1999, Theorem 2.7)
- the subsequence principle (Billingsley, 1999, Theorem 2.6)
- Slutsky's theorem (Billingsley, 1999, Theorem 3.1)

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.

3.2 The Polish Spaces $C(\mathbb{R}^p, \mathbb{R}^d)$

Let $C(\mathbb{R}^p, \mathbb{R}^d)$ denote the family of all continuous functions from $\mathbb{R}^p$ to $\mathbb{R}^d$ with the topology of uniform convergence on compact sets. We also use $C(\mathbb{R}^p)$ as shorthand for $C(\mathbb{R}^p, \mathbb{R}^1)$. This is a Polish space (Bourbaki, 1998, Ch. 10, Sec. 3.3, Corollary, part (b)).

We will consider a log likelihood whose parameter space is all of $\mathbb{R}^p$ and which is (almost surely) finite-valued and continuous, to be a random element of $C(\mathbb{R}^p)$, and we shall restrict ourselves to such log likelihoods (the measurability details involved in this are relegated to Appendix A).

At first sight the parameter space being all of $\mathbb{R}^p$ seems a serious restriction. Conventional theory only requires the parameter space to be a neighborhood of the true parameter value, but this takes "n goes to infinity" too seriously. In the "no n" view, the actual log likelihood for the actual

data should be nearly quadratic and it cannot be if it is infinite or undefined or discontinuous on some part of $\mathbb{R}^p$. Moreover, if the support of P_θ does not depend on θ (one of the "usual regularity conditions"), one can usually reparameterize so that the parameter space is all of $\mathbb{R}^p$ and the log likelihood is everywhere finite. This reparameterization is an essential part of good asymptotic approximation.

We will be interested not only in the log likelihood $\theta \mapsto l(\theta)$, but also its first derivative $\theta \mapsto \nabla l(\theta)$, the *score function*, and its second derivative $\theta \mapsto \nabla^2 l(\theta)$, the *Hessian function*. The score function we take to be a random element of $C(\mathbb{R}^p, \mathbb{R}^p)$, the score $\nabla l(\theta)$ being the p vector of first partial derivatives. The Hessian function we take to be a random element of $C(\mathbb{R}^p, \mathbb{R}^{p \times p})$, the Hessian $\nabla^2 l(\theta)$ being the $p \times p$ matrix of second partial derivatives. The random function

$$\theta \mapsto \big(l(\theta), \nabla l(\theta), \nabla^2 l(\theta)\big)$$

is a random element of $C(\mathbb{R}^p, \mathbb{R}^{1+p+p \times p})$.

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 these sequences from the discussion if we wanted to. The models have log likelihoods l_n and true parameter values ψ_n.

Merely for comparison with conventional theory (Appendix C) we also introduce a "rate" τ_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 τ_n to play. Our theory is unchanged if we let $\tau_n = 1$ for all n.

3.4 The Key Assumption

Define random functions q_n by

$$q_n(\delta) = l_n(\psi_n + \tau_n^{-1}\delta) - l_n(\psi_n), \qquad \delta \in \mathbb{R}^p. \tag{3.1}$$

We assume that q_n and its two-term Maclaurin series converge to the log likelihood of an LAMN model and its two-term Maclaurin series. For this reason we assume the log likelihoods l_n are twice continuously differentiable,

so the derivatives of q_n are

$$\begin{aligned}\nabla q_n(\delta) &= \tau_n^{-1} \nabla l_n(\psi_n + \tau_n^{-1}\delta) \\ \nabla^2 q_n(\delta) &= \tau_n^{-2} \nabla^2 l_n(\psi_n + \tau_n^{-1}\delta)\end{aligned}$$

Our assumption can then be written

$$\big(q_n, \nabla q_n, \nabla^2 q_n\big) \xrightarrow{\mathcal{L}} \big(q, \nabla q, \nabla^2 q\big), \tag{3.2}$$

this being weak convergence in the Polish space $C(\mathbb{R}^p, \mathbb{R}^{1+p+p\times p})$, where

$$q(\delta) = \delta' Z - \tfrac{1}{2}\delta' K \delta, \qquad \delta \in \mathbb{R}^p \tag{3.3}$$

is the log likelihood of an LAMN model.

The single assumption (3.2) contains the following separate assumptions about the asymptotic LAMN model with log likelihood (3.3).

(a) Its log likelihood q is quadratic (with probability one).

(b) Its true parameter value is $\delta = 0$.

(c) The distribution of K does not depend on the parameter δ. This is the second condition of the LAMN theorem (Theorem 2.1).

(d) K is almost surely positive definite. This is our additional assumption for LAMN models.

(e) The log likelihood q is the log likelihood of an actual probability model, hence

$$E\big\{e^{q(\delta)}\big\} = 1, \qquad \text{for all } \delta. \tag{3.4}$$

Some of these are pure assumptions. Some have reasoning behind them.

First we note that, since $q_n(0)$ is almost surely zero for all n by (3.1), we must also have the same property for $q(0)$, hence no constant term in (3.3).

Second we note that $e^{q_n(\delta)}$ is a probability density with respect to the true distribution (with parameter ψ_n), so

$$E_{\psi_n}\big\{e^{q_n(\delta)}\big\} = 1, \qquad \text{for all } \delta.$$

This allows to conclude by Fatou's lemma for convergence in law (Billingsley, 1999, Theorem 3.4)

$$E\big\{e^{q(\delta)}\big\} \le 1, \qquad \text{for all } \delta. \tag{3.5a}$$

On the other hand from the asymptotic model being LAMN we conclude

$$E_\theta\big\{e^{q(\delta)} \;\big|\; K\big\} = E_\theta\big\{e^{\delta' Z - \frac{1}{2}\delta' K \delta} \;\big|\; K\big\} = e^{\delta' K \theta} \tag{3.5b}$$

and hence

$$E_\theta\big\{ \tfrac{1}{2} e^{q(\delta)} + \tfrac{1}{2} e^{q(-\delta)} \bigm| K \big\} = \cosh(\delta' K \theta)$$

so finally

$$E_\theta\big\{ \tfrac{1}{2} e^{q(\delta)} + \tfrac{1}{2} e^{q(-\delta)} \big\} = E_\theta\big\{ \cosh(\delta' K \theta) \big\} \ge 1, \qquad \text{for all } \delta \text{ and } \theta. \tag{3.5c}$$

But if the inequality (3.5a) were strict for any δ, this would contradict (3.5c). Hence (3.4) holds. So (e) is not an additional assumption. It follows from the rest of our setup. Moreover, if θ is the true parameter value in the asymptotic model, comparison of (3.5a) and (3.5c) shows that for this θ (3.5c) must hold with equality for all δ. But this implies $\delta' K \theta = 0$ almost surely for all δ, hence $K\theta = 0$ almost surely, hence by the positive definiteness of K we have $\theta = 0$. So (b) is not an additional assumption either.

Property (3.4) is what is referred to in the literature as *contiguity* of the sequence of probability measures having parameter values $\psi_n + \tau_n^{-1}\delta$ to the sequence having parameter values ψ_n, rather it is one of many equivalent conditions which collectively are referred to as contiguity (Le Cam and Yang, 2000, Theorem 1 of Chapter 3). As we saw, contiguity follows from the convergence in law (3.2) plus the asymptotic model being a proper probability model (having densities that integrate to one). Thus we see that contiguity has no role to play in "no n" asymptotics, since "improper" asymptotic models make no sense.

Please note that, despite our assuming LAMN with "N" standing for normal, we can say we are not actually assuming asymptotic normality. The 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.

3.5 The Starting Point for Newton's Method

For comparison with classical results, we assume a so-called τ_n consistent starting point for maximization, a sequence $\tilde{\theta}_n$ such that

$$\tilde{\delta}_n = \tau_n(\tilde{\theta}_n - \psi_n) \tag{3.6}$$

is a tight sequence of random elements of $\mathbb{R}^p$. In the "no n" view, where we may have $\tau_n = 1$ for all n, we may as well have (3.6) be a constant sequence of random variables, in which case it is is automatically tight. Then this section adds no additional assumptions, so our only assumption is our "key assumption" (3.2) of Section 3.4.

3.6 The Newton Update

In our main theorem (Theorem 3.1 below), we consider the MLE be a so-called one-step-Newton update of the starting point. In Appendix B we show that Newton iterated to convergence also works. The one-step-Newton update is

$$\hat{\theta}_n = \tilde{\theta}_n + \bigl(-\nabla^2 l_n(\tilde{\theta}_n)\bigr)^{-1} \nabla l_n(\tilde{\theta}_n) \tag{3.7}$$

unless the matrix being inverted is positive definite. Otherwise Newton's method makes no sense (as an attempt at maximization), and we say $\hat{\theta}_n$ is undefined. (Formally, we add a point "undefined" to $\mathbb{R}^d$ and say this value propagates through all operations, like "not a number" in computer arithmetic. We could also identify "undefined" with a particular point of $\mathbb{R}^d$, say zero.) Defining

$$\hat{\delta}_n = \tau_n(\hat{\theta}_n - \psi_n) \tag{3.8}$$

and putting everything on the "τ_n scale" gives

$$\begin{aligned} \hat{\delta}_n &= \tilde{\delta}_n + \tau_n \bigl(-\nabla^2 l_n(\tilde{\theta}_n)\bigr)^{-1} \nabla l_n(\tilde{\theta}_n) \\ &= \tilde{\delta}_n + \bigl(-\nabla^2 q_n(\tilde{\delta}_n)\bigr)^{-1} \nabla q_n(\tilde{\delta}_n) \end{aligned} \tag{3.9}$$

3.7 The Main Theorem

Theorem 3.1. *Assume the "key assumption" of Section 3.4, that* (3.2) *holds with q the log likelihood of an LAMN model* (3.3)*. In addition, assume that $K = -\nabla^2 q$ is almost surely positive-definite-valued. Assume tightness of the sequence* (3.6) *of starting points. Then*

$$-\tau_n^{-2} \nabla^2 l_n(\tilde{\theta}_n) \xrightarrow{\mathcal{L}} K \tag{3.10a}$$

$$-\tau_n^{-2} \nabla^2 l_n(\hat{\theta}_n) \xrightarrow{\mathcal{L}} K \tag{3.10b}$$

$$\tau_n(\hat{\theta}_n - \psi_n) \xrightarrow{\mathcal{L}} K^{-1} Z \tag{3.10c}$$

and

$$\bigl(-\nabla^2 l_n(\hat{\theta}_n)\bigr)^{1/2} (\hat{\theta}_n - \psi_n) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I). \tag{3.10d}$$

where $\hat{\theta}_n$ is defined by (3.7) *or is undefined when the matrix in* (3.7) *is not positive definite, and where the matrix square root in* (3.10d) *is the symmetric square root when the matrix is positive definite and undefined otherwise.*

The "I" on the right hand side of (3.10d) indicates the $p \times p$ identity matrix. The fact that the random variables on the right hand sides of all the

limits in the theorem are never undefined implies that the left hand sides are undefined with probability converging to zero.

Those who like to talk about asymptotic normality of estimators would rewrite (3.10d) as

$$\hat{\theta}_n \approx \mathcal{N}\left(\psi_n, \left(-\nabla^2 l_n(\hat{\theta}_n)\right)^{-1}\right), \tag{3.11}$$

where the double wiggle means "approximately distributed" or something of the sort. Strictly speaking, (3.11) is mathematical nonsense, having no mathematical content except by allusion to (3.10d), but it is similar to

$$\hat{\theta}_n \approx \mathcal{N}\left(\psi, I_n(\hat{\theta}_n)^{-1}\right) \tag{3.12}$$

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

Proof. The random sequences $\tilde{\delta}_n$, q_n, ∇q_n, and $\nabla^2 q_n$ are marginally tight (either by assumption or by the converse half of Prohorov's theorem). Hence they are jointly tight (Billingsley, 1999, Problem 5.9) in the Polish space $\mathbb{R}^p \times C(\mathbb{R}^p, \mathbb{R}^{1+p+p\times p})$, the product of Polish spaces being a Polish space (Billingsley, 1999, Appendix M6). Hence by the direct half of Prohorov's theorem there exist jointly convergent subsequences. Hence by Skorohod's theorem there exist "starred" analogs that converge almost surely and have the same laws.

Temporarily suppressing subsubscripts (we will restore them at the end of the proof), we get

$$\left(\tilde{\delta}_n^*, q_n^*, \nabla q_n^*, \nabla^2 q_n^*\right) \xrightarrow{\text{as}} \left(\tilde{\delta}^*, q^*, \nabla q^*, \nabla^2 q^*\right).$$

In order for q^* to have the same law as q it must be defined just as in (3.3) except with stars

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

where (Z^*, K^*) has the same law as (Z, K).

Now we use the property that uniform convergence on compact sets is the same as continuous convergence (Rockafellar and Wets, 1998, Theorem 7.17), so

$$-\nabla^2 q_n^*(\tilde{\delta}_n^*) \xrightarrow{\text{as}} -\nabla^2 q^*(\tilde{\delta}^*) = K^* \tag{3.13a}$$

and

$$\begin{aligned}
\hat{\delta}_n^* &= \tilde{\delta}_n^* + \left(-\nabla^2 q_n^*(\tilde{\delta}_n^*)\right)^{-1} \nabla q_n^*(\tilde{\delta}_n^*) \\
&\xrightarrow{\text{as}} \tilde{\delta}^* + (K^*)^{-1}(Z^* - K^* \tilde{\delta}^*) \\
&= (K^*)^{-1} Z^*
\end{aligned}$$

(we have used here the fact that the matrix inverse function is continuous on the set of positive definite matrices and that K^* is almost surely positive definite). Thus we conclude

$$\hat{\delta}_n^* \xrightarrow{\text{as}} (K^*)^{-1} Z^*, \tag{3.13b}$$

and hence

$$-\nabla^2 q_n^*(\hat{\delta}_n^*) \xrightarrow{\text{as}} K^*. \tag{3.13c}$$

Combining (3.13b) with (3.13c) and the fact that matrix square root is a continuous operation on the set of positive definite matrices (Horn and Johnson, 1985, Problem 7.2.18) and that K^* is almost surely positive definite gives

$$\bigl(-\nabla^2 q_n^*(\hat{\delta}_n^*)\bigr)^{1/2} \hat{\delta}_n^* \xrightarrow{\text{as}} (K^*)^{-1/2} Z^* \tag{3.13d}$$

and, since almost sure convergence implies convergence in law (by dominated convergence)

$$\bigl(-\nabla^2 q_n^*(\hat{\delta}_n^*)\bigr)^{1/2} \hat{\delta}_n^* \xrightarrow{\mathcal{L}} (K^*)^{-1/2} Z^*.$$

Moreover,

$$\mathcal{L}\left((K^*)^{-1/2} Z^*\right) = \mathcal{L}\left(K^{-1/2} Z\right) = \mathcal{N}(0, I)$$

because $K^{-1/2} Z$ is conditionally multivariate standard normal given K, hence also has this distribution unconditionally. Hence, since

$$\mathcal{L}\left(\bigl(-\nabla^2 q_n^*(\hat{\delta}_n^*)\bigr)^{1/2} \hat{\delta}_n^*\right) = \mathcal{L}\left(\bigl(-\nabla^2 q_n(\hat{\delta}_n)\bigr)^{1/2} \hat{\delta}_n\right), \qquad \text{for all } n,$$

we have

$$\bigl(-\nabla^2 q_n(\hat{\delta}_n)\bigr)^{1/2} \hat{\delta}_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, I). \tag{3.14a}$$

By similar arguments we remove the stars from (3.13a), (3.13c), and (3.13b) to get

$$-\nabla^2 q_n(\tilde{\delta}_n) \xrightarrow{\mathcal{L}} K \tag{3.14b}$$

$$-\nabla^2 q_n(\hat{\delta}_n) \xrightarrow{\mathcal{L}} K \tag{3.14c}$$

$$\hat{\delta}_n \xrightarrow{\mathcal{L}} K^{-1} Z \tag{3.14d}$$

And (3.14b), (3.14c), (3.14d), and (3.14a) would be what was to be proved were it not for the suppressed subsubscripts.

What we have actually proved so far is that for every subsequence

$$\bigl(\tilde{\delta}_{n_k}, q_{n_k}, \nabla q_{n_k}, \nabla^2 q_{n_k}\bigr), \qquad k = 1, 2, \ldots$$

there is a subsubsequence

$$\big(\tilde{\delta}_{n_{k_l}}, q_{n_{k_l}}, \nabla q_{n_{k_l}}, \nabla^2 q_{n_{k_l}}\big), \qquad l = 1, 2, \ldots$$

such that

$$\big(-\nabla^2 q_{n_{k_l}}(\hat{\delta}_{n_{k_l}})\big)^{1/2} \hat{\delta}_{n_{k_l}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, I)$$

but since every such subsubsequence converges to the same limit, the whole sequence converges by the subsequence principle (Billingsley, 1999, Theorem 2.6), that is, (3.14a) holds (as is with subscripts rather than subsubsubscripts). A similar argument shows that (3.14b), (3.14c), and (3.14d) hold as is, and that finishes the proof of the theorem. □

Corollary 3.2. *If K in the theorem is constant, then* (3.10c) *can be replaced by*

$$\tau_n(\hat{\theta}_n - \psi_n) \xrightarrow{\mathcal{L}} \mathcal{N}(0, K^{-1}) \tag{3.15}$$

The proof is obvious. Those who like to talk about asymptotic normality of estimators would rewrite (3.15) as

$$\hat{\theta}_n \approx \mathcal{N}\left(\psi_n, (\tau_n^2 K)^{-1}\right), \tag{3.16}$$

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.

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 to the widest possible audience. We are not advocating political correctness in asymptotics.

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 transfered 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 and in other areas, such as statistical genetics, 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" 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 was justified in Section 3.2. 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.).

We hope that readers who share Le Cam's enthusiasm for weak regularity conditions will not scorn our decision to use Cramér style conditions for our first explication of the "no n" view. As Geyer (1994) shows, similar arguments can be made using weaker topologies, but a lot is lost. Our theorems are much simpler and much more relevant to actual statistical practice. Our "key assumption" (3.2) is about the weakest that still permits treatment of Newton's method.

A Measurability

When is a map $F : \Omega \to C(\mathbb{R}^p, \mathbb{R}^d)$ measurable? Each $F(\omega)$ is a function $\mathbb{R}^p \to \mathbb{R}^d$, and its value at a point θ is written $F(\omega)(\theta)$. This notation being cumbersome, we introduce $f : \Omega \times \mathbb{R}^p \to \mathbb{R}^d$ defined by $f(\omega, \theta) = F(\omega)(\theta)$.

The answer is that F is measurable if and only if f is a Carathéodory function, meaning

- $\theta \mapsto f(\omega, \theta)$ is continuous for each ω, and
- $\omega \mapsto f(\omega, \theta)$ is measurable for each θ.

Theorem 4.54 in Aliprantis and Border (1999) gives almost this result, the only difference being that they have $F_B : \Omega \to C(B, \mathbb{R}^d)$ where B is a compact set in place of our F (and analogously f_B in place of our f).

The map $r_B : g \mapsto g|B$ is continuous from $C(\mathbb{R}^p, \mathbb{R}^d)$ to $C(B, \mathbb{R}^d)$ when $B \subset \mathbb{R}^p$ and $g|B$ denotes the restriction of g to B. Moreover, $F_B = r_B \circ F$. Hence if F is measurable, then each F_B is measurable, hence each f_B is Carathéodory, and this implies f is Carathéodory.

Conversely, if f is Carathéodory, then each f_B is Carathéodory, hence each F_B is measurable. Let W be a measurable subset of $C(\mathbb{R}^p, \mathbb{R}^d)$ and write $W_B = r_B^{-1}(W)$. Then $F_B^{-1}(W_B)$ is measurable for each B. Now let B_1, B_2, ... be a sequence of compact sets such that $\bigcup_{n=1}^{\infty} B_n = \mathbb{R}^p$. Then

$$F^{-1}(W) = \bigcap_{n=1}^{\infty} F_{B_n}^{-1}\big(W_{B_n}\big)$$

So F is measurable.

We also note that for any compact set B the maps

$$g \mapsto g(\theta)$$

and

$$g \mapsto \sup_{\theta \in B} \|g(\theta)\|$$

from $C(\mathbb{R}^p, \mathbb{R}^d)$ to $\mathbb{R}$ are continuous. No measurability issues arise in taking such suprema.

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. Write $\hat{\theta}_n^{(0)}$ instead of $\tilde{\theta}_n$ (same starting point, different notation). And for each i define recursively, just like (3.7),

$$\hat{\theta}_n^{(i+1)} = \hat{\theta}_n^{(i)} + \big(-\nabla^2 l_n(\hat{\theta}_n^{(i)})\big)^{-1} \nabla l_n(\hat{\theta}_n^{(i)}) \tag{B.1}$$

so the $\hat{\theta}_n^{(i)}$ are successive iterates of Newton's method starting at $\hat{\theta}_n^{(0)} = \tilde{\theta}_n$. If for any i the matrix being inverted in (B.1) fails to be positive definite, then we say Newton's method fails. Otherwise, if the infinite sequence of iterates $\hat{\theta}_n^{(i+1)}$, $i = 1, 2, \ldots$ fails to converge, then we also say Newton's method fails. If Newton's method does not fail, then we define $\hat{\theta}_n$ to be the limit of the sequence of iterates, otherwise we say $\hat{\theta}_n$ is undefined (and as in Section 3 say that "undefined" propagates through all operations).

Theorem B.1. *Under the assumptions of Theorem 3.1, define (as discussed above) $\hat{\theta}_n$ to be the point to which Newton's method started at $\tilde{\theta}_n$ converges, if all Newton iterations are well defined and the sequence of iterates converges, and undefined otherwise. Then all conclusions of Theorem 3.1 hold.*

Proof. The the proof starts with a Prohorov-Skorohod construction just like the beginning of the proof of Theorem 3.1. Our argument is omega-by-omega, but we do not indicate this in our notation.

First note that if B is a compact convex set and C is the unit sphere, then $(\delta, t) \mapsto -t'\nabla^2 q_n^*(\delta)t$ converges continuously on $B \times C$ to $(\delta, t) \mapsto t'Kt$. Hence

$$\inf_{\delta \in B, t \in C} -t'\nabla^2 q_n^*(\delta)t \xrightarrow{\text{as}} \inf_{t \in C} t'Kt$$

and by assumption the right hand side is almost surely positive. Hence q_n^* is eventually (meaning there exists an N, which may depend on B and ω, such that the property holds for $n \geq N$) strictly concave on B.

Next consider the Newton map (that produces the next Newton iterate from the current one)

$$G_n^* : \delta \mapsto \delta + \bigl(-\nabla^2 q_n^*(\delta)\bigr)^{-1} \nabla q_n^*(\delta)$$

This is seen (just as in the proof of Theorem 3.1) to converge continuously to the constant map G^* having the value $(K^*)^{-1}Z^*$ (constant as a function of δ, the constant is different for each ω). Hence if we denote the ball of radius ϵ centered at δ by $B(\delta, \epsilon)$, we see for any $\epsilon > 0$ and $\eta > \epsilon$ that G_n^* eventually maps $B\bigl((K^*)^{-1}Z^*, \eta\bigr)$ into $B\bigl((K^*)^{-1}Z^*, \epsilon\bigr)$.

For any $\epsilon > 0$ choose η and N (which may depend on ω and ϵ) such that for $n \geq N$ we have (1) $\tilde{\delta}_n^*$ in $B\bigl((K^*)^{-1}Z^*, \eta\bigr)$, which is possible because $\tilde{\delta}_n^*$ converges, (2) q_n^* strictly concave on $B\bigl((K^*)^{-1}Z^*, \eta\bigr)$, and (3) G_n^* mapping $B\bigl((K^*)^{-1}Z^*, \eta\bigr)$ into $B\bigl((K^*)^{-1}Z^*, \epsilon\bigr)$. For such n, by Brower's fixed point theorem (Ortega and Rheinboldt, 2000, Theorem 6.3.2), G_n^* has a fixed point $\check{\delta}_n^*$ in $B\bigl((K^*)^{-1}Z^*, \epsilon\bigr)$. This implies $\nabla q_n^*(\check{\delta}_n^*) = 0$, hence that $\check{\delta}_n^*$ is the unique global maximizer of q_n^* over $B\bigl((K^*)^{-1}Z^*, \eta\bigr)$, hence that $\check{\delta}_n^*$ is the unique fixed point of G_n^*, hence that Newton's method does not fail and the sequence of iterates converges to $\check{\delta}_n^*$, which we may now denote $\hat{\delta}_n^*$.

Hence every cluster point of the sequence $\hat{\delta}_n^*$ must be in $B\big((K^*)^{-1}Z^*, \epsilon\big)$. Since ϵ was arbitrary we must have (3.13b). Now the rest of the proof proceeds just like that of Theorem 3.1. □

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 ψ 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

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

(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,* (3.2) *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

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

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

$$(r_n, \nabla r_n, \nabla^2 r_n) \xrightarrow{\mathcal{L}} (q, \nabla q, \nabla^2 q) \tag{C.2a}$$

in the Polish space $C(\mathbb{R}^p, \mathbb{R}^{1+p+p\times p})$ is a consequence of the mapping theorem (Billingsley, 1999, Theorem 2.7) and (C.1). Define a function F from $\mathbb{R}^p$ to $C(\mathbb{R}^p, \mathbb{R}^{1+p+p\times p})$

$$F : z \mapsto \bigl(\delta \mapsto (z'\delta - \tfrac{1}{2}\delta' K \delta, z - K\delta, -K)\bigr).$$

We claim that F is continuous, which (using the equivalence of continuous convergence and uniform convergence on compact sets) says no more than that

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

imply

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

which are obvious. Then an application of the mapping theorem in conjunction with (C.1) says

$$F\bigl(\nabla q_n(0)\bigr) \xrightarrow{\mathcal{L}} F(Z) \tag{C.2b}$$

where Z is a random vector that has the distribution on the right hand side of (C.1), and that is the desired conclusion: (C.2a) in different notation.

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

$$(q_n, \nabla q_n, \nabla^2 q_n) - (r_n, \nabla r_n, \nabla^2 r_n) \xrightarrow{P} 0,$$

this being a convergence in probability statement in $C(\mathbb{R}^p, \mathbb{R}^{1+p+p\times p})$, then we are done. By another application of Slutsky, it is enough to show the separate convergences

$$q_n - r_n \xrightarrow{P} 0 \tag{C.3a}$$

$$\nabla q_n - \nabla r_n \xrightarrow{P} 0 \tag{C.3b}$$

$$\nabla^2 q_n - \nabla^2 r_n \xrightarrow{P} 0 \tag{C.3c}$$

which take place in $C(\mathbb{R}^p, \mathbb{R}^1)$, in $C(\mathbb{R}^p, \mathbb{R}^p)$, and in $C(\mathbb{R}^p, \mathbb{R}^{p\times p})$, respectively. But these are equivalent to

$$\sup_{\delta\in B} |q_n(\delta) - r_n(\delta)| \xrightarrow{P} 0 \tag{C.4a}$$

$$\sup_{\delta\in B} \|\nabla q_n(\delta) - \nabla r_n(\delta)\| \xrightarrow{P} 0 \tag{C.4b}$$

$$\sup_{\delta\in B} \left\|\nabla^2 q_n(\delta) - \nabla^2 r_n(\delta)\right\| \xrightarrow{P} 0 \tag{C.4c}$$

holding for every compact set B, which can be seen using a metric for $C(\mathbb{R}^p, \mathbb{R}^d)$, which can be constructed using any sequence $B_1 \subset B_2 \subset \ldots$ of compact sets such that $\bigcup_{n=1}^{\infty} B_n = \mathbb{R}^p$, the metric d being defined by

$$d(f, g) = \sum_{n=1}^{\infty} \frac{2^{-n} \|f - g\|_{B_n}}{1 + \|f - g\|_{B_n}} \tag{C.5}$$

(Rudin, 1973, Section 1.44), where we have introduced the notation

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

Then it is clear from (C.5) that

$$d(f, g) \le \|f - g\|_{B_n} + 2^{-n}$$

and hence $d(f, g)$ can be made as small as one pleases by controlling $\|f-g\|_{B_n}$ for sufficiently large n.

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

$$\left\|\nabla^2 h_i(\theta)\right\| \le M, \qquad \theta \in B(\psi, \epsilon). \tag{C.6}$$

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

$$\sup_{\theta \in B(\psi,\epsilon)} \|H_n(\theta) - H(\theta)\| \xrightarrow{\text{as}} 0 \tag{C.7}$$

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

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

and the first term on the right hand side is dominated by the left hand side of (C.7) for n such that $n^{-1/2}\eta \le \epsilon$ and hence converges in probability to zero by (C.7), and the second term on the right hand side goes to zero by the continuity of H. Since this argument works for arbitrarily large η, it proves (C.4c).

Now using some of the facts established in the preceding section 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

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

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 η, however large, and that proves (C.4b).

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. □

References

Aliprantis, C. D. and Border, K. C. (1999). *Infinite Dimensional Analysis: A Hitchhicker's Guide*, 2nd edition. Springer-Verlag, Berlin.

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*, 2nd edition. 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.

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.

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

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.*, volume I, pp. 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.

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*, 2nd edition. Springer-Verlag, New York.

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.

Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative Solutions of Nonlinear Equations in Several Variables*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1970 original.

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

Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational Analysis*. Springer-Verlag, Berlin.

Rudin, W. (1973). *Functional Analysis*. 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.