Stat 8112 Lecture Notes

The Wald Consistency Theorem

Charles J. Geyer

April 29, 2012

1 Analyticity Assumptions

Let $\{ f_\theta : \theta \in \Theta \}$ be a family of subprobability densities[1] with respect to a measure μ on a measurable space $\mathcal{S}$. We start with the following assumptions.

(a) Both the sample space $\mathcal{S}$ and the parameter space Θ are Borel spaces.

(b) The map $(x, \theta) \mapsto f_\theta(x)$ is upper semianalytic.

A topological space is a *Borel space* if it is homeomorphic to a Borel subset of a Polish space (Bertsekas and Shreve, 1978, Definition 7.7). Examples of Borel spaces include any Borel subset of a Euclidean space $\mathbb{R}^d$ and, more generally, any Borel subset of a Polish space (Bertsekas and Shreve, 1978, Proposition 7.12). If $\mathcal{X}$ and $\mathcal{Y}$ are Borel spaces, a function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is *upper semianalytic* if for every $c \in \mathbb{R}$ the set

$$\{ (x, y) \in \mathcal{X} \times \mathcal{Y} : g(x, y) > c \}$$

is an analytic subset of $\mathcal{X} \times \mathcal{Y}$. This is not the place to define analytic sets, see Bertsekas and Shreve (1978, pp. 156 ff.). A much stronger condition that implies upper semianalyticity is simply that the map $(x, \theta) \mapsto f_\theta(x)$ be (jointly) Borel measurable, i. e., that

$$\{ (x, \theta) \in \mathcal{S} \times \Theta : f_\theta(x) > c \}$$

is a Borel subset of $\mathcal{S} \times \Theta$ for each $c \in \mathbb{R}$ (Bertsekas and Shreve, 1978, Proposition 7.36) This always holds in practical applications.

[1] A real-valued function f is a subprobability density with respect to μ if $f(x) \geq 0$ for μ almost all x and $\int f \, d\mu \leq 1$. A subprobability density is actually a probability density if $\int f \, d\mu = 1$. We are mostly interested in families of probability densities, but subprobability densities enter in the process of compactification (Section 3).

2 Topological Assumptions

Let θ_0 be a point in Θ, the "true" parameter value, and denote the probability measure having density f_{θ_0} with respect to μ by P_{θ_0}. We assume, without loss of generality, that P_{θ_0} is *complete,* i. e., that every subset of a P_{θ_0}-null set is measurable.[2] Suppose X_1, X_2, ... is an independent identically distributed sequence of observations having the distribution P_{θ_0}. The log likelihood corresponding to a sample of size n is

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \tag{1}$$

(this is actually log likelihood divided by n rather than log likelihood, but maximizing one maximizes the other and we shall not distinguish them).

Wald's approach makes further topological assumptions about the densities and the parameter space.

(c) The parameter space Θ is a compact metric space.

(d) For every $\theta \in \Theta$ and every $x \in \mathcal{S}$ except for x in a P_{θ_0} nullset that may depend on θ, the map $\phi \mapsto f_\phi(x)$ is upper semicontinuous at θ.

Spelled out in detail, (d) says that for every $\theta \in \Theta$ there is a measurable subset N_θ of $\mathcal{S}$ such that $P_{\theta_0}(N_\theta) = 0$ and for all $x \in \mathcal{S} \setminus N_\theta$ and every sequence $\{\phi_n\}$ converging to θ

$$\limsup_{n \to \infty} f_{\phi_n}(x) \le f_\theta(x). \tag{2}$$

This has the following consequence. Let $d(\,\cdot\,,\,\cdot\,)$ be a metric for Θ — being a Borel space, Θ is metrizable (Bertsekas and Shreve, 1978, p. 118). Then for every $\theta \in \Theta$ and every x not in the null set N_θ described above

$$\lim_{\epsilon \downarrow 0} \sup_{\substack{\phi \in \Theta \\ d(\theta, \phi) < \epsilon}} f_\phi(x) = f_\theta(x) \tag{3}$$

because if (3) does not hold, there is an $r > f_\theta(x)$ and a sequence $\{\phi_n\}$ such that $d(\theta, \phi_n) < 1/n$ and $f_{\phi_n}(x) > r$, and then (2) cannot hold either.

[2] Any measure space $(\Omega, \mathcal{A}, \mu)$ can be *completed* producing a measure space $(\Omega, \mathcal{A}^*, \mu^*)$ as follows. Let $\mathcal{A}^*$ consist of all subsets A of Ω such that there exist sets A_1 and A_2 in $\mathcal{A}$ such that $A_1 \subset A \subset A_2$ and $P(A_1) = P(A_2)$, and define $P^*(A) = P(A_1)$. Then it is easily verified that (a) $\mathcal{A}^*$ is a σ-field, (b) μ^* is countably additive on $\mathcal{A}^*$ and (c) $A \subset N$ and $P(N) = 0$ implies $A \in \mathcal{A}^*$.

3 Compactification

Assumption (c), the compactness assumption, usually does not hold in applications. The original parameter space must be compactified in a way such that the maps $\theta \mapsto f_\theta(x)$ remain upper semicontinuous for the points θ added to the original parameter space in the compactification and the Wald integrability condition (Section 6 below) also holds for the points added in the compactification.

When one compactifies the parameter space, one must devise new subdensities in the family corresponding to these new points. The most obvious way to do this is by taking limits of the subdensities in the original model. If $\theta_n \to \theta$ with θ_n in the original model and θ a new point added in the compactification, then the pointwise limit defined by

$$f_\theta(x) = \lim_{n \to \infty} f_{\theta_n}(x), \qquad \text{for all } x \tag{4}$$

is an obvious candidate provided the limit exists. Even if the f_{θ_n} are all probability densities, Fatou's lemma only guarantees that f_θ is a subprobability density.

Compactification has been much discussed in the literature, see, for example Kiefer and Wolfowitz (1956) and Bahadur (1971). Bahadur gives a very general recipe for compactifying the parameter space. On p. 34 of Bahadur (1971) two conditions are required, that $g_M(x, r)$ defined by his equation (9.5) be Borel measurable and that his equation (9.6) hold. The measurability condition is unnecessary under the usual regularity conditions involving analytic sets, because then $g_M(x, r)$ is measurable with respect to the completion of P_{θ_0} and that is all that is required. Hence if we use analytic set theory, only Bahadur's (9.6) is required for a suitable compactification.

4 Identifiability

A parameterization of a family of probability distributions is *identifiable* if there do not exist two distinct parameter values that correspond to the same distribution. This will be assumed in this handout.

(e) The parameterization is identifiable.

This assumption can be dropped at the cost of some additional complication in the statement of theorems (Redner, 1981).

5 Kullback-Leibler Information

Define the Kullback-Leibler information function

$$\lambda(\theta) = E_{\theta_0} \log \frac{f_\theta(X)}{f_{\theta_0}(X)} = \int \log \frac{f_\theta(x)}{f_{\theta_0}(x)} P_{\theta_0}(dx). \tag{5}$$

By Jensen's inequality $\lambda(\theta) \leq 0$ for all θ, and by the conditions for equality in Jensen's inequality, $\lambda(\theta) = \lambda(\theta_0)$ only if $f_\theta(x) = f_{\theta_0}(x)$ for almost all x $[P_{\theta_0}]$.

Hence by the identifiablity assumption (e), λ is a non-positive function on Θ that achieves its unique maximum at θ_0 where it takes the value zero. Note that the value $-\infty$ is allowed for λ. This does not cause any difficulties.

6 The Wald Integrability Condition

For each $\theta \in \Theta$, the strong law of large numbers implies

$$l_n(\theta) \xrightarrow{\text{a.s.}} \lambda(\theta), \qquad \text{as } n \to \infty,$$

but this does not imply consistency of maximum likelihood. Pointwise convergence of a sequence of functions does not necessarily imply convergence of a sequence of maximizers of those functions. Some uniformity of the functional convergence is necessary.

To get this uniformity, Wald used dominated convergence, and to get that, Wald had to make an additional strong assumption, which will be referred to as "Wald's integrability condition"

(f) For every $\theta \in \Theta$ there exists an $\epsilon > 0$ such that

$$E_{\theta_0} \sup_{\substack{\phi \in \Theta \\ d(\theta,\phi) < \epsilon}} \log \frac{f_\phi(X)}{f_{\theta_0}(X)} < \infty \tag{6}$$

In order for (6) to make sense, the integrand must be measurable, that is,

$$w_{\epsilon,\theta}(x) = \sup_{\substack{\phi \in \Theta \\ d(\theta,\phi) < \epsilon}} \log \frac{f_\phi(x)}{f_{\theta_0}(x)} \tag{7}$$

must be a measurable function of x. Under the usual regularity assumptions involving analytic sets, $w_{\epsilon,\theta}$ is not, in general, Borel measurable. However it is upper semianalytic (Bertsekas and Shreve, 1978, Proposition 7.47) and hence measurable with respect to the P_{θ_0} completion of the Borel σ-field (Bertsekas and Shreve, 1978, p. 167 and Corollary 7.42.1). Thus the integral in (6) is well-defined.

7 Alternative Assumptions

If one does not assume (a) and (b), then it is necessary to impose further continuity conditions on the densities to obtain Borel measurability of $w_{\epsilon,\theta}$. The following condition is taken from Wang (1985), but it is similar to all other conditions in this literature (Wald, 1949; Kiefer and Wolfowitz, 1956; Perleman, 1972; Geyer, 1994).

(g) For every $\theta \in \Theta$ and every $x \in \mathcal{S}$ except for x in a P_{θ_0} nullset that does not depend on θ, the map $\phi \mapsto f_\phi(x)$ is lower semicontinuous at θ.

This condition is used only to obtain measurability so that the integral in (6) is well-defined. Hence it is completely superfluous when the usual regularity conditions involving analytic sets hold.

8 Upper Semicontinuity

Wald's integrability condition (f) and the upper semicontinuity condition (d) imply that λ is an upper semicontinuous function. For any sequence $\{\phi_n\}$ converging to θ

$$\begin{aligned}
\limsup_{n\to\infty} \lambda(\phi_n) &= \limsup_{n\to\infty} E_{\theta_0} \log \frac{f_{\phi_n}(X)}{f_{\theta_0}(X)} \\
&\le E_{\theta_0} \limsup_{n\to\infty} \log \frac{f_{\phi_n}(X)}{f_{\theta_0}(X)} \\
&= E_{\theta_0} \log \frac{f_\theta(X)}{f_{\theta_0}(X)} \\
&= \lambda(\theta)
\end{aligned}$$

by dominated convergence, since the integrand is eventually dominated by $w_{\epsilon,\theta}$ given by (7).

A closed subset of a compact set is compact (Browder, 1996, Proposition 9.52). Hence for every $\eta > 0$, the set

$$K_\eta = \{\, \theta \in \Theta : d(\theta, \theta_0) \ge \eta \,\} \tag{8}$$

is compact. An upper semicontinuous function achieves its maximum over a compact set.[3] Hence λ achieves its maximum over each K_η, call that $m(\eta)$. By the identifiability assumption (e), we cannot have $m(\eta) = 0$ because that would imply there is a $\theta \in K_\eta$ such that $\lambda(\theta) = \lambda(\theta_0)$.

[3] This is true in general topology, but in metric spaces which are our concern here, the

9 Theorem

Wald's theorem can now be roughly stated (a more precise statement follows). Wald proved that the log likelihood l_n converges to λ uniformly from above in the sense that with probability one

$$\limsup_{n \to \infty} \sup_{\theta \in K_\eta} l_n(\theta) \le \sup_{\theta \in K_\eta} \lambda(\theta), \qquad \text{almost surely,} \tag{9}$$

where λ is defined by (5) and K_η is defined by (8). Let δ_n be any sequence decreasing to zero. Call any sequence of estimators $\hat{\theta}_n(X_1, \ldots, X_n)$ satisfying

$$l_n(\hat{\theta}_n(X_1, \ldots, X_n)) > \sup_{\theta \in \Theta} l_n(\theta) - \delta_n$$

an *approximate maximum likelihood estimator* (AMLE). For such a sequence we have $d(\hat{\theta}_n(X_1, \ldots, X_n), \theta_0)$ is eventually less than η, because there exists an $N_1 \in \mathbb{N}$ such that

$$\sup_{\substack{\theta \in \Theta \\ d(\theta, \theta_0) \ge \eta}} l_n(\theta) \le m(\eta)/2, \qquad n \ge N_1,$$

where $m(\eta)$ denotes the right-hand side of (9), and there exists $N_2 > N_1$ such that $\delta_n < -m(\eta)/2$ when $n \ge N_2$, from which we infer

$$l_n(\hat{\theta}_n(X_1, \ldots, X_n)) > m(\eta)/2, \qquad n \ge N_2$$

which together imply

$$d\big(\hat{\theta}_n(X_1, \ldots, X_n), \theta_0\big) < \eta, \qquad n \ge N_2. \tag{10}$$

Since for every $\eta > 0$ there exists an N_2 (which may depend on η) such that (10) holds, this says that the AMLE converges almost surely to θ_0.

The reason for introducing AMLE is that the maximum of the log likelihood need not be achieved, in which case the maximum likelihood estimator (MLE) does not exist, but an AMLE always exists. If the MLE always exists, then the MLE is also an AMLE and is covered by the theorem.

proof is trivial: if K is a compact set and f is an upper semicontinuous function on K, then for every n there is a is a θ_n such that $f(\theta_n) \ge \sup f - 1/n$, and, since compact is the same as sequentially compact for metric spaces, there is a convergent subsequence $\theta_{n_k} \to \theta$, and $\theta \in K$ because K is closed. Finally, $\limsup f(\theta_{n_k}) \le f(\theta)$ by upper semicontinuity of f, but $f(\theta_n) \to \sup f$ by construction. Thus $f(\theta) = \sup f$.

10 Proof

Now let us go though a careful statement and proof of Wald's theorem, the main point being to establish (9).

Theorem 1. *Suppose assumptions (a) through (f) hold. Then (9) holds and consequently any approximate maximum likelihood estimator is strongly consistent.*

The part of the proof showing that (9) implies strong consistency of AMLE has already been discussed. It remains only to be shown that the regularity conditions (a) through (f) imply (9). For any $\theta \in \Theta$ and any ϵ small enough so that (6) holds, the strong law of large numbers implies that

$$\begin{aligned} \sup_{\substack{\phi \in \Theta \\ d(\theta, \phi) < \epsilon}} l_n(\phi) &\le \frac{1}{n} \sum_{i=1}^{n} \sup_{\substack{\phi \in \Theta \\ d(\theta, \phi) < \epsilon}} \log \frac{f_\phi(X_i)}{f_{\theta_0}(X_i)} \\ &\xrightarrow{\text{a.s.}} E_{\theta_0} \sup_{\substack{\phi \in \Theta \\ d(\theta, \phi) < \epsilon}} \log \frac{f_\phi(X)}{f_{\theta_0}(X)} \end{aligned} \tag{11}$$

As we have seen, the upper semicontinuity assumption (d) implies (3), and that implies that the integrand of the right hand side of (11) converges to $\log(f_\theta / f_{\theta_0})$ as $\epsilon \downarrow 0$. The integrand is a decreasing function of ϵ, since a sup over a smaller set is less than a sup over a larger set, and the integrand is integrable by Wald's integrability assumption (f) for all small enough ϵ. Hence by dominated convergence

$$\lim_{\epsilon \downarrow 0} E_{\theta_0} \sup_{\substack{\phi \in \Theta \\ d(\theta, \phi) < \epsilon}} \log \frac{f_\phi(X)}{f_{\theta_0}(X)} = \lambda(\theta)$$

Thus for any $\theta \in K_\eta$ and any $\gamma > 0$ there is an ϵ_θ such that

$$E_{\theta_0} \sup_{\substack{\phi \in \Theta \\ d(\theta, \phi) < \epsilon_\theta}} \log \frac{f_\phi(X)}{f_{\theta_0}(X)} < m(\eta) + \gamma,$$

where, as in the preceding sections, K_η is defined by (8) and $m(\eta)$ denotes the right-hand side of (9). For each θ define the open set

$$V_\theta = \{ \phi \in \Theta : d(\theta, \phi) < \epsilon_\theta \}.$$

The family $\{ V_\theta : \theta \in K_\eta \}$ is an open cover of K_η and hence has a finite subcover V_{θ_1}, V_{θ_2}, $\ldots$, V_{θ_d} (this uses the definition from general topology:

a set is *compact* if every open cover has a finite subcover). Now another application of the strong law of large numbers says that for almost all sample paths X_1, X_2, ...

$$\limsup_{n \to \infty} \sup_{\phi \in V_{\theta_i}} l_n(\phi) < m(\eta) + 2\gamma, \qquad i = 1, \ldots d. \tag{12}$$

Since the V_{θ_i} cover K_η this implies

$$\limsup_{n \to \infty} \sup_{\theta \in K_\eta} l_n(\theta) < m(\eta) + 2\gamma,$$

which, since $\gamma > 0$ was arbitrary, implies (9).

11 Example

The following example is taken from DeGroot and Schervish (2002, Example 6.5.8) where it purports to be as an example of what is problematic about maximum likelihood (it is no such thing, as we shall see). The data is generated by the following curious mechanism: a coin is flipped and the data is standard normal if the coin comes up heads and is general normal if the coin comes up tails. The results of the coin flip are not observed, so the distribution of the data is a mixture of two normals. The density is given by

$$f_{\mu,\nu}(x) = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \left[\exp\left(-\frac{x^2}{2} \right) + \frac{1}{\sqrt{\nu}} \exp\left(-\frac{(x - \mu)^2}{2\nu} \right) \right]. \tag{13}$$

Admittedly, this is a toy problem. But it illustrates all the issues that arise with the general normal mixture problem, which has been studied for more than a century and has a rich literature.

If we observe independent and identically distributed (IID) data having probability density function (PDF) (13), then the likelihood is

$$L(\mu, \nu) = \prod_{i=1}^{n} f_{\mu,\nu}(x_i).$$

If we set $\mu = x_i$ and let $\nu \to 0$, then we see that the i-th term goes to infinity and the other terms are bounded. Hence

$$\begin{aligned} \hat{\mu}_n &= x_n \\ \hat{\nu}_n &= 0 \end{aligned}$$

maximizes the likelihood (at infinity) but does not converge to the true parameter value. The same issue arises with the general normal mixture problem.

Here we follow Hathaway (1985) in proposing to constrain the variance parameter ν away from zero. Fix $\epsilon > 0$. It may be chosen arbitrarily small, say $\epsilon = 10^{-10^{10}}$, so small that no question arises that the true variance is larger than that. Any ϵ will do, so long as $\epsilon > 0$. If we maximize the likelihood subject to the constraint $\nu \geq \epsilon$, the problem discussed above does not arise.

To compactify the parameter space, we must add the values $+\infty$ and $-\infty$ for μ and $+\infty$ for ν. This makes the compactified parameter space $[-\infty, +\infty] \times [\epsilon, +\infty]$. It is clear that

$$f_{\mu,\nu}(x) \to \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad \mu \to \pm\infty,$$

so we assign this subprobability density to parameter points in the compactification having $\mu = \pm\infty$. We get the same limit when $\nu \to \infty$, so we assign this subprobability density to those parameter points in the compactification too.

The likelihood is continuous (not just upper semicontinuous) so the upper semicontinuity condition (d) holds.

We now need to check the integrability condition (e). In this example it turns out we can sup over the whole compactified parameter space. Let (μ_0, ν_0) be the true unknown parameter vector. For any (μ^*, ν^*) in the compactified parameter space we have

$$\begin{aligned} \frac{f_{\mu^*,\nu^*}(x)}{f_{\mu_0,\nu_0}(x)} &\leq \frac{\exp\left(-\frac{x^2}{2}\right) + \frac{1}{\sqrt{\epsilon}}}{\exp\left(-\frac{x^2}{2}\right) + \frac{1}{\sqrt{\nu}} \exp\left(-\frac{(x-\mu_0)^2}{2\nu_0}\right)} \\ &\leq \frac{\exp\left(-\frac{x^2}{2}\right) + \frac{1}{\sqrt{\epsilon}}}{\exp\left(-\frac{x^2}{2}\right)} \\ &\leq 1 + \frac{1}{\sqrt{\epsilon}} \exp\left(\frac{x^2}{2}\right) \end{aligned}$$

and, if we have chosen $\varepsilon \leq 1$ (which we may always do)

$$\frac{f_{\mu^*,\nu^*}(x)}{f_{\mu_0,\nu_0}(x)} \leq \frac{2}{\sqrt{\epsilon}} \exp\left(\frac{x^2}{2}\right)$$

so

$$\log \frac{f_{\mu^*,\nu^*}(x)}{f_{\mu_0,\nu_0}(x)} \leq \log(2) - \frac{1}{2}\log(\epsilon) + \frac{x^2}{2}$$

and this is clearly integrable because second moments of the normal distribution exist.

Thus we have satisfied the conditions for Wald's theorem, and the maximum likelihood estimator is strongly consistent for this model if we impose the constraint $\nu \geq \epsilon$.

Hathaway (1985) actually looked at the general normal mixture model with d components to the mixture for an arbitrary (but known d). This model has $3d$ parameters, the d means and d variances of the components (none of which are assumed to be mean zero and variance one like in the our toy example) and the d probabilities of the components of the mixture (none of which are assumed to be 1/2 like in our toy example). Instead of putting a lower bound on the variances, Hathaway (1985) imposes the constraints

$$\frac{\nu_i}{\nu_j} \geq \epsilon, \qquad i = 1, \ldots, d \text{ and } j = 1, \ldots, d \tag{14}$$

So the variances have no lower bound or upper bound, but are not allowed to be too different from each other (but $\epsilon = 10^{-10^{10}}$ is still o. k., so they can be *very* different but not *arbitrarily* different. Hathaway (1985) proves strong consistency of the constrained maximum likelihood estimator for the general normal mixture problem, the constraints being (14). His proof uses Wald's theorem.

12 A Trick due to Kiefer and Wolfowitz

Kiefer and Wolfowitz (1956) noted that Wald's method does not work as stated even for location-scale families, but a simple modification does work. Given a PDF f, the *location-scale family with standard density* f is the family of distributions having densities given by

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} \cdot f\left(\frac{x - \mu}{\sigma}\right) \tag{15}$$

where μ, the *location parameter*, is real-valued and σ, the *scale parameter*, is positive-real-valued. The "standard density" f is often chosen to have mean zero and variance one, in which case μ is the mean and σ is the standard deviation of the distribution having density (15). But f need not have moments (for example, f could be the standard Cauchy PDF), in which case μ is not the mean and σ is not the standard deviation (the mean and standard deviation don't exist), and even if f does have moments, it need not have mean zero and variance one, in which case μ is not the mean and σ is not the standard deviation (the mean and standard deviation do exist but aren't μ and σ).

The same problem arises with the location-scale problem as arises with the normal mixture problem: when $\mu = x$ the limit as $\sigma \to 0$ in (15) is infinite. This gives problems both in trying to compactify the model and in verifying the integrability condition. To compactify the model, observe that

$$\lim_{\mu \to \pm\infty} f_{\mu,\sigma}(x) = 0, \qquad \text{for all } x \text{ and all } \sigma \tag{16a}$$

and

$$\lim_{\sigma \to +\infty} f_{\mu,\sigma}(x) = 0, \qquad \text{for all } x \text{ and all } \mu \tag{16b}$$

and

$$\lim_{\sigma \to 0} f_{\mu,\sigma}(x) = 0, \qquad \text{for all } x \text{ and all } \mu \text{ such that } x \neq \mu \tag{16c}$$

(16a) and (16b) holding simply because we must have $f(x) \to 0$ as $x \to \pm\infty$ in order for f to be integrable, but (16c) needing verification in each case because it requires $f(x) = o(1/|x|)$ as $x \to \pm\infty$, which is typical behavior for PDF but not necessary. If all three of these hold, we see that the subprobability densities that are equal to zero almost everywhere are the appropriate densities to assign to all points added to the original parameter space in making the compactification.

We run into trouble verifying the integrability condition. Consider verifying it at a parameter point $(\mu, 0)$ in the compactification. For convenience we take the supremum over a box-shaped neighborhood $B = (\mu - \eta, \mu + \eta) \times [0, \delta)$. Let μ_0 and σ_0 be the true unknown parameter values. Then

$$\sup_{(\mu^*,\sigma^*) \in B} \log \frac{f_{\mu^*,\sigma^*}(x)}{f_{\mu_0,\sigma_0}(x)} = \infty, \qquad \mu - \eta < x < \mu + \eta,$$

and this cannot be integrable for all μ no matter how small we take η to be (so long as $\eta > 0$).

The trick due to Kiefer and Wolfowitz (1956) is to simply consider IID pairs (X_1, X_2) having marginal distributions (15). Because this distribution is continuous we have $X_1 \neq X_2$ almost surely, and we may (or may not) have

$$E_{\mu_0,\sigma_0} \sup_{(\mu^*,\sigma^*) \in B} \log \frac{f_{\mu^*,\sigma^*}(X_1) f_{\mu^*,\sigma^*}(X_2)}{f_{\mu_0,\sigma_0}(X_1) f_{\mu_0,\sigma_0}(X_2)} < \infty \tag{17}$$

(the Wald integrability condition applied to the distribution of IID pairs), and if (17) does hold, we conclude that

$$(\hat{\mu}_{2n}, \hat{\sigma}_{2n}) \xrightarrow{\text{a.s.}} (\mu_0, \sigma_0),$$

where $\hat{\mu}_n$ and $\hat{\sigma}_n$ are the AMLE's for sample size n, by direct application of the Wald theorem.

This seems to say only that the even-index terms of the sequence of estimators converge, but there is a simple argument that says that for every $\eta > 0$ we still have (9), from which we conclude AMLE are strongly consistent. The argument involving pairs of data points says that (9) holds with $l_n(\theta)$ replaced by $l_{2n}(\theta)$, but also we can write

$$l_{2n+1}(\theta) = \frac{1}{2n+1} \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} + \frac{1}{2n+1} \sum_{i=2}^{2n+1} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}$$

hence

$$\begin{aligned} &\sup_{\theta \in K_\eta} l_{2n+1}(\theta) \\ &\qquad \le \frac{1}{2n+1} \sup_{\theta \in K_\eta} \log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} + \sup_{\theta \in K_\eta} \frac{1}{2n+1} \sum_{i=2}^{2n+1} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \qquad (18) \end{aligned}$$

by the upper semicontinuity assumption $\log[f_\theta(X_1)/f_{\theta_0}(X_1)]$ achieves its supremum over K_η and this supremum, being a value of the function, is finite. Hence the first term on the right-hand side of (18) converges to zero, not just almost surely, but for all ω. The second term on the right-hand side of (18) has the same distribution as $2n/(2n+1)$ times the left-hand side of (9) with n replaced by $2n$, hence it has $\limsup$ equal to the right-hand side of (9). If we have (9) when the $\limsup$ is taken over even numbered terms and when it is taken over odd numbered terms, then we have it when it is taken over all terms.

References

Bahadur, R. R. (1971). *Some Limit Theorems in Statistics.* Philadelphia: SIAM.

Bertsekas, D. P. and Shreve S. E. (1978). *Stochastic Optimal Control: The Discrete Time Case.* New York: Academic Press.

Browder, A. (1996). *Mathematical Analysis: An Introduction.* New York: Springer.

DeGroot, M. H. and Schervish, M. J. (2002). *Probability and Statistics*, third edition. Redding, MA: Addison-Wesley.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 261–274.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, **13**, 795–800.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27**, 887–906.

Perleman, M. D. (1972). On the strong consistency of approximate maximum likelihood estimates. *Proceedings of the Sixth Berkley Symposium on Mathematical Statistics and Probability*, **1**, 263–281, University of California Press.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Annals of Statistics*, **9**, 225–228.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.

Wang, J.-L. (1985). Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics. *Annals of Statistics*, **13**, 932–946.