On the Convergence of Monte Carlo Maximum Likelihood Calculations

By
Charles J. Geyer[1].
Technical Report No. 571
School of Statistics
University of Minnesota
February 18, 1992
Revised September 24, 1992

[1]Research supported in part by grant DMS-9007833 from the National Science Foundation

Abstract

Monte Carlo maximum likelihood for normalized families of distributions (Geyer and Thompson, 1992) can be used for an extremely broad class of models. Given any family $\{ h_\theta : \theta \in \Theta \}$ of nonnegative integrable functions, maximum likelihood estimates in the family obtained by normalizing the the functions to integrate to one can be approximated by Monte Carlo, the only regularity conditions being a compactification of the parameter space such that the the evaluation maps $\theta \mapsto h_\theta(x)$ remain continuous. Then with probability one the Monte Carlo approximant to the log likelihood hypoconverges to the exact log likelihood, its maximizer converges to the exact maximum likelihood estimate, approximations to profile likelihoods hypoconverge to the exact profile, and level sets of the approximate likelihood (support regions) converge to the exact sets (in Painlevé-Kuratowski set convergence). The same results hold when there are missing data (Thompson and Guo, 1991, Gelfand and Carlin, 1991) if a Wald-type integrability condition is satisfied. Asymptotic normality of the Monte Carlo error and convergence of the Monte Carlo approximation to the observed Fisher information are also shown.

1 Monte Carlo Maximum Likelihood

1.1 Normalized Families of Densities

Suppose we have a family of nonnegative functions $\{ h_\theta : \theta \in \Theta \}$ on a probability space, all of which are integrable with respect to a measure μ and none integrating to zero. Let the integrals be denoted $c(\theta) = \int h_\theta \, d\mu$. Then for each θ in Θ the function f_θ defined by

$$f_\theta(x) = \frac{1}{c(\theta)} h_\theta(x)$$

is a probability density with respect to μ. We we call a family $\{ f_\theta : \theta \in \Theta \}$ of this form a *normalized family* of densities. The function $\theta \mapsto c(\theta)$ is the *normalizer* of the family, and the functions h_θ are the *unnormalized densities* of the family. We denote the distribution corresponding to θ by P_θ and expectation with respect to P_θ by E_θ, i. e. $P_\theta(A) = \int_A f_\theta \, d\mu$ and $E_\theta g(X) = \int g f_\theta \, d\mu$.

Normalized families are interesting because they include the important special cases of exponential families and Gibbs distributions and the conditional families arising in conditional likelihood inference (Geyer and Thompson, 1992). They also have two important mathematical properties. For arbitrary functions h_θ realizations X_1, X_2, ... from P_θ can be simulated without knowledge of the normalizer $c(\theta)$ by the Metropolis-Hastings algorithm (Metropolis, et al., 1953; Hastings, 1970). Moreover, maximum likelihood estimation can be carried out, again without knowledge of the normalizer or its derivatives, using these Monte Carlo simulations (Geyer and Thompson, 1992). Somewhat surprisingly, since there is so little mathematical structure to work with, Monte Carlo maximum likelihood converges for any such family under continuity of the maps $\theta \mapsto h_\theta(x)$.

The log likelihood corresponding to an observation x we take for convenience to be the log likelihood ratio against an arbitrary fixed parameter point ψ

$$l(\theta) = \log \frac{h_\theta(x)}{h_\psi(x)} - \log \frac{c(\theta)}{c(\psi)} = \log \frac{h_\theta(x)}{h_\psi(x)} - \log E_\psi \frac{h_\theta(X)}{h_\psi(X)} \tag{1}$$

since

$$E_\psi \frac{h_\theta(X)}{h_\psi(X)} = \int \frac{h_\theta(x)}{h_\psi(x)} f_\psi(x) \, d\mu(x) = \frac{1}{c(\psi)} \int h_\theta(x) \, d\mu(x) = \frac{c(\theta)}{c(\psi)}. \tag{2}$$

Although the notation suggests that ψ is a point in the parameter space of interest, this is not necessary. h_ψ can be any nonnegative integrable function such that for any $\theta \in \Theta$, if $h_\psi(x) = 0$ then $h_\theta(x) = 0$ except perhaps for x in a null set that may depend on θ. This domination condition is necessary so that the set of points where $h_\psi(x) = 0$ can be ignored in the integrals in (2). Similar domination condition conditions will be assumed without explicit statement throughout the paper.

Given a sample X_1, ..., X_n from P_ψ generated by the Metropolis-Hastings algorithm, the natural Monte Carlo approximation of the log likelihood is

$$l_n(\theta) = \log \frac{h_\theta(x)}{h_\psi(x)} - \log E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)} \tag{3}$$

where $E_{n,\psi}$ denotes the 'empirical' expectation with respect to P_ψ defined by

$$E_{n,\psi} g(X) = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

If the Markov chain $X_1, X_2, \ldots$ generated by the Metropolis-Hastings algorithm is irreducible, then $E_{n,\psi} g(X)$ converges almost surely to $E_\psi g(X)$ for any integrable function g. In particular, $l_n(\theta)$ converges almost surely to $l(\theta)$, for any fixed θ. The 'almost surely' here means for almost all sample paths of the Monte Carlo simulation; the observation x is considered fixed. Note that the nullset of sample paths for which convergence fails may depend on θ.

Let $\hat{\theta}$ be the maximizer of l and let $\hat{\theta}_n$ be a maximizer of l_n. Geyer and Thompson (1992) show that if the normalized family is an exponential family, then $\hat{\theta}_n$ converges to $\hat{\theta}$ almost surely. They remark that an analogous result should hold outside of exponential families. Section 2 gives such a theorem.

1.2 Missing Data

A similar but subtly different application of Monte Carlo maximum likelihood occurs with missing data (Thompson and Guo, 1991), which includes ordinary (non-Bayes) empirical Bayes as a special case. If $f_\theta(x, y)$ is the joint density with x missing and y observed, then the normalizing constant for the conditional distribution of x given y is the likelihood $f_\theta(y)$. Again for convenience we use the likelihood ratio against the fixed parameter point ψ, then the log likelihood is

$$l(\theta) = \log \frac{f_\theta(y)}{f_\psi(y)} = \log E_\psi \left\{ \frac{f_\theta(X, Y)}{f_\psi(X, Y)} \,\middle|\, Y = y \right\} \tag{4}$$

The natural Monte Carlo approximation of the log likelihood is now

$$l_n(\theta) = \log E_{n,\psi} \left\{ \frac{f_\theta(X, Y)}{f_\psi(X, Y)} \,\middle|\, Y = y \right\} = \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{f_\theta(X_i, y)}{f_\psi(X_i, y)} \right) \tag{5}$$

where $X_1, X_2, \ldots$ are realizations from the conditional distribution of X given $Y = y$, typically simulated using the Metropolis-Hastings algorithm when the normalizing constant $f_\theta(y)$ is unknown.

The subtle difference between (3) and (5) relates not to the conditioning—in either case we need to simulate from a density known up to a constant of proportionality—but to the minus sign in (3). To get convergence results, we need to bound l_n uniformly from above on neighborhoods, so in (5) the Monte Carlo average must be bounded *above*, whereas in (3) the average must (because of the minus sign) be bounded *below*. The former requires an integrability assumption like that imposed by Wald (1949) to obtain consistency of maximum likelihood; the latter does not.

1.3 Missing Data in Normalized Families

A generalization that includes both of the preceding cases has been proposed by Gelfand and Carlin (1991) for estimation in normalizing constant families with missing data. Now the unnormalized densities are $h_\theta(x, y)$ with x missing and y observed. Then the log likelihood, obtained by integrating over the missing data, is

$$l(\theta) = \log E_\psi\left(\frac{h_\theta(X,Y)}{h_\psi(X,Y)} \,\middle|\, Y = y\right) - \log E_\psi \frac{h_\theta(X,Y)}{h_\psi(X,Y)} \tag{6}$$

and its natural Monte Carlo approximation is

$$\begin{aligned} l_n(\theta) &= \log E_{n,\psi}\left(\frac{h_\theta(X,Y)}{h_\psi(X,Y)} \,\middle|\, Y = y\right) - \log E_{n,\psi} \frac{h_\theta(X,Y)}{h_\psi(X,Y)} \\ &= \log\left(\frac{1}{n}\sum_{i=1}^n \frac{h_\theta(X_i^\star, y)}{h_\psi(X_i^\star, y)}\right) - \log\left(\frac{1}{n}\sum_{j=1}^n \frac{h_\theta(X_j, Y_j)}{h_\psi(X_j, Y_j)}\right) \end{aligned} \tag{7}$$

where $X_1^\star$, $X_2^\star$, ... are samples from the conditional distribution of X given $Y = y$ and (X_1, Y_1), (X_2, Y_2), ... are samples from the unconditional distribution (both for the parameter value ψ). Gelfand and Carlin suggest maximizing (7) to obtain an approximation to the MLE. As in the simple missing data problems of the preceding section, a Wald-type integrability condition seems to be required to assure convergence.

This double sampling is necessary only when the first term in (6) cannot be calculated exactly. When it can be, it is better to do so (Geyer et al., 1993). Then the situation is the same as in Section 1.1. No Wald-type condition is needed for convergence.

2 Likelihood Convergence

2.1 Hypoconvergence of the Monte Carlo Likelihood

Our treatment of the convergence of Monte Carlo likelihood for normalized families begins with a proof that the Monte Carlo log likelihood (3) hypoconverges to the exact log likelihood (1). Hypoconvergence is a type of convergence of functions that is useful in optimization theory (essentially a one-sided locally uniform convergence). The basics of the theory are given in the appendix (or the reader may just take equations 8a and 8b as a definition).

Theorem 1 *For a normalized family of densities (Section 1.1), if the parameter set Θ is a separable metric space (e. g., $\mathbb{R}^d$), if the evaluation maps $\theta \mapsto h_\theta(x)$ are*

(a) *lower semicontinuous at each θ except for x in a P_ψ nullset that may depend on θ,*

(b) *upper semicontinuous for the observed x and for x not in a P_ψ nullset (that does not depend on θ),*

and if the Metropolis-Hastings algorithm is irreducible, then the Monte Carlo log likelihood (3) hypoconverges to the exact log likelihood (1) with probability one. Also the exact log likelihood is upper semicontinuous and the normalizer of the family is lower semicontinuous.

PROOF. What is to be shown is that $l \le \text{h-}\liminf_n l_n \le \text{h-}\limsup_n l_n \le l$ which from (22) in the appendix is equivalent to

$$l(\theta) \le \inf_{B \in \mathcal{N}(\theta)} \liminf_{n\to\infty} \sup_{\varphi \in B} l_n(\varphi) \tag{8a}$$

$$l(\theta) \ge \inf_{B \in \mathcal{N}(\theta)} \limsup_{n\to\infty} \sup_{\varphi \in B} l_n(\varphi) \tag{8b}$$

where $\mathcal{N}(\theta)$ denotes the set of neighborhoods of the point θ.

By assumption there is a countable base $\mathcal{B} = \{B_1, B_2, \ldots\}$ for the topology of Θ. For any point θ, let $\mathcal{N}_c(\theta) = \mathcal{B} \cap \mathcal{N}(\theta)$. Note that the infima over the uncountable set $\mathcal{N}(\theta)$ in (8) can be replaced by infima over the countable set $\mathcal{N}_c(\theta)$. Choose a countable dense subset $\Theta_c = \{\theta_1, \theta_2, \ldots\}$ as follows. For each n let θ_n be a point of B_n satisfying

$$l(\theta_n) \ge \sup_{\varphi \in B_n} l(\varphi) - \frac{1}{n}$$

We will need

$$\lim_{n\to\infty} E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)} = E_\psi \frac{h_\theta(X)}{h_\psi(X)} = \frac{c(\theta)}{c(\psi)} \tag{9}$$

and

$$\lim_{n\to\infty} E_{n,\psi} \inf_{\varphi \in B} \frac{h_\varphi(X)}{h_\psi(X)} = E_\psi \inf_{\varphi \in B} \frac{h_\varphi(X)}{h_\psi(X)} \tag{10}$$

to hold simultaneously for all $\theta \in \Theta_c$ and all $B \in \mathcal{B}$. This follows from the irreducibility assumption, since the union of a countable number of nullsets (one exception set for each limit) is still a nullset. The infima in (10) are measurable because of assumption (b) in the theorem.

First we tackle (8a). If $B \in \mathcal{B}$ and $\theta \in B \cap \Theta_c$

$$l(\theta) = \lim_{n\to\infty} l_n(\theta) \le \liminf_{n\to\infty} \sup_{\varphi \in B} l_n(\varphi)$$

by (9). So

$$\sup_{\varphi \in B \cap \Theta_c} l(\varphi) \le \liminf_{n\to\infty} \sup_{\varphi \in B} l_n(\varphi)$$

and

$$\inf_{B \in \mathcal{N}_c(\theta)} \sup_{\varphi \in B \cap \Theta_c} l(\varphi) \le \inf_{B \in \mathcal{N}_c(\theta)} \liminf_{n\to\infty} \sup_{\varphi \in B} l_n(\varphi)$$

The left hand side is equal to $l(\theta)$ if l is upper semicontinuous by the construction of Θ_c. Hence upper semicontinuity of l implies (8a). Since $\theta \mapsto h_\theta(x)$ is upper semicontinuous and since a sum of upper semicontinuous functions is upper semicontinuous, it remains only to be shown that $-\log[c(\theta)/c(\psi)]$ is upper semicontinuous, which

is true if the normalizer $c(\theta)$ is lower semicontinuous, which follows from Fatou's lemma and the lower semicontinuity of $\theta \mapsto h_\theta(X)$: if $\theta_k \to \theta$

$$c(\theta) \le \int \left(\liminf_{k\to\infty} h_{\theta_k}(x) \right) d\mu(x) \le \liminf_{k\to\infty} \int h_{\theta_k}(x)\, d\mu(x) = \liminf_{k\to\infty} c(\theta_k)$$

This establishes (8a) and the assertions about upper and lower semicontinuity of the log likelihood and the normalizer.

Now

$$\begin{aligned}
\inf_{B\in\mathcal{N}_c(\theta)} \limsup_{n\to\infty} \sup_{\varphi\in B} l_n(\varphi) &\le \inf_{B\in\mathcal{N}_c(\theta)} \left(\sup_{\varphi\in B} \log \frac{h_\varphi(x)}{h_\psi(x)} - \log \liminf_{n\to\infty} E_{n,\psi} \inf_{\varphi\in B} \frac{h_\varphi(X)}{h_\psi(X)} \right) \\
&= \log \frac{h_\theta(x)}{h_\psi(x)} - \log \sup_{B\in\mathcal{N}_c(\theta)} \lim_{n\to\infty} E_{n,\psi} \inf_{\varphi\in B} \frac{h_\varphi(X)}{h_\psi(X)} \\
&= \log \frac{h_\theta(x)}{h_\psi(x)} - \log \sup_{B\in\mathcal{N}_c(\theta)} E_\psi \inf_{\varphi\in B} \frac{h_\varphi(X)}{h_\psi(X)}
\end{aligned}$$

where the inequality follows from the continuity and monotonicity of the logarithm function and because of superadditivity of the supremum operation (and subadditivity of the infimum operation), and the equalities follow from the upper semicontinuity of $\theta \mapsto h_\theta(x)$ and from (10). The limit will be equal to $l(\theta)$ and establish (8b) if

$$\sup_{B\in\mathcal{N}_c(\theta)} E_\psi \inf_{\varphi\in B} \frac{h_\varphi(X)}{h_\psi(X)} = \frac{c(\theta)}{c(\psi)}$$

Now the integrand here satisfies

$$0 \le \inf_{\varphi\in B} \frac{h_\varphi(x)}{h_\psi(x)} \le \frac{h_\theta(x)}{h_\psi(x)}, \qquad \forall\ x \tag{11}$$

(since $\theta \in B$). Since the right hand side is integrable by (2) and the evaluation maps are assumed lower semicontinuous, dominated convergence implies

$$\sup_{B\in\mathcal{N}_c(\theta)} E_\psi \inf_{\varphi\in B} \frac{h_\varphi(X)}{h_\psi(X)} \to E_\psi \sup_{B\in\mathcal{N}_c(\theta)} \inf_{\varphi\in B} \frac{h_\varphi(X)}{h_\psi(X)} = E_\psi \frac{h_\theta(X)}{h_\psi(X)} = \frac{c(\theta)}{c(\psi)} \tag{12}$$

This completes the proof. □

If we attempt to apply the program of the preceding section to either of the missing data models (Sections 1.2 and 1.3), we find it doesn't work without additional assumptions. To get a theorem we impose a Wald-type integrability condition following Wald (1949).

Theorem 2 *For the simple missing data problem (Section 1.2), if Θ is a separable metric space and evaluation maps $\theta \mapsto f_\theta(x, y)$ are*

(a) *upper semicontinuous at each θ except for x in a $P_\psi(X|Y = y)$ nullset that may depend on θ,*

(b) *lower semicontinuous except for x in a $P_\psi(X|Y = y)$ nullset (that does not depend on θ),*

if the Metropolis-Hastings algorithm is irreducible, and if for every in $\theta \in \Theta$ there is a neighborhood B of θ such that

$$E_\psi\left(\sup_{\varphi \in B} \frac{f_\varphi(X,Y)}{f_\psi(X,Y)} \,\middle|\, Y = y\right) < \infty \tag{13}$$

then the Monte Carlo log likelihood (5) hypoconverges to the exact log likelihood (4) with probability one. Also the exact log likelihood is continuous.

If the evaluation maps are actually continuous except for x in a $P_\psi(X|Y = y)$ nullset (that does not depend on θ), then then (5) also epiconverges to (4) with probability one.

PROOF. The argument establishing (8a) remains the same except for the invocation of Fatou's lemma. Now dominated convergence is used to prove $l(\theta_k) \to l(\theta)$, the dominating function being provided by (13), and this gives continuity of l rather than just lower semicontinuity. The argument establishing (8b) remains the same, except that infima become suprema and vice versa (because of the change of sign of the random term) and (11) must be replaced by (13) in justifying dominated convergence.

If the evaluation maps are almost surely continuous, then the argument in (11) and (12) is still valid and proves

$$l(\theta) \le \sup_{B \in \mathcal{N}(\theta)} \liminf_{n \to \infty} \inf_{\varphi \in B} l_n(\varphi)$$

which together with (8b) implies epiconvergence. □

REMARK. Simultaneous hypo- and epiconvergence is equivalent to continuous convergence, i. e., $\theta_n \to \theta$ implies $l_n(\theta_n) \to l(\theta)$. In a locally compact space (e. g. $\mathbb{R}^d$) it is also equivalent to continuity of l plus convergence of l_n to l uniformly on compact sets (Rockafellar and Wets, forthcoming, Theorem 3D.7).

Theorem 3 *For a missing data problem in a normalized family (Section 1.3) if the evaluation maps $\theta \mapsto h_\theta(x,y)$ are*

(a) *lower semicontinuous at each θ except for (x,y) in a P_ψ nullset that may depend on θ,*

(b) *upper semicontinuous except for (x,y) in a P_ψ nullset (that does not depend on θ),*

(c) *continuous for the observed y and for x not in a $P_\psi(X|Y = y)$ nullset (that does not depend on θ),*

if the Metropolis-Hastings algorithm is irreducible, and if (13) holds with f_θ replaced by h_θ, then the Monte Carlo log likelihood (7) hypoconverges to the exact log likelihood (6) with probability one. Also the exact log likelihood is upper semicontinuous.

PROOF. This is just a combination of the two preceding proofs. The proof of Theorem 1 shows that the second term in the log likelihood (7) hypoconverges, and the proof of Theorem 2 shows that the first term simultaneously epiconverges and hypoconverges. The sum thus hypoconverges (see the proof of Theorem 2.15 in Attouch, 1984). □

2.2 Convergence of the MLE Calculation

Theorem 4 *If $l_n \xrightarrow{h} l$ with probability one, if a sequence $\{\hat{\theta}_n\}$ satisfies*

$$l_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} l_n(\theta) - \epsilon_n$$

with $\epsilon_n \to 0$, and if $\{\hat{\theta}_n\}$ is contained in a compact set almost surely (resp. in probability), and if there is a unique maximum likelihood estimate $\hat{\theta}$, then $\hat{\theta}_n \to \hat{\theta}$ and $l_n(\hat{\theta}_n) \to l(\hat{\theta})$ almost surely (resp. in probability).

PROOF. The assertion about almost sure convergence follows directly from the theorem and Proposition 1 in the appendix. If $\{\hat{\theta}_n\}$ is contained in a compact set, then every subsequence has a convergent subsubsequence, and each such subsubsequence must converge to $\hat{\theta}$. Hence the whole sequence converges to $\hat{\theta}$. Moreover, the optimal values must converge as well.

The assertion about convergence in probability follows by almost the same argument. A sequence bounded in probability is tight, hence every subsequence has a subsubsequence which converges in distribution by Prohorov's theorem. By Skorohod representation, the convergence can be considered almost sure, in which case the only possible limit is $\hat{\theta}$. Hence the whole sequence and the optimal values converge in distribution to point masses at $\hat{\theta}$ and $l(\hat{\theta})$ (which is the same as convergence in probability). □

The theorem applies trivially when the whole parameter space Θ is a compact set. This is the usual way in which proofs of this sort proceed, following Wald (1949), who used the one-point compactification, Kiefer and Wolfowitz, (1956), who used more general compactifications, and Bahadur (1971), who gives a very general formulation, showing that most models are compactifiable in the appropriate topology (the one induced by vague convergence of the associated probability measures). Lacking a suitable compactification, it would be necessary to establish a uniform bound on the estimator by ad hoc methods.

2.3 Convergence of Profile Likelihoods

Suppose g is a continuous mapping from the original parameter space Θ to a new parameter space Φ (both metric spaces). The *profile likelihood* is the function on Φ defined by

$$l_p(\phi) = \sup_{\theta \in g^{-1}(\phi)} l(\theta).$$

Theorem 5 *If the Monte Carlo log likelihood hypoconverges to the exact log likelihood, and the parameter space Θ is compact, then the Monte Carlo profile log likelihood hypoconverges to the exact profile log likelihood.*

PROOF. What is to be established is the analogue of (8) with l and l_n replaced by l_p and $l_{p,n}$. For (8a) we may assume $l_p(\phi) > -\infty$. Then for any $R < l_p(\phi)$ there is a $\theta \in g^{-1}(\phi)$ such that

$$R \leq l(\theta) \leq \inf_{B \in \mathcal{N}(\theta)} \liminf_{n \to \infty} \sup_{\eta \in B} l_n(\eta) \leq \inf_{B \in \mathcal{N}(\phi)} \liminf_{n \to \infty} \sup_{\eta \in g^{-1}(B)} l_n(\eta)$$

where the second inequality is just (8a) and the third inequality is true because the infimum is over a smaller set, each $g^{-1}(B)$ being a neighborhood of θ. Since right hand side is h-$\liminf_n l_{p,n}$, this establishes the analogue of (8a).

For (8b) we may assume $l_p(\phi) < +\infty$. Hence for every $\epsilon > 0$ and $\theta \in g^{-1}(\phi)$, there is by (8b) a neighborhood $B_\epsilon(\theta)$ of θ such that

$$l(\theta) + \epsilon \geq \limsup_{n \to \infty} \sup_{\eta \in B_\epsilon(\theta)} l_n(\eta).$$

By the compactness assumption there are $\theta_1, \ldots, \theta_m$ such that $W = \bigcup_{i=1}^m B_\epsilon(\theta_i)$ covers $g^{-1}(\phi)$. Also by compactness there is a neighborhood B of ϕ, such that $g^{-1}(B) \subset W$. Then

$$\limsup_{n \to \infty} \sup_{\eta \in g^{-1}(B)} l_n(\eta) \leq \limsup_{n \to \infty} \sup_{\eta \in W} l_n(\eta) \leq \sup_{i=1,\ldots,m} l(\theta_i) + \epsilon \leq l_p(\phi) + \epsilon.$$

This establishes the analog of (8b). □

2.4 Convergence of Level Sets

Hypoconvergence also implies Painlevé-Kuratowski set convergence (Appendix A.1) for level sets of the of the log likelihood $\operatorname{lev}_\alpha l = \{\, \theta : l(\theta) \geq \alpha \,\}$ which are used in forming likelihood-based interval estimates (called *support regions* in Edwards, 1972).

We may look either at a fixed level α or at a fixed distance γ down from the maximum. The latter case makes no sense unless $l_n(\hat{\theta}_n) \to \sup l$, which need not happen, though it must under the assumptions for Theorem 4.

Theorem 6 *If the $l_n \xrightarrow{h} l$, then*

$$\begin{aligned} \limsup_n \operatorname{lev}_\alpha l_n &\subset \operatorname{lev}_\alpha l \\ \liminf_n \operatorname{lev}_\alpha l_n &\supset \operatorname{lev}_\beta l, \qquad \beta > \alpha, \end{aligned}$$

and if

$$\operatorname{cl}\left(\bigcup_{\beta > \alpha} \operatorname{lev}_\beta l \right) = \operatorname{lev}_\alpha l, \tag{14}$$

also holds, then

$$\lim_n \operatorname{lev}_\alpha l_n = \operatorname{lev}_\alpha l. \tag{15}$$

If, in addition, $l_n(\hat\theta_n) \to \sup l$, *then*

$$\begin{aligned} \limsup_n \operatorname{lev}_{l_n(\hat\theta_n)-\gamma} l_n &\subset \operatorname{lev}_{\sup l-\gamma} l \\ \liminf_n \operatorname{lev}_{l_n(\hat\theta_n)-\gamma} l_n &\supset \operatorname{lev}_{\sup l-\delta} l, \qquad \delta < \gamma, \end{aligned}$$

and if (14) also holds for $\alpha = \sup l - \gamma$, *then*

$$\lim_n \operatorname{lev}_{l_n(\hat\theta_n)-\gamma} l_n = \operatorname{lev}_{\sup l-\gamma} l \tag{16}$$

PROOF. The assertions about limits inferior and superior are direct consequences of Theorem 3.1 in Beer, et al. (1992), which says that $\limsup_n \operatorname{lev}_{\alpha_n} l_n \subset \operatorname{lev}_\alpha l$ holds for every sequence $\alpha_n \to \alpha$ and $\liminf_n \operatorname{lev}_{\alpha_n} l_n \supset \operatorname{lev}_\alpha l$ holds for some sequence $\alpha_n \to \alpha$. The assertions about limits follow from the nesting of level sets and the fact that set limits are closed ($\operatorname{lev}_\alpha l$ is closed because a hypo-limit is always upper semicontinuous). □

Before leaving the subject of likelihood convergence it is perhaps worth pausing for a moment and comparing the results obtained here with the results that are obtainable for the exponential family case (Geyer, 1990, Geyer and Thompson, 1992). There the log likelihood and its Monte Carlo approximation are concave, and this has several consequences that improve the preceding results. First, if the exact log likelihood has a unique maximizer, the boundedness assumptions of Theorem 4 can be dropped, because then a hypoconvergent sequence of concave functions is *equi-level-bounded* (eventually dominated by a function with compact level sets) (Rockafellar and Wets, forthcoming, Propositions 3C.21 and 3C.22). For the same reason the compactness assumption in Theorem 5 can be dropped. Finally, (14) is automatically true for any level below the maximum (Rockafellar, 1970, Theorem 7.6). So (15) and (16) hold for $\alpha < \sup l$.

3 Asymptotic Normality

Asymptotic normality of $\sqrt{n}(\hat\theta_n - \hat\theta)$ is very similar to the asymptotics of maximum likelihood.

Theorem 7 *Suppose the following assumptions hold*

(a) The MLE $\hat\theta$ *is unique and the parameter space* Θ *contains an open neighborhood of* $\hat\theta$ *in* $\mathbb{R}^d$.

(b) The Monte Carlo MLE $\hat\theta_n$ *converges in probability to* $\hat\theta$.

(c) $c(\theta) = \int h_\theta \, d\mu$ *can be differentiated twice under the integral sign.*

(d) $\sqrt{n}\nabla l_n(\hat\theta) \xrightarrow{\mathcal{L}} N(0, A)$ *for some covariance matrix* A.

(e) $B = -\nabla^2 l(\hat{\theta})$ *is positive definite.*

(f) $\nabla^3 l_n(\theta)$ *is bounded in probability uniformly in a neighborhood of* $\hat{\theta}$.

then

$$-\nabla^2 l_n(\hat{\theta}_n) \to B, \quad \textit{in probability} \tag{17}$$

and

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}) \xrightarrow{\mathcal{L}} N(0, B^{-1}AB^{-1}) \tag{18}$$

A proof would be entirely classical and is omitted.

All of the conditions except (d) are fairly straightforward, and one can imagine verifying them (if they hold) by standard methods. Condition (e) can be verified using dominated convergence and ergodicity if an integrable function can be found that dominates third partial derivatives with respect to theta of h_θ / h_ψ.

The conclusion (17) is particularly interesting, since it gives an estimate of the observed Fisher information, which may be of interest aside from its use in (18). This point has also been made by Gelfand and Carlin (1991), Guo and Thompson (1992), and several discussants of Geyer and Thompson (1992).

Condition (d) is hard, if Markov chain Monte Carlo is being used for the simulations, because it involves a Markov chain central limit theorem. General Markov chain central limit theorems do exist (Nummelin, 1984; Kipnis and Varadhan, 1986), but can be difficult to apply in practice, except when the state space is finite and the CLT is automatic (Chung, 1967, p. 99 ff.) The Kipnis-Varadahn theorem is the simplest for general state spaces, requiring only reversibility and summability of the autocovariances. A Metropolis-Hastings algorithm can always be arranged so that the Markov chain is reversible, a point attributed to P. Green in Besag (1986), but the summability condition is difficult. For related work in the specific context of Markov chain Monte Carlo see Shervish and Carlin (1992), Chan (1993), Liu et al. (1991), Tierney (1991), and Geyer (1993).

Assuming that (d) holds, the variance A typically cannot be calculated theoretically and must be estimated by Monte Carlo.

$$\nabla l_n(\theta) = \frac{\nabla h_\theta(x)}{h_\theta(x)} - \frac{E_{n,\psi} \frac{\nabla h_\theta(X)}{h_\psi(X)}}{E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)}} = \frac{E_{n,\psi} \left[\Big(t_\theta(x) - t_\theta(X) \Big) \frac{h_\theta(X)}{h_\psi(X)} \right]}{E_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)}} \tag{19}$$

where $t_\theta(X) = \nabla h_\theta(X)/h_\theta(X)$. Using assumption (c) to differentiate under the integral sign

$$\begin{aligned} \nabla l(\theta) &= \frac{\nabla h_\theta(x)}{h_\theta(x)} - \frac{\nabla c(\theta)}{c(\theta)} \\ &= \frac{\nabla h_\theta(x)}{h_\theta(x)} - \int \frac{\nabla h_\theta(x)}{h_\theta(x)} \frac{h_\theta(x)}{c(\theta)} \, d\mu(x) \\ &= t_\theta(x) - E_\theta t_\theta(X), \end{aligned}$$

and this is zero when $\theta = \hat{\theta}$. The denominator in (19) converges to $c(\theta)/c(\psi)$; the expectation of the numerator with respect to P_ψ is

$$\begin{aligned} E_\psi \left\{ \Big(t_\theta(x) - t_\theta(X)\Big) \frac{h_\theta(X)}{h_\psi(X)} \right\} &= \frac{c(\theta)}{c(\psi)} \int \Big(t_\theta(x) - t_\theta(y)\Big) f_\theta(y)\, d\mu(y) \\ &= \frac{c(\theta)}{c(\psi)} \Big(t_\theta(x) - E_\theta t_\theta(X)\Big), \end{aligned}$$

which is also zero when $\theta = \hat{\theta}$. Thus the numerator is the sample mean for a functional of the Markov chain

$$z_\theta(X) = \Big(t_\theta(x) - t_\theta(X)\Big) \frac{h_\theta(X)}{h_\psi(X)}$$

which has expectation zero under the stationary distribution. Hence by the continuous mapping theorem

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_\theta(X_i) \xrightarrow{\mathcal{L}} \frac{c(\psi)}{c(\theta)} N(0, A)$$

Let $\gamma(t) = \gamma(-t)$ be the lag t autocovariance of $z_\theta(X_i)$ at stationarity, i. e.

$$\gamma(t) = \operatorname{Cov}\left(z_\theta(X_0), z_\theta(X_t)\right)$$

when the starting position X_0 of the Markov chain is a realization from P_ψ, then for reversible chains (Kipnis and Varadhan, 1986)

$$A = \frac{c(\theta)^2}{c(\psi)^2} \sum_{t=-\infty}^{+\infty} \gamma(t) \tag{20}$$

Both factors in (20) can be estimated, $c(\theta)/c(\psi)$ by the denominator in (19), and the sum by standard time series methods (for a review see Geyer, 1993; see also Hastings, 1970; Geweke, 1992; Han, 1991; and Green and Han, 1992).

4 Discussion

'Normalized families of densities' are an important class of statistical models. We now have two interesting properties that hold for the whole class. The Metropolis-Hastings algorithm can be used to simulate realizations from any distribution in the model, and Monte Carlo likelihood approximation can be used to do likelihood-based statistical inference. When there are no missing data, mere continuity is enough to guarantee convergence. With missing data, Wald-type integrability conditions are required. This class is extremely flexible, allowing a very wide scope for modeling and supporting the notion of a 'model liberation movement' called for by Professor A. F. M. Smith in his discussion of Geyer and Thompson (1992).

Monte Carlo likelihood may be useful even in missing data problems where where the EM algorithm can be used to calculate the MLE, since the Monte Carlo approximates the whole likelihood surface. The use of (17) to approximate the observed Fisher information may be useful in problems where analytical methods (Sundberg, 1974; Louis, 1982) are intractable. It is especially useful in conjunction with Monte Carlo EM (Tanner and Wei, 1990; Guo and Thompson, 1992), but may also be a competitor for the SEM algorithm (Meng and Rubin, 1991).

Acknowledgements

Conversations with Elizabeth Thompson, Julian Besag, and Michael Newton helped change my focus from exponential families to the general 'normalized families' of Section 1. The whole approach to convergence of optimization problems used in this paper comes from a course taught by Terry Rockafellar in 1990 at the University of Washington using a draft of the book (Rockafellar and Wets, forthcoming). Roger Wets provided the reference to Beer et al. (1992), and suggested the approach used in proving Theorem 5. Xiaotong Shen found a mistake in my first proof of Theorem 1.

A Appendix

A.1 Set Convergence

At several points the concept of Painlevé-Kuratowski set convergence (Sec 1.4.1 in Attouch, 1984) was needed. Given a sequence of sets C_n, the set limit superior is the set

$$\limsup_{n\to\infty} C_n = \bigcap_{m=1}^{\infty} \operatorname{cl}\left(\bigcup_{n=m}^{\infty} C_n\right)$$

and the limit inferior is the set

$$\liminf_{n\to\infty} C_n = \operatorname{cl}\left(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \operatorname{cl} C_n\right)$$

Note that these are topological convergence notions, different from the set theoretic notions commonly used in probability theory (defined by the same formulas without the closure operations). In a metric space the following definitions are equivalent to the preceding ones (Proposition 1.34 in Attouch, 1984). The set limit superior is the set of points x such that there is a subsequence $x_{n_k} \to x$ with $x_{n_k} \in C_{n_k}$, and the set limit inferior is the set of points x such that there is a sequence $x_n \to x$ with $x_n \in C_n$ for all n after some n_0. In short, the limit superior is the set of cluster points and the limit inferior is the set of limit points. If the set limits superior and inferior agree, then their common value is said to be the limit of the sequence.

A.2 Epiconvergence and Hypoconvergence

Epiconvergence and hypoconvergence are types of convergence of sequences of functions that are useful in optimization problems. If a sequence of functions g_n epiconverges to a limit g (written $g_n \xrightarrow{e} g$) and x_n minimizes g_n then any cluster point of the sequence $\{x_n\}$ is a minimizer of g. Hypoconvergence is the analogous notion for maximization problems. Since x maximizes g if and only if it minimizes $-g$, hypoconvergence (written $g_n \xrightarrow{h} g$) is defined by $g_n \xrightarrow{h} g$ if and only if $(-g_n) \xrightarrow{e} (-g)$.

Epiconvergence is related to set convergence in the following way. The *epigraph* of a function an extended-real-valued ($\pm\infty$ allowed) function g with domain S is the set

$$\operatorname{epi} g = \{\, (x, \lambda) \in S \times \mathbb{R} : g(x) \le \lambda \,\}$$

of points lying on or above the graph. A sequence of functions g_n *epiconverges* to a function g if and only if the sequence of sets $\operatorname{epi} g_n$ converges to the set $\operatorname{epi} g$.

There are several equivalent characterizations that are sometimes more useful. Given a sequence of functions g_n, the epi-limits inferior and superior are the functions (Attouch, 1984, p. 26)

$$(\text{e-}\liminf_n g_n)(x) = \sup_{B \in \mathcal{N}(x)} \liminf_{n \to \infty} \inf_{y \in B} g_n(y) \tag{21a}$$

$$(\text{e-}\limsup_n g_n)(x) = \sup_{B \in \mathcal{N}(x)} \limsup_{n \to \infty} \inf_{y \in B} g_n(y) \tag{21b}$$

where $\mathcal{N}(x)$ denotes the set of neighborhoods of the point x. The sequence g_n epiconverges to a function $\text{e-}\lim_n g_n = g$ if and only if the the epi-limits inferior and superior agree and are equal to g. Similar notation with the prefix e- replaced by h- is used for hypoconvergence.

$$(\text{h-}\liminf_n g_n)(x) = \inf_{B \in \mathcal{N}(x)} \liminf_{n \to \infty} \sup_{y \in B} g_n(y) \tag{22a}$$

$$(\text{h-}\limsup_n g_n)(x) = \inf_{B \in \mathcal{N}(x)} \limsup_{n \to \infty} \sup_{y \in B} g_n(y) \tag{22b}$$

Another pair of conditions that are equivalent for functions on metric spaces are the following (Attouch, 1984, p. 30). A sequence of functions g_n epiconverges to a function g if the following two conditions hold at every point x

(a) $\liminf_n g_n(x_n) \ge g(x)$ for every sequence $x_n \to x$.

(b) $\limsup_n g_n(x_n) \le g(x)$ for some sequence $x_n \to x$.

This says that epiconvergence is a combination of one-sided locally uniform convergence (Condition (a)), with something weaker than pointwise convergence from the other side (Condition (b)).

The main reason for the importance of epiconvergence is the following proposition, which is Theorem 1.10 in Attouch (1984).

Proposition 1 *Suppose* $g_n \xrightarrow{e} g$, $x_n \to x$ *and* $g_n(x_n) - \inf g_n \to 0$ *then*

$$g(x) = \inf g = \lim_{n \to \infty} g_n(x_n).$$

That is, if x_n is an ϵ_n-minimizer of g_n with $\epsilon_n \to 0$, then any convergent subsequence of $\{x_n\}$ must converge to a point x which minimizes g and the optimal values $g_n(x_n)$ must also converge to the asymptotic optimal value $g(x)$. Two points are worth comment here. First, there is no requirement that the minimizers be unique. If g has a unique minimizer x, then x is the only cluster point of the sequence $\{x_n\}$. Otherwise, there may be many cluster points, but all of them must minimize g. Second, the proposition does not rule out escape to infinity; it only describes what happens if $x_n \to x$. It does say that if the sequence $\{x_n\}$ is confined to a compact set and if g has a unique minimizer, then x_n converges to that minimizer.

References

Attouch, H. (1984) *Variational Convergence of Functions and Operators.* Boston: Pitman.

Bahadur, R. R. (1958) Examples of inconsistency of maximum likelihood estimates. *Sankhyā*, **20**, 207–210.

—— (1971) *Some Limit Theorems in Statistics.* Philadelphia: SIAM.

Beer, G., Rockafellar, R. T. and Wets, R. J.-B. (1992) A characterization of epi-convergence in terms of convergence of level sets. *Proc. Amer. Math. Soc.* To appear.

Besag, J. (1986) On the statistical analysis of dirty pictures (with discussion). *J. R. Statist. Soc.* B, **48**, 259–302.

Chan, K. S. (1993). Asymptotic behavior of the Gibbs sampler. *J. Am. Statist. Ass.* To appear.

Chung, K. L. (1967) *Markov Chains with Stationary Transition Probabilities*, second edition. Berlin: Springer-Verlag.

Edwards, A. W. F. (1972) *Likelihood.* Cambridge: Cambridge University Press.

Gelfand, A. E. and Carlin, B. P. (1991) Maximum likelihood estimation for constrained or missing data models. Research Report 91–002, Division of Biostatistics, University of Minnesota.

Gelfand, A. E. and Smith A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, ed. J. Bernado, Oxford University Press, ????

Geyer, C. J. (1990) *Likelihood and Exponential Families.* Ph. D. Dissertation, University of Washington.

Geyer, C. J. (1991) Reweighting Monte Carlo mixtures. Technical Report No. 568, School of Statistics, University of Minnesota.

Geyer, C. J. (1993) Practical Markov chain Monte Carlo (with discussion). *Statistical Science*. To appear.

Geyer, C. J., Ryder, O. A., Chemnick, L. G. and Thompson, E. A. (1993) Analysis of relatedness in the California condors from DNA fingerprints. *Mol. Biol. Evol.* To appear.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc. B*, **54**, 657–699.

Green, P. J. and Han, X.-L. (1992). Metropolis methods, Gaussian proposals, and antithetic variables. *Lect. Notes. Statist.* **74** 142–164. Berlin: Springer-Verlag.

Guo, S. W. and Thompson, E. A. (1992) Monte Carlo estimation of mixed models for large complex pedigrees. In Technical Report No. 229, Department of Statistics, University of Washington.

Han, X.-L. (1991). Spectral window estimation of integrated autocorrelation time. Research Report, University of Bristol.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

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

Kipnis, C. and Varadhan, S. R. S. (1986) Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Commun. Math. Phys.*, **104**, 1–19.

Liu, J., Wong, W. H., and Kong, A. (1991) Correlation structure and convergence rate of the Gibbs sampler with various scans. Technical Report No. 304, Department of Statistics, University of Chicago.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B*, **44**, 226–233.

Meng, X.-L. and Rubin, D. B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Ass.*, **86**, 899–909.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Nummelin, E. (1984) *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge: Cambridge University Press.

Priestly, M. B. (1981) *Spectral Analysis and Time Series*. London: Academic Press.

Rockafellar, R. T. (1970) *Convex Analysis*. Princeton: Princeton University Press.

Rockafellar, R. T. and Wets, R. J. B. (forthcoming) *Variational Analysis*. New York: Springer-Verlag.

Sheehan, N. and Thomas, A. (1991) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics*. To appear.

Shervish, M. J. and Carlin, B. P. (1992) On the convergence rate of successive substitution sampling. *J. Comp. Graphical Statist.* To appear.

Sundberg, R. (1974) Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.*, **1**, 49–58.

Tanner, M. A. and Wei, G. C. G. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Ass.*, **85**, 699–704.

Thompson, E. A. and Guo, S. W. (1991) Evaluation of likelihood ratios for complex genetic models.

Tierney, L. (1991) Markov chains for exploring posterior distributions. Technical Report No. 560, School of Statistics, University of Minnesota.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 595–601.