On the strong consistency of asymptotic M-estimators

Djalil Chafaï and Didier Concordet

Preprint. September 2006.
Accepted for publication in Journal of Statistical Planning and Inference.

Abstract

The aim of this article is to simplify Pfanzagl's proof of consistency for asymptotic maximum likelihood estimators, and to extend it to more general asymptotic M-estimators. The method relies on the existence of a sort of contraction of the parameter space which admits the true parameter as a fixed point. The proofs are short and elementary.

1 Introduction

After the seminal work[1] of Fisher, the asymptotic properties of maximum likelihood estimators, and in particular their consistency, were studied by various authors, including Doob [Doo34], Cramér [Cra46], and Huzurbazar [Huz48]. Nowadays, one of the best known result regarding consistency goes back to Wald, who gave in [Wal49] a short and elegant proof of strong consistency of parametric maximum likelihood estimators. Since that time, several authors studied various versions of such consistency problems, including among others, Le Cam [LC53], Kiefer and Wolfowitz [KW56], Bahadur [Bah67, Bah71], Huber [Hub67], Perlman [Per72], Wang [Wan85], and Pfanzagl [Pfa88, Pfa90].

Wald's original proof relies roughly on local compactness of the parameter space, on continuity and coercivity[2] of the log-likelihood, on the law of large numbers, and last but not least on local uniform integrability of the log-likelihood. It does not require differentiability, and makes extensive use of likelihood ratios. The integrability assumption has been weakened by many authors, including for instance Kiefer and Wolfowitz in [KW56] and Perlman in [Per72], see also [Bah71]. One can find a modern presentation of Wald's method for M-estimators in van der Vaart's monograph [vdV98].

Pfanzagl gave in [Pfa88, Pfa90] a proof of strong consistency of asymptotic maximum likelihood estimators for nonparametric "concave models" with respect to the

[1] The interested reader may find a quite recent account in [Ald97] and references therein.

[2] By coercivity we mean that the log-likelihood tends to $-\infty$ when the parameter tends to ∞.

estimated parameter, including nonparametric mixtures. His approach relies in particular on a simplification of an earlier work of Wang in [Wan85] based on uniform local bound of the likelihood ratio.

The present work was initially motivated by the inverse problems considered in [CL06]. Our aim is to simplify Pfanzagl's approach, and to extend the framework from asymptotic maximum likelihood to more general asymptotic M-estimators. In particular, log-likelihood ratios are replaced by contrast differences. The hypotheses appearing in our main Theorem are unnecessarily strong. However, they allow a simple and short presentation. We emphasize the role played by a sort of contraction map a^* defined on the parameter space. We do not assume any coercivity of the contrast as in [Wal49]. However, we require the compactness of the space of the estimated parameter, as in [KW56] and [vdV98] for example. This compactness comes usually for free in the case of fully nonparametric models. We do not make use of any Uniform Law of Large Numbers. Our method does not belong to the Glivenko-Cantelli approaches of consistency, as in [Dud98], [Fio00], [AK94], [vdV98] and [vdG03, vdG00] and references therein.

Let Θ be a separable Hausdorff topological space with countable base. Let $(P_\theta)_{\theta\in\Theta}$ be a known family of Borel measures on a measurable space $\mathcal{X}$. Let $\theta^* \in \Theta$ be some unknown point of Θ such that $P^* := P_{\theta^*}$ is a probability measure. Let $(X_n)_{n\in\mathbb{N}}$ be an i.i.d. sequence of observed random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking their values in $\mathcal{X}$, with common law P^*. Let $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ be a sequence of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, taking their values in Θ, and such that $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ is $\mathcal{F}_n$-measurable for any $n \in \mathbb{N}$, where $\mathcal{F}_n := \sigma(X_0, \ldots, X_n)$. We say that $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ is *strongly consistent* if and only if

$$\mathbb{P} - \text{a.s.} \quad \lim_{n\to+\infty} \widehat{\theta}_n = \theta^*. \tag{1}$$

We use in the sequel the abbreviations "a.s." for *almost sure*, "a.a." for *almost all*, and "a.e." for *almost everywhere*. Let $\Theta \times \mathcal{X} \ni (\theta, x) \mapsto m(\theta, x) \in \mathbb{R}$ be a known function such that $m_\theta := m(\theta, \cdot)$ is measurable for any $\theta \in \Theta$. For any n, we define the random function $M_n : \Theta \to \mathbb{R}$ by

$$M_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} m(\theta, X_i).$$

This can be written also $M_n(\theta) = \mathbb{P}_n m_\theta$ where $\mathbb{P}_n := \frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n})$ is the empirical measure. We say that $(\widehat{\theta}_n)_n$ is a *sequence of asymptotic M-estimators* if and only if

$$\mathbb{P} - \text{a.s.} \quad \overline{\lim_{n\to+\infty}} \left(\sup_{\Theta} M_n - M_n(\widehat{\theta}_n) \right) = 0. \tag{2}$$

The term *asymptotic* is used for the same notion (with the likelihood) by Pfanzagl in [Pfa88]. In the literature, some authors, including Wald and Perlman, use the term *approximate* rather than *asymptotic*. However, the term *approximate* has been used by Bahadur in a different sense in [Bah71, page 34].

For example, if for large enough n, there exists an $\mathcal{F}_n$-measurable $\widehat{\theta}_n$ in Θ such that $M_n(\widehat{\theta}_n) = \sup_\Theta M_n$, then such a random sequence $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ fulfils (2).

For any probability measure P on $\mathcal{X}$, let $\mathrm{L}^1_+(\mathcal{X}, P)$ (resp. $\mathrm{L}^1_-(\mathcal{X}, P)$) be the set of random variables $Z : \mathcal{X} \to \mathbb{R}$ such that $Z^+ := \max(+Z, 0)$ (resp. $Z^- := \max(-Z, 0)$) is in $\mathrm{L}^1(\mathcal{X}, P)$. On $E(\mathcal{X}, P) := \mathrm{L}^1_-(\mathcal{X}, P) \cup \mathrm{L}^1_+(\mathcal{X}, P)$, the expectation $P(Z) = P(Z^+) - P(Z^-)$ makes sense and takes its values in $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. For any $\theta \in \Theta$ such that $m_\theta \in E(\mathcal{X}, P^*)$, we define the *contrast* $M^*(\theta) \in \overline{\mathbb{R}}$ by

$$M^*(\theta) := P^* m_\theta. \tag{3}$$

In the sequel, we say that the model is *identifiable* when for any $\theta \in \Theta$, the condition $P_\theta = P^*$ implies that $\theta = \theta^*$.

Example 1.1 (Log-Likelihood). *Assume that for some fixed Borel measure Q on $\mathcal{X}$, one has $P_\theta \ll Q$ for any $\theta \in \Theta$. Let $f_\theta := dP_\theta/dQ$ and assume that $f_\theta > 0$ on $\mathcal{X}$ for any $\theta \in \Theta$. Define $m(\theta, x) := \log(f_\theta(x))$. Then $M_n : \Theta \to \mathbb{R}$ is the* log-likelihood *random functional given by $M_n(\theta) = \mathbb{P}_n m_\theta = \mathbb{P}_n \log(f_\theta)$. We will speak about sequences of "asymptotic maximum likelihood estimators". The log-likelihood ratio is*

$$M_n(\theta_1) - M_n(\theta_2) = \mathbb{P}_n \log(f_{\theta_1}/f_{\theta_2}).$$

As usual for the log-likelihood, when $M^*(\theta^*)$ is finite, one can write for any θ

$$M^*(\theta) - M^*(\theta^*) = -\mathbf{Ent}\left(P_{\theta^*} \,|\, P_\theta\right),$$

where $\mathbf{Ent}(P_{\theta_1} \,|\, P_{\theta_2})$ is the Kullback-Leibler relative entropy of P_{θ_1} with respect to P_{θ_2}. In particular, $M^*(\theta) \le M^*(\theta^*)$ with equality if and only if $P_\theta = P_{\theta^*}$, which implies $\theta = \theta^*$ if the model is identifiable. Notice that when Q is the Lebesgue measure on $\mathcal{X} = \mathbb{R}^n$, then $-M^*(\theta^*) = -\int_{\mathcal{X}} f_{\theta^*}(x) \log(f_{\theta^*}(x))\, dx$ is the Shannon entropy of f_{θ^*}.

Example 1.2 (Beyond the log-likelihood). *Assume that for some fixed Borel measure Q on $\mathcal{X}$, one has $P_\theta \ll Q$ for any $\theta \in \Theta$, with $P_\theta(\mathcal{X}) \le 1$ and $f_\theta := dP_\theta/dQ$. Let $\Phi, \Psi : (0, +\infty) \to \mathbb{R}$ be two smooth functions. Assume that $\Psi(f_\theta) \in \mathrm{L}^1(\mathcal{X}, Q)$ for any $\theta \in \Theta$. Define m_θ by*

$$m_\theta = \Phi(f_\theta) - \int_{\mathcal{X}} \Psi(f_\theta)\, dQ + P_\theta(\mathcal{X}).$$

This gives rise the the following empirical contrast

$$M_n(\theta) = \mathbb{P}_n(\Phi(f_\theta)) - \int_{\mathcal{X}} \Psi(f_\theta)\, dQ + P_\theta(\mathcal{X}).$$

In particular, if $\theta \in \Theta$ is such that $\Phi(f_\theta) \in \mathrm{L}^1(\mathcal{X}, P^*)$ where here again $P^* := P_{\theta^*}$,

$$M^*(\theta) = P^*(\Phi(f_\theta)) - \int_{\mathcal{X}} \Psi(f_\theta)\, dQ + P_\theta(\mathcal{X}).$$

Assume now that $u \mapsto u\Phi'(u)$ is locally integrable on $\mathbb{R}_+$, and consider the case where Ψ is the Φ-transform given for any $u \in (0, +\infty)$ by

$$\Psi(u) = \int_0^u v\Phi'(v)\, dv.$$

For $\Phi : u \mapsto \log(u)$, one has $\Psi : u \mapsto u$ and we recover the log-likelihood contrast

$$M^*(\theta) = P^*(\log(f_\theta)).$$

For $\Phi : u \mapsto u$, one has $\Psi : u \mapsto \frac{1}{2}u^2$, and we get the quadratic contrast

$$M^*(\theta) = -\tfrac{1}{2}\|f_\theta - f_{\theta^*}\|^2_{\mathrm{L}^2(\mathcal{X},Q)} + \tfrac{1}{2}\|f_{\theta^*}\|^2_{\mathrm{L}^2(\mathcal{X},Q)} + P_\theta(\mathcal{X}).$$

In both cases, the map $\theta \mapsto M^*(\theta)$ admits θ^* as unique maximum provided that the model is identifiable. More generally, define the Φ-transform $\Theta : (0, +\infty)^2 \to \mathbb{R}$ by

$$\begin{aligned}\Theta(u, v) : &= u\Phi(v) - \Psi(v)\\ &= u\Phi(v) - \int_0^v w\Phi'(w)\, dw.\end{aligned}$$

When θ and θ^* are such that both $\Theta(f_{\theta^*}, f_{\theta^*})$ and $\Theta(f_{\theta^*}, f_\theta)$ belong to $\mathrm{L}^1(\mathcal{X}, Q)$,

$$M^*(\theta) = \int_{\mathcal{X}} (\Theta(f_{\theta^*}, f_\theta) - \Theta(f_{\theta^*}, f_{\theta^*}))\, dQ + \int_{\mathcal{X}} \Theta(f_{\theta^*}, f_{\theta^*})\, dQ + P_\theta(\mathcal{X}).$$

Notice that Θ is linear in Φ. One can consider useful examples for which the function Φ is bounded, in such a way that m_θ is bounded for any $\theta \in \Theta$. For instance, let us examine the case where $\Phi : u \mapsto -(1+u)^{-2}$. Then, $\Psi : u \mapsto -u^2(1+u)^{-2}$, and the map $\theta \mapsto M^*(\theta)$ admits θ^* as unique maximum, provided identifiability holds, since for any $(u, v) \in \mathbb{R}^2_+$,

$$\Theta(u, v) = -\frac{u + v^2}{(1+v)^2} \quad \textit{and} \quad \Theta(u, v) - \Theta(u, u) = -\frac{(v-u)^2}{(1+u)(1+v)^2}.$$

The function Ψ is additionally bounded here. The similar case $\Phi : u \mapsto -(1+u^2)^{-1}$ is also quite interesting. Notice that $\Theta(u, \cdot)$ is concave on $(0, +\infty)$ as soon as Φ is concave, non decreasing, with $\Phi'(v) + v\Phi''(v) \geq 0$ for any $v > 0$. Observe that this is not the approach of Pfanzagl in [Pfa90], which is more related to the log-likelihood ratio. Notice that in the case of the log-likelihood, one has $\Phi : u \mapsto \log(u)$, which gives $\Psi : u \mapsto u$ and $\Theta : (u, v) \mapsto -u\log(v) - v$, and thus $\Theta(u, v) - \Theta(u, u) = u\log(u/v) + u - v$. It might be possible to extensively study such "Φ-estimators", in the spirit of the "Φ-calculus" developed in [Cha04, Cha06]. This is however outside the scope of this short article.

One can notice that the observation of Lindsay in [Lin83a, Lin83b] regarding the nature of maximum likelihood for nonparametric mixture models remains valid for more general models provided that m is concave.

2 Main result and Corollaries

With the settings given in the Introduction, the following Theorem holds.

Theorem 2.1. *Assume that Θ is compact and that the following assumptions hold.*

(A1) *For P^*-a.a. $x \in \mathcal{X}$, the map $m(\cdot, x)$ is continuous on Θ;*

(A2) *There exists a continuous map $a^* : \Theta \to \Theta$ which may depend on θ^* such that for any $\theta \neq \theta^*$, there exists a neighborhood $V \subset \Theta$ of θ for which $\sup_V (m - m_{a^*}) \in \mathrm{L}^1_+(\mathcal{X}, P^*)$ and $P^*(m_\theta - m_{a^*(\theta)}) < 0$.*

Then any sequence $(\widehat{\theta}_n)_n$ of asymptotic M-estimators is strongly consistent.

Proof. Postponed to section 4. □

The quantity $P^*(m_\theta - m_{a^*(\theta)})$ in **(A2)** has a meaning in $\overline{\mathbb{R}}$ since the first part of **(A2)** ensures that $m_\theta - m_{a^*(\theta)} \in \mathrm{L}^1_+(\mathcal{X}, P^*)$. Moreover, $P^*(m_\theta - m_{a^*(\theta)})$ reads $M^*(\theta) - M^*(a^*(\theta))$ when the couple $(m_\theta, m_{a^*(\theta)})$ is in $\mathrm{L}^1_-(\mathcal{X}, P^*) \times \mathrm{L}^1_+(\mathcal{X}, P^*)$ or in $\mathrm{L}^1_+(\mathcal{X}, P^*) \times \mathrm{L}^1_-(\mathcal{X}, P^*)$.

Since θ^* is unknown in practice, each assumption in Theorem 2.1 must hold for any $\theta^* \in \Theta$ such that P_{θ^*} is a probability measure, in order to make the result useful.

Remark 2.2 (Assumptions). *The first part of **(A2)** is in a way an M-estimator version of the integrability condition considered by Kiefer and Wolfowitz for the log-likelihood in [KW56]. The assumptions **(A1)** and **(A2)** required by Theorem 2.1 can be weakened. However, they permit a streamlined presentation. In particular, only lower semi-continuity is needed in **(A1)**, see for instance [Pfa88]. Additionally, and following for example [Per72, page 266], the uniform integrability assumption **(A2)** can be weakened, by considering blocks of $k > 1$ observations instead of one observation, see also [vdV98, comments following Theorem 5.14].*

As stated in the following Corollary, Theorem 2.1 implies a version of Wald consistency Theorem for asymptotic M-estimators, see [Wal49], [Per72, Section 2 page 269], and [vdV98, Theorem 5.14].

Corollary 2.3 (Perlman-Wald). *Assume that Θ is compact, and that for P^*-a.a. $x \in \mathcal{X}$, the map $m(\cdot, x)$ is continuous on Θ. Assume that for any θ in Θ, there exists a neighborhood V such that $\sup_V m \in \mathrm{L}^1(\mathcal{X}, P^*)$. Assume in addition that M^* achieves its supremum over Θ at θ^*, and only at θ^*. Then, any sequence of asymptotic M-estimators is strongly consistent.*

Proof. One has $m_\theta \in \mathrm{L}^1(\mathcal{X}, P^*)$ for any θ in Θ, and thus $M^* : \Theta \to \mathbb{R}$ is well defined. Moreover, **(A2)** holds with a constant map $a^* \equiv \theta^*$. Namely, for any $\theta \neq \theta^*$, one has on one hand $P^*(m_\theta - m_{\theta^*}) < 0$ since $M^*(\theta) < M^*(\theta^*)$, and on the other hand

$$\sup_V (m - m_{a^*}) = -m_{\theta^*} + \sup_V m \in \mathrm{L}^1(\mathcal{X}, P^*).$$

□

As stated in the following Corollary, Theorem 2.1 implies the main result of Pfanzagl in [Pfa88] for concave models, itself based on an earlier result of Wang in [Wan85]. This is typically the case for mixtures models, for which Θ is a convex set of probability measures on some measurable space, cf. section 3.

Corollary 2.4 (Pfanzagl-Wang). *Let Q be a reference Borel measure on $\mathcal{X}$. Consider the case where Θ is a convex compact subset of a linear space such that for any $\theta \in \Theta$, $P_\theta(\mathcal{X}) \leq 1$ and $P_\theta \ll Q$ with $f_\theta := dP_\theta/dQ > 0$ on $\mathcal{X}$. Suppose that Q-a.e. on $\mathcal{X}$, the map $\theta \mapsto f_\theta(x)$ is concave and continuous on Θ. Assume that the model is identifiable. Consider $m_\theta := \log(f_\theta)$ and the related log-likelihood M_n. Then any sequence of asymptotic log-likelihood estimators is strongly consistent.*

Proof. First of all, we notice that it is not possible to take $a^* \equiv \theta^*$ since we cannot ensure that the condition $m_{\theta^*} - m_\theta = \log(f_{\theta^*}/f_\theta) \in \mathrm{L}^1_+(\mathcal{X}, P^*)$ of **(A2)** is true. However, the concavity of the model allows to take a map a^* which is a strict contraction around θ^*. Namely, for an arbitrary $\lambda \in (0,1)$, let us take

$$a^*(\theta) := \lambda\theta^* + (1-\lambda)\theta.$$

The concavity of the model yields

$$m_{a^*(\theta)} - m_\theta = \log\left(\frac{f_{\lambda\theta^*+(1-\lambda)\theta}}{f_\theta}\right) \geq \log\left(\frac{\lambda f_{\theta^*} + (1-\lambda) f_\theta}{f_\theta}\right) \geq \log(1-\lambda).$$

Now, we have $\log(1-\lambda) \in \mathrm{L}^1(\mathcal{X}, P^*)$ since $\lambda < 1$. Define the function $\Phi : \mathbb{R}_+ \to \mathbb{R}$ by $\Phi(u) := u\log(\lambda u + (1-\lambda))$. The concavity of the model yields

$$P^*(m_{a^*(\theta)} - m_\theta) \geq \int_{\mathcal{X}} f_{\theta^*} \log\left(\frac{\lambda f_{\theta^*} + (1-\lambda) f_\theta}{f_\theta}\right) dQ = \int_{\mathcal{X}} \Phi\left(\frac{f_{\theta^*}}{f_\theta}\right) f_\theta \, dQ.$$

Let us show that the right hand side of the inequality above is strictly positive when $\theta \neq \theta^*$. One has $P_\theta(\mathcal{X}) > 0$ since $f_\theta > 0$. Define $\Psi(u) := u\Phi(1/u)$. Jensen's inequality for the probability measure $P_\theta(\mathcal{X})^{-1} P_\theta$ and the convex function Φ yields

$$\int_{\mathcal{X}} \Phi\left(\frac{f_{\theta^*}}{f_\theta}\right) f_\theta \, dQ \geq \Psi(P_\theta(\mathcal{X})). \tag{4}$$

It is enough to show that either (4) is strict or the right hand side of (4) is strictly positive. Since $\lambda > 0$, the function Φ is strictly convex. Thus equality holds in (4) if and only if $P_\theta(f_{\theta^*} = \alpha f_\theta) = 1$ for some $\alpha \in \mathbb{R}_+$. The only admissible case is $\alpha = P_\theta(\mathcal{X})^{-1} > 1$ since $P_{\theta^*}(\mathcal{X}) = 1$ and since identifiability forbids $P_\theta(f_{\theta^*} = f_\theta) = 1$. Therefore, if $P_\theta(\mathcal{X}) = 1$, inequality (4) is necessarily strict. On the other hand, $\Psi(1) = 0$ and $\Psi(u) > 0$ when $u < 1$. Thus the right hand side of (4) is always non negative, and is strictly positive as soon as $P_\theta(\mathcal{X}) < 1$. We conclude that $P^*(m_{a^*(\theta)} - m_\theta) > 0$ as soon as $\theta \neq \theta^*$. This shows that **(A2)** holds with $V = \Theta$, and the proof is thus complete. □

Remark 2.5 (About the map a^*). *Let $a^* : \Theta \to \Theta$ be a map which satisfies the condition $P^*(m_\theta - m_{a^*(\theta)}) < 0$ for any $\theta \neq \theta^*$ of **(A2)**. Then, the impossibility of $P^*(m_\theta - m_\theta) < 0$ for any θ yields that*

- *$a^*(\theta) \neq \theta$ for any $\theta \neq \theta^*$. In particular,*
  - *the map a^* cannot be the identity map ;*
  - *if a^* is constant, then $a^* \equiv \theta^*$;*
  - *the point θ^* is the only possible fixed point for a^*.*

*The proof of Corollary 2.3 gives an example where $a^* \equiv \theta^*$ works and fulfills **(A2)**. In contrast, Corollary 2.4 provides a situation where a constant a^* does not fulfill **(A2)**. However, we have shown in the proof of Corollary 2.4 that an a^* map which is a strict contraction around θ^* fulfills **(A2)**. Actually, when Θ has the structure of a convex subset of a vector space, any strict contraction around θ^* fulfills the properties of a^* listed above. The existence of a fixed point can be related to Brouwer-like fixed point Theorems. For instance, any continuous mapping of a non-empty compact convex subset of $\mathbb{R}^d$ into itself contains at least one fixed point. Consequently, when Θ is a non-empty compact and convex subset of $\mathbb{R}^d$, any continuous a^* map admits θ^* as a unique fixed point. There exists numerous dimension free Brouwer-like fixed points theorems, due to Schauder, Tikhonov, Kakutani, ..., see for instance [Zei86] and [Goe02].*

Remark 2.6 (Infinite values of m). *Theorem 2.1 does not allow m to take the value $-\infty$. This limitation is due to the fact that differences of the form $m_\theta - m_{\theta'}$ do not make sense if m is allowed to take the value $-\infty$. The consistency proof of Wald does not suffer from such a limitation since it does not rely on m differences, but it requires however strong uniform integrability assumptions. A careful reading of the proof of Theorem 2.1 shows that only differences of the form $m_\theta - m_{a^*(\theta)}$ are involved. On the other hand, according to Remark 2.5, $a^*(\theta) \neq \theta$ for any $\theta \neq \theta^*$. Consequently, one may allow, in Theorem 2.1, the map $m(\theta, x)$ to take the value $-\infty$ for at most one value of θ. For the log-likelihood, $m_\theta = \log(f_\theta)$ and one has $m_\theta(x) = -\infty$ if and only if $f_\theta(x) = 0$. One may allow $f_\theta \equiv 0$ for at most one value of θ in Corollary 2.4.*

Remark 2.7. *Let $\theta \in \Theta$ such that $m_\theta \in E(\mathcal{X}, P^*)$. Then, the law of large numbers applies and gives that P^*-a.s., $\lim_n M_n(\theta) = M^*(\theta) \in \overline{\mathbb{R}}$, and the a.s. subset of $\mathcal{X}$ may depend on θ. In particular $M_n(\theta) = M^*(\theta) + o_P(1)$. For a sequence $(\widehat{\theta}_n)_n$ satisfying (2), one can write for any $\theta \in \Theta$ with finite $M_n(\theta)$*

$$\begin{aligned} M_n(\widehat{\theta}_n) &= M_n(\widehat{\theta}_n) - M_n(\theta) + M_n(\theta) \\ &\geq -\left(\sup_\Theta M_n - M_n(\widehat{\theta}_n) \right) + M_n(\theta) \\ &= o_P(1) + M(\theta) \end{aligned}$$

where the last step follows by (2) and the law of large numbers.

3 Log-Likelihood and mixtures models

For any topological space $\mathcal{Z}$ equipped with its Borel σ-field, we denote by $\mathcal{M}_1(\mathcal{Z})$ the set of probability measures on $\mathcal{Z}$, and by $\mathcal{C}_b(\mathcal{Z})$ the set of bounded real valued continuous functions on $\mathcal{Z}$. The Prohorov topology on $\mathcal{M}_1(\mathcal{Z})$ is defined as follows: $\theta_n \to \theta$ in $\mathcal{M}_1(\mathcal{Z})$ if and only if $\int_{\mathcal{Z}} f\, d\theta_n \to \int_{\mathcal{Z}} f\, d\theta$ for any $f \in \mathcal{C}_b(\mathcal{Z})$. It is known that a subset of $\mathcal{M}_1(\mathcal{Z})$ is compact if and only if it is tight. As a consequence, $\mathcal{M}_1(\mathcal{Z})$ is not compact in general. Following [Pfa88, section 5 page 149], the set sub-probabilities provides a compactification which allows the following consistency result for asymptotic log-likelihood estimators of nonparametric mixture models.

Corollary 3.1 (Pfanzagl). *Let $\mathcal{Z}$ be a locally compact Hausdorff topological space with countable base. Let Q be a measure on a measurable space $\mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{Z} \to (0, +\infty)$ be such that $\int k(x,z) dQ(x) = 1$ for any $z \in \mathcal{Z}$ and $k(x,\cdot) \in \mathcal{C}_b(\mathcal{Z})$ for any $x \in \mathcal{X}$. Let $\Theta := \mathcal{M}_1(\mathcal{Z})$ and consider the family $(P_\theta)_{\theta\in\Theta}$ of probability measures on $\mathcal{X}$ defined by $dP_\theta = f_\theta dQ$ with $f_\theta(x) := \int k(x,z)\, d\theta(z)$. Assume that the model is identifiable. Let $m : \Theta \times \mathcal{X} \to \mathbb{R}$ be the map defined by $m(\theta, x) := \log f_\theta(x)$, and M_n be the corresponding log-likelihood. Then any sequence of asymptotic maximum likelihood estimators is strongly consistent for the Prohorov topology.*

Proof. As explained above, $\Theta = \mathcal{M}_1(\mathcal{Z})$ is not compact for the Prohorov topology, and one must consider a suitable compactification, as in [Bah71] for instance. Let $\mathcal{C}_0(\mathcal{Z})$ be the set of real valued continuous functions on $\mathcal{Z}$ which vanish at infinity. Let $\overline{\Theta}$ be the set of Borel measures θ on $\mathcal{Z}$ such that $\theta(\mathcal{Z}) \leq 1$ (i.e. sub-probabilities), equipped with the vague topology related to $\mathcal{C}_0(\mathcal{Z})$. Namely, $\theta_n \to \theta$ in $\overline{\Theta}$ if and only if $\int_{\mathcal{Z}} f\, d\theta_n \to \int_{\mathcal{Z}} f\, d\theta$ for any $f \in \mathcal{C}_0(\mathcal{Z})$. The injection $\Theta \subset \overline{\Theta}$ is continuous; $\overline{\Theta}$ is a compact metrizable topological space, and thus has a countable base. Moreover, $\overline{\Theta}$ is convex, and for any $\theta \in \overline{\Theta}$, there exists $\theta' \in \Theta$ and $\alpha \in [0,1]$ such that $\theta = \alpha\theta'$.

We extend the set of probability measures $(P_\theta)_{\theta\in\Theta}$ on $\mathcal{X}$ to the set of sub-probability measures $(P_\theta)_{\theta\in\overline{\Theta}}$ on $\mathcal{X}$, where $dP_\theta = f_\theta dQ$ and $f_\theta(x) := \int k(x,z)\, d\theta(z)$. One has by virtue of Fubini-Tonelli Theorem that $P_\theta(\mathcal{X}) = \theta(\mathcal{Z})$, and thus $P_\theta \in \mathcal{M}_1(\mathcal{X})$ if and only if $\theta \in \Theta := \mathcal{M}_1(\mathcal{Z})$. Notice that θ^* is taken in Θ.

Let $\theta \in \overline{\Theta}$ such that $P_\theta = P_{\theta^*}$. Since θ^* is taken in Θ, one has that $P_\theta \in \mathcal{M}_1(\mathcal{X})$, therefore $\theta \in \Theta$ and thus $\theta = \theta^*$ by identifiability in Θ. Notice that $\overline{\Theta}$ is the convex envelope of $\Theta \cup \{0\}$. The set $\overline{\Theta}$ contains the null measure 0, for which $f_0 \equiv 0$ and thus $m_0 \equiv -\infty$. If $\theta \in \overline{\Theta}$ with $\theta \neq 0$, then $f_\theta > 0$ on $\mathcal{X}$ since $k > 0$, and thus $m_\theta(x) := \log f_\theta(x)$ is finite for any $x \in \mathcal{X}$. For any $x \in \mathcal{X}$, the map $\theta \in \overline{\Theta} \mapsto m_\theta(x)$ is continuous since $k(x,\cdot)$ is in $\mathcal{C}_0(\mathcal{Z})$.

For any $\theta \in \overline{\Theta}$ with $\theta \neq 0$, one can write $\theta = \alpha\theta'$ with $\theta' \in \Theta$ and $\alpha := \theta(\mathcal{Z}) \in [0,1]$. One has then $f_\theta = \alpha f_{\theta'}$ and thus $m_\theta = \log\alpha + m_{\theta'}$. Therefore,

$$M_n(\theta) = \log\alpha + M_n(\theta') \leq M_n(\theta').$$

As a consequence, $\sup_{\theta\in\Theta} M_n(\theta) = \sup_{\theta\in\overline{\Theta}} M_n(\theta)$, and one may substitute Θ by $\overline{\Theta}$ in the definition (2). Now, let $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ be a sequence in Θ of asymptotic maximum

likelihood estimators. Corollary 2.4 and Remark 2.6 for $(P_\theta)_{\theta\in\overline{\Theta}}$ apply and give the P^*-a.s. convergence for the vague topology of $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ towards θ^*. Since both the sequence and the limit are in Θ, the convergence holds for the Prohorov topology, and the desired result is established. □

Remark 3.2. *A mixture model can always be seen as a conditional model. The observed random variables X with values in $\mathcal{X}$ is the first component of the couple (X, Z) with values in $\mathcal{X} \times \mathcal{Z}$. The component Z is not observed. However, the conditional law $\mathcal{L}(X \mid Z = z)$ is known, and has density $k(\cdot, z)$ with respect to Q on $\mathcal{X}$. If $\theta = \mathcal{L}(Z)$, then $\mathcal{L}(X)$ has density f_θ with respect to Q on $\mathcal{X}$.*

4 Proof of main result

Lemma 4.1 (Reformulation). *The random sequence $(\widehat{\theta}_n)_n$ is a sequence of asymptotic M-estimators if and only if*

$$\mathbb{P}\text{–}a.s., \quad \forall(\theta_n)_n \in \Theta^{\mathbb{N}}, \quad \overline{\lim_{n\to+\infty}} \left(M_n(\theta_n) - M_n(\widehat{\theta}_n)\right) \leq 0. \tag{5}$$

Proof. The proof is done "ω by ω", and the a.s. sets in (2) and (5) are the same. Recall that $(\widehat{\theta}_n)_n$ is a sequence of asymptotic M-estimators if and only if (2) holds. Actually, the definition of the supremum gives $\sup_{\theta\in\Theta} M_n(\theta) - M_n(\widehat{\theta}_n) \geq 0$. Therefore, (2) is equivalent to

$$\mathbb{P}\text{–a.s.}, \quad \overline{\lim_{n\to+\infty}} \left(\sup_{\theta\in\Theta} M_n(\theta) - M_n(\widehat{\theta}_n)\right) \leq 0. \tag{6}$$

The Lemma is thus reduced to the equivalence between (6) and (5). We begin by the proof of the implication (6) $\Rightarrow$ (5). Let A be some $\mathbb{P}$–a.s. set such that (6) holds. We proceed by fixing $\omega \in A$. We hide the dependency on ω in the notation of M_n and $\widehat{\theta}_n$ to lightweight the expressions. Let $(\theta_n)_n$ be a sequence in Θ. By definition of the supremum, we have $M_n(\theta_n) \leq \sup_{\theta\in\Theta} M_n(\theta)$. Thus, we get

$$M_n(\theta_n) - M_n(\widehat{\theta}_n) \leq \sup_{\theta\in\Theta} M_n(\theta) - M_n(\widehat{\theta}_n).$$

Taking the $\overline{\lim}_{n\to+\infty}$ of both sides and using (6) provides the expected result (5). It remains to establish the implication (5) $\Rightarrow$ (6). Let A be some $\mathbb{P}$–a.s. set such that (5) holds. Here again, we proceed by fixing $\omega \in A$, and we hide the dependency on ω in the notation of the random objects like M_n and $\widehat{\theta}_n$. By definition of the supremum, there exists, for any n, an element $\theta_n \in \Theta$ such that

$$\sup_{\theta\in\Theta} M_n(\theta) - M_n(\theta_n) - \frac{1}{n} \leq 0.$$

Notice that θ_n depends on ω since M_n depends on ω. This yields

$$\overline{\lim_{n\to+\infty}} \left(\sup_{\theta\in\Theta} M_n(\theta) - M_n(\theta_n) \right) \leq 0. \tag{7}$$

Now we write the telescopic sum

$$\sup_{\theta\in\Theta} M_n(\theta) - M_n(\widehat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - M_n(\theta_n) + M_n(\theta_n) - M_n(\widehat{\theta}_n),$$

which gives

$$\begin{aligned}\overline{\lim_{n\to+\infty}} \left(\sup_{\theta\in\Theta} M_n(\theta) - M_n(\widehat{\theta}_n) \right) \\ \leq \overline{\lim_{n\to+\infty}} \left(\sup_{\theta\in\Theta} M_n(\theta) - M_n(\theta_n) \right) + \overline{\lim_{n\to+\infty}} \Big(M_n(\theta_n) - M_n(\widehat{\theta}_n) \Big).\end{aligned}$$

The two terms of the right hand side are "≤ 0" by virtue of (7) and (5) respectively. This provides the desired result (6), as expected. □

Lemma 4.2 (Separation). *Assume that $\mathbb{P}$–a.s., for any neighborhood U of θ^*, for any sequence $(\theta_n)_n$ in U^c, there exists a sequence $(\theta'_n)_n$ in Θ such that*

$$\underline{\lim_{n\to+\infty}} \left(M_n(\theta'_n) - M_n(\theta_n) \right) > 0. \tag{8}$$

Then, any asymptotic M-estimators sequence $(\widehat{\theta}_n)_n$ is strongly consistent.

Proof. Suppose that (8) holds for some a.s. set A, and that $(\widehat{\theta}_n)_n$ is a sequence of asymptotic M-estimators which is not strongly consistent. Saying that $(\widehat{\theta}_n)_n$ is not strongly consistent means that for any $\mathbb{P}$–a.s. set, there exists a neighborhood U of θ^* and a subsequence $(\widehat{\theta}_{n_k})_k$ in U^c. In particular, on the a.s. set A, this gives a neighborhood U of θ^* and a subsequence $(\widehat{\theta}_{n_k})_k$ in U^c. Now, by virtue of (8),

$$\mathbb{P}\text{–a.s.}, \quad \exists (\theta'_{n_k})_k \in \Theta^{\mathbb{N}}, \quad \underline{\lim_{k\to+\infty}} \Big(M_{n_k}(\theta'_{n_k}) - M_{n_k}(\widehat{\theta}_{n_k}) \Big) > 0,$$

where the a.s. set is A. This contradicts (5) which holds $\mathbb{P}$–a.s. too. □

Lemma 4.3 (The a^* map). *Assume that Θ is compact and that there exists a map $a^* : \Theta \to \Theta$ such that for any $\theta \neq \theta^*$, there exists a neighborhood U_θ of θ such that*

$$\mathbb{P}\text{–a.s.}, \quad \underline{\lim_{n\to+\infty}} \inf_{U_\theta} \left(M_n(a^*) - M_n \right) > 0. \tag{9}$$

Then, any asymptotic M-estimators sequence $(\widehat{\theta}_n)_n$ is strongly consistent.

Proof. Let us show that the assumptions of Lemma 4.2 are fulfilled. We will establish (8) for an a.s. set A which does not depend on the neighborhood U of θ^*. Namely, let U be an open neighborhood of θ^*. For any $\theta \in U^c$, let U_θ and A_θ be the neighborhood of θ and the $\mathbb{P}$–a.s. set for which (9) holds. Notice that A_θ depends on U_θ. The set $U^c \subset \cup_{\theta \in U^c} U_\theta$ is compact as a closed subset of the compact set Θ. We can thus extract a finite sub-covering $U^c \subset \cup_{i=1}^k U_{\theta_i}$, and write

$$\begin{aligned}\varliminf_n \inf_{U^c} (M_n(a^*) - M_n) &\geq \varliminf_n \min_{1\leq i\leq k} \inf_{U_{\theta_i}} (M_n(a^*) - M_n) \\ &= \min_{1\leq i\leq k} \varliminf_n \inf_{U_{\theta_i}} (M_n(a^*) - M_n).\end{aligned}$$

By virtue of (9) we get from the above that

$$\mathbb{P}\text{–a.s., } \varliminf_n \inf_{U^c} (M_n(a^*) - M_n) > 0, \tag{10}$$

where the $\mathbb{P}$–a.s. set is $A_U := \cap_{i=1}^k A_{\theta_i}$. Recall that U was a freely chosen neighborhood of θ^*. Consider now a countable base $(U_k)_k$ for θ^*. Then (10) holds on the $\mathbb{P}$–a.s. set $A := \cap_{i=1}^\infty A_{U_k}$, which does not depend on U. Notice at this step that

$$M_n(a^*(\theta_n)) - M_n(\theta_n) \geq \inf_{U^c} (M_n(a^*) - M_n)$$

as soon as $\theta_n \in U^c$ by definition of the infimum. This gives (8) from (10) on the $\mathbb{P}$–a.s. set A defined above, with $(\theta'_n)_{n\in\mathbb{N}} = (a^*(\theta_n))_{n\in\mathbb{N}}$. □

Proof of Theorem 2.1. The desired result follows from Lemma 4.3. Namely, let us show that (9) is a consequence of **(A1)** and **(A2)**. Let $\theta \neq \theta^*$ and let a^* and V as in **(A2)**. Let $V_k \searrow \{\theta\}$ be a decreasing local base with $V_0 \subset V$. Let $Z := \inf_V(m_{a^*} - m)$ and $Z_k := \inf_{V_k}(m_{a^*} - m)$ and $Z_\infty := m_{a^*(\theta)} - m_\theta$. By **(A1)** and the continuity of a^* and the separability of Θ, we get that $Z_k : \mathcal{X} \to \overline{\mathbb{R}}$ is measurable, and that

$$\mathbb{P}^*\text{–a.s., } Z \leq Z_k \nearrow Z_\infty.$$

Now, by **(A2)**, we get that $Z \in \mathrm{L}^1_-(\mathcal{X}, P^*)$ and $Z_\infty \in \mathrm{L}^1_-(\mathcal{X}, P^*)$ and $P^*(Z_\infty) > 0$. Observe that $Z \geq -Z^- \in \mathrm{L}^1(\mathcal{X}, P^*)$. Thus, by the monotone convergence Theorem,

$$\lim_k P^*(Z_k) = P^*(Z_\infty) > 0.$$

Therefore, $P^*(Z_k) > 0$ for some k (actually for k large enough). Let us denote $U_\theta := V_k$. Now, by the law of large numbers

$$\mathbb{P}\text{–a.s., } \lim_n \mathbb{P}_n\left(\inf_{U_\theta}(m_{a^*} - m)\right) = P^*\left(\inf_{U_\theta}(m_{a^*} - m)\right) > 0.$$

This gives finally (9) since for any n

$$\inf_{U_\theta}(M_n(a^*) - M_n) = \inf_{U_\theta} \mathbb{P}_n(m_{a^*} - m) \geq \mathbb{P}_n\left(\inf_{U_\theta}(m_{a^*} - m)\right).$$

□

Acknowledgements. The article benefited from the comments and criticism of the Advisory Editor and two anonymous referees. The authors would like also to sincerely thank Professor Jon A. Wellner who has kindly answered to their questions during his visit in Toulouse.

References

[AK94] M. Akahira et H. Kashima – "On the consistency of the maximum likelihood estimator through its uniform consistency", *Statistics* **25** (1994), no. 4, p. 333–341.

[Ald97] J. Aldrich – "R. A. Fisher and the making of maximum likelihood 1912–1922", *Statist. Sci.* **12** (1997), no. 3, p. 162–176.

[Bah67] R. R. Bahadur – "Rates of convergence of estimates and test statistics", *Ann. Math. Statist.* **38** (1967), p. 303–324.

[Bah71] — , *Some limit theorems in statistics*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1971, Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 4.

[Cha04] D. Chafaï – "Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities", *J. Math. Kyoto Univ.* **44** (2004), no. 2, p. 325–363.

[Cha06] D. Chafaï – "Binomial-poisson entropic inequalities and the M/M/∞ queue", to appear in ESAIM/PS, ArXiv.org:math.PR/0510488 and CNRS-HAL/ccsd-00012429, march 2006.

[CL06] D. Chafaï et J.-M. Loubes – "On nonparametric maximum likelihood for a class of stochastic inverse problems", *Statistics and Probability Letters* **76** (2006), no. 12, p. 1225–1237.

[Cra46] H. Cramér – *Mathematical Methods of Statistics*, Princeton Mathematical Series, vol. 9, Princeton University Press, Princeton, N. J., 1946.

[Doo34] J. L. Doob – "Probability and statistics", *Trans. Amer. Math. Soc.* **36** (1934), no. 4, p. 759–775.

[Dud98] R. M. Dudley – "Consistency of M-estimators and one-sided bracketing", High dimensional probability (Oberwolfach, 1996), Progr. Probab., vol. 43, Birkhäuser, Basel, 1998, p. 33–58.

[Fio00] S. Fiorin – "The strong consistency for maximum likelihood estimates: a proof not based on the likelihood ratio", *C. R. Acad. Sci. Paris Sér. I Math.* **331** (2000), no. 9, p. 721–726.

[Goe02] K. Goebel – *Concise course on fixed point theorems*, Yokohama Publishers, Yokohama, 2002.

[Hub67] P. J. Huber – "The behavior of maximum likelihood estimates under nonstandard conditions", Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics, Univ. California Press, Berkeley, Calif., 1967, p. 221–233.

[Huz48] V. S. Huzurbazar – "The likelihood equation, consistency and the maxima of the likelihood function", *Ann. Eugenics* **14** (1948), p. 185–200.

[KW56] J. Kiefer et J. Wolfowitz – "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters", *Ann. Math. Statist.* **27** (1956), p. 887–906.

[LC53] L. Le Cam – "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates", *Univ. California Publ. Statist.* **1** (1953), p. 277–329.

[Lin83a] B. G. Lindsay – "The geometry of mixture likelihoods: a general theory", *Ann. Statist.* **11** (1983), no. 1, p. 86–94.

[Lin83b] — , "The geometry of mixture likelihoods. II. The exponential family", *Ann. Statist.* **11** (1983), no. 3, p. 783–792.

[Per72] M. D. Perlman – "On the strong consistency of approximate maximum likelihood estimators", *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics* (Berkeley, Calif.), Univ. California Press, 1972, p. 263–281.

[Pfa88] J. Pfanzagl – "Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures", *J. Statist. Plann. Inference* **19** (1988), no. 2, p. 137–158.

[Pfa90] — , "Large deviation probabilities for certain nonparametric maximum likelihood estimators", *Ann. Statist.* **18** (1990), no. 4, p. 1868–1877.

[vdG00] S. A. van de Geer – *Empirical Processes in m-Estimation*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2000.

[vdG03] — , "Asymptotic theory for maximum likelihood in nonparametric mixture models", *Comput. Statist. Data Anal.* **41** (2003), no. 3-4, p. 453–464.

[vdV98] A. W. van der Vaart – *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 1998.

[Wal49] A. Wald – "Note on the consistency of the maximum likelihood estimate", *Ann. Math. Statistics* **20** (1949), p. 595–601.

[Wan85] J.-L. Wang – "Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics", *Ann. Statist.* **13** (1985), no. 3, p. 932–946.

[Zei86] E. Zeidler – *Nonlinear functional analysis and its applications. I*, Springer-Verlag, New York, 1986, Fixed-point theorems, Translated from the German by Peter R. Wadsack.

Djalil Chafaï, corresponding author.
Address: UMR 181 INRA/ENVT Physiopathologie et Toxicologie Expérimentales,
École Nationale Vétérinaire de Toulouse,
23 Chemin des Capelles, F-31076, Toulouse Cedex 3, France.
E-mail: mailto:d.chafai(AT)envt.fr
Address: UMR 5583 CNRS/UPS Laboratoire de Statistique et Probabilités,
Institut de Mathématiques de Toulouse, Université Paul Sabatier,
118 route de Narbonne, F-31062, Toulouse, Cedex 4, France.
E-mail: mailto:chafai(AT)math.ups-tlse.fr
Web: http://www.lsp.ups-tlse.fr/Chafai/