数理解析研究所講究録 1391 巻 2004 年 47-67

Nonparametric Maximum Likelihood Estimation of Probability Measures: Existence and Consistency*

Nobusumi SAGARA (佐柄信純)

Faculty of Economics, Hosei University (法政大学経済学部)
4342, Aihara, Machida, Tokyo, 194–0298 Japan
e-mail: nsagara@mt.tama.hosei.ac.jp

February 2004

Abstract

This paper formulates the nonparametric maximum likelihood estimation of probability measures and generalizes the consistency result on the maximum likelihood estimator (MLE). We drop the independence assumption on the underlying stochastic process and replace it with the assumption that the stochastic process is stationary and ergodic. The present proof employs Birkhoff's ergodic theorem and the martingale convergence theorem. The main result is applied to the parametric and nonparametric maximum likelihood estimation of density functions.

Mathematics Subject Classification (*2000*): Primary 62G20; Secondary 62G07, 62F12.

Key words: MLE; Consistency; Nonparametric maximum likelihood estimation; Stationarity; Ergodicity.

*This is a condensed version of the paper with the same title. The full version will be mailed electronically on request. This paper was presented at the Mathematical Economics Workshop at Research Institute for Mathematical Sciences, Kyoto University (RIMS) and at the workshop at Keio University. The author thanks Professor Toru Maruyama for helpful comments. This research was partly supported by Grant-in-Aid for Scientific Research (No. 14730021) from the Japan Society for the Promotion of Science.

1 Introduction

This paper formulates the nonparametric maximum likelihood estimation of probability measures and generalizes the consistency result on the maximum likelihood estimator (MLE) by introducing a parameterized family of density functions corresponding to a nonparametric family of probability measures. The extensions are twofold: First, we drop the independence assumption on the underlying stochastic process and replace it with the assumption that the stochastic process is stationary and ergodic. In previous proofs of consistency, the independence assumption is used to apply the strong law of large numbers (SLLN), which plays a crucial role. The present proof employs Birkhoff's ergodic theorem and the martingale convergence theorem. Second, we present a proof that does not resort directly to the integrability of the log-likelihood or the log-likelihood ratio, which has been imposed since Wald (1949), under the compactness assumption on the estimation set of probability measures. The technical assumptions on a parameterized family of density functions imposed in the previous studies are expressed systematically in terms of a nonparametric family of probability measures. Consequently, consistency is formulated in terms of the convergence of probability measures in total variation.

The consideration of a parameterized family of density functions corresponding to a nonparametric family of probability measures with respect to maximum likelihood estimation is not the standard framework. One of the prominent merits of our approach lies in the fact that the nonparametric maximum penalized likelihood estimation of density functions along the lines of De Montricher, Tapia, and Thompson (1975), Dong and Wets (2000), Good and Gaskins (1971), Klonias (1982), and Silverman (1982) is embedded into our framework. Thus, the existence and consistency of the maximum penalized likelihood estimator are formulated within our framework by the constructive method. Moreover, the approach developed here can be subsumed into the standard framework with the maximum likelihood estimation of the parameters in density functions.

Since the pioneering work of Wald (1949), who first gave a rigorous proof of the consistency of the MLE, several other proofs have appeared under less restrictive hypotheses. Generalizations of this result are found in, for example, Bahadur (1967), Huber (1967), Kiefer and Wolfowitz (1956), Le Cam (1953), Perlman (1972), and Wang (1985). While the consistency of the MLE has been studied as a traditional issue in statistical inference, a recent development in stochastic programming exemplifies the fact that the problem of statistical estimation can also be formulated as the approximation of minimizers in a general stochastic optimization problem, based on the

epi-convergence approach (Arstein and Wets (1995), Dupačová and Wets (1988), and King and Wets (1991)). The epi-convergence of an stochastic objective function to the *epi-limit* function (*epi-consistency*) guarantees the convergence of minimizers of the objective function to a minimizer of the epi-limit. Therefore, the problem of maximum likelihood estimation is formalized as an application of a stochastic optimization problem and the consistency of the MLE is proved under weaker hypotheses. The proof of consistency of the MLE along this line is found in Dong and Wets (2000), Geyer (1994), and Hess (1996) for independent processes, and Choirat, Hess, and Seri (2003) for stationary ergodic processes. The idea of this approach in statistical estimation originated with Huber (1967), while Hoffman-Jørgensen (1992) and Peškir (1998) elaborated Huber's method of approximation of the MLE employing essentially the same method of epi-convergence.

This paper presents another approach for proving the consistency of the MLE without the independence assumption. A difficulty in treating a dependent process lies in the fact that the likelihood function is not the product of the density functions. This forces Choirat, Hess, and Seri (2003) to maximize a *psuedolikelihood* to obtain the consistency of the M-estimator. In this paper, we maximize the 'exact' likelihood function for dependent processes to obtain the MLE, which seems novel in the literature.

The paper is organized as follows. In Section 2, we introduce the probability model in which the analysis is carried out, and we formulate the nonparametric maximum likelihood estimation of probability measures. In Section 3, we present the main result of this paper, Theorem 3.1: the existence and consistency of the MLE. We also apply the main result to the parametric and nonparametric maximum likelihood estimation of density functions. Section 4 briefly summarizes the related literature regarding Birkhoff's ergodic theorem and the martingale convergence theorem, which are the techniques developed in this paper. (Section 5 collects mathematical results on the measurability of correspondences and the martingale property of a likelihood ratio, which are required to prove the main results. Section 6 is devoted to the proofs of the theorems).[1]

2 Description of the Problem

Let $(\Omega, \mathscr{F}, P)$ be a probability space with the σ-field $\mathscr{F}$ of a sample space Ω and the probability measure P on Ω. The set of probability measures on Ω is denoted by $\mathscr{P}$ and metrized by the total variation $\|\mu-\lambda\| := \sup_{A\in\mathscr{F}} |\mu(A)-\lambda(A)|$, $\mu, \lambda \in \mathscr{P}$. Note that $\mathscr{P}$ is a complete metric space under the total

[1] Sections 5 and 6 are omitted here, which are available in the full version of the paper.

variation metric (Dunford and Schwartz (1958, p. 161)). We endow $\mathscr{P}$ with the Borel σ-field.

Let $(\boldsymbol{X}, \mathscr{T})$ be a measurable space with the σ-field $\mathscr{T}$ of $\boldsymbol{X}$. For $s = 1, 2, \ldots$, let $(\boldsymbol{X}^s, \mathscr{T}^s)$ be the s-fold product of $(\boldsymbol{X}, \mathscr{T})$ equipped with the product σ-field $\mathscr{T}^s$ of the product space $\boldsymbol{X}^s$. For $s = \infty$, we then write $(\boldsymbol{X}^\infty, \mathscr{T}^\infty)$. We denote an $\{\mathscr{F}_s\}_{s=1}^\infty$-adapted stochastic process on Ω with values in $\boldsymbol{X}$ by $\{X_s\}_{s=1}^\infty$, where $\{\mathscr{F}_s\}_{s=1}^\infty$ is a filtration on Ω such that $\mathscr{F}_s$ is the sub-σ-field of $\mathscr{F}$ generated by the observations $(X_1, \ldots, X_s)$ and $\mathscr{F}_\infty$ is the σ-field generated by $\bigcup_{s=1}^\infty \mathscr{F}_s$.

A stochastic process $\{X_s\}_{s=1}^\infty$ is *stationary* under $\mu \in \mathscr{P}$, if for any $B_s \in \mathscr{T}$, $s = 1, 2, \ldots$, the equality $\mu(\bigcap_{s=1}^\infty X_s^{-1}(B_s)) = \mu(\bigcap_{s=1}^\infty X_{s+t}^{-1}(B_s))$ holds for each t. A set $A \in \mathscr{F}$ is said to be $\{X_s\}_{s=1}^\infty$-*invariant* if there exist $B_s \in \mathscr{T}$, $s = 1, 2, \ldots$, such that $A = \bigcap_{s=0}^\infty X_{s+t}^{-1}(B_s)$ for each t. The collection of all $\{X_s\}_{s=1}^\infty$-invariant sets constitutes a sub-σ-field of $\mathscr{F}$, and we denote it by $\mathscr{F}^*$. A stochastic process $\{X_s\}_{s=1}^\infty$ is called *ergodic* under $\mu \in \mathscr{P}$ if any $\{X_s\}_{s=1}^\infty$-invariant set has a μ-measure of zero or one.

2.1 Nonparametric Maximum Likelihood Estimation of Probability Measures

A *probability model* M is a triplet consisting of a probability space $(\Omega, \mathscr{F}, P)$, a measurable space $(\boldsymbol{X}, \mathscr{T})$, and an $\{\mathscr{F}_s\}_{s=1}^\infty$-adapted, $\boldsymbol{X}$-valued stochastic process $\{X_s\}_{s=1}^\infty$ on Ω, which is denoted by

$$M = \langle (\Omega, \mathscr{F}, P), (\boldsymbol{X}, \mathscr{T}), \{X_s\}_{s=1}^\infty \rangle.$$

In probability theory, Ω is the set of states, $\mathscr{F}$ is the σ-field of events, P is the objective probability of events, and the measurable space $(\boldsymbol{X}, \mathscr{T})$ is the set of observations. In what follows, we assume that P is the unknown (or true) probability measure to be estimated. Let a subset $\mathscr{A}$ of $\mathscr{P}$ be an *estimation set*. For each $\mu \in \mathscr{A}$, let $f_s(\cdot\,; \mu)$ be a measurable function on $\boldsymbol{X}^s$. An estimation set $\mathscr{A}$ is *represented* by a family of density functions $\{(f_s(\cdot\,; \mu))_{s=1}^\infty \mid \mu \in \mathscr{A}\}$ if for $t = 1, \ldots, \infty$ there exists a σ-finite measure ν_t on $\boldsymbol{X}^t$ such that

$$\mu((X_1, \ldots, X_t) \in B^t) = \int_{B^t} f_t(x_1, \ldots, x_t; \mu) \nu_t(dx_1 \ldots dx_t) \qquad (2.1)$$

for any $B^t \in \mathscr{T}^t$. Thus, when $t \neq \infty$, the probability measure on $\boldsymbol{X}^t$ defined by the right-hand side of (2.1) is a finite-dimensional distribution of $\{X_s\}_{s=1}^\infty$.

The problem of statistical estimation is defined for a given probability model M and an estimation set $\mathscr{A}$ that is represented by a family of density

functions. The problem under investigation consists of obtaining an appropriate estimator of P from $\mathscr{A}$ after observing a sample of data $x_1, \ldots, x_t \in \boldsymbol{X}$ of a realization of $(X_1, \ldots, X_t)$, where t is the sample size. An *estimator* is provided by a map from $\boldsymbol{X}^t$ into $\mathscr{A}$. Nonparametric maximum likelihood estimation of probability measures is formulated as the following optimization problem:

$$\sup_{\mu \in \mathscr{A}} f_t(x_1, \ldots, x_t; \mu). \qquad (\mathbb{P}_t)$$

An estimator $\mu_t : \boldsymbol{X}^t \to \mathscr{A}$ that is a solution to $(\mathbb{P}_t)$, ν_t-a.e. $(x_1, \ldots, x_t) \in \boldsymbol{X}^t$, is called an *MLE*. A sequence $\{\mu_t\}_{t=1}^{\infty}$ of MLEs is said to be *consistent* if $\mu_t(x_1 \ldots, x_t) \to P$, ν_∞-a.e. $(x_1, x_2, \ldots) \in \boldsymbol{X}^\infty$.

2.2 Regular Probability Model

Nonparametric maximum likelihood estimation is formulated in a probability model with an estimation set that is represented by a parameterized family of density functions corresponding to a nonparametric family of probability measures. Obviously, it is not guaranteed that any probability model admits such representation. This motivates the following definitions.

Definition 2.1. A probability model $M = \langle (\Omega, \mathscr{F}, P), (\boldsymbol{X}, \mathscr{T}), \{X_s\}_{s=1}^{\infty} \rangle$ is *regular* if $\{X_s\}_{s=1}^{\infty}$ is stationary and ergodic under P.

Note that an independent and identically distributed (*i.i.d.*) process is stationary and ergodic (Breiman (1968, Corollary 6.33)). If a Markov chain with stationary transition probabilities is *indecomposable* and has a stationary initial distribution, it is also stationary and ergodic (Breiman (1968, Proposition 7.11 and Theorem 7.16)). Therefore, any probability model admits two types of stochastic processes as special cases. One is an *i.i.d.* process, the other is a Markov chain that has the above properties.

Definition 2.2. An estimation set $\mathscr{A}$ is *admissible* for a probability model $M = \langle (\Omega, \mathscr{F}, P), (\boldsymbol{X}, \mathscr{T}), \{X_s\}_{s=1}^{\infty} \rangle$ if the following conditions are satisfied:

(i) $P \in \mathscr{A}$.

(ii) $\mathscr{A}$ is relatively compact.

(iii) There exists a σ-finite measure Q such that $\mu \ll Q$ for any $\mu \in \mathscr{A}$.

(iv) $\{X_s\}_{s=1}^{\infty}$ is stationary and ergodic under any $\mu \in \mathscr{A}$.

(v) For any $\mu \in \mathscr{A} \setminus \{P\}$ there exists a measurable set $B \in \mathscr{T}$ such that $\mu(X_s^{-1}(B)) \neq P(X_s^{-1}(B))$ for some s.

We engage ourselves with a regular probability model with an admissible estimation set. The following theorem is a starting point for the analysis in the sequel.

Theorem 2.1. *For any regular probability model M, there exist an estimation set $\mathscr{A}$ and a family of density functions $\{(f_s(\cdot\,;\mu))_{s=1}^{\infty} | \mu \in \mathscr{A}\}$ such that $\mathscr{A}$ is admissible for M and represented by $\{(f_s(\cdot\,;\mu))_{s=1}^{\infty} | \mu \in \mathscr{A}\}$.*

3 Main Results

We show that in a regular probability model with an admissible estimation set, an MLE exists and converges in total variation to the true probability measure with a probability of one. This is a main result of this paper. This result is applied to the nonparametric maximum likelihood estimation of density functions. We also apply it to the standard framework in which the maximum likelihood estimation is formulated in terms of the density functions with a general parameter space.

3.1 Existence and Consistency of the MLE

Theorem 3.1. *Let $M = \langle(\Omega, \mathscr{F}, P), (\boldsymbol{X}, \mathscr{T}), \{X_s\}_{s=1}^{\infty}\rangle$ be a regular probability model, and let $\mathscr{A}$ be admissible for M and represented by a family of density functions $\{(f_s(\cdot\,;\mu))_{s=1}^{\infty} | \mu \in \mathscr{A}\}$. Then for each t there exists a measurable map $\mu_t : \boldsymbol{X}^t \to \mathscr{A}$ with the following properties:*

(i) $f_t(X_1(\omega), \ldots, X_t(\omega); \mu_t(X_1(\omega), \ldots, X_t(\omega)))$
$= \sup_{\mu \in \mathscr{A}} f_t(X_1(\omega), \ldots, X_t(\omega); \mu)$ μ-*a.e.* $\omega \in \Omega$ *for any* $\mu \in \mathscr{A}$.

(ii) $\|\mu_t(X_1(\omega), \ldots, X_t(\omega)) - P\| \to 0$ P-*a.e.* $\omega \in \Omega$.

Remark 3.1. We have dealt with a general sample space Ω that does not have any topological structure. When it is a metric space, there exists another mode of amenable topologies on $\mathscr{P}$: the *topology of weak convergence*. Because this topology is weaker than the metric topology with total variation, the compactness in Definition 2.2(ii) implies the compactness in the topology of weak convergence, and the convergence of the MLE in Theorem 3.1 implies the weak convergence of the MLE. The topology of weak convergence is metrized by the *Prohorov metric*, but in general, it is not equivalent to the total variation metric. Therefore the topological requirement in the present paper might be more stringent than the topology of weak convergence. Under some restrictive hypotheses, the consistency result in the standard framework can be translated into our nonparametric framework by

considering the continuous relation between the parameter space and the set of probability measures endowed with the topology of weak convergence (see Bahadur (1971, Section 9)).

3.2 Nonparametric Maximum Likelihood Estimation of Density Functions

We apply the main result to the nonparametric estimation of density functions. We transform the problem of the nonparametric maximum likelihood estimation of density functions into problem $(\mathbb{P}_t)$ in a regular probability model with an admissible estimation set and show that Theorem 3.1 can be applied also in the nonparametric framework to prove consistency. To this end, the point is the construction of an admissible estimation set from a nonparametric family of density functions.

Let $(\boldsymbol{X}, \mathscr{T}, \nu)$ be a σ-finite measure space. Define the set of density functions on $\boldsymbol{X}$ with respect to ν by

$$\mathscr{D}_\nu := \left\{ f \in L^1(\boldsymbol{X}, \mathscr{T}, \nu) \,\middle|\, \int f(x)\nu(dx) = 1, f \geq 0 \right\}.$$

Let $\mathscr{E} \subset \mathscr{D}_\nu$ be a nonparametric family of density functions. The problem is, given sample observations $x_1, \ldots, x_t \in \boldsymbol{X}$, to find an estimator $\hat{f} \in \mathscr{E}$ of the true density function $f^* \in \mathscr{E}$ associated with an *i.i.d.* stochastic process such that its distribution is given by the probability measure ν^* with $\nu^*(B) = \int_B f^*(x)\nu(dx)$, $B \in \mathscr{T}$. Then the nonparametric maximum likelihood estimation of density functions is given by

$$\sup_{f \in \mathscr{E}} f(x_1) \ldots f(x_t). \qquad (\mathbb{N}_t)$$

For $t = 1, \ldots, \infty$, let ν_t be the product measure on $\boldsymbol{X}^t$ induced by ν. A map $\hat{f}_t : \boldsymbol{X}^t \to \mathscr{D}_\nu$ is an MLE if $\hat{f}_t(x_1, \ldots, x_t)$ is a solution to $(\mathbb{N}_t)$, ν_t-a.e. $(x_1, \ldots, x_t) \in \boldsymbol{X}^t$. Consistency of the MLE is defined in terms of the convergence $\hat{f}_t(x_1, \ldots, x_t) \to f^*$ in $L^1(\boldsymbol{X}, \mathscr{T}, \nu)$, ν_∞-a.e. $(x_1, x_2, \ldots) \in \boldsymbol{X}^\infty$. If there is no constraint on $\mathscr{E}$, which is the case for $\mathscr{E} = \mathscr{D}_\nu$, then the MLE is simply the sum of the Dirac functions that assigns equal mass to each sample point: $\hat{f}_t(x_1, \ldots, x_t)(x) = t^{-1} \sum_{s=1}^t \mathbb{1}_{x_s}(x)$. Therefore, the really important case is $\mathscr{E} \subsetneq \mathscr{D}_\nu$.

The problem under consideration is to construct a regular probability model with an admissible estimation set from this nonparametric framework and to subsume $(\mathbb{N}_t)$ into $(\mathbb{P}_t)$. The following theorem is a basis for the analysis in the sequel: an immediate consequence of *Kolmogorov's existence theorem* (Shiryaev (1995, Corollary II.9.1)).

Proposition 3.1 (Kolmogorov). *There exist a probability space* $(\Omega, \mathscr{F}, P)$ *and an* $\boldsymbol{X}$*-valued stochastic process* $\{X_s\}_{s=1}^{\infty}$ *on* Ω *such that* $\{X_s\}_{s=1}^{\infty}$ *is i.i.d. under* P *and its distribution is given by* ν^*.

Let $\{\mathscr{F}_s\}_{s=1}^{\infty}$ be a filtration on Ω such that $\mathscr{F}_s$ is a sub-σ-field of $\mathscr{F}$ generated by the observations $(X_1, \ldots, X_s)$ and $\mathscr{F}_\infty$ is the σ-field generated by $\bigcup_{s=1}^{\infty} \mathscr{F}_s$. Then $\{X_s\}_{s=1}^{\infty}$ is $\{\mathscr{F}_s\}_{s=1}^{\infty}$-adapted. Thus, Proposition 3.1 validates our involvement with the regular probability model

$$\tilde{M} = \langle (\Omega, \mathscr{F}, P), (\boldsymbol{X}, \mathscr{T}), \{X_s\}_{s=1}^{\infty} \rangle.$$

The next step is to construct an admissible estimation set from $\mathscr{E}$.

Denote the set of all probability measures on $\boldsymbol{X}$ by $\mathscr{P}^{\boldsymbol{X}}$. For each $f \in \mathscr{E}$, define $\nu_f \in \mathscr{P}^{\boldsymbol{X}}$ by $\nu_f(B) := \int_B f(x)\nu(dx)$, $B \in \mathscr{T}$. Let $\sigma(X)$ be the σ-field on Ω generated by the random variable $X : \Omega \to \boldsymbol{X}$.

Assumption 3.1. There exists an $\boldsymbol{X}$-valued random variable X on Ω with the following properties:

(i) $\sigma(X) = \mathscr{F}$.

(ii) $P(X^{-1}(B)) = \nu^*(B)$ for any $B \in \mathscr{T}$.

(iii) $B \in \mathscr{T}$ and $B \subset \boldsymbol{X} \setminus X(\Omega)$ implies $\nu(B) = 0$.

Assumption 3.2. $\mathscr{E}$ is relatively compact in $L^1(\boldsymbol{X}, \mathscr{T}, \nu)$.

Assumption 3.3. For each $f \in \mathscr{E} \setminus \{f^*\}$ there exists a $\mu \in \mathscr{P} \setminus \{P\}$ such that $\{X_s\}_{s=1}^{\infty}$ is *i.i.d.* under μ and its distribution is given by ν_f.

Let $\mathscr{P}_\nu^{\boldsymbol{X}} := \{\nu' \in \mathscr{P}^{\boldsymbol{X}} \mid \nu' \ll \nu\}$ and $\mathscr{P}_Q := \{\mu \in \mathscr{P} \mid \mu \ll Q\}$. Denote the topological equivalence relation in terms of a homeomorphism by $\cong$. The underlying probabilistic structure has the following important property:

Theorem 3.2. *Under Assumption 3.1, there exists a* σ*-finite measure* Q *such that* $\mathscr{D}_\nu \cong \mathscr{P}_\nu^{\boldsymbol{X}} \cong \mathscr{P}_Q$.

Define the estimation set $\mathscr{A}$ of probability measures by

$$\mathscr{A} := \left\{ \mu \in \mathscr{P} \;\middle|\; \begin{array}{l} \exists f \in \mathscr{E} : \ \mu(X_s^{-1}(B)) = \mu(X^{-1}(B)) \\ \qquad\qquad\qquad = \nu_f(B) \ \forall B \in \mathscr{T} \ \forall s \\ \qquad \mu\Big(\bigcap_{s=1}^{t} X_s^{-1}(B_s)\Big) = \prod_{s=1}^{t} \nu_f(B_s) \ \forall B_s \in \mathscr{T} \ \forall s \end{array} \right\}.$$

Note that by construction, $P \in \mathscr{A}$ in view of Proposition 3.1 and Assumption 3.1, and $\{X_s\}_{s=1}^{\infty}$ is *i.i.d.* under any $\mu \in \mathscr{A}$.

Theorem 3.3. *Suppose that Assumptions 3.1–3.3 hold. Then $\mathscr{A}$ is admissible for $\tilde{M}$, $\mathscr{A} \subset \mathscr{P}_Q$, and $\mathscr{A} \cong \mathscr{E}$.*

Theorem 3.3 ensures that $\mathscr{A}$ is identified with $\mathscr{E}$ and each element in $\mathscr{E}$ is parameterized by an element in $\mathscr{A}$. Thus, $\mathscr{A}$ is represented by the family of density functions $\{f(\cdot;\mu) | \mu \in \mathscr{A}\} = \mathscr{E}$. Therefore, we can apply Theorem 3.1. Let $\mu_t : \boldsymbol{X}^t \to \mathscr{A}$ be the MLE in Theorem 3.1. It is obvious that $f(\cdot;\mu_t(x_1,\ldots,x_t))$ is a solution to $(\mathbb{N}_t)$, a.e. $(x_1,\ldots,x_t)$. Define $\hat{f}_t : \Omega \to \mathscr{E}$ by $\hat{f}_t(\omega) := f(\cdot;\mu_t(\omega))$. Then the convergence $\mu_t(\omega) \to P$ a.e. ω in total variation is equivalent to $f_t(\omega) \to f^*$ a.e. ω in $L^1(\boldsymbol{X}, \mathscr{T}, \nu)$.

As concerns the nonparametric estimation of density functions, Dong and Wets (2000) develops a constrained maximum likelihood estimation in a general framework by employing the epi-convergence approach, which can be effectively applied to the nonparametric maximum penalized likelihood estimation. They investigate the following problem of the form:

$$\begin{aligned} &\sup \sum_{s=1}^{t} \log f(x_s) \\ &\text{s.t. } f \in G \subset H := L^2(\boldsymbol{X}, \mathscr{T}, \nu). \end{aligned} \tag{$\mathbb{L}_t$}$$

In particular, when G is restricted to

$$G = \{f \in H | \Phi(f) \leq \beta,\ f \in C\},\ \beta \in \mathbb{R},$$

where the functional Φ describes a constraint on density functions, with the suitable choice of $\alpha \geq 0$, the problem $(\mathbb{L}_t)$ can be transformed into the following equivalent problem (Dong and Wets (2000, Lemma 4.1)):

$$\sup_{f \in C} \left[\sum_{s=1}^{t} \log f(x_s) - \alpha \Phi(f) \right]. \tag{$\mathbb{M}_t$}$$

Here, $-\alpha\Phi(f)$ is a penalized term. They provide the existence and consistency for $(\mathbb{L}_t)$, which broadly encompasses the previous result on the maximum penalized likelihood estimation problem $(\mathbb{M}_t)$ with the specific functional forms of Φ similar to those in Good and Gaskins (1974), Klonias (1982), Montricher, Tapia, and Thompson (1975), and Silverman (1982). It is obvious that the problems $(\mathbb{L}_t)$ and $(\mathbb{M}_t)$ are subsumed into our framework by considering $\mathscr{E} = G$ and $\mathscr{E}_\Phi = \{f \in \mathscr{D}_\nu | f \in H \cap C,\ \Phi(f) \leq \beta\}$.

Roughly speaking, the consistency result established by Dong and Wets (2000) is based on the following observation. Under certain conditions, the SLLN guarantees the following.

- The objective function $L_t(f) := -t^{-1}\sum_{s=1}^{t}\log f(x_s)$ epi-converges to the epi-limit $L(f) := -\int \log f(x)\nu(dx)$ as $t \to \infty$, ν_∞-a.e. $(x_1, x_2, \dots)$. This epi-limit is a nonstochastic function.
- A cluster point $\hat{f}_0$ of the sequence of minimizers $\{\hat{f}_t\}_{t=1}^{\infty}$, where $\hat{f}_t \in \arg\min_{f\in S} L_t(f)$, is a minimizer of the epi-limit: $\hat{f}_0 \in \arg\min_{f\in S} L(f)$, ν_∞-a.e. $(x_1, x_2, \dots)$.
- The set of minimizers $\arg\min_{f\in S} L(f)$ contains the true density function f^*.

As Dong and Wets note, their consistency result does not guarantee the existence of an arg min-estimator $\hat{f}_t$ and a cluster point $\hat{f}_0$ of $\{\hat{f}_t\}_{t=1}^{\infty}$. These will, of course, exist if G is compact. Thus, Assumption 3.2 is implicit in their structure.

Note that the family of density functions under consideration is broader in our framework than in theirs, because L^2-space is included in L^1-space. Because the epi-convergence approach works for Hilbert spaces, their method is not valid when considering a family of density functions in L^1. One of the merits of using the epi-convergence together with the L^2-space lies in the fact that their consistency result also works in the weak topology in L^2 (Dong and Wets (2000, Theorem 7.12)). Although our consistency result also implies weak convergence in L^1, if $\mathscr{E}$ is only assumed to be relatively weakly compact in L^1 and not to be relatively norm compact, then consistency in terms of the weak convergence in L^1 is not obtained by our method. The weak compactness is more easily achieved than the norm compactness. In some families of density functions, however, the norm compactness in L^1 can be verified in a simple manner (see Remark 3.2 and Examples 3.1–3.4 below).

Assumption 3.1 is innocuous. Conditions (i) and (ii) in Assumption 3.1 will be required also in their framework, when considering explicitly the a.e.-convergence on the sample space Ω. As in the proof of Theorem 3.2, conditions (i) and (iii) in Assumption 3.1 together enable us to establish the one-to-one correspondence between $\mathscr{P}^{\boldsymbol{X}}$ and $\mathscr{P}$, which is not so stringent and is implicit in many statistical models.

Assumption 3.3 is our point of departure from the assumptions in Dong and Wets. Although it is indeed unnecessary for them, in the present framework it plays an important role in the proof of Theorem 3.3. This seems, however, a plausible assumption because it requires that any density function in $\mathscr{E}$ generates an *i.i.d.* process under some probability measure on the sample space, which is implicitly assumed in many statistical models.

On the contrary, Dong and Wets (2000, Theorems 7.9 and 7.12) assume the local integrability of the log-likelihood: for any $g \in S$ there exists a

neighborhood $V(g)$ of g such that $\int \log f(x)\nu(dx) < \infty$ for any $f \in V(g)$. We do not require this condition.

Remark 3.2. To apply Theorem 3.1, we have imposed the relative compactness of $\mathscr{E}$ in Assumption 3.2. It may not be a easy task to check whether or not a given family of density functions is relatively compact in L^1. However, when the observation space $\boldsymbol{X}$ is finite-dimensional, which is the case in most applications, we might be more hopeful. Let $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}, \nu)$ be a measure space on the n-dimensional Euclidean space $\mathbb{R}^n$ with the Borel σ-field $\mathscr{B}_{\mathbb{R}^n}$ and the Lebesgue measure ν on $\mathbb{R}^n$. The following characterization of the compactness of $\mathscr{E}$ in $L^1(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}, \nu)$ provides a useful criterion for many cases.

Theorem 3.4. *A subset $\mathscr{E}$ of $\mathscr{D}_\nu$ is compact in $L^1(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}, \nu)$ if and only if it is closed and the following condition is satisfied:*

$$\lim_{r\to\infty} \sup_{f\in\mathscr{E}} \int_{\|x\|>r} f(x)\nu(dx) = 0. \tag{3.1}$$

Condition (3.1) is easy to check. For example, when each density function in $\mathscr{E}$ is defined on some bounded set in $\mathbb{R}^n$, it holds obviously by assigning value zero to each element in $\mathscr{E}$ outside the bounded set.

Remark 3.3. Because $L^p \subset L^1$ for $p > 1$ and the relative topology of L^p inherited from the L^1-norm is finer than the L^p-norm topology (Aliprantis and Border (1999, Corollary 12.3)), an L^p-norm closed set is obviously L^1-norm closed. Therefore, in view of Theorem 3.4, when we limit ourselves to a family of density functions in L^p, it suffices to show the L^p-norm closeness to ensure the L^1-norm compactness.

Consider some applications of Theorem 3.4 to the nonparametric maximum penalized likelihood estimation.

Example 3.1 (De Montricher, Tapia, and Thompson (1975)). Let $[a, b]$ be a closed interval in $\mathbb{R}$. For each integer $k = 1, 2, \ldots$, denote the Sobolev space $H_0^k([a, b])$ of functions $f \in L^2([a, b])$ with $(i-1)$-th generalized derivatives $f^{(i-1)}$ vanishing at the end points $\{a, b\}$, and with i-th generalized derivatives $f^{(i)}$ in $L^2([a, b])$, $i = 1, \ldots, s$:

$$H_0^k([a,b]) := \left\{ f \in L^2([a,b]) \;\middle|\; \begin{array}{l} f^{(i-1)}(a) = f^{(i-1)}(b) = 0 \\ f^{(i)} \in L^2([a,b]),\ i = 1, \ldots, k \end{array} \right\}.$$

The inner product in $H_0^k([a, b])$ is given by $\langle f, g\rangle_{H_0^k} := \int_a^b f^{(k)}(x) g^{(k)(x)} \nu(dx)$, which makes $H_0^k([a, b])$ a Hilbert space, and the norm in $H_0^k([a, b])$ is given

by $\|f\|_{H_0^k} = \langle f, f\rangle_{H_0^k}^{\frac{1}{2}}$. By Schwartz's inequality, we have

$$|f^{(i-1)}(x)| = \left|\int_a^x f^{(i)}(y)\nu(dy)\right| \leq (b-a)\|f^{(i)}\|_{L^2([a,b])} \quad \text{for any } x \in [a,b]$$

for each $i = 1, \ldots, k$. Therefore, $\|f_n - f\|_{H_0^k} \to 0$ implies $\|f_n - f\|_{L^2([a,b])} \to 0$. By assigning zero to each element in $H_0^k([a,b])$ outside the interval $[a,b]$, we can regard $H_0^k([a,b])$ as a subset of $L^2(\mathbb{R}, \mathscr{B}_\mathbb{R}, \nu)$. Consider the functional Φ on $H_0^k([a,b])$ of the form

$$\Phi(f) := \|f\|_{H_0^k}.$$

As mentioned previously, the constraint $\Phi(f) \leq \beta$ in problem $(\mathbb{L}_t)$ is replaced by the penalty term $-\alpha\Phi(f)$ for some $\alpha \geq 0$ in problem $(\mathbb{M}_t)$. Define the estimation set of density functions by

$$\mathscr{E}_\Phi := \{f \in \mathscr{D}_\nu \mid f \in H_0^k([a,b]),\ \Phi(f) \leq \beta\}.$$

We claim that $\mathscr{E}_\Phi$ is compact in $L^1(\mathbb{R}, \mathscr{B}_\mathbb{R}, \nu)$. Because Φ is continuous in the H_0^k-norm, $\mathscr{E}_\Phi$ is closed in $H_0^k([a,b])$. Thus, $\mathscr{E}_\Phi$ is closed in $L^2([a,b])$, and hence closed also in $L^1(\mathbb{R}, \mathscr{B}_\mathbb{R}, \nu)$ by Remark 3.3. Therefore, it suffices to show that $\mathscr{E}_\Phi$ satisfies (3.1), but this is evident because each element in $\mathscr{E}_\Phi$ takes value zero outside $[a,b]$. If we replace $L^2([a,b])$ with $L^1([a,b])$ in the above argument, we obtain the Sobolev space $W_0^{k,1}([a,b])$ with the norm $\|f\|_{W_0^{k,1}} := \int |f^{(k)}(x)|\nu(dx)$, which makes $W_0^{k,1}([a,b])$ a Banach space. It is obvious that the compactness in $L^1(\mathbb{R}, \mathscr{B}_\mathbb{R}, \nu)$ is obtained also for $W_0^{k,1}([a,b])$.

Example 3.2 (Good and Gaskins (1971)). Good and Gaskins proposed two penalty functions in the penalized maximum likelihood estimation of density functions. Let $H^p([a,b])$ be the Sobolev space of functions $f \in L^2([a,b])$ with the i-th order generalized derivatives $f^{(i)}$ in $L^2([a,b])$, $i = 1, \ldots, p$. The norm in $H^p([a,b])$ is given by $\|f\|_{H^p} := \|f\|_{L^2([a,b])} + \sum_{i=1}^p \|f^{(i)}\|_{L^2([a,b])}$. The first problem involves a penalty function Φ on $H^1([a,b])$ with the following form:

$$\Phi(f) = \int_a^b f'(x)^2\nu(dx),$$

together with the constraint $\int_a^b f(x)^2\nu(dx) = 1$. Define

$$\mathscr{E}_\Phi := \left\{f \in \mathscr{D}_\nu \,\middle|\, f \in H^1([a,b]),\ \int_a^b f(x)^2\nu(dx) = 1,\ \Phi(f) \leq \beta\right\}.$$

By Schwartz's inequality, we have

$$\begin{aligned}|\Phi(f)-\Phi(g)| &= \left|\int (f'(x)^2 - g'(x)^2)\nu(dx)\right| \\ &= \left|\int (f'(x)-g'(x))(f'(x)+g'(x))\nu(dx)\right| \\ &\le \|f'-g'\|_{L^2([a,b])}\,\|f'+g'\|_{L^2([a,b])}.\end{aligned}$$

Therefore, Φ is continuous. It is obvious that $\mathscr{E}_\Phi$ is closed in the H^1-norm. Thus, $\mathscr{E}_\Phi$ is closed in $L^2([a,b])$, and hence closed also in $L^1(\mathbb{R},\mathscr{B}_\mathbb{R},\nu)$ by Remark 3.3. Therefore, $\mathscr{E}_\Phi$ is compact in $L^1(\mathbb{R},\mathscr{B}_\mathbb{R},\nu)$. If we replace $L^2([a,b])$ with $L^1([a,b])$ in the above argument, we obtain the Sobolev space $W^{1,1}([a,b])$ with the norm $\|f\|_{W^{1,1}} := \|f\|_{L^1([a,b])} + \|f'\|_{L^1([a,b])}$, which makes $W^{1,1}([a,b])$ a Banach space. It is obvious that the compactness in $L^1(\mathbb{R},\mathscr{B}_\mathbb{R},\nu)$ is obtained also for $W^{1,1}([a,b])$.

Example 3.3 (Good and Gaskins (1971)). The second problem in Good and Gaskins treats the following form of the penalty function on $H^2([a,b])$

$$\Phi(f) := \int_a^b f'(x)^2\nu(dx) + \gamma\int_a^b f''(x)^2\nu(dx),\ \gamma\in\mathbb{R}$$

together with the constraint $\int_a^b f(x)^2\nu(dx)=1$. Let

$$\mathscr{E}_\Phi := \left\{ f\in\mathscr{D}_\nu \,\middle|\, f\in H^2([a,b]),\ \int_a^b f(x)^2\nu(dx)=1,\ \Phi(f)\le\beta \right\}.$$

As in Example 3.1, it suffices to show that Φ is continuous in the H^2-norm to demonstrate the compactness of $\mathscr{E}_\Phi$ in $L^1(\mathbb{R},\mathscr{B}_\mathbb{R},\nu)$. The continuity of Φ follows from the same argument as in Example 3.2. If we replace $L^2([a,b])$ with $L^1([a,b])$ in the above argument, we obtain the Sobolev space $W^{2,1}([a,b])$ with the norm $\|f\|_{W^{2,1}} := \|f\|_{L^1([a,b])} + \|f'\|_{L^1([a,b])} + \|f''\|_{L^1([a,b])}$, which makes $W^{2,1}([a,b])$ a Banach space. It is obvious that the compactness in $L^1(\mathbb{R},\mathscr{B}_\mathbb{R},\nu)$ is obtained also for $W^{2,1}([a,b])$.

Example 3.4 (Silverman (1982)). Let S be a bounded open set in $\mathbb{R}^n$. Suppose that Φ on the Hilbert space $H^p(S)$ is given by

$$\Phi(f) := \|f\|_{H^p},$$

together with the constraint $\int_S e^{f(x)}\nu(dx)=1$. Define

$$\mathscr{E}_\Phi := \left\{ f\in\mathscr{D}_\nu \,\middle|\, f\in H^p(S),\ \int_S e^{f(x)}\nu(dx)=1,\ \Phi(f)\le\beta \right\}.$$

As in the above examples, it is evident that $\mathscr{E}_\Phi$ is compact in $L^1(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}, \nu)$. If we replace $H^p(S)$ with the Banach space $W^{p,1}(S)$, then the compactness in $L^1(\mathbb{R}, \mathscr{B}_{\mathbb{R}}, \nu)$ is obtained also for $W^{p,1}(S)$.

3.3 Maximum Likelihood Estimation of the Parameters in Density Functions

Bahadur (1971, Section 9) transformed the standard statistical model into the maximum likelihood estimation of probability measures by considering the continuous relation between the parameter space and the family of probability measures. In this section, we present a converse approach. We transform the present framework into the standard statistical model in which parameters in density functions are estimated.

Note that the nonparametric estimation of density functions in Section 3.2 has a straightforward translation into the parametric estimation. Let Ξ be a *parameter space* and let Θ be a subset of Ξ. Let $f : \boldsymbol{X} \times \Xi \to \mathbb{R}$ be a function such that $f(\cdot\,; \theta)$ is a density function in $\mathscr{D}_\nu$ for each $\theta \in \Xi$. Suppose that $\theta^* \in \Theta$ is the (unknown) true parameter. Then $f^* = f(\cdot\,; \theta^*)$ is the true density function. If $\mathscr{E} = \{f(\cdot\,; \theta) | \, \theta \in \Theta\}$ is relatively compact in $L^1(\boldsymbol{X}, \mathscr{T}, \nu)$, then the result in Section 3.2 is obviously applicable. This covers the consistency result in the standard statistical models such as Bahadur (1967), Hess (1996), Huber (1967), Kiefer and Wolfowitz (1956), Le Cam (1953), Perlman (1972), Wald (1949), and Wang (1985). With this transformation, the compactness of Θ and the continuity of $\theta \mapsto f(\cdot\,; \theta)$ are not required. By this method, however, it seems difficult to obtain the consistency of the MLE without the independence assumption.

On the contrary, the nonparametric estimation of probability measures in Section 3.1 is transformed into the parametric estimation for dependent processes. Let $M = \langle (\Omega, \mathscr{F}, P), (\boldsymbol{X}, \mathscr{T}), \{X_s\}_{s=1}^\infty \rangle$ be a regular probability model, and let $\mathscr{A}$ be admissible for M and represented by a family of density functions $\{(f_s(\cdot\,; \mu))_{s=1}^\infty | \, \mu \in \mathscr{A}\}$. Suppose that Ξ is endowed with a Hausdorff topology and $\Xi \ni \theta \overset{\Psi}{\mapsto} \mu_\theta \in \mathscr{P}$ is a homeomorphism. Then $\{\mu_\theta | \, \theta \in \Xi\} = \mathscr{P}$ is a parameterized family of probability measures. Put $\Theta := \Psi^{-1}(\mathscr{A})$ and $\theta^* := \Psi^{-1}(P)$. Thus, θ^* is the true parameter. Define $f_t : \boldsymbol{X}^t \times \Xi \to \mathbb{R}$ by $f_t(x_1, \ldots, x_t; \theta) := f_t(x_1, \ldots, x_t; \mu_\theta)$. Then the likelihood maximization problem $(\mathbb{P}_t)$ is transformed to

$$\sup_{\theta \in \Theta} f_t(x_1, \ldots, x_t; \theta). \tag{$\mathbb{Q}_t$}$$

Let $\mu_t : \boldsymbol{X}^t \to \mathscr{A}$ be the MLE stated in Theorem 3.1. It is obvious that $\theta_t(x_1, \ldots, x_t) := \Psi^{-1}(\mu_t(x_1, \ldots, x_t))$ is a solution to $(\mathbb{Q}_t)$, a.e. $(x_1 \ldots, x_t) \in$

$\boldsymbol{X}^t$. Define $\theta_t(\omega) := \theta_t(X_1(\omega), \ldots, X_t(\omega))$. The convergence $\|\mu_t(\omega) - P\| \to 0$ a.e. ω is equivalent to $\theta_t(\omega) \to \theta^*$ a.e. ω.

The result stated here is the consistency of the MLE in the standard framework without the independence assumption. Among previous authors, Hess (1996) presented the general result on consistency that does not assume the compactness of the parameter space and the continuity of density functions with respect to a parameter, under the hypothesis that the underlying stochastic process is pairwise independent identically distributed. We now compare our assumptions with those in the existing literature by reducing the present framework to the case of independent processes.

Let $(\boldsymbol{X}, \mathscr{T}, Q^{\boldsymbol{X}})$ be a σ-finite measure space with $Q^{\boldsymbol{X}}(B) := Q(X_1^{-1}(B))$ for any $B \in \mathscr{T}$. Suppose that $\{X_s\}_{s=1}^{\infty}$ is *i.i.d.* under $\{\mu^\theta | \, \theta \in \Theta\}$. Define the probability measure on $\boldsymbol{X}$ by $\mu_\theta^{\boldsymbol{X}}(B) := \mu^\theta(X_s^{-1}(B))$ for any $B \in \mathscr{T}$. Note that the definition of $\mu_\theta^{\boldsymbol{X}}$ does not depend on the choice of s. Then $\mu_\theta^{\boldsymbol{X}} \ll Q^{\boldsymbol{X}}$ for any $\theta \in \Theta$ and f_t is of the form $f_t(x_1, \ldots, x_t; \theta) = f(x_1; \theta) \ldots f(x_t; \theta)$, where $f(\,\cdot\,; \theta)$ is a density function on $\boldsymbol{X}$ with $\mu_\theta^{\boldsymbol{X}}(B) = \int_B f(x; \theta) Q^{\boldsymbol{X}}(dx)$ for any $B \in \mathscr{T}$.

In comparison with Hess (1996), the conditions in Definition 2.2 correspond to the following conditions.

(D-1) Θ is relatively compact.

(D-2) $\theta \mapsto f(x; \theta)$ is continuous for $Q^{\boldsymbol{X}}$-a.e. $x \in \boldsymbol{X}$.

(D-3) $Q^{\boldsymbol{X}}\{x \in \boldsymbol{X} | \, f(x; \theta) \neq f(x; \theta^*)\} > 0$ for any $\theta \in \Theta \setminus \{\theta^*\}$.

Conditions (D-1) and (D-2) are somewhat stronger than the *sup-compactness* in Hess (1996, Condition (c)) when Θ is compact. Hess assumes Θ to be a Suslin space. Θ is also a Suslin space in our formulation, because $\mathscr{P}$ is a complete metric space and $\Theta \cong \mathscr{A}$. Condition (D-3) is the *identifiability* condition imposed by many authors (Bahadur (1967, Condition (c)), Kiefer and Wolfowitz (1956, Assumption 4), Hess (1996, Condition (d)), and Wald (1949, Assumption 4)).

Note that we do not impose the integrability of the log-density function:

$$\text{(D-4)} \quad \int |\log f(x; \theta^*)| Q^{\boldsymbol{X}}(dx) < \infty,$$

which is assumed in Hoffman-Jørgensen (1992, Condition (1.2.2)), Huber (1967, (A-3)), and Wald (1949, Assumption 6)). Nor do we require the *global integrability* of the log-likelihood ratio:

$$\text{(D-5)} \quad \int \left(\log \frac{g(x)}{f(x; \theta^*)} \right) f(x; \theta^*) Q^{\boldsymbol{X}}(dx) < \infty, \text{ where } g(x) := \sup_{\theta \in \Theta} f(x; \theta),$$

which is imposed in Bahadur (1967, Condition (b)), Hess (1996, Condition (H)), Huber (1967, (A-5)), and Kiefer and Wolfowitz (1956, Assumption 5). There is common understanding among statisticians that condition (D-5) is problematic and is not satisfied in many common statistical models. Indeed, it is not necessary for an MLE to be consistent (see Hess (1996, Remark 6.5)). Perlman (1972) clarifies the role of (D-5) in proving consistency when Θ is not compact and shows that if it is compact, then the *local integrability* of the log-likelihood ratio implies (D-5), where the local integrability requires that for any $\theta' \in \Theta$ there exists a neighborhood $V(\theta')$ of θ' such that

$$\text{(D-6)} \quad \int \left(\log \frac{f(x;\theta)}{f(x;\theta^*)} \right) f(x;\theta^*) Q^{\boldsymbol{X}}(dx) < \infty \quad \text{on } V(\theta'),$$

which is satisfied in many reasonable statistical models. While we assume the relative compactness of Θ, we do not employ condition (D-6) explicitly. Wang (1985, pp. 932–933) illustrates a nonparametric estimation problem of a certain family of density functions in which the consistency of the MLE holds even if (D-6) is violated.

For each $\theta \in \Theta$, let ν_∞^θ be a probability measure on $\boldsymbol{X}^\infty$ induced by μ_θ as in the proof of Lemma ??, and let $\nu_\infty^* := \nu_\infty^{\theta^*}$. Define $\hat{\mathscr{P}}^{\boldsymbol{X}^\infty} := \{\nu_\infty^\theta | \theta \in \Theta\}$. Let $V_\varepsilon(\theta^*)$ be the ε-neighborhood of θ^*. Examining the proof by Perlman (1972), we notice that the following condition plays an important role in proving the consistency of the MLE:

$$\text{(D-7)} \quad \nu_\infty \left(\limsup_{t\to\infty} \sup_{\theta \in \Theta \setminus V_\varepsilon(\theta^*)} \log \frac{f(x_1;\theta)\dots f(x_t;\theta)}{f(x_1;\theta^*)\dots f(x_t;\theta^*)} < 0 \right) = 1$$

for any $\nu_\infty \in \hat{\mathscr{P}}^{\boldsymbol{X}^\infty}$ and $\varepsilon > 0$. This condition is implied by the *dominance by zero* condition of $\log f(x;\theta)/f(x;\theta^*)$ on $\Theta \setminus V_\varepsilon(\theta^*)$ for any $\varepsilon > 0$ (see Definition 1 and Theorem 2.1 of Perlman (1972)). Wang (1985) extends condition (D-7) and apply the consistency result to the nonparametric family of concave continuous distribution functions. Returning to the case for dependent processes, in view of Theorem ??, we obtain the following sufficient condition for consistency, which is weaker than (D-7) in the present framework:

$$\text{(D-8)} \quad \nu_\infty^* \left(\lim_{t\to\infty} \log \frac{f_t(x_1,\dots,x_t;\theta)}{f_t(x_1,\dots,x_t;\theta^*)} < 0 \right) = 1 \quad \text{for any } \theta \in \Theta.$$

Examining the proof of Theorem 3.1(ii), we notice that (D-8) is implicitly employed. The proof of Theorem 3.1(ii) suggests that it is possible to extend the consistency result under the more general hypothesis (D-8) in the standard framework, at least for relatively compact Θ.

Remark 3.4. Le Cam (1990) collects many tricky examples in which MLEs misbehave. In particular, Examples 1 to 4 of Le Cam demonstrate the nonexistence of an MLE. Note that these examples do not satisfy the relative compactness of the parameter space Θ and they are based on the fact that for some parameter value $\theta_0 \in \Theta$, the density function $f(x;\theta)$ tends to infinity as $\theta \to \theta_0$. Le Cam (1990) also illustrates density functions, adapted from Bahadur (1958) and Ferguson (1982), in which an MLE is inconsistent. The example given by Bahadur lacks the relative compactness of Θ. The density function in Ferguson is continuous on the interval $\Theta = [0,1]$, but does not satisfy condition (D-8).

4 Comments on the Technique

The proof of Theorem 3.1 relies heavily on the use of Birkhoff's ergodic theorem and the martingale convergence theorem. We briefly mention the application of these two theorems to the consistency of the MLE in the literature.

Choirat, Hess, and Seri (2003, Corollary 2.4) prove that for any stationary ergodic stochastic process $\{X_s\}_{s=1}^{\infty}$ and random lower semicontinuous (LSC) function $g : \boldsymbol{X} \times \Xi \to \mathbb{R}$ with certain integrability conditions, the sample average $t^{-1}\sum_{s=1}^{t} g(X_s(\omega),\cdot)$ epi-converges to the epi-limit $\int g(X_s(\omega),\cdot)Q(d\omega)$ a.e. ω. This is a generalization of Birkhoff's ergodic theorem for random (LSC) functions, in which the underlying integrand depends continuously on a parameter, which is also an extension of the epi-convergence approach to the SLLN for random LSC functions developed by Arstein and Wets (1996, Theorem 2.3), and King and Wets (1991, Theorems 2.3 and 2.4). Birkhoff's ergodic theorem for random LSC functions using this approach is obtained also by Korf and Wets (2001, Theorem 8.2) under somewhat restrictive hypotheses, but Korf and Wets present a systematic treatment for endowing the space of random LSC functions with a suitable σ-field. The consistency of the M-estimator for stationary ergodic processes is derived in Choirat, Hess, and Seri (2003, Theorem 2.8) by putting $g(x,\theta) := -\log f(x;\theta)$. This proof of consistency is now standard in the literature, and is essentially the same as Hess (1996, Theorem 5.1) with the integrability condition (D-5).

In this paper, we apply the standard version of Birkhoff's ergodic theorem to the empirical process $\{\mathbb{1}_B(X_s(\omega))\}_{s=1}^{\infty}$ with $B \in \mathscr{T}$ to obtain the a.e. convergence $t^{-1}\sum_{s=1}^{t} \mathbb{1}_B(X_s(\omega)) \to P(X_1^{-1}(B))$, as is shown in the proof of Lemma 5.1. Our use of Birkhoff's ergodic theorem seems more subsidiary than that by Choirat, Hess, and Seri (2003), as it is employed only to show the mutual singularity of the underlying probability measures, but this prop-

erty is essential for the martingale convergence of a likelihood ratio (Theorem 5.1).

Peškir (1998) investigates the following stochastic optimization problem: $\sup_{\theta\in\Theta} h_t(\omega,\theta)$, where $\{(h_t(\cdot,\theta),\mathscr{F}_t)_{t=1}^{\infty} | \theta \in \Theta\}$ constitutes a family of reversed submartingales. The martingale convergence theorem implies $h_t(\omega,\theta) \to E[h_\infty(\theta)] := \int h_\infty(\omega,\theta)Q(d\omega)$ a.e. ω for each $\theta \in \Theta$ if the tail σ-field $\bigcap_{t=1}^{\infty} \mathscr{F}_t$ has the zero-one property. Under certain measurability and continuity assumptions on h_t, Peškir (1998, Theorem 3.3) shows that $\theta_t(\omega) \in \arg\min_{\theta\in\Theta} h_t(\omega,\theta)$ converges to an element in $\arg\min_{\theta\in\Theta} E[h_\infty(\theta)]$ a.e. ω. This type of problem originates in Huber (1967, pp. 224–226) and is elaborated by Hoffman-Jørgensen (1992). The result of Peškir is an extension of Hoffman-Jørgensen (1992, Section 1.12) and obviously contains a partial generalization of the consistency of the MLE for dependent processes such as exchangeable processes (Peškir (1996, p. 314)), but the result does not seem to be applicable to stationary ergodic processes.

In Theorem 5.1 of this paper, we show that a likelihood ratio is a martingale and converges to zero with probability one by the martingale convergence theorem. This fact is well known when the underlying stochastic process is *i.i.d.* We drop the independence assumption and generalize the result when the stochastic process is assumed to be only stationary and ergodic. The fact that the likelihood ratio converges to zero was proved by Wald (1949, Theorem 1) for *i.i.d.* processes. The martingale property of a likelihood ratio was established originally by Doob (1953, p. 93) under the absolute continuity hypothesis of the underlying probability measures. Doob (1953, pp. 348–350) also demonstrated the consistency of the MLE by applying the martingale convergence theorem for *i.i.d.* processes. Billingsley (1986, Example 35.4) gives another proof of the convergence of a likelihood ratio by effective use of the strong law of large numbers, and Williams (1991, Section 14.17) presents a different approach, applying Kakutani's theorem on product martingales (Williams (1991, Section 14.12)).

References

[1] Arstein, Z., Wets, R. J.-B., 1995. Consistency of minimizers and the SLLN for stochastic programs. J. Convex Anal. 2, 1–17.

[2] Aliprantis, C. D., Border, K. C., 1999. Infinite Dimensional Analysis: a Hitchhiker's Guide, 2nd ed. Springer-Verlag, Berlin.

[3] Bahadur, R. R., 1958. Examples of inconsistency of maximum likelihood estimates. Sankhyā Ser. A 20, 207–210.

[4] Bahadur, R. R., 1967. Rates of convergence of estimates and test statistics. Ann. Math. Statist. 38, 303–324.

[5] Bahadur, R. R., 1971. Some Limit Theorems in Statistics. SIAM, Philadelphia.

[6] Billingsley, P., 1986. Probability and Measure, 2nd ed. John Wiley & Sons, New York.

[7] Breiman, L., 1968. Probability. Addison-Wesley, Reading, Massachusetts.

[8] Choirat, C., Hess, C., Seri, R., 2003. A functional version of the Birkhoff ergodic theorem for a normal integrand: a variational approach. Ann. Statist. 31, 63–92.

[9] De Montricher, G., Tapia, R. A., Thompson, J. R., 1975. Nonparametric maximum likelihood estimation of probability densities by penalty function methods. Ann. Statist. 3, 1329–1348.

[10] Dong, M. X., Wets, R. J.-B., 2000. Estimating density functions: a constrained maximum likelihood approach. J. Nonparametr. Statist. 12, 549–595.

[11] Doob, J. L., 1953. Stochastic Processes. John Wiley & Sons, New York.

[12] Dunford, N., Schwartz, J. T., 1958. Linear Operators, Part I: General Theory. John Wiley & Sons, New York.

[13] Dupačová, J., Wets, R. J.-B., 1988. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. Ann. Statist. 16, 1517–1549.

[14] Ferguson, T. S., 1982. An inconsistent maximum likelihood estimate. J. Amer. Statist. Assoc. 77, 831–834.

[15] Geyer, C. J., 1994. On the convergence of Monte Carlo maximum likelihood calculations. J. Roy. Statist. Soc. Ser. B 56, 261–274.

[16] Good, I. J., Gaskins, R. A., 1971. Nonparametric roughness penalties for probability densities. Biometrica 58, 255–277.

[17] Hess, C., 1996. Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator. Ann. Statist. 24, 1298–1315.

[18] Hoffman-Jørgensen, J., 1992. Asymptotic likelihood theory. In Butković, D, Kraljević, H., Kurepa, S., Hoffman-Jørgensen, J., (Eds.), Functional Analysis III, Various Publications Series 40, Matematisk Institut, Aarhus Universitet, 5–192.

[19] Huber, P. J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1, Univ. California Press, Berkeley, 221–233.

[20] Kiefer, J., Wolfowitz, J., 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many identical parameters. Ann. Math. Statist. 27, 887–906.

[21] King, A. J., Wets, R. J.-B., 1991. Epi-consistency of convex stochastic programs. Stochastics Stochastics Rep. 34, 83–92.

[22] Klonias, V. K., 1982. Consistency of two nonparametric penalized likelihood estimators of the probability densities, Ann. Statist. 10, 811–824.

[23] Korf, L., Wets, R. J.-B., 2001. Random lsc functions: an ergodic theorem. Math. Oper. Res. 26, 421–445.

[24] Le Cam, L., 1953. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. In Neyman, J., Loève, M., Struve, O., (Eds.), Univ. California Publ. Statist. 1, Univ. California Press, Berkeley, 277–329.

[25] Le Cam, L., 1990. Maximum likelihood: an introduction. Internat. Statist. Rev. 58, 153–171.

[26] Perlman, M. D., 1972. On the strong consistency of approximate maximum likelihood estimators. Proc. Sixth Berkeley Symp. Math. Statist. Probab. 1, Univ. California Press, Berkeley, 263–281.

[27] Peškir, G., 1998. Consistency of statistical models described by families of reversed martingales. Probab. Math. Statist. 18, 289–318.

[28] Shiryaev, A. N., 1995. Probability, 2nd ed. Springer-Verlag, Berlin.

[29] Silverman, B. W., 1982. On the estimation of a probability density function by the maximum penalized likelihood method. Ann. Statist. 10, 795–810.

[30] Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20, 595–601.

[31] Wang, J.-L., 1985. Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics. Ann. Statist. 13, 932–946.

[32] Williams, D., 1991. Probability with Martingales. Cambridge Univ. Press, Cambridge.