Probab. Theory Relat. Fields (2007) 138:411–449
DOI 10.1007/s00440-006-0031-4

Donsker-type theorems for nonparametric maximum likelihood estimators

Richard Nickl

Received: 24 September 2005 / Revised: 14 September 2006 / Published online: 24 October 2006

Abstract Let $\mathcal{P}$ be a nonparametric probability model consisting of smooth probability densities and let $\hat{p}_n$ be the corresponding maximum likelihood estimator based on n independent observations each distributed according to the law $\mathbb{P}$. With $\hat{\mathbb{P}}_n$ denoting the measure induced by the density $\hat{p}_n$, define the stochastic process $\hat{v}_n : f \longmapsto \sqrt{n} \int f d(\hat{\mathbb{P}}_n - \mathbb{P})$ where f ranges over some function class $\mathcal{F}$. We give a general condition for Donsker classes $\mathcal{F}$ implying that the stochastic process $\hat{v}_n$ is asymptotically equivalent to the empirical process in the space $\ell^\infty(\mathcal{F})$ of bounded functions on $\mathcal{F}$. This implies in particular that $\hat{v}_n$ converges in law in $\ell^\infty(\mathcal{F})$ to a mean zero Gaussian process. We verify the general condition for a large family of Donsker classes $\mathcal{F}$. We give a number of applications: convergence of the probability measure $\hat{\mathbb{P}}_n$ to $\mathbb{P}$ at rate $\sqrt{n}$ in certain metrics metrizing the topology of weak(-star) convergence; a unified treatment of convergence rates of the MLE in a continuous scale of Sobolev-norms; $\sqrt{n}$-efficient estimation of nonlinear functionals defined on $\mathcal{P}$; limit theorems at rate $\sqrt{n}$ for the maximum likelihood estimator of the convolution product $\mathbb{P} * \mathbb{P}$.

Keywords Nonparametric maximum likelihood estimator · Uniform central limit theorem · Plug-in property · Differentiable functionals · Convolution products

R. Nickl
University of Vienna, Vienna, Austria

R. Nickl (✉)
Department of Mathematics, University of Connecticut, 196, Auditorium Road, Storrs, CT 06269-3009, USA
e-mail: nickl@math.uconn.edu

Mathematical Subject Classification (2000) Primary: 60F05 · 62G07;
Secondary: 62F12 · 46F05

1 Introduction

Let $X_1, ..., X_n$ be independent and identically distributed (i.i.d.) according to some law $\mathbb{P}$ on a Borel set $\Omega \subseteq \mathbb{R}$. Denote by $\mathbb{P}_n$ the usual empirical measure on Ω induced by the sample. If $\mathcal{P}$ is a given probability model consisting of smooth densities on Ω, the (nonparametric) maximum likelihood estimator (MLE) $\hat{p}_n$ is defined as the element of $\mathcal{P}$ for which the supremum $\sup_{p\in\mathcal{P}} \mathbb{P}_n \log p = \sup_{p\in\mathcal{P}} n^{-1} \sum \log p(X_i)$ is attained. Convergence properties of the MLE in the Hellinger distance (or metrics related to the L^2-norm) in the case where the set $\mathcal{P}$ is genuinely infinite-dimensional have been studied in [4,35,40,41]; see also the monograph [36]. The literature has focussed on *strong* limiting properties of these estimators, in particular, on (optimal) convergence rates of the estimation error measured in distances that typically dominate the total variation metric on the set (of probability measures corresponding to) $\mathcal{P}$. A *weak* limit theory seems also to be of interest, since it is typically the case that 'weak' (rather than 'strong') theorems are used for statistical inference. In particular, it is of interest whether or not the rate $\sqrt{n}$ occurring in the (uniform) central limit theorem for empirical measures can be achieved. Thus under *weak* limit theory we understand here that $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P})(\cdot)$ converges in law in the Banach space $\ell^\infty(\mathcal{F})$ of bounded functions on some class $\mathcal{F}$ of measurable functions where $\hat{\mathbb{P}}_n$ is the (random) measure induced by the maximum likelihood estimator $\hat{p}_n$. For concrete choices of $\mathcal{F}$, such a general result then delivers many special weak limit theorems of inferential importance as corollaries. To our knowledge, weak convergence properties of the nonparametric maximum likelihood estimator at this level of generality have not been derived in the literature. The only result in this direction that we are aware of is Kiefer and Wolfowitz [16]. Translating their particular result into the general terminology of the present paper, they show for $\mathcal{P}$ the set of monotone increasing (decreasing) densities on the positive half-line and $\mathcal{F}$ the set of indicator functions of all intervals of the form $(0, x]$, that the difference between the measure induced by the maximum likelihood estimator and the empirical measure is of order $o_{\mathbb{P}}(n^{-1/2})$ (in fact, even of smaller order) in the norm of $\ell^\infty(\mathcal{F})$. A result similar to the one in [16] was obtained recently in [29] for the MLE of a log-concave density.

In Sect.2 of this paper we consider the MLE over a set of probability densities $\mathcal{P}_t$ contained in a (fractional) Sobolev space of order t (defined over Ω). In Theorem 1 we show for $\mathcal{F} = \mathcal{U}_t$, a ball in the same Sobolev space that contains $\mathcal{P}_t$, that the $\ell^\infty(\mathcal{U}_t)$-norm of the difference between the probability measure $\hat{\mathbb{P}}_n$ induced by the MLE and the empirical measure $\mathbb{P}_n$ decreases at a certain rate faster than $n^{-1/2}$, the rate being in accordance with the one obtained in [16]. In particular, the random variable $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P})$ converges in law in $\ell^\infty(\mathcal{U}_t)$. In Theorem 2 we move on to give a general approximation condition (Condition 2) on function

classes $\mathcal{F}$ under which the $\ell^\infty(\mathcal{F})$-norm of the difference between $\hat{\mathbb{P}}_n$ and $\mathbb{P}_n$ decreases at rate $o_{\mathbb{P}}(n^{-1/2})$. Hence, if $\mathcal{F}$ is—in addition—also a Donsker class, $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P})$ converges in law in $\ell^\infty(\mathcal{F})$. In Proposition 1 we then verify Condition 2 for a large family of (Donsker) classes $\mathcal{F}$ consisting of smooth functions. We discuss in detail in Sect. 2.1 the connection with related results for kernel density estimators (smoothed empirical measures), as well as the relationship to the 'plug-in property' recently introduced by Bickel and Ritov [3].

In Sect. 3 we exploit the general results from Sect. 2 to obtain a number of interesting consequences. First, let $\mathfrak{P}(\Omega)$ denote the set of all (Borel-) probability measures on Ω. We give (fast) rates for $d(\hat{\mathbb{P}}_n, \mathbb{P}_n)$ in certain metrics d metrizing the weak(-star) topology on $\mathfrak{P}(\Omega)$, implying in particular $d(\hat{\mathbb{P}}_n, \mathbb{P}) = O_{\mathbb{P}}(n^{-1/2})$. Second, we give a unified treatment of the convergence rate of the MLE in a continuous scale of Sobolev-norms. Third, the above results are useful for establishing $\sqrt{n}$-asymptotic normality and efficiency of the (plug-in) MLE for a large class of nonlinear functionals Φ defined on $\mathcal{P}_t$. This improves substantially upon related work by [40], and can be applied, in particular, to the integral functionals considered in [2,5,15,17,18]. Finally, we derive limit theorems at rate $\sqrt{n}$ for the nonparametric maximum likelihood estimator of the convolution product $\mathbb{P} * \mathbb{P}$. Similar results have recently been obtained by [13,30] for kernel density estimators.

1.1 Notation and definitions

We collect here the main notation and definitions used in the paper. For $(B, \|\cdot\|)$ a normed space, B' denotes the topological dual space. The operator norm of a continuous linear functional L on B will be denoted by $\|L\|'$. For an arbitrary (non-empty) set M, let $\ell^\infty(M)$ denote the Banach space of bounded real-valued functions H on M normed by

$$\|H\|_{\infty,M} := \sup_{m \in M} |H(m)| .$$

We denote by $\mathcal{B}_S$ the Borel-σ-algebra of a topological space S. Throughout the paper, we shall at least assume that the set Ω satisfies $\varnothing \neq \Omega \in \mathcal{B}_{\mathbb{R}}$. The symbol $\mathsf{L}^\infty(\Omega)$ denotes the Banach space of $\mathcal{B}_\Omega$-*measurable* bounded real-valued functions on Ω normed by the usual sup-norm $\|\cdot\|_\infty$, and $\mathsf{C}(\Omega)$ denotes the closed subspace of bounded real-valued continuous functions on Ω with the induced norm.

For Ω an open set in $\mathcal{B}_{\mathbb{R}}$, we define Hölder and Lipschitz classes, that is, sets of the form

$$\mathcal{F}_{s,\infty,C} = \left\{ f \in \mathsf{C}(\Omega) : \|f\|_{s,\infty} = \sum_{\alpha=0}^{[s]} \left\|D^\alpha f\right\|_\infty + \sup_{x \neq y} \frac{\left|D^{[s]} f(x) - D^{[s]} f(y)\right|}{|x-y|^{s-[s]}} \leq C \right\} \tag{1}$$

where D^α denotes the (classical) derivative of order α, where $0 < C < \infty$, $s > 0$, and where $[s] = s - \{s\}$ with $[s]$ integer and $0 < \{s\} \leq 1$.

We shall furthermore denote by $\mathcal{L}^0(\Omega)$ the set of real-valued $\mathcal{B}_\Omega$-measurable functions on Ω. For $h \in \mathcal{L}^0(\Omega)$ and some Borel measure μ on Ω, we set $\mu h := \int_\Omega h \, \mathrm{d}\mu$ and $\|h\|_{r,\mu} := (\int_\Omega |h|^r \, \mathrm{d}\mu)^{1/r}$ for $1 \leq r \leq \infty$ (where $\|h\|_{\infty,\mu}$ denotes the μ-essential supremum of $|h|$). As usual, we denote by $\mathcal{L}^r(\Omega, \mu)$ the vector space of all $h \in \mathcal{L}^0(\Omega)$ that satisfy $\|h\|_{r,\mu} < \infty$. In accordance, $L^r(\Omega, \mu)$, denotes the corresponding Banach spaces of equivalence classes $[h]_\mu$ (modulo equality μ -almost everywhere), $h \in \mathcal{L}^r(\Omega, \mu)$. The symbol λ will always denote Lebesgue-measure on Ω, and we shall occasionally write $\lambda\,|\Omega$ to specify the underlying support.

The Sobolev spaces over some open set $\Omega \subseteq \mathbb{R}$ and for integer $m \geq 0$ are given by

$$\mathcal{W}_2^m(\Omega, \lambda\,|\Omega) := \{f \in \mathcal{L}^0(\Omega) : \|f\|_{m,2,\lambda|\Omega} < \infty\}, \tag{2}$$

where the Sobolev seminorm is given by $\|f\|_{m,2,\lambda|\Omega} = \sum_{0 \leq \alpha \leq m} \left\|D_w^\alpha f\right\|_{2,\lambda|\Omega}$. Here $D_w^\alpha f$ denotes the *weak* (or generalized) derivative of integer order α (see, e.g., [1], 1.62). We denote by $W_2^m(\Omega, \lambda)$ the Hilbert space of equivalence classes of functions $[f]_\lambda$ obtained by taking the quotient of $\mathcal{W}_2^m(\Omega, \lambda)$ w.r.t. the set $\{f \in \mathcal{L}^0(\Omega) : \|f\|_{m,2,\lambda} = 0\}$. For positive non-integer s, Sobolev spaces on an open set $\Omega \subseteq \mathbb{R}$ can be defined by interpolation: we set $W_2^s(\Omega, \lambda) := [W_2^m(\Omega, \lambda), L^2(\Omega, \lambda)]_\theta$ where $(1-\theta)m = s$ for m integer and $0 < \theta < 1$, and where the interpolation couple $[\cdot, \cdot]_\theta$ is defined in the usual way, cf., e.g., Definition 1.2.1 in [20]. The norm on the Hilbert space $W_2^s(\Omega, \lambda)$ is again denoted by $\|\cdot\|_{s,2,\lambda}$. The definition via interpolation is used in much of the literature, e.g., in ([1], 7.30–7.32), and is equivalent to other common definitions (e.g., the one in ([34], 3.4.2), cf. Parts 1 and 2 of Proposition 2). Clearly, $W_2^r(\Omega, \lambda) \subseteq W_2^s(\Omega, \lambda)$ holds for $r \geq s$ with continuous injection. For $s > 1/2$ and Ω a bounded C^∞-domain in $\mathbb{R}$ (for a definition see ([34], 3.2.1)), every $\left[f\right]_\lambda \in W_2^s(\Omega, \lambda)$ contains exactly one bounded continuous function (see Part 3 of Proposition 2 below); hence, in that case one can define the Hilbert space $\mathsf{W}_2^s(\Omega, \lambda) = \{f \in \mathsf{C}(\Omega) : [f]_\lambda \in W_2^s(\Omega, \lambda)\}$ again equipped with the norm $\|\cdot\|_{s,2,\lambda}$.

For a sequence of i.i.d. random variables $X_1, \ldots, X_n$ distributed according to the law $\mathbb{P}$ on Ω, define the empirical measure $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$. Given a subset $\mathcal{F}$ of $\mathcal{L}^0(\Omega)$, define the $\mathcal{F}$-indexed *empirical process* ν_n by

$$f \longmapsto \nu_n(f) := \sqrt{n}\,(\mathbb{P}_n - \mathbb{P})f = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - \mathbb{P}f) \quad (f \in \mathcal{F}). \tag{3}$$

We use the symbol $\rightsquigarrow_S$ to denote convergence in law of random elements in a metric space S in the generalized sense of Hoffmann-Jorgensen, see Chap. 3 in [8]. A function class $\mathcal{F} \subseteq \mathcal{L}^2(\Omega, \mathbb{P})$ is said to be $\mathbb{P}$-Donsker if it is $\mathbb{P}$-pregaussian and if $\nu_n \rightsquigarrow_{\ell^\infty(\mathcal{F})} \mathbb{G}$ where $\mathbb{G}$ is a zero-mean Gaussian process indexed by $\mathcal{F}$ with covariance function $\mathbb{P}((f - \mathbb{P}f)(g - \mathbb{P}g))$ for $f, g \in \mathcal{F}$, and with almost all

its sample paths bounded and uniformly continuous, see p. 94 in [8] for details. We note that ν_n need not be $\mathcal{B}_{\ell^\infty(\mathcal{F})}$ -measurable, but convergence in law of ν_n still implies $\|\nu_n\|_{\infty,\mathcal{F}} = O_{\mathbb{P}^*}(1)$ by Prohorov's theorem, where $\mathbb{P}^*$ denotes outer probability. [All random elements in this paper can be viewed as being defined on $(\Omega^\infty, \mathcal{B}_{\Omega^\infty}, \mathbb{P}^\infty)$. For real-valued random elements Z_n we use the notation $Z_n = O_{\mathbb{P}^*}(a_n)$ to denote $\limsup_{n\to\infty} \mathbb{P}^{\infty *}(|Z_n|/a_n > M) \to 0$ for $M \to \infty$, where $\mathbb{P}^{\infty *}$ is the outer measure corresponding to $\mathbb{P}^\infty$. A similar remark applies to the symbols $o_{\mathbb{P}^*}$, $O_{\mathbb{P}}$, and $o_{\mathbb{P}}$.]

2 Weak convergence of the MLE in $\ell^\infty(\mathcal{F})$

Let $\{X_i\}_{i=1}^n$ be i.i.d. according to the law $\mathbb{P}$ on the open set $\Omega \subseteq \mathbb{R}$. Unless otherwise stated, we shall assume throughout the rest of the paper that Ω is a bounded C^∞-domain; for a definition see ([34], 3.2.1). [We note that for the one-dimensional case considered here, this is tantamount to assuming that Ω is a finite union of bounded open intervals that are separated, that is, they are at a positive distance of each other. Nevertheless, we shall use the shorter term 'bounded C^∞-domain' in the sequel, also because it is the appropriate concept for generalization of our results to higher dimensions.] The (log-)likelihood function is given by

$$L_n(p) := \mathbb{P}_n \log p = \frac{1}{n}\sum_{i=1}^n \log p(X_i) \tag{4}$$

where the density function $p : \Omega \to \mathbb{R}$ belongs to the probability model

$$\mathcal{P}_t := \mathcal{P}_{t,\zeta,D}(\Omega) = \left\{ p \in \mathsf{W}_2^t(\Omega,\lambda) : \|p\|_{1,\lambda} = 1, \quad \inf_{x\in\Omega} p(x) \geq \zeta, \quad \|p\|_{t,2,\lambda} \leq D \right\}. \tag{5}$$

Here $t > 1/2$, $\zeta > 0$, and $0 < D < \infty$, are given constants specifying the model. We note that $\mathcal{P}_t$ consists of bounded continuous functions. The uniform lower bound ζ is common in the literature when considering maximum likelihood estimators. It is possible to generalize all subsequent results to many other function classes, see Remark 3.5.2.

The maximum likelihood estimator is defined as an element $\widehat{p}_n \in \mathcal{P}_t$ which satisfies

$$L_n(\widehat{p}_n) = \sup_{p\in\mathcal{P}_t} L_n(p). \tag{6}$$

As shown in Sect. 4.2, $\mathcal{P}_t$ is a compact subset of $\mathsf{C}(\Omega)$ and the function $L_n(\cdot)$ is continuous on $\mathcal{P}_t$ w.r.t. the sup-norm topology; hence, the supremum in (6) is attained, that is, an element $\widehat{p}_n$ satisfying (6) exists. Viewed as a map from Ω^n to the metric space $(\mathcal{P}_t, \|\cdot\|_\infty)$, the MLE $\widehat{p}_n$ can in fact (and will) be chosen to

be $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{(\mathcal{P}_t,\|\cdot\|_\infty)}$-measurable. [We note here once and for all that all results of this paper hold not only for a particular, but for any $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{(\mathcal{P}_t,\|\cdot\|_\infty)}$-measurable selection. For further details on measurability see Sect. 4.2.1.] It is possible to generalize our results to approximate MLEs which attain the maximum only up to a term which is of sufficiently small order. The case of sieved MLEs is somewhat different, see Remark 3.5.1.

We wish to derive a 'Donsker-type' theorem where the empirical measure is replaced by the measure induced by the maximum likelihood estimator defined in (6). To this end, for $\mathcal{F}$ some class of $\mathcal{B}_\Omega$-measurable real-valued functions, define the mapping

$$f \longmapsto \hat{\nu}_n(f) = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P})f \quad (f \in \mathcal{F}) \tag{7}$$

where the (random) measure $\hat{\mathbb{P}}_n$ is defined by

$$\hat{\mathbb{P}}_n(A) = \int_A \widehat{p}_n \mathrm{d}\lambda \quad (A \in \mathcal{B}_\Omega). \tag{8}$$

We wish to study the behaviour of $\hat{\nu}_n$ as a random element in $\ell^\infty(\mathcal{F})$; in particular, we wish to provide general conditions under which $\hat{\nu}_n \rightsquigarrow_{\ell^\infty(\mathcal{F})} \mathbb{G}$ holds, where $\mathbb{G}$ is a centered Gaussian process indexed by $\mathcal{F}$.

Condition 1 *The probability model $\mathcal{P}_t = \mathcal{P}_{t,\zeta,D}(\Omega)$ is given by* (5) *above where $\Omega \subseteq \mathbb{R}$ is a bounded C^∞-domain and where $t > 1/2$, and $\zeta > 0$, as well as $0 < D < \infty$ are given constants.* (*To ensure that $\mathcal{P}_{t,\zeta,D}(\Omega)$ is non-empty, we assume that $\zeta \leq \lambda(\Omega)^{-1} \leq D^2$.*)

1. *The random variables $X_1, \ldots, X_n$ are independent identically distributed according to the law $\mathbb{P}$ on $\mathcal{B}_\Omega$; in fact, they are the coordinate projections of the infinite product probability space $(\Omega^\infty, \mathcal{B}_{\Omega^\infty}, \mathbb{P}^\infty)$. The Radon–Nikodym derivative $\mathrm{d}\mathbb{P}/\mathrm{d}\lambda$ exists and is almost everywhere equal to an element $p_0 \in \mathcal{P}_t$.*
2. *The function p_0 satisfies the strict inequalities $\inf_{x\in\Omega} p_0(x) > \zeta$ and $\|p_0\|_{t,2,\lambda} < D$.*

Note that we are estimating the continuous representative p_0 of $\mathrm{d}\mathbb{P}/\mathrm{d}\lambda$. The condition $t > 1/2$ is crucial for many reasons, see Remark 3.5.3.

The following remark on Condition 1.2 is in order: when deriving weak convergence properties of M-estimators in the 'parametric' (i.e., finite-dimensional) case, it is standard to assume that the true parameter is in the interior of the parameter space (w.r.t. the Euclidean topology). The point here is that the parameter space contains a neighborhood of the true parameter that is an open set in the *same* topology in which the M-estimator is consistent, a fact that is central to the classical proof in the finite-dimensional case. This approach to proving asymptotic normality of M-estimators is not directly viable in the infinite-dimensional setup: there the usual requirement of (relative) compactness of the 'parameter space' in any norm-topology in which an M-estimator

is consistent prohibits the assumption that the true parameter is an interior point w.r.t. this norm topology. Condition 1.2 acts as a substitute for such an assumption. Note that it is related, but not identical to the (non-topological) concept of an *internal* point in the sense of V.1.6 in [10].

Taking this, as well as a number of other subtle issues, into account, it is possible to modify and extend the classical asymptotic normality proof for MLEs to obtain the following result. For the proof of the theorem as well as a discussion of our proof strategy we refer to Sect. 4.3.1. We note that in the following theorem ν_n and $\hat{\nu}_n$ take values in $\ell^\infty(\mathcal{U}_{t,B})$ since $\mathcal{U}_{t,B}$ is uniformly bounded (cf. Part 3 of Proposition 2).

Theorem 1 *Assume that Condition* 1 *is satisfied. For* $\mathcal{U}_{t,B} = \{f \in \mathsf{W}_2^t(\Omega,\lambda) : \|f\|_{t,2,\lambda} \leq B\}$, $0 < B < \infty$, *and* $\hat{\nu}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P})$ *we have that*

$$\left\|\hat{\nu}_n - \nu_n\right\|_{\infty,\mathcal{U}_{t,B}} = o_{\mathbb{P}}\left(n^{-(t-k)/(2t+1)}\right) \tag{9}$$

holds for every real $k > 1/2$. *Furthermore,*

$$\hat{\nu}_n \rightsquigarrow_{\ell^\infty(\mathcal{U}_{t,B})} \mathbb{G},$$

where $\mathbb{G}$ *is a mean zero Gaussian process indexed by* $\mathcal{U}_{t,B}$ *with covariance function* $\Psi(f,g) = \mathbb{P}((f - \mathbb{P}f)(g - \mathbb{P}g))$ *for* $f, g \in \mathcal{U}_{t,B}$, *and with almost all its sample paths bounded and uniformly continuous.*

Theorem 1 is similar in spirit to the main result in [16]. As mentioned in Sect. 1, they show for $\mathcal{P}$ the set of monotone increasing (decreasing) densities on the positive half-line and $\mathcal{F}$ the set of indicator functions of all intervals of the form $(0,x]$, that the quantity $\left\|\hat{\nu}_n - \nu_n\right\|_{\infty,\mathcal{F}}$ is—up to logarithmic terms—of the order $o_{\mathbb{P}}(n^{-1/6})$. This rate corresponds to the one obtained in case $t = 1$ in the above theorem. [But note that the above theorem does *not* imply the result by [16]]. We furthermore note that the Donsker class $\mathcal{U}_{t,B}$ featured in Theorem 1 is connected to the 'parameter space' $\mathcal{P}_t$ in the sense that both sets have the same smoothness index t, which is essential in the proof of that theorem. A similar connection also exists between the corresponding classes $\mathcal{P}$ and $\mathcal{F}$ in the set-up of [16]. Furthermore, Theorem 1 also contains the results on functional estimation in [40], Example 1 as a special case; see Corollary 4 below and Remark 3.5.4 for details.

The question now arises whether the above mentioned connection between $\mathcal{P}_t$ and $\mathcal{F}$ is intrinsically necessary for a weak convergence result to hold or whether it can be relaxed to some extent: as the index t increases, i.e., as the MLE is constrained to a smoother class of functions, one would expect the MLE $\hat{\mathbb{P}}_n$ to be 'farther away' from $\mathbb{P}_n$. At first sight, this intuition seems to be confirmed by Theorem 1 in that the set $\mathcal{U}_{t,B}$ on which the two processes are (asymptotically) close to each other becomes smaller as t increases. However, at the same time Theorem 1 shows that the rate at which the norm $\left\|\hat{\nu}_n - \nu_n\right\|_{\infty,\mathcal{U}_{t,B}}$

decreases becomes *faster* as t increases, raising the question whether one might be able to replace $\mathcal{U}_{t,B}$ by a general (Donsker) class $\mathcal{F}$ *independent* of t—e.g., by $\mathcal{U}_{s,B}$ for *arbitrary* $s > 1/2$—and still be able to obtain a weak convergence result for $\hat{\mathbb{P}}_n - \mathbb{P}$ at rate $n^{1/2}$. We shall answer this question in the positive in Theorem 3.

To this end we now formulate a general approximation condition for function classes $\mathcal{F}$, which will be sufficient to prove such a more general result in Theorem 2.

Condition 2 *Let $\mathcal{P}_t$ and p_0 be as in Condition 1, and let $\widehat{p}_n$ be the MLE. Let $\mathcal{F}$ be a (non-empty) subset of $\mathcal{L}^1(\Omega, \lambda)$. Assume that for every $f \in \mathcal{F}$, there exists a sequence $u_n(f)$ in $\mathsf{W}_2^t(\Omega, \lambda)$ such that*

$$a_n := \sup_{f \in \mathcal{F}} \left| \int_\Omega (\widehat{p}_n - p_0)(u_n(f) - f) \mathrm{d}\lambda \right| = o_{\mathbb{P}^*}(n^{-1/2})$$

holds as well as that

$$b_n := \sup_{f \in \mathcal{F}} \|u_n(f)\|_{t,2,\lambda} = O(n^{(t-k^*)/(2t+1)})$$

holds for some real $k^* > 1/2$. Assume further that

$$c_n := \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - \mathbb{P})(u_n(f) - f)| = o_{\mathbb{P}^*}(n^{-1/2}).$$

Sufficient conditions for and further discussion of Condition 2 will be given after Theorem 2. For the proof of the following theorem as well as a discussion of our proof strategy see Sect. 4.3.2.

Theorem 2 *Assume that Conditions 1 and 2 are satisfied. For $\hat{\nu}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P})$ we have that*

$$\left\| \hat{\nu}_n - \nu_n \right\|_{\infty,\mathcal{F}} \leq n^{1/2}(a_n + c_n) + b_n o_{\mathbb{P}^*}(n^{-(t-k)/(2t+1)}) \tag{10}$$

holds for every real $k > 1/2$; in particular,

$$\left\| \hat{\nu}_n - \nu_n \right\|_{\infty,\mathcal{F}} = o_{\mathbb{P}^*}(1). \tag{11}$$

If, in addition, $\mathcal{F}$ is a $\mathbb{P}$-Donsker class then also

$$\hat{\nu}_n \rightsquigarrow_{\ell^\infty(\mathcal{F})} \mathbb{G}, \tag{12}$$

where $\mathbb{G}$ is a mean zero Gaussian process indexed by $\mathcal{F}$ with covariance function $\Psi(f,g) = \mathbb{P}((f - \mathbb{P}f)(g - \mathbb{P}g))$ for $f, g \in \mathcal{F}$, and with almost all its sample paths bounded and uniformly continuous.

Note that in the context of the above theorem $\hat{\nu}_n$ is not guaranteed to always take its values in $\ell^\infty(\mathcal{F})$. However, since ν_n is a random element in $\ell^\infty(\mathcal{F})$ whenever $\mathcal{F}$ is a $\mathbb{P}$-Donsker class, the result (11) shows that $\hat{\nu}_n$ is then so on a set A_n whose complement has (outer) probability converging to zero as sample size increases. The result in (12) is hence to be interpreted accordingly.

We conjecture that Condition 2 covers many Donsker classes $\mathcal{F}$. In fact, it covers most Donsker classes of *smooth* functions as the following proposition, which is proved in Sect. 4.3.3, shows.

Proposition 1 *Let Ω be a bounded C^∞-domain and let $t > 1/2$ be given. Furthermore, let $s > 1/2$ be arbitrary. Then any bounded (non-empty) subset $\mathcal{F}$ of $\mathsf{W}_2^s(\Omega, \lambda)$ satisfies Condition 2 for the given t and for any $1/2 < k^* \leq \min(s, t)$. More precisely, if $s < t$ holds, Condition 2 is satisfied with $a_n = O_{\mathbb{P}^*}(n^{(-t-s)/(2t+1)})$, with $b_n = O(n^{(t-s)/(2t+1)})$, and with $c_n = O_{\mathbb{P}}(n^{-1/2+(k'-s)/(2t+1)})$ for every $1/2 < k' < s$. If $s \geq t$ holds, then $a_n = c_n = 0$ and $b_n = O(1)$.*

This proposition covers Hölder and Lipschitz classes $\mathcal{F}_{s,\infty,C}$ (cf. definition (1)) with $s > 1/2$. [This follows from the imbeddings used in the proof of Corollary 1.] By using imbedding theorems for function spaces on Ω, the proposition also covers bounded subsets of Besov- and Triebel spaces, including in particular, Sobolev spaces W_r^s with either $s > 1/2$ and $2 \leq r < \infty$ or with $s > 1/r$ and $1 \leq r < 2$. See, e.g., Sect. 3.3.1 in [34] for details.

Condition 2 requires the function class $\mathcal{F}$ to be sufficiently well-approximable by smooth functions. In particular, to obtain appropriate bounds for a_n and b_n in Condition 2 for a given class $\mathcal{F}$ one needs to examine its approximation-theoretic properties (together with a rate for the MLE). We note that such an approximation result will typically also imply that the L^2-norm of the approximation error $u_n(f) - f$ tends to zero (uniformly in f). As a consequence, the requirement for c_n in Condition 2 is then automatically satisfied, provided, e.g., the sets of approximation errors $\{u_n(f) - f : f \in \mathcal{F}\}$ are contained in some $\mathbb{P}$-Donsker class $\mathcal{H}$ that satisfies Pollard's or Ossiander's empirical CLT (see [38], p. 220f.).

Function classes $\mathcal{F}$ that are not well-approximable by smooth functions, for example $\mathcal{F} = \{f\}$, with $f \in \mathsf{L}^\infty(\Omega)$ arbitrary, are not covered by Theorem 2. [This seems to be a general phenomenon in the weak convergence theory of 'plug-in estimators' in the sense of [3], see the last but one paragraph in Sect. 2.1 below.]

Finally, Theorem 2 together with Proposition 1 give an affirmative answer to the question raised in the second paragraph following Theorem 1. This is summarized in Theorem 3 which generalizes Theorem 1. [Again, ν_n and $\hat{\nu}_n$ take values in $\ell^\infty(\mathcal{U}_{s,B})$ since $\mathcal{U}_{s,B}$ is uniformly bounded (cf. Part 3 of Proposition 2).]

Theorem 3 *Assume that Condition 1 is satisfied and that $s > 1/2$. For $\mathcal{U}_{s,B} = \{f \in \mathsf{W}_2^s(\Omega, \lambda) : \|f\|_{s,2,\lambda} \leq B\}$, $0 < B < \infty$, and $\hat{\nu}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P})$ we have*

$$\left\| \hat{\nu}_n - \nu_n \right\|_{\infty, \mathcal{U}_{s,B}} = o_{\mathbb{P}}\left(n^{-(\min(s,t)-k)/(2t+1)}\right) \tag{13}$$

for every real $k > 1/2$. Furthermore,

$$\hat{\nu}_n \rightsquigarrow_{\ell^\infty(\mathcal{U}_{s,B})} \mathbb{G},$$

where $\mathbb{G}$ is a mean zero Gaussian process indexed by $\mathcal{U}_{s,B}$ with covariance function $\Psi(f,g) = \mathbb{P}((f-\mathbb{P}f)(g-\mathbb{P}g))$ for $f, g \in \mathcal{U}_{s,B}$, and with almost all its sample paths bounded and uniformly continuous.

Proof Follows immediately from (10) in Theorem 2, Proposition 1, straightforward rate calculations, and Part 5 of Proposition 2. □

Theorem 3 shows that the difference between the MLE $\hat{\mathbb{P}}_n$ and the empirical measure $\mathbb{P}_n$ in $\ell^\infty(\mathcal{U}_{s,B})$ is always at least of order $o_{\mathbb{P}}(n^{-1/2})$ regardless of how large the smoothness index t of the underlying probability model $\mathcal{P}_t$ is. It is important to note that here—in contrast to Theorem 1—the indexing function class $\mathcal{F} = \mathcal{U}_{s,B}$ on which the processes are asymptotically close to each other (but not the rate in the above theorem) is *independent* of t. [One might ask whether $\min(s,t)$ can be replaced by s in (13). We do not know whether the appearance of $\min(s,t)$ in (13) is a genuine feature of MLEs or an artefact of our proof strategy.]

2.1 Discussion

Since $\mathcal{P}_t$ is a (genuinely) nonparametric model, the empirical measure $\mathbb{P}_n$ is an efficient estimator for $\mathbb{P}$ in the Banach space $\ell^\infty(\mathcal{F})$ as soon as $\mathcal{F}$ is a (universal) Donsker class (cf., e.g., [38], p. 420) Therefore, by Theorems 2 and 3, the MLE is also an efficient estimator for $\mathbb{P}$ in $\ell^\infty(\mathcal{F})$. This is achieved by showing that the MLE is asymptotically closer to $\mathbb{P}_n$ than to $\mathbb{P}$ in $\ell^\infty(\mathcal{F})$. Hence—as concluded also by Kiefer and Wolfowitz [16] in their more specific setup—from a pure efficiency point of view (in $\ell^\infty(\mathcal{F})$), there is little justification for using the MLE (or any other estimator) instead of $\mathbb{P}_n$. Does this imply that the conventional wisdom that MLEs *do* take advantage of additional information on the probability model—in the present context given by the information that $\mathbb{P}$ possesses a smooth density p_0 contained in $\mathcal{P}_t$—is unfounded in infinite-dimensional models?

When estimating parameters in infinite-dimensional spaces, the 'value' of additional information on the underlying probability model typically will depend on the metric (or rather, topology) w.r.t. which one assesses the properties of an estimator. In the previous paragraph, we only considered convergence in the metric of $\ell^\infty(\mathcal{F})$ for Donsker classes $\mathcal{F}$. Clearly, changing the class $\mathcal{F}$ amounts to changing the metric (or topology). As we shall argue in the subsequent paragraphs, an estimator that uses all information will typically have optimal properties in *all* relevant metrics *simultaneously*.

It is shown on the one hand in Theorem 2 that the MLE achieves the convergence rate $1/\sqrt{n}$ in the metric $\|\cdot\|_{\infty,\mathcal{F}}$ for certain Donsker classes $\mathcal{F}$. On the other hand, if one takes $\mathcal{F}$ equal to the unit ball in $L^2(\Omega,\lambda)$, then the MLE

achieves the minimax (over $\mathcal{P}_t$ with $t > 1/2$) rate of convergence $n^{-t/(2t+1)}$ in the norm $\|\cdot\|_{\infty,\mathcal{F}} = \|\cdot\|_{2,\lambda}$. [This follows from results in [4,35], see Corollary 3 and Proposition 6.] In contrast, the empirical measure $\mathbb{P}_n$ is certainly an inconsistent estimator in the L^2-metric. Corollaries 2 and 3 below highlight the superiority of the MLE over $\mathbb{P}_n$ in some detail: they show that the MLE performs optimal in a continuous range of metrics (induced by certain Sobolev norms), ranging from metrics for the weak topology on $\mathcal{P}_t$ to stronger metrics that dominate, e.g., the total variation norm on $\mathcal{P}_t$. This has interesting statistical implications discussed below.

Another interesting question would be whether the MLE achieves the rate of convergence $1/\sqrt{n}$ in $\ell^\infty(\mathcal{F})$ for $\mathbb{P}$ -pregaussian classes $\mathcal{F}$ that are *not* $\mathbb{P}$-Donsker, in which case $\mathbb{P}_n$ fails to be a ($\sqrt{n}$-) consistent estimator. It is indeed possible to construct smoothed empirical measures $\tilde{\mathbb{P}}_n$ (i.e., kernel or histogram-based estimators with non-standard bandwidths) so that $\sqrt{n}(\tilde{\mathbb{P}}_n - \mathbb{P})$ converges in law in the space $\ell^\infty(\mathcal{F})$ to a sample-continuous Gaussian process $\mathbb{G}$ over $\mathcal{F}$ for $\mathbb{P}$-pregaussian classes $\mathcal{F}$ that are *not* $\mathbb{P}$-Donsker. Results of this kind are proved in [26,27] under the assumption that $\mathbb{P}$ possesses a twofold differentiable density with compact support. We do not know whether such results can also be shown for the maximum likelihood estimator (but this is of course an interesting open question). From a practical point of view, we are not aware of many concrete examples for pregaussian classes that are *not* Donsker on the sample space $\mathbb{R}^d$ (or relevant subsets thereof). [We refer, however, to [21], where it is shown that balls in certain Besov spaces on $\mathbb{R}^d$ are pregaussian but not Donsker.]

In light of these facts we conjecture here that—in infinite dimensional models with convergent Hellinger-bracketing integral (e.g., $\mathcal{P}_t$ with $t > 1/2$)—the usual wisdom that MLEs use all information of the probability model is mirrored in the fact that the MLE achieves the minimax rate of convergence in most (if not all) metrics. We note that this simultaneous optimality property is *not* necessarily shared by other common density estimators. For example, while the nonstandard-bandwidth-kernel density estimators discussed above can be tuned to converge in law in $\ell^\infty(\mathcal{F})$ at rate $\sqrt{n}$ for Donsker (or even pregaussian) classes $\mathcal{F}$, this comes at the expense of a rate of convergence of the density estimator in stronger norms (e.g., in $\|\cdot\|_{2,\lambda}$) *which is slower than the minimax rate*. See [28,37,42] as well as [26,27] for results of this kind. Related results for density estimators based on series expansions are derived in [23].

It transpires from the recent paper Bickel and Ritov [3], that the simultaneous optimality of an estimator in various metrics is of statistical importance. Bickel and Ritov [3] defined an estimator $\tilde{\mathbb{P}}_n$ to possess the *uniform plug-in property* for the class of linear functionals arising from $f \in \mathcal{F}$ via $\phi_f(\cdot) = \int_\Omega f \mathrm{d}(\cdot)$ if

$$\sup_{f\in\mathcal{F}} \left|\sqrt{n}(\phi_f(\tilde{\mathbb{P}}_n) - \phi_f(\mathbb{P}))\right| = \sup_{f\in\mathcal{F}} \left|\sqrt{n}\int f \mathrm{d}(\tilde{\mathbb{P}}_n - \mathbb{P})\right| = O_{\mathbb{P}}(1) \qquad (14)$$

and if— *simultaneously*— $\tilde{\mathbb{P}}_n$ achieves the minimax rate of convergence over the underlying probability model in L^2-loss. It was already discussed in the previous

paragraph that the MLE achieves the minimax rate of convergence over $\mathcal{P}_t$ with $t > 1/2$ in L^2-loss. Consequently, Theorem 2 above implies that the MLE possesses the 'uniform plug-in property' for Donsker classes satisfying Condition 2.

Bickel and Ritov [3] already gave a number of statistical examples where one can take advantage of density estimators possessing the plug-in property. In Corollaries 4–6 below, we will provide some further examples in order to show how one may apply our results to efficiently estimate parameters that cannot be estimated via the empirical measure.

3 Some implications of Theorems 2 and 3

3.1 Convergence of the MLE in the weak topology on $\mathfrak{P}(\Omega)$

Let $\mathfrak{P}(\Omega)$ denote the set of all (Borel) probability measures on Ω. In this section we consider the weak topology on $\mathfrak{P}(\Omega)$. [That is, we view $\mathfrak{P}(\Omega)$ as a bounded subset of $\mathsf{C}(\Omega)'$ and equip it with the weak-star topology.] An application of the general Donsker-theorem for empirical processes gives convergence at rate $\sqrt{n}$ of $\mathbb{P}_n$ to $\mathbb{P}$ in certain metrics metrizing the weak topology on $\mathfrak{P}(\Omega)$, see, e.g., [12,33], and also [14]. Theorem 3 allows one to state similar results for the maximum likelihood estimator. Consider first the (dual) bounded Lipschitz metric on $\mathfrak{P}(\Omega)$ given by

$$\beta(\mu, \upsilon) = \sup_{f \in \mathcal{F}_{1,\infty,1}} \left| \int_\Omega f \mathrm{d}(\mu - \upsilon) \right| \tag{15}$$

for $\mu,\upsilon \in \mathfrak{P}(\Omega)$ where $\mathcal{F}_{1,\infty,1}$ is the unit ball in the space of bounded Lipschitz-functions on Ω (see (1)). As is well-known, the bounded Lipschitz metric β metrizes the weak topology on $\mathfrak{P}(\Omega)$ ([9], Proposition 11.3.2).

Corollary 1 *Assume that Condition 1 holds. We then have*

$$\beta(\hat{\mathbb{P}}_n, \mathbb{P}_n) = o_{\mathbb{P}}\big(n^{-1/2-(\min(1,t)-k)/(2t+1)}\big) \tag{16}$$

for every real $k > 1/2$. Furthermore,

$$\beta(\hat{\mathbb{P}}_n, \mathbb{P}) = O_{\mathbb{P}}(n^{-1/2}).$$

Next define the Sobolev-norm metric d_s on $\mathfrak{P}(\Omega)$ (for $s > 1/2$) by

$$d_s(\mu, \upsilon) := \sup_{f \in \mathcal{U}_{s,1}} \left| \int_\Omega f \mathrm{d}(\mu - \upsilon) \right|$$

for every $\mu,\upsilon \in \mathfrak{P}(\Omega)$, where $\mathcal{U}_{s,1} = \{f \in \mathsf{W}_2^s(\Omega,\lambda) : \|f\|_{s,2,\lambda} \leq 1\}$. [Note that $d_s(\mu,\upsilon)$ is always finite since $\mathcal{U}_{s,1}$ is uniformly bounded.] For $1/2 < s < 1$, this metric is stronger than the dual bounded Lipschitz metric β, that is,

$$\beta(\mu,\upsilon) \leq K d_s(\mu,\upsilon)$$

holds for every $\mu,\upsilon \in \mathfrak{P}(\Omega)$ and some constant K (possibly depending on s); see the proof of Corollary 1. We give the following metrization lemma which was proved in Theorem 2.2 of [12] for the slightly different case where Ω is a compact Riemannian manifold.

Lemma 1 *Let $\Omega \subseteq \mathbb{R}$ be a bounded C^∞-domain and suppose that $s > 1/2$. Then the metric d_s metrizes the topology of weak convergence on $\mathfrak{P}(\Omega)$.*

Corollary 2 *Assume that Condition 1 holds and that $s > 1/2$. We then have*

$$d_s(\hat{\mathbb{P}}_n, \mathbb{P}_n) = o_{\mathbb{P}}\big(n^{-1/2-(\min(s,t)-k)/(2t+1)}\big)$$

for every real $k > 1/2$. Furthermore,

$$d_s(\hat{\mathbb{P}}_n, \mathbb{P}) = O_{\mathbb{P}}(n^{-1/2}).$$

3.2 Convergence rates of the MLE in Sobolev norms

Whereas the maximum likelihood estimator $\hat{\mathbb{P}}_n$ does not improve upon $\mathbb{P}_n$ in the the metric d_s for $s > 1/2$, it does so in stronger topologies. To elucidate this fact, it is interesting to extend the definition of the Sobolev-norm metric d_s in the previous subsection to the case $0 \leq s < \infty$. The functional

$$\|\cdot\|_{-s,2,\lambda} = \sup_{[f]_\lambda \in U_{s,1}} \left| \int_\Omega (\cdot) f \, \mathrm{d}\lambda \right| \tag{17}$$

with $U_{s,1}$ the unit ball of $W_2^s(\Omega,\lambda)$ $(0 \leq s < \infty)$ induces a norm on $L^2(\Omega,\lambda)$. [In fact, $\|\cdot\|_{-s,2,\lambda}$ is just the restriction of the operator norm of the dual space $(W_2^s(\Omega,\lambda))'$ to $(L^2(\Omega,\lambda))' = L^2(\Omega,\lambda)$.] By Theorem 3, the convergence rate of the MLE in the norm $\|\cdot\|_{-s,2,\lambda}, s > 1/2$, is of order $1/\sqrt{n}$. On the other hand, the convergence rate of the MLE in the norm $\|\cdot\|_{0,2,\lambda} = \|\cdot\|_{2,\lambda}$ is of order $n^{-t/(2t+1)}$ by results due to [4,35], see also Proposition 6 below. The following corollary shows that these seemingly unrelated convergence rates in the norms $\|\cdot\|_{-s,2,\lambda}$, $s > 1/2$, and $\|\cdot\|_{2,\lambda}$ are in fact related by a 'continuous transition' of rates of convergence in intermediate Sobolev norms. [Recall that $\|\cdot\|_{r,2,\lambda}$ for $r \geq 0$ was defined after (2) above.]

Corollary 3 *Assume that Condition 1 holds and let $\delta > 0$ be arbitrary. We then have*

$$\left\|\hat{p}_n - p_0\right\|_{r,2,\lambda} = \begin{cases} O_{\mathbb{P}}(1) & \text{if } r = t \\ O_{\mathbb{P}^*}(n^{-(t-r)/(2t+1)}) & \text{if } t > r \geq 0 \\ O_{\mathbb{P}}(n^{-(t-r)/(2t+1)+\delta}) & \text{if } 0 > r \geq -1/2 \\ O_{\mathbb{P}}(n^{-1/2}) & \text{if } -1/2 > r \end{cases}.$$

For $r = 0$, the rate of convergence is well-known to be best possible in the minimax sense. The same can be shown to be true in case of integer $r > 0$, cf. the results in [32], and it is reasonable to expect this also for noninteger $r \in [-1/2, t)$ (possibly up to the δ-term). [For measurability issues in case $t > r \geq 0$ see the discussion following Proposition 6 below.] In case of the empirical measure $\mathbb{P}_n$ with $-1/2 > r$, we have $\sup_{f \in \mathcal{U}_{-r,1}} \left|\int_\Omega f \mathrm{d}(\mathbb{P}_n - \mathbb{P})\right| = O_{\mathbb{P}}(n^{-1/2})$ where $\mathcal{U}_{-r,1}$ is the unit ball of $\mathsf{W}_2^{-r}(\Omega, \lambda)$, but for $r \geq 1/2$, the norm $\|\cdot\|_{r,2,\lambda}$ cannot even be applied to $\mathbb{P}_n$. In fact, in case $r \geq 1/2$, $\sup_{f \in \mathcal{U}_{-r,1}} \left|\int_\Omega f \mathrm{d}(\mathbb{P}_n - \mathbb{P})\right| = \infty$ a.s. can be shown to hold for *any* set $\mathcal{U}_{-r,1}$ of representatives of elements of $U_{-r,1}$ by a similar reasoning as in Theorem 7 in [21].

3.3 Estimation of functionals defined on $\mathcal{P}_t$

Many statistical problems can be formulated as the problem of estimating a (given) functional Φ defined on some set of probability measures. An asymptotic estimation theory for this framework was first established by [39], who considered functionals defined on cumulative distribution functions. Alternatively, one can consider functionals Φ defined on sets of probability *densities*. We wish to apply Theorem 2 (and Corollary 3) in this context. First, we consider general functionals defined on (open subsets of) the Banach spaces $W_2^r(\Omega, \lambda)$ with arbitrary $-\infty < r < t$. Here we use the convention that $W_2^r(\Omega, \lambda) = (W_2^{-r}(\Omega, \lambda))'$ in case of negative r. [It is easy to see that $L^2(\Omega, \lambda)$ and hence $\mathcal{P}_t$ is contained in $W_2^r(\Omega, \lambda)$ for negative r if one views elements f of $L^2(\Omega, \lambda)$ as linear functionals $\phi_f(\cdot) = \int_\Omega (\cdot) f \mathrm{d}\lambda$ acting on $W_2^{-r}(\Omega, \lambda)$ via integration, see also Sect. 2.2 in [22].] This includes the important case where the functional is defined on $L^2(\Omega, \lambda)$ (corresponding to $r = 0$). The following corollary provides general conditions such that $\Phi(p_0)$ is efficiently estimable at rate $\sqrt{n}$ by the plug-in MLE.

Corollary 4 *Suppose that Condition 1 holds. Let A be an open subset of $W_2^r(\Omega, \lambda)$ with $-1/2 \leq r < t$ containing p_0. Let $\Phi : A \to \mathbb{R}$ be a given real-valued functional. Assume that Φ is Fréchet differentiable at the point $p_0 \in A$ with Fréchet derivative $D\Phi(p_0)$ and suppose that*

$$|\Phi(p_0 + h) - \Phi(p_0) - D\Phi(p_0)(h)| = O(\|h\|_{r,2,\lambda}^{\omega}) \tag{18}$$

holds for all $h \in A$ and some $\omega > (2t+1)/2(t-r)$. Assume further that $\sqrt{n} D\Phi(p_0)(\hat{p}_n - p_0) = \hat{\nu}_n(u_{\Phi,\mathbb{P}})$ holds for some $u_{\Phi,\mathbb{P}} \in \mathcal{L}^2(\Omega, \mathbb{P})$ [which is,

e.g., satisfied if $D\Phi(p_0) \in (L^2(\Omega,\lambda))'$] *and that* $\mathcal{F} = \{u_{\Phi,\mathbb{P}}\}$ *satisfies Condition 2. We then have*

$$\sqrt{n}(\Phi(\hat{p}_n) - \Phi(p_0)) \rightsquigarrow_{\mathbb{R}} N(0, \|u_{\Phi,\mathbb{P}} - \mathbb{P}u_{\Phi,\mathbb{P}}\|_{2,\mathbb{P}}^2). \tag{19}$$

We note that, if Φ is Fréchet differentiable at $p_0 \in W_2^r(\Omega,\lambda)$ for some $r \leq 0$, a $L^2(\Omega,\lambda)$-Riesz-representer $u_{\Phi,\mathbb{P}}$ of $D\Phi(T_{\mathbb{P}})$ always exists. Note further that in the remaining cases $-\infty < r < -1/2$ not covered by the corollary, the result (19) follows from a standard delta method argument under the sole condition that Φ is Hadamard differentiable at the point p_0 in A with Hadamard derivative $D\Phi(p_0) \in (W_2^r(\Omega,\lambda))'$, since in this case the MLE-process $\hat{v}_n$ converges in law in $(W_2^r(\Omega,\lambda))'$ by Theorem 3. [Clearly, this argument would also work for the empirical process v_n.]

In case $r = 0$, it is easily seen how Corollary 4 exploits the virtues of Bickel and Ritov's [3] plug-in property: density estimators $\tilde{p}_n$ that possess this property (such as the MLE $\hat{p}_n$) achieve optimal convergence rates for $\|\tilde{p}_n - p_0\|_{2,\lambda}$. This admits the minimal condition $\omega > (2t+1)/2t$ in (18) necessary to imply that the remainder term $\|h\|_{2,\lambda}^{\omega} = \|\tilde{p}_n - p_0\|_{2,\lambda}^{\omega}$ is of stochastic order $o_{\mathbb{P}}(n^{-1/2})$. Simultaneously, the 'plug-in property' guarantees convergence in law of the linearization term $\sqrt{n}D\Phi(p_0)(\tilde{p}_n - p_0)$ for a large class of functionals Φ.

Corollary 4 substantially generalizes upon results by [40], see Remark 3.5.4 for more discussion. An example to which Corollary 4 can be applied is the entropy functional. Another example is integrated squared density derivatives, which were considered, e.g., in [2,5,15,17,18]. These authors are interested in the parameter $\Phi(p_0) = \|D^{\alpha}p_0\|_{2,\lambda}^2$ where D^{α} denotes the classical (Fréchet-) differential operator of order α. We consider weak derivatives D_w^{α}. [Clearly, $D_w^{\alpha}p_0$ coincides with the Fréchet-derivative $D^{\alpha}p_0$ if the latter exists. Otherwise, considering weak derivatives is more general.]

Corollary 5 *Let* α *be a nonnegative integer. Suppose that Condition 1 holds and that* $t - 2\alpha - 1/2 > 0$ *is satisfied. Assume further that in case* $\alpha \geq 1$ *the condition* $\lim_{x\in\Omega, x\to x^*} D_w^j p_0(x) = 0$ *holds for every point* x^* *in the boundary of* Ω *and for* $\alpha \leq j \leq 2\alpha - 1$. *Then*

$$\sqrt{n}\left(\|D_w^{\alpha}\hat{p}_n\|_{2,\lambda}^2 - \|D_w^{\alpha}p_0\|_{2,\lambda}^2\right) \rightsquigarrow_{\mathbb{R}} N\left(0, \|g - \mathbb{P}g\|_{2,\mathbb{P}}^2\right)$$

with $g = 2D_w^{2\alpha}p_0$.

Observe that $[D_w^j p_0] \in W_2^{t-j}(\Omega,\lambda)$ and $t - j > 1/2$ hold for $\alpha \leq j \leq 2\alpha - 1$. Hence, a continuous representative of $D_w^j p_0$ exists, and the condition in the corollary involving the limit of the weak derivative of p_0 refers to this representative. These boundary conditions parallel similar assumptions used in [5,17,18]).

We note that $\|D_w^{\alpha}\hat{p}_n\|_{2,\lambda}^2$ is asymptotically efficient, that is, its limiting variance can be shown to achieve the semiparametric lower variance bound for

estimating $\|D_w^\alpha p_0\|_{2,\lambda}^2$. We further note that, in light of the results in [2,5,17,18], one may also be interested in the case $1/4 < t - 2\alpha \leq 1/2$. We conjecture that the rate $\sqrt{n}$ is *not* attained by the plug-in MLE for these parameters. We also note that the proof of Corollary 5 can be adapted to deal with the more general class of integral functionals considered in [5,17,18].

3.4 Estimation of $\mathbb{P} * \mathbb{P}$

Given an i.i.d. sample from a random variable taking values in $\mathbb{R}$ with density φ, [30] as well as [13] constructed kernel density estimators $\tilde{\varphi}_n$ for φ such that the convolution products $\tilde{\varphi}_n * \tilde{\varphi}_n - \varphi * \varphi$ converge in law at rate $\sqrt{n}$ in $L^p(\mathbb{R}, \lambda)$ spaces ($1 \leq p \leq \infty$). Such results are of interest, e.g., for estimating the density of a moving average process $y_t = \sigma \epsilon_{t-1} + \epsilon_t$ with $(\epsilon_t)_{t=1}^n$ i.i.d. according to the density φ. See also [11] for further statistical applications. We show that similar results can be proved for the maximum likelihood estimator by using our Theorem 3.

Let X be a random variable with unknown density $p_0 \in \mathcal{P}_{t,\zeta,D}((0,1))$ satisfying, in addition, $\lim_{x\to 0} p_0(x) = \lim_{x\to 1} p_0(x) < \infty$. [Note that any $p \in \mathcal{P}_{t,\zeta,D}((0,1))$ with $t > 1/2$ is uniformly continuous on $(0,1)$ by Part 5 of Proposition 2 below.] Then p_0 can be viewed as a continuous function on the one-dimensional torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, and the convolution $p_0 * p_0$ is always well defined. Given an i.i.d. sample of size n from X, we construct the maximum likelihood estimator $\hat{p}_n$ over $\mathcal{P}_{t,\zeta,D}((0,1))$. Setting $\hat{p}_n(0) = \lim_{x\to 1} \hat{p}_n(x) < \infty$, the function $\hat{p}_n$ can be viewed as an element of the space of bounded Borel-measurable (and hence also integrable) functions on $\mathbb{T}$. The convolution product $\hat{p}_n * \hat{p}_n$ is then again well defined. In particular, $\hat{p}_n * \hat{p}_n$ is contained, for every n, in $\mathsf{C}((0,1))$.

Corollary 6 *Suppose that Condition 1 holds with* $\Omega = (0,1)$ *and that* $\lim_{x\to 0} p_0(x) = \lim_{x\to 1} p_0(x)$ *is satisfied. We then have that* $\sqrt{n}(\hat{p}_n * \hat{p}_n - p_0 * p_0)$ *converges in law in the space* $\mathsf{C}((0,1))$.

It can be furthermore shown that the limiting variable equals the one of $2\sqrt{n}(\mathbb{P}_n - \mathbb{P}) * \mathbb{P}$, which is easily seen to be mean zero Gaussian. [Note however, that $\mathbb{P}_n * \mathbb{P}$ is an infeasible 'estimator' and the naive plug-in estimator $\mathbb{P}_n * \mathbb{P}_n$ is not even contained in $\mathsf{C}((0,1))$ for any n.] Clearly, convergence in law in $\mathsf{C}((0,1))$ implies convergence in law in $L^p((0,1),\lambda)$ for $1 \leq p \leq \infty$. Also, the proof method of the corollary easily generalizes to estimation of the density of $X + Y$, where X and Y are independent random variables on $(0,1)$, see Sect. 4.1 in [22] for more details.

3.5 Remarks and extensions

We collect some remarks on the results of the paper in this section.

1. (*Other density estimators*) Suppose the MLE defined by (6) were replaced by a sieved maximum likelihood estimator where the maximization is undertaken over a (suitably fast) growing sequence of (possibly finite-dimensional)

compact subsets $\mathcal{P}_{t,H(n)}$ of $\mathcal{P}_t$ where $\mathcal{P}_t$ is contained in the closure (in some relevant topology) of $\bigcup_{n=1}^{\infty} \mathcal{P}_{t,H(n)}$. Our proof strategy needs considerable (but natural) adaptation in this case, see [23]. Likewise, it seems reasonable to expect that similar weak convergence results can be proved for regularization MLEs. It is of course of interest to investigate whether results of this kind can be proved for other density estimators. We refer to the discussion in Sect. 2.1, and also to [23] where similar results are proved for trigonometric series estimators.

2. (*Parameter space*) The results in this paper are given for the parameter space $\mathcal{P}_t$ contained in the Sobolev–Hilbert space $\mathsf{W}_2^t(\Omega, \lambda)$ on the bounded set $\Omega \subseteq \mathbb{R}$. They can be generalized to the spaces $\mathsf{W}_p^t(\Omega, \lambda)$ with $1 < p < \infty$ and real-valued $t > d/p$ where Ω is a bounded C^∞-domain in $\mathbb{R}^d$ quite straightforwardly. Similarly, they can be generalized to Hölder, Besov or Triebel function classes (defined over such domains). Also, the case where additional restrictions are being imposed on the parameter space $\mathcal{P}_t$ is of interest: as long as these restrictions specify subsets of $\mathcal{P}_t$ that are convex and closed (in the sup-norm topology), we expect no major difficulties in proving results similar to Theorems 1–3.

3. (*Requirement of* $t > 1/2$) The condition $t > 1/2$ is necessary in many respects. It implies containment of $\mathcal{P}_t$ in the space of bounded continuous functions which is used throughout the proofs. More importantly, it delivers asymptotic equicontinuity of the empirical process indexed by $\mathcal{P}_t$, which implies the—in many ways essential—optimal convergence rates in Proposition 6. [It is known that for $t \leq 1/2$, minimum contrast estimators such as the MLE do not necessarily achieve optimal convergence rates, see Sect. 4 in [4].] The condition $t > 1/2$ would also be necessary to show *uniform* (in $\mathcal{P}_t$) *consistency* at rate $\sqrt{n}$ of the MLE (e.g., for estimating certain functionals $\Phi(p_0)$), which is necessary to make confidently inferential use of asymptotic results. It is well known that even for fixed functionals Φ and 'too large' nonparametric models (for example $\mathcal{P}_t$ with $t = 0$), *no* uniformly consistent estimators exist, the minimax risk converges to infinity and the rate $\sqrt{n}$ is unattainable by any estimator: see, e.g., [7] or Sect. 3 in [19] for a general study of such problems.

4. (*Corollary 4*) The asymptotic distribution of nonlinear real-valued functionals of the nonparametric MLE has also been considered in the paper Wong and Severini [40], who give high-level conditions on general (not necessarily density-) MLEs. In Example 1, [40] consider the case of (log-)density estimation over a parameter space of (log-)densities constrained by a Lipschitz condition of order $t \geq 2$. [Roughly speaking, they assume $\log p_0 \in \mathcal{F}_{t,\infty,C}$ for $t \geq 2$, cf. (1) above.] Wong and Severini [40] show $\sqrt{n}$-asymptotic normality of the plug-in MLE $\Phi(\hat{p}_n) - \Phi(p_0)$ if $\Phi : L^2(\Omega, \lambda) \to \mathbb{R}$ is twice-Frechet differentiable, and if, in addition, the Riesz-representer $u_{\Phi,\mathbb{P}}$ of the first derivative of Φ is orthogonal to p_0 and contained in a Lipschitz space of order $t \geq 2$. These conditions are much stronger than those in Corollary 4 (with $r = 0$). In particular, Wong and Severini [40] require that the order of the smoothness constraint t on the Riesz-representer has to be equal to the order of the smoothness constraint on the class of admissible densities. This undesirable dependence of the class

of admissible functionals Φ on the 'parameter space' $\mathcal{P}_t$ in [40] result is highly restrictive in applications. It is removed by Corollary 4 above.

4 Proofs and preliminary results

We need to establish a number of preliminary results before proving Theorems 1 and 2 as well as Proposition 1 in Sect. 4.3 below. Section 4.4 contains the proofs for Sect. 3.

The following proposition summarizes some facts on Sobolev spaces that will be used throughout the proofs. For $s \geq 0$, denote by $W_2^s(\mathbb{R}, \lambda\,|\mathbb{R})\,|\Omega$ the Banach space of restrictions to Ω of elements of $W_2^s(\mathbb{R}, \lambda\,|\mathbb{R})$ (which was defined after (2)) normed by

$$\|f\|_{s,2,\Omega} = \inf\{\|g\|_{s,2,\lambda|\mathbb{R}} : [g]_{\lambda|\mathbb{R}} \in W_2^s(\mathbb{R}, \lambda\,|\mathbb{R}) : [g\,|\Omega]_{\lambda|\Omega} = [f]_{\lambda|\Omega}\}. \quad (20)$$

Let F denote the usual Fourier–Plancherel transform acting on $L^2(\mathbb{R}, \lambda\,|\mathbb{R})$ scaled with $(2\pi)^{-1/2}$. Define for real $s \geq 0$

$$H^s(\mathbb{R}, \lambda\,|\mathbb{R}) = \{[h]_{\lambda|\mathbb{R}} \in L^2(\mathbb{R}, \lambda\,|\mathbb{R}) : \|h\|_{\wedge,s,2,\lambda|\mathbb{R}} := \left\|\langle u\rangle^s Fh\right\|_{2,\lambda|\mathbb{R}} < \infty\}$$

with the notation $\langle u\rangle^s = (1 + |u|^2)^{s/2}$ (and where we use the obvious generalization of $\|f\|_{2,\lambda|\mathbb{R}}$ for complex-valued functions f).

Proposition 2 *Let Ω be a bounded C^∞-domain and let $s \geq 0$.*

1. *We have $W_2^s(\mathbb{R}, \lambda\,|\mathbb{R}) = H^s(\mathbb{R}, \lambda\,|\mathbb{R})$ and the norms $\|\cdot\|_{s,2,\lambda|\mathbb{R}}$ and $\|\cdot\|_{\wedge,s,2,\lambda|\mathbb{R}}$ are equivalent.*
2. *We have $W_2^s(\Omega, \lambda) = W_2^s(\mathbb{R}, \lambda\,|\mathbb{R})\,|\Omega$ and the norms $\|\cdot\|_{s,2,\lambda|\Omega}$ and $\|\cdot\|_{s,2,\Omega}$ are equivalent.*
3. *Let $s > 1/2$. The imbeddings $W_2^s(\Omega, \lambda) \hookrightarrow \mathsf{C}(\Omega)$ as well as $\mathsf{W}_2^s(\Omega, \lambda) \hookrightarrow \mathsf{C}(\Omega)$ hold. In particular,*

$$\|g\|_\infty \leq C_s \|g\|_{s,2,\lambda}$$

holds for all $g \in \mathsf{W}_2^s(\Omega, \lambda)$ with imbedding constant $0 < C_s < \infty$.
4. *If $t > 1/2$, the set $\mathcal{P}_t$ defined in (5) is contained in $\{f \in \mathsf{C}(\Omega) : \zeta \leq f(x) \leq C_t D$ for all $x \in \Omega\}$.*
5. *Let $s > 1/2$ and let $\mathcal{U}$ be a bounded subset of $\mathsf{W}_2^s(\Omega, \lambda)$. Then $\mathcal{U}$ is uniformly equicontinuous on Ω. Furthermore, $\mathcal{U}$ is a $\mathbb{P}$ -Donsker class for every Borel probability measure $\mathbb{P}$ on Ω.*
6. *Let $s > 1/2$. $\mathsf{W}_2^s(\Omega, \lambda)$ is a multiplication algebra, that is, $\|fg\|_{s,2,\lambda} \leq M \|f\|_{s,2,\lambda} \|g\|_{s,2,\lambda}$ holds for some positive finite constant $0 < M < \infty$ and all $f, g \in \mathsf{W}_2^s(\Omega, \lambda)$.*

Proof These results are known and we only collect references. Part 1 follows from Theorem 1.7.1 in [20]. Part 2 follows from Part 1 and Theorems 1.9/1-2

in [20], noting that any bounded C^∞-domain satisfies Condition 7.10 in that monograph. The first imbedding of Part 3 is proved, e.g., in Theorem 1.9.8 in [20], which immediately implies the second imbedding and also Part 4 of the proposition by definition of $\mathcal{P}_t$. For Part 5, we infer uniform equicontinuity of $\mathcal{U}$ from 2.7.1/12 in [34]. Furthermore, the $L^2(\mathbb{P})$-bracketing metric entropy of any bounded subset $\mathcal{U}$ of $\mathsf{W}_2^s(\Omega,\lambda)$ is seen to be of order $\varepsilon^{-1/s}$ for every $\mathbb{P}\in\mathfrak{P}(\Omega)$ (where $\varepsilon\to 0$ denotes the bracket-size) upon noting that $W_2^s(\mathbb{R},\lambda\,|\mathbb{R})$ coincides with the Besov space $B_{22}^s(\mathbb{R},\lambda\,|\mathbb{R})$ defined in [24] and upon using Part 2 of Corollary 2 in [24] with $\beta=0, d=1, \mu=\mathbb{P}, r=2, p=q=2$. This implies that for $s>1/2$, the set $\mathcal{U}$ is $\mathbb{P}$-Donsker for every $\mathbb{P}\in\mathfrak{P}(\Omega)$ in view of Ossiander's CLT (Theorem 7.2.1 in [8]). Finally, $\mathsf{W}_2^s(\Omega,\lambda)$ is a multiplication algebra for $s>1/2$, since $W_2^s(\mathbb{R},\lambda\,|\mathbb{R})$ is one by ([34], 2.8.3) and by using Part 2 together with the fact that the restricted norm inherits multiplicativity. □

4.1 Differential calculus and limit theory for the likelihood function

We shall denote by $L_{(i)}(p)$ the likelihood per observation, that is $L_{(i)}(p)=\log p(X_i)$. We will derive the Fréchet derivatives of the likelihood function $p\longmapsto L_n(p)=n^{-1}\sum_{i=1}^n L_{(i)}(p)$ as well as of its limiting function $p\longmapsto \mathbb{P}L_{(i)}(p)=\int_\Omega \log p(x)\mathrm{d}\mathbb{P}(x)$ both viewed as mappings defined on a suitable open subset $\mathcal{V}$ of the Banach space $\mathsf{L}^\infty(\Omega)$. This is convenient as the set $\mathcal{P}_t$ will be seen to be contained in this open set. Recall that first derivatives are to be understood as elements of $\mathsf{L}^\infty(\Omega)'$, whereas second derivatives are continuous bilinear, real-valued functionals defined on $\mathsf{L}^\infty(\Omega)\times\mathsf{L}^\infty(\Omega)$ (with obvious extension for higher derivatives).

Proposition 3 *For $\Omega\in\mathcal{B}_\mathbb{R}$, let $\mathcal{V}=\{d\in\mathsf{L}^\infty(\Omega): d(x)>\zeta/2$ for all $x\in\Omega\}$ where $0<\zeta<\infty$. For $f_1,...,f_\alpha\in\mathsf{L}^\infty(\Omega)$, $\alpha\geq 1$, the multilinear mapping representing the α-th Fréchet-derivative of $L_n:\mathcal{V}\to\mathbb{R}$ at the point $d\in\mathcal{V}$ is given by*

$$D^\alpha L_n(d)(f_1,\ldots,f_\alpha)=n^{-1}(\alpha-1)!(-1)^{\alpha-1}\sum_{i=1}^n d^{-\alpha}(X_i)f_1(X_i)\cdots f_\alpha(X_i).$$

*Furthermore, the multilinear mapping representing the α-*th *Fréchet-derivative of $\mathbb{P}L_{(i)}(\cdot)$ at the point $d\in\mathcal{V}$ is given by*

$$\begin{aligned}D^\alpha\mathbb{P}L_{(i)}(d)(f_1,\ldots,f_\alpha) &= \mathbb{P}D^\alpha L_{(i)}(d)(f_1,\ldots,f_\alpha)\\ &=(\alpha-1)!(-1)^{\alpha-1}\int_\Omega d^{-\alpha}f_1\cdots f_\alpha\mathrm{d}\mathbb{P}.\end{aligned}$$

Proof For the first part it is sufficient to consider the likelihood per observation $L_{(i)}(p)=\log p(X_i)$ which is the composite mapping consisting of the logarithm and the evaluation map δ_{X_i} on $\mathsf{L}^\infty(\Omega)$. Since $\delta_{X_i}\in(\mathsf{L}^\infty(\Omega))'$, we

have $D\delta_{X_i}(d)(f) = \delta_{X_i}(f)$ for every $f \in \mathsf{L}^\infty(\Omega)$. Since $\delta_{X_i}(d) > \zeta/2$ holds for all $d \in \mathcal{V}$, and since the logarithm is differentiable on $\mathbb{R}^+$ we have by the chain rule on Banach spaces (8.2.1 in [6]) that

$$D(\log d(X_i))(f) = D((\log \circ \delta_{X_i})(d))(f) = d^{-1}(X_i)f(X_i).$$

By a similar reasoning, we have for the second derivative

$$D^2(\log d(X_i))(f,g) = -d^{-2}(X_i)f(X_i)\,g(X_i)$$

by using 8.12.1 in [6], and this reasoning is easily iterated to give the α-th derivative. The second part of the proposition follows from the fact that differentiation and integration of $L_{(i)}(\cdot)$ can be interchanged as a consequence of Proposition 4 below. We verify the conditions of this proposition with $E = \mathsf{L}^\infty(\Omega)$, $V = \mathcal{V}$, $\mu = \mathbb{P}$, and $f(v,s) = \log d(X_i)$ for the first derivative; higher order derivatives following in a similar manner. Notice first that $\log d$ is contained in $\mathsf{L}^\infty(\Omega)$, and thus in $\mathcal{L}^1(\Omega,\mathbb{P})$, for every $d \in \mathcal{V}$. Also, by Part 1 of this proposition, for every $d \in \mathcal{V}$, the derivative $DL_{(i)}(d)(f) = d^{-1}(X_i)f(X_i)$ of $L_{(i)}(p)$ at d exists, and is continuous as a map from $\mathcal{V}$ to $\mathsf{L}^\infty(\Omega)'$. Note finally that $\sup_{d\in\mathcal{V}} \|DL_{(i)}(d)\|'_\infty \le \sup_{d\in\mathcal{V}} \|d^{-1}\|_\infty = 2/\zeta < \infty$ where $2/\zeta \in \mathcal{L}^1(\Omega,\mathbb{P})$ is the dominating function. □

By Part 4 of Proposition 2, the set $\mathcal{P}_t$ is contained in the $\mathsf{L}^\infty(\Omega)$-open set $\mathcal{V}$. Thus the above result shows that the likelihood function and its limiting counterpart are Fréchet-differentiable at each $p \in \mathcal{P}_t$ (the former for all $(X_1,\dots,X_n)^T \in \Omega^n$).

The following proposition gives sufficient conditions for interchanging the order of differentiation and integration in a general setting. It was used in the proof of Part 2 of Proposition 3 above. Here, for $(S,\mathcal{A},\mu)$ some measure space, $\mathcal{L}^1(S,\mathcal{A},\mu)$ denotes the vector space of $\mathcal{A}$-measurable μ-integrable real-valued functions defined on S.

Proposition 4 *Let V be an open subset of some Banach space E and let $(S,\mathcal{A},\mu)$ be a measure space. Suppose that the function $f(v,s) : V \times S \to \mathbb{R}$ is contained in $\mathcal{L}^1(S,\mathcal{A},\mu)$ for every $v \in V$. Assume that for every $v \in V$ and every $s \in S$ the Fréchet-derivative $D_1 f(v,s)$ w.r.t. the first variable exists. Furthermore, assume that for every $s \in S$, the map $v \longmapsto D_1 f(v,s)$ from V to E' is continuous. Suppose further that there exists a function $g \in \mathcal{L}^1(S,\mathcal{A},\mu)$ such that $\|D_1 f(v,s)\|'_E \le g(s)$ for every $v \in V$ and $s \in S$. Then, the function $\varphi : v \longmapsto \int_S f(v,s)\mathrm{d}\mu(s)$ from $V \subseteq E \to \mathbb{R}$ is Fréchet differentiable with derivative $D\varphi(v)(h) = \int_S D_1 f(v,s)(h)\mathrm{d}\mu(s)$ for $h \in E$.*

Proof We need to show for every $v \in V$ that

$$\left|\varphi(v+h) - \varphi(v) - \int_S D_1 f(v,s)(h)\mathrm{d}\mu(s)\right| = o(\|h\|_E) \tag{21}$$

as $\|h\|_E \to 0$, and that $h \mapsto \int_S D_1 f(v,s)(h)\mathrm{d}\mu(s)$ is norm-continuous on E. Without loss of generality we may assume that $\|h\|_E$ is sufficiently small such that the line connecting v and $v+h$ is contained in V. Inserting the definition of φ, we have

$$\left|\int_S (f(v+h,s) - f(v,s) - D_1 f(v,s)(h))\mathrm{d}\mu(s)\right| \leq \int_S |(f(v+h,s) - f(v,s) - D_1 f(v,s)(h))|\,\mathrm{d}\mu(s)$$

where the integrand is $o(\|h\|_E)$ by assumption. Hence (21) follows from the dominated convergence theorem if we show that there exists a function $g^* \in \mathcal{L}^1(S,\mathcal{A},\mu)$ which dominates the integrand divided by $\|h\|_E$. Apply the (pathwise) mean value theorem and the chain rule on Banach spaces (8.2.1 in [6]) to obtain

$$\begin{aligned}|f(v+h,s) - f(v,s) - D_1 f(v,s)(h)| &\leq \left\|D_1 f(\tilde{v},s) - D_1 f(v,s)\right\|'_E \|h\|_E \\ &\leq 2g(s)\,\|h\|_E = g^*(s)\,\|h\|_E\end{aligned}$$

under the conditions of the proposition. Here, $\tilde{v} = v + \xi h$ $(0 \leq \xi \leq 1)$ are mean values contained in V. Continuity of $h \mapsto \int_S D_1 f(v,s)(h)\mathrm{d}\mu$ on E follows from

$$|D_1 f(v,s)(h)| \leq \|D_1 f(v,x)\|'_E \|h\|_E \leq |g(s)|\,\|h\|_E$$

which gives $|D\varphi(v)(h)| \leq C\|h\|_E$ for $C = \int_S |g|\,\mathrm{d}\mu$. □

In what follows, for a (possibly random) symmetric bilinear functional Ψ defined on $\mathcal{L}^0(\Omega)$, we shall use the following notation where $\mathcal{H}$ and $\mathcal{G}$ are subsets of $\mathcal{L}^0(\Omega)$:

$$\|\Psi\|_{\infty,\mathcal{H},\mathcal{G}} := \sup_{h\in\mathcal{H}} \sup_{g\in\mathcal{G}} |\Psi(h,g)|\,. \tag{22}$$

If $\mathcal{U}_E = \mathcal{H} = \mathcal{G}$ is the unit ball of some Banach space E contained in $\mathcal{L}^0(\Omega)$, the norm $\|\Psi\|_{\infty,\mathcal{U}_E,\mathcal{U}_E}$ just equals the usual operator norm of the restriction of Ψ to $E \times E$. We use the same notation for multilinear functionals. The following lemma establishes stochastic bounds (with rate) for the likelihood derivatives which are (symmetric multi-) linear functionals on $\mathsf{L}^\infty(\Omega) \subseteq \mathcal{L}^0(\Omega)$.

Lemma 2 *Let Condition 1.1 hold and let* $1 \leq \alpha < \infty$. *Let* $\mathcal{H}_j$, $j = 1,\ldots,\alpha$, *be bounded subsets of* $\mathsf{L}^\infty(\Omega)$ *that are* $\mathbb{P}$*-Donsker. We then have*

$$\sup_{p\in\mathcal{P}_t} \left\|D^\alpha L_n(p) - \mathbb{P}D^\alpha L_{(i)}(p)\right\|_{\infty,\mathcal{H}_1,\ldots,\mathcal{H}_\alpha} = O_{\mathbb{P}^*}(n^{-1/2}). \tag{23}$$

Proof Note that

$$\begin{aligned}
&\sup_{p\in\mathcal{P}_t} \left\| D^\alpha L_n(p) - \mathbb{P}D^\alpha L_{(i)}(p) \right\|_{\infty,\mathcal{H}_1,\ldots,\mathcal{H}_\alpha} \\
&\quad = \sup_{p\in\mathcal{P}_t} \sup_{h_1\in\mathcal{H}_1} \cdots \sup_{h_\alpha\in\mathcal{H}_\alpha} |(\alpha-1)!| \\
&\qquad \times \left| n^{-1}\sum_{i=1}^n (p^{-\alpha}h_1\cdots h_\alpha)(X_i) - \int_\Omega (p^{-\alpha}h_1\cdots h_\alpha)\mathrm{d}\mathbb{P} \right| \\
&\quad = \sup_{p\in\mathcal{P}_t} \sup_{h_1\in\mathcal{H}_1} \cdots \sup_{h_\alpha\in\mathcal{H}_\alpha} |(\alpha-1)!| \left| (\mathbb{P}_n - \mathbb{P})(p^{-\alpha}h_1\cdots h_\alpha) \right|
\end{aligned}$$

by definition of the operator 'norm' (22). If we show that

$$\mathcal{C} := \left\{ p^{-\alpha}h_1\cdots h_\alpha : \ p\in\mathcal{P}_t,\ h_j\in\mathcal{H}_j \text{ for } j=1,\ldots,\alpha \right\}$$

is a $\mathbb{P}$-Donsker class, we have proved (23): this is so since then the empirical process $n^{1/2}(\mathbb{P}_n - \mathbb{P})$ indexed by $\mathcal{C}$ converges in law in $(\ell^\infty(\mathcal{C}), \|\cdot\|_{\infty,\mathcal{C}})$ to a measurable limit which, by Prohorov's Theorem (1.3.8 in [38]), entails that $n^{1/2}(\mathbb{P}_n - \mathbb{P})$ is uniformly tight in that space, and hence

$$\|\mathbb{P}_n - \mathbb{P}\|_{\infty,\mathcal{C}} = O_{\mathbb{P}^*}\left(n^{-1/2}\right).$$

Since finite products of uniformly bounded Donsker classes are again Donsker (by [38], 2.10.8), it is sufficient to show that $1/\mathcal{P}_t$ is $\mathbb{P}$-Donsker to prove that $\mathcal{C}$ is a $\mathbb{P}$-Donsker class. For $1/\mathcal{P}_t$ to be $\mathbb{P}$ -Donsker, it is sufficient (by [38], 2.10.9) to show that $\mathcal{P}_t$ is so, since the functions in $\mathcal{P}_t$ are uniformly bounded away from zero. The set $\mathcal{P}_t$ is bounded in $\mathsf{W}_2^t(\Omega,\lambda)$ and is thus $\mathbb{P}$-Donsker by Part 5 of Proposition 2. □

4.2 Existence and 'Strong' convergence properties of the MLE

Lemma 3 *Let Ω be a bounded C^∞-domain and let $t > s > 1/2$. Then $\mathcal{P}_t$ is a compact subset of $\mathsf{W}_2^s(\Omega,\lambda)$ as well as of $\mathsf{C}(\Omega)$.*

Proof By Part 3 of Proposition 2, it is sufficient to prove that $\mathcal{P}_t = \mathcal{P}_{t,\zeta,D}(\Omega)$ is a compact subset of $\mathsf{W}_2^s(\Omega,\lambda)$. Observe first that the imbedding

$$\mathrm{id} : \mathsf{W}_2^t(\Omega,\lambda) \hookrightarrow \mathsf{W}_2^s(\Omega,\lambda)$$

is a compact imbedding since $t > s$ holds (see Theorem 1.16.1 in [20]). In other words, the Sobolev ball

$$\mathcal{U}_{t,D} = \{ f\in\mathsf{W}_2^t(\Omega,\lambda) : \|f\|_{t,2,\lambda} \le D \} \quad (0 < D < \infty)$$

is relatively compact in the Banach space $\mathsf{W}_2^s(\Omega,\lambda)$. We now show that $\mathcal{U}_{t,D}$ is in fact compact: let f_m be any sequence in $\mathcal{U}_{t,D}$. By relative compactness of $\mathrm{id}(\mathcal{U}_{t,D})$, the sequence $\mathrm{id}(f_m)$ converges (by passing to a subsequence if necessary) to some $g \in \mathsf{W}_2^s(\Omega,\lambda)$ in the $\|\cdot\|_{s,2,\lambda}$-norm and hence, by Part 3 of Proposition 2, also in the $\|\cdot\|_\infty$-norm. Since the Banach space $\mathsf{W}_2^t(\Omega,\lambda)$ is separable and reflexive, the weak topology of $\mathsf{W}_2^t(\Omega,\lambda)$ on the ball $\mathcal{U}_{t,D}$ is metrizable and $\mathcal{U}_{t,D}$ is compact in this topology, see V.4.7 and V.5.1 in [10]. Consequently, there exists some weak accumulation point $f \in \mathcal{U}_{t,D}$ so that $L(f_m)$ (or, if necessary, a subsequence independent of L) converges to $L(f)$ for all $L \in \left(\mathsf{W}_2^t(\Omega,\lambda)\right)'$. Since the evaluation functional δ_x is contained in $\left(\mathsf{W}_2^t(\Omega,\lambda)\right)'$ for any $x \in \Omega$ (by Part 3 of Proposition 2), we have $\delta_x(f_m) \to \delta_x(f)$ for all $x \in \Omega$, that is, $f_m \to f$ pointwise. This implies that $g = \mathrm{id}(f) \in \mathrm{id}(\mathcal{U}_{t,D})$ and thus $\mathcal{U}_{t,D}$, more precisely $\mathrm{id}(\mathcal{U}_{t,D})$, is compact in $\mathsf{W}_2^s(\Omega,\lambda)$.

Fix $x \in \Omega$ and let the set $\mathcal{P}_x^\zeta$ stand for the set of all functions f in $\mathsf{W}_2^s(\Omega,\lambda)$ that satisfy $\zeta \le f(x)$ at the point $x \in \Omega$. We then have

$$\mathcal{P}_x^\zeta = \left\{f \in \mathsf{W}_2^s(\Omega,\lambda) : \zeta \le f(x) < \infty\right\} = \delta_x^{-1}([\zeta,\infty))$$

which is a closed set since the inverse image of the closed set $[\zeta,\infty)$ under the continuous evaluation map δ_x is closed. We therefore have that the set

$$\mathcal{P}^\zeta := \left\{f \in \mathsf{W}_2^s(\Omega,\lambda) : \zeta \le f(x) < \infty \ \forall x \in \Omega\right\} = \bigcap_{x\in\Omega} \mathcal{P}_x^\zeta$$

is closed since arbitrary intersections of closed sets are again closed. We next show that also

$$\mathcal{P}^{(1)} = \{g \in \mathsf{W}_2^s(\Omega,\lambda) : \ \|g\|_{1,\lambda} = 1\}$$

is closed. Suppose for $f \in \mathsf{W}_2^s(\Omega,\lambda)$ there exists a sequence g_m in $\mathcal{P}^{(1)}$ such that $\|f - g_m\|_{s,2,\lambda} \to 0$ as $m \to \infty$. Since Ω is bounded, we have $\|h\|_{1,\lambda} \le K\|h\|_\infty \le KC_s\|h\|_{s,2,\lambda}$ for all $h \in \mathsf{W}_2^s(\Omega,\lambda)$ by the imbedding $\mathsf{C}(\Omega) \hookrightarrow L^1(\Omega,\lambda)$ with imbedding constant K and by Part 3 of Proposition 2. Consequently $\|f - g_m\|_{1,\lambda} \to 0$ holds for $m \to \infty$. By $\left|\|f\|_{1,\lambda} - \|g_m\|_{1,\lambda}\right| \le \|f - g_m\|_{1,\lambda}$, we have $\|f\|_{1,\lambda} = 1$ and hence $\mathcal{P}^{(1)}$ is closed in $\mathsf{W}_2^s(\Omega,\lambda)$. Observe therefore that $\mathcal{P}_t = \mathcal{P}^{(1)} \cap \mathcal{P}^\zeta \cap \mathcal{U}_{t,D}$ is compact in $\mathsf{W}_2^s(\Omega,\lambda)$ since it is the intersection of a compact and two closed subsets. □

4.2.1 Existence and measurability

For any bounded subset $\mathcal{U}$ of $\mathsf{W}_2^s(\Omega,\lambda)$ with $s > 1/2$ and Ω a bounded C^∞-domain, the map

$$(X_1,\ldots,X_n)^T \longmapsto n^{1/2}(\mathbb{P}_n - \mathbb{P})$$

from Ω^n to $(\ell^\infty(\mathcal{U}), \|\cdot\|_{\infty,\mathcal{U}})$ is $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{\ell^\infty(\mathcal{U})}$ measurable by using Part 5 of Proposition 2 (equicontinuity) and Theorem 5.3.8 in [8]. [Inspection of the proof of Lemma 2 then shows that the l.h.s. of (23) is $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{\mathbb{R}}$ measurable whenever the classes $\mathcal{H}_1, \ldots, \mathcal{H}_\alpha$ are equicontinuous on Ω.]

The following proposition establishes measurability of certain suprema and is taken from Lemma A3 in [25].

Proposition 5 *Let $(S, \mathcal{A})$ be a (non-empty) measurable space, let (Θ, d) be a (non-empty) compact metric space and let $u : S \times \Theta \to \mathbb{R}$ be a function that is $\mathcal{A}$-measurable in its first argument for each $\theta \in \Theta$ and that is continuous on Θ in its second argument for each $s \in S$. Then there exists an $\mathcal{A}$-$\mathcal{B}_{(\Theta,d)}$ measurable function $\tilde{\theta} : S \to \Theta$ such that*

$$u(s, \tilde{\theta}(s)) = \sup_{\theta \in \Theta} u(s, \theta)$$

holds for each $s \in S$.

Identify $(S, \mathcal{A})$ with the sample space $(\Omega^n, \mathcal{B}_{\Omega^n})$ and (Θ, d) with the compact metric space $(\mathcal{P}_t, \|\cdot\|_\infty)$ (see Lemma 3) and note that the likelihood function (4) is continuous on $\mathcal{P}_t$ ($t > 1/2$, $\zeta > 0$) in the $\|\cdot\|_\infty$-topology and measurable for given p. Proposition 5 then establishes the existence of $\widehat{p}_n$ as a $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{(\mathcal{P}_t, \|\cdot\|_\infty)}$-measurable selection of the optimization problem (6). [We note that all results in the paper clearly hold for *every* $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{(\mathcal{P}_t, \|\cdot\|_\infty)}$-measurable selection. In fact, they hold for *any* selection (measurable or not) if one formulates all results in terms of outer probability $\mathbb{P}^*$.] Note further that the map $p \mapsto \int f p \mathrm{d}\lambda$ from $(\mathcal{P}_t, \|\cdot\|_\infty)$ to $(\ell^\infty(\mathcal{F}), \|\cdot\|_{\infty,\mathcal{F}})$ is continuous if $\mathcal{F}$ is a bounded subset of $\mathcal{L}^1(\Omega, \lambda)$. Consequently, the mapping $\hat{v}_n$ is $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{\ell^\infty(\mathcal{F})}$-measurable. Observe finally that the map associating with each probability density $p \in \mathcal{P}_t$ the corresponding probability measure is continuous viewed as a mapping from $(\mathcal{P}_t, \|\cdot\|_\infty)$ to $\mathfrak{P}(\Omega)$ equipped with the weak topology τ_{weak}. [To see this note that $\|\cdot\|_\infty$-convergence implies $\|\cdot\|_{1,\lambda}$-convergence (for bounded Ω) which implies convergence in the total variation metric on $\mathfrak{P}(\Omega)$. This in turn implies weak convergence on $\mathfrak{P}(\Omega)$.] Hence, the map $(X_1, \ldots, X_n) \longmapsto \hat{\mathbb{P}}_n$ from $\Omega^n \to \mathfrak{P}(\Omega)$ is $\mathcal{B}_{\Omega^n}$- $\mathcal{B}_{(\mathfrak{P}(\Omega), \tau_{\text{weak}})}$-measurable.

4.2.2 Consistency and rates of convergence

In the following proposition, we derive the rate of convergence of the MLE in the L^2-norm by using results due to [35]. By a suitable interpolation inequality, these results also imply convergence rates for 'intermediate (fractional) derivatives' with $0 \leq s \leq t$. We also give almost sure consistency in corresponding Sobolev-norms.

Proposition 6 *Suppose that Condition* 1.1 *holds. Then*

$$\|\widehat{p}_n - p_0\|_{s,2,\lambda} = O_{\mathbb{P}^*}\left(n^{-(t-s)/(2t+1)}\right) \tag{24}$$

holds for $0 \leq s \leq t$. *Furthermore,*

$$\|\widehat{p}_n - p_0\|_{s,2,\lambda} \to 0 \quad \mathbb{P}\text{-}a.s. \tag{25}$$

holds as $n \to \infty$ *for every* $0 \leq s < t$.

Proof To prove (25), we may restrict ourselves to $t > s > 1/2$. Note that $(\mathcal{P}_t, \|\cdot\|_{s,2,\lambda})$ is a compact metric space by Lemma 3 above. Therefore, the space $\mathsf{C}((\mathcal{P}_t, \|\cdot\|_{s,2,\lambda}))$ of bounded real-valued functions on $\mathcal{P}_t$ that are $\|\cdot\|_{s,2,\lambda}$-continuous is a separable Banach space normed by the sup-norm $\|\cdot\|_{\infty,\mathcal{P}_t}$, see, e.g., 11.2.5 in [9]. Observe that both $\log p(x)$ and $\int_\Omega \log p \mathrm{d}\mathbb{P}$ are bounded real-valued continuous functions on $(\mathcal{P}_t, \|\cdot\|_{s,2,\lambda})$ for every $x \in \Omega$: boundedness of both functions follows from $\zeta \leq p(x) \leq C_t D < \infty$ for all $p \in \mathcal{P}_t$ and $x \in \Omega$ (by Part 4 of Proposition 2) and boundedness of the logarithm on $[\zeta, C_t D]$. Continuity of $\log p(x)$ follows from the fact that $p_m \to p$ in the $\|\cdot\|_{s,2,\lambda}$-topology implies $p_m \to p$ in the $\|\cdot\|_\infty$-topology (Part 3 of Proposition 2) and the fact that the logarithm is continuous on $[\zeta, C_t D]$, which implies $\log p_m(x) \to \log p(x)$ for all (in fact uniformly in) $x \in \Omega$. For $\int_\Omega \log p \mathrm{d}\mathbb{P}$ continuity follows from the same reasoning and the dominated convergence theorem. Therefore, the sum of i.i.d random variables $S_n(p) = \sum_{i=1}^n (\log p(X_i) - \int_\Omega \log p \mathrm{d}\mathbb{P})$ satisfies Mourier's SLLN in $\mathsf{C}((\mathcal{P}_t, \|\cdot\|_{s,2,\lambda}))$ (see Corollary 7.1.8. in [8]), that is

$$\sup_{p \in \mathcal{P}_t} |S_n(p)/n| = \sup_{p \in \mathcal{P}_t} \left| L_n(p) - \int_\Omega \log p \mathrm{d}\mathbb{P} \right| \to 0 \quad \mathbb{P}\text{-a.s.}$$

holds as n goes to infinity. This gives almost sure consistency in the metric induced by the $\|\cdot\|_{s,2,\lambda}$-norm by standard arguments, see, e.g., Lemma 3.1 in [25]: Both $L_n(p)$ and the limiting function $\int_\Omega \log p \mathrm{d}\mathbb{P}$ are continuous on $(\mathcal{P}_t, \|\cdot\|_{s,2,\lambda})$ and the unique maximizer of the limiting function is the 'true' density function p_0, i.e., $\sup_{p \in \mathcal{P}_t} \int_\Omega \log p \mathrm{d}\mathbb{P} = \int_\Omega \log p_0 \mathrm{d}\mathbb{P}$ under Condition 1.1.

We next prove (24). Note first that the Hellinger-bracketing-metric entropy $H_{[\,]}(\varepsilon, \mathcal{P}_t^{1/2}, \|\cdot\|_{2,\lambda})$ is of order $\varepsilon^{-1/t}$ where $\varepsilon \to 0$ denotes the bracket size: since the functions in $\mathcal{P}_t$ are bounded from below by ζ, it follows from standard arguments that $H_{[\,]}(\varepsilon, \mathcal{P}_t^{1/2}, \|\cdot\|_{2,\lambda|\Omega})$ can be bounded (up to irrelevant constants) by the bracketing metric entropy $H_{[\,]}(\varepsilon, \mathcal{P}_t, \|\cdot\|_{2,\lambda|\Omega})$, which is seen to be of order $\varepsilon^{-1/t}$ by using Part 2 of Corollary 2 in [24] with $\beta = 0$, $d = 1$, $\mu = \lambda\,|\Omega$, $r = 2, p = q = 2$, and upon noting that $W_2^s(\mathbb{R}, \lambda\,|\mathbb{R})$ coincides with the Besov space $B_{22}^s(\mathbb{R}, \lambda\,|\mathbb{R})$ considered in [24]. This gives, by (7.26) in [36], that $\int_\Omega (\widehat{p}_n^{1/2} - p_0^{1/2})^2 \mathrm{d}\lambda = O_\mathbb{P}(n^{-2t/(2t+1)})$. Then

$$O_\mathbb{P}\left(n^{-2t/(2t+1)}\right) = \int_\Omega \left(\widehat{p}_n^{1/2} - p_0^{1/2}\right)^2 \mathrm{d}\lambda$$

$$= \int_\Omega \left((\widehat{p}_n - p_0)\,(\widehat{p}_n^{1/2} + p_0^{1/2})^{-1} \right)^2 \mathrm{d}\lambda$$
$$\geq C \cdot \|\widehat{p}_n - p_0\|_{2,\lambda}^2$$

for a suitable positive real number C, since the densities in $\mathcal{P}_t$ are uniformly bounded by Part 4 of Proposition 2. Taking the square root delivers (24) for the case $s = 0$. The case $0 < s \leq t$ follows from the following interpolation inequality (see Remark 1.9.1 and Theorem 1.9.6 in [20])

$$\|f\|_{s,2,\lambda} \leq C \|f\|_{t,2,\lambda}^{s/t} \|f\|_{2,\lambda}^{(t-s)/t}$$

for $f \in \mathsf{W}_2^t(\Omega, \lambda)$ and $0 < C < \infty$, and the fact that $\|\widehat{p}_n - p_0\|_{t,2,\lambda} \leq 2D$ by Condition 1.1. □

For given $1/2 < s < t$, Lemma 3 and Proposition 5 can be used to show that a $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{(\mathcal{P}_t, \|\cdot\|_{s,2,\lambda})}$-measurable selection of the optimization problem (6) exists. For this selection, expression (24) can then be formulated avoiding outer probability. [We note that such a selection depends on s and is not guaranteed to be $\mathcal{B}_{\Omega^n}$-$\mathcal{B}_{(\mathcal{P}_t, \|\cdot\|_{r,2,\lambda})}$-measurable for $r > s$.]

4.3 Proofs for Sect. 2

4.3.1 Proof of Theorem 1

The classical proof of asymptotic normality of MLEs in finite-dimensional models exploits the fact that the first derivative of the likelihood function (score) evaluated at the maximizer (MLE) is zero and then proceeds by a suitable Taylor-approximation of the score around the true value; see, e.g., Chap. 8 in [25] for a detailed recent account and references. This classical approach, however, is not viable in the infinite-dimensional setup of this paper. Here, the derivative $DL_n(\widehat{p}_n)(h)$ of the likelihood-function at the maximum $\widehat{p}_n$ evaluated at h belonging to

$$\mathcal{T}_t = \left\{ h \in \mathcal{L}^0(\Omega) : \|h\|_{t,2,\lambda} < \infty, \int_\Omega h \mathrm{d}\lambda = 0 \right\},$$

which is the tangent space of $\mathcal{P}_t$ at p_0 (or any other 'internal' point of $\mathcal{P}_t$), is generally *non-zero* for every given n, even under Condition 1 (which implies that p_0 is an 'internal' point), since there is no guarantee whatsoever that $\widehat{p}_n$ is an 'internal' point: the 'internality' condition $\|\widehat{p}_n\|_{t,2,\lambda} < D$ for large n could be inferred from $\|p_0\|_{t,2,\lambda} < D$ if the MLE were consistent in the norm-topology of $\mathsf{W}_2^t(\Omega, \lambda)$, which, however, is not to be expected since $\mathcal{P}_t$ is non-compact in this topology. However, closeness of the MLE $\widehat{p}_n$ to p_0 in the norm-topology of $\mathsf{W}_2^t(\Omega, \lambda)$ is not crucial in the weak convergence context, but rather it is essential

to establish closeness of $DL_n(\widehat{p}_n)(h)$ to $\mathbb{P}DL_{(i)}(p_0)(h)$ for $h \in \mathcal{T}_t$ at the order $o_{\mathbb{P}}(n^{-1/2})$. Since $\mathbb{P}DL_{(i)}(p_0)(h) = 0$, this is equivalent to showing that

$$DL_n(\widehat{p}_n)(h) = o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{26}$$

for every $h \in \mathcal{T}_t$. Showing (26) even uniformly in $\|\cdot\|_{t,2,\lambda}$-bounded subsets of $\mathcal{T}_t$ then delivers $\left\|\hat{\nu}_n - \nu_n\right\|_{\infty,\mathcal{U}_{t,B}} = o_{\mathbb{P}}(1)$ by using the close connection between the empirical process and the likelihood derivatives.

Throughout Sect. 4.3.1, we shall assume that the conditions of Theorem 1 hold. We define the operator $\Pi : \mathsf{L}^{\infty}(\Omega) \to \mathsf{L}^{\infty}(\Omega) \cap \{g : \int_{\Omega} g\mathrm{d}\lambda = 0\}$ given by $\Pi(f) = (f - \mathbb{P}f)p_0$. Note the connection to the empirical process, i.e., that $DL_n(p_0)(\Pi(f)) = (\mathbb{P}_n - \mathbb{P})f$ holds.

Lemma 4 *We have that*

$$\sup_{f \in \mathcal{U}_{t,B}} \left|DL_n\big(\widehat{p}_n\big)(\Pi(f))\right| = o_{\mathbb{P}^*}\left(n^{-1/2-(t-k)/(2t+1)}\right)$$

holds for every real $k > 1/2$.

Proof Recall that we use $(\Omega^{\infty}, \mathcal{B}_{\Omega^{\infty}}, \mathbb{P}^{\infty})$ as the underlying probability space on which the data $X_1, X_2, \ldots$ are defined. We shall denote elements of Ω^{∞} by $\omega = (x_1, x_2, \ldots)$. We shall say that a sequence $A_n \subseteq \Omega^{\infty}$ is *eventual*, if for $\mathbb{P}^{\infty}$-almost all $\omega \in \Omega^{\infty}$ there exists an index $N(\omega) \in \mathbb{N}$ such that the relation $\omega \in A_n$ holds for $n \geq N(\omega)$.

Step 1: Throughout the proof, by a point *internal* to $\mathcal{P}_t$ we mean a probability density function $p \in \mathcal{P}_t$ that satisfies $\|p\|_{t,2,\lambda} < D$ as well as $\inf_{x \in \Omega} p(x) > \zeta$. Observe that, for $0 < \varepsilon \leq 1$, the quantity $(1-\varepsilon)\widehat{p}_n + \varepsilon p_0$ is always an internal point of $\mathcal{P}_t$ as a consequence of Condition 1. Set

$$\mathcal{U}_{t,\eta,0} = \left\{ w \in \mathsf{W}_2^t(\Omega, \lambda) : \|w\|_{t,2,\lambda} < \eta, \int_{\Omega} w\mathrm{d}\lambda = 0 \right\}$$

where $\eta = D - \|p_0\|_{t,2,\lambda}$ which is positive since p_0 is an internal point. Define

$$\hat{h}_n(w) := (1-\varepsilon)\widehat{p}_n + \varepsilon p_0 + \varepsilon w \quad (w \in \mathcal{U}_{t,\eta,0}). \tag{27}$$

We now show that, for ε small enough,

$$\left\{\omega \in \Omega^{\infty} : \left\{\hat{h}_n(w) : w \in \mathcal{U}_{t,\eta,0}\right\} \subseteq \mathcal{P}_t\right\} \tag{28}$$

is eventual. To see this, observe the following three facts: first, by the triangle inequality

$$\begin{aligned}\left\|\hat{h}_n(w)\right\|_{t,2,\lambda} &\leq (1-\varepsilon)\, D + \varepsilon \left\|p_0\right\|_{t,2,\lambda} + \varepsilon\eta \\ &\leq D\end{aligned}$$

holds for every $0 < \varepsilon \leq 1$, every n and every $w \in \mathcal{U}_{t,\eta,0}$ by definition of η. This verifies the Sobolev-norm condition for containment of $\hat{h}_n(w)$ in $\mathcal{P}_t$. Second, by Condition 1.2, there exists a $\beta > 0$ such that $\inf_{x\in\Omega} p_0(x) > \zeta + \beta$ is satisfied. Note that by Part 3 of Proposition 2 and Part 2 of Proposition 6, the MLE is almost surely $\|\cdot\|_\infty$-consistent, and therefore $\inf_{x\in\Omega} \widehat{p}_n(x) \geq \zeta + \beta$ is eventual. Since $\|w\|_\infty \leq C_t \|w\|_{t,2,\lambda} \leq C_t\eta$ by Part 3 of Proposition 2, it follows that

$$\begin{aligned}\hat{h}_n(w) &= (1-\varepsilon)\,\widehat{p}_n + \varepsilon p_0 + \varepsilon w \\ &\geq \zeta + \beta - \varepsilon C_t\eta \quad \text{for every } w \in \mathcal{U}_{t,\eta,0}\end{aligned} \tag{29}$$

holds eventually. Thus, for $\varepsilon \leq \beta/(C_t\eta)$, the inequality $\hat{h}_n(w) \geq \zeta$, every $w \in \mathcal{U}_{t,\eta,0}$, holds eventually. Third, since $\hat{h}_n(w) > 0$ for every $w \in \mathcal{U}_{t,\eta,0}$ and $\varepsilon \leq \beta/(C_t\eta)$ holds eventually by (29), and since w integrates to zero, $\hat{h}_n(w)$ is a density for all $w \in \mathcal{U}_{t,\eta,0}$ eventually.

Let now ε satisfy $0 < \varepsilon \leq \min(1, \beta/(C_t\eta))$. In view of (28), since $\widehat{p}_n$ is a maximizer of $L_n(\cdot)$ over $\mathcal{P}_t$, and since $L_n(\cdot)$ is Fréchet differentiable at $\widehat{p}_n$ by Proposition 3, the derivative of $L_n(\cdot)$ at $\widehat{p}_n$ in the direction of $\hat{h}_n(w)$, $w \in \mathcal{U}_{t,\eta,0}$, has to equal zero eventually: that is

$$DL_n\left(\widehat{p}_n\right)\left(w - \widehat{p}_n + p_0\right) = 0 \quad \text{for every } w \in \mathcal{U}_{t,\eta,0} \tag{30}$$

holds eventually (where we have divided by $\varepsilon > 0$). [We note that a convex combination similar to (27) has been used in [40] in their framework.]

Step 2: For every $f \in \mathcal{U}_{t,B}$, we have by Condition 1, Parts 3 and 6 of Proposition 2

$$\begin{aligned}\|\Pi(f)\|_{t,2,\lambda} &\leq M \|f - \mathbb{P}f\|_{t,2,\lambda} \|p_0\|_{t,2,\lambda} \\ &\leq MD(\|f\|_{t,2,\lambda} + \|\mathbb{P}f\|_{t,2,\lambda}) \\ &\leq MD(B + \|C_tB\|_{t,2,\lambda}) \\ &\leq MDB(1 + C_tC') < \infty\end{aligned} \tag{31}$$

where $C' = \|\mathbf{1}_\Omega\|_{t,2,\lambda} = \|\mathbf{1}_\Omega\|_{2,\lambda} < \infty$. Now with η as in Step 1 define

$$s(\Pi(f)) = \eta \|\Pi(f)\|_{t,2,\lambda}^{-1} \Pi(f)$$

if $\Pi(f) \neq 0$, and set $s(\Pi(f)) = 0$ otherwise. Then it follows that $s(\Pi(f)) \in \mathcal{U}_{t,\eta,0}$ for every $f \in \mathcal{U}_{t,B}$.

Step 3: Inserting $s(\Pi(f))$ for w in (30) we obtain that

$$DL_n(\widehat{p}_n)(s(\Pi(f)) - \widehat{p}_n + p_0) = 0 \quad \text{for every } f \in \mathcal{U}_{t,B}$$

holds eventually. By linearity of $DL_n(\widehat{p}_n)(\cdot)$ we hence have that

$$DL_n(\widehat{p}_n)(s(\Pi(f))) = DL_n(\widehat{p}_n)(\widehat{p}_n - p_0) \quad \text{for every } f \in \mathcal{U}_{t,B} \tag{32}$$

holds eventually. Using Proposition 3, we see that the expected value of the likelihood derivative at p_0 equals zero along the directions $\{h : \int_\Omega h \mathrm{d}\lambda = 0\}$ and thus in particular along the direction $\widehat{p}_n - p_0$:

$$\mathbb{P}DL_{(i)}(p_0)(\widehat{p}_n - p_0) = \int_\Omega (\widehat{p}_n - p_0) p_0^{-1} \mathrm{d}\mathbb{P} = \|\widehat{p}_n\|_{1,\lambda} - \|p_0\|_{1,\lambda} = 0 \tag{33}$$

since both $\widehat{p}_n$ and p_0 are probability densities. Thus, we have from (32) that

$$DL_n(\widehat{p}_n)(s(\Pi(f))) = (DL_n(\widehat{p}_n) - \mathbb{P}DL_{(i)}(p_0))(\widehat{p}_n - p_0) \quad \text{for every } f \in \mathcal{U}_{t,B}$$

holds eventually. Let now k be as in the lemma. W.l.o.g. we may restrict ourselves to the case $k \leq t$. Choose a real j, $1/2 < j < k$. Let $\mathcal{U}_{j,1}$ denote the unit ball of $\mathsf{W}_2^j(\Omega, \lambda)$ which is a $\mathbb{P}$-Donsker class by Part 5 of Proposition 2. By Proposition 3 we obtain

$$\begin{aligned}
\left|(DL_n(\widehat{p}_n) - \mathbb{P}DL_{(i)}(p_0))(\widehat{p}_n - p_0)\right| &\leq \left|(DL_n(\widehat{p}_n) - \mathbb{P}DL_{(i)}(\widehat{p}_n))(\widehat{p}_n - p_0)\right| \\
&\quad + \left|(\mathbb{P}DL_{(i)}(\widehat{p}_n) - \mathbb{P}DL_{(i)}(p_0))(\widehat{p}_n - p_0)\right| \\
&\leq \sup_{p \in \mathcal{P}_t} \left\|DL_n(p) - \mathbb{P}DL_{(i)}(p)\right\|_{\infty,\mathcal{U}_j} \\
&\quad \|\widehat{p}_n - p_0\|_{j,2,\lambda} + \zeta^{-1} \|\widehat{p}_n - p_0\|_{2,\lambda}^2 =: Z_n.
\end{aligned}$$

[We note here once and for all that expressions like $\mathbb{P}DL_{(i)}(\widehat{p}_n)$ are to be understood as $\mathbb{P}DL_{(i)}(p)$ evaluated at $p = \widehat{p}_n$.] It follows from Lemma 2 and Proposition 6 (and the results from Sect. 4.2.1) that

$$\begin{aligned}
Z_n &= O_{\mathbb{P}}(n^{-1/2}) O_{\mathbb{P}^*}(n^{-(t-j)/(2t+1)}) + O_{\mathbb{P}}(n^{-2t/(2t+1)}) \\
&= o_{\mathbb{P}^*}(n^{-1/2-(t-k)/(2t+1)}).
\end{aligned} \tag{34}$$

Summarizing we obtain that

$$DL_n(\widehat{p}_n)(s(\Pi(f))) = Z_n \quad \text{for every } f \in \mathcal{U}_{t,B}$$

holds eventually. Multiplying by $\eta^{-1} \|\Pi(f)\|_{t,2,\lambda}$ we have that

$$DL_n(\widehat{p}_n)(\Pi(f)) = \eta^{-1} \|\Pi(f)\|_{t,2,\lambda} Z_n \quad \text{for every } f \in \mathcal{U}_{t,B}$$

holds eventually. Using (31), (34) and the fact that Z_n does not depend on $f \in \mathcal{U}_{t,B}$, we arrive at

$$\sup_{f \in \mathcal{U}_{t,B}} |DL_n(\widehat{p}_n)(\Pi(f))| = o_{\mathbb{P}^*}(n^{-1/2-(t-k)/(2t+1)})$$

which completes the proof. □

Proof (Theorem 1) Inserting for the definitions (3) and (7), we have

$$\begin{aligned}\left\|\hat{v}_n - v_n\right\|_{\infty,\mathcal{U}_{t,B}} &= \sqrt{n} \sup_{f \in \mathcal{U}_{t,B}} \left|(\hat{\mathbb{P}}_n - \mathbb{P})f - (\mathbb{P}_n - \mathbb{P})f\right| \\ &= \sqrt{n} \sup_{f \in \mathcal{U}_{t,B}} \left|\int_\Omega (f - \int_\Omega f \mathrm{d}\mathbb{P})(\widehat{p}_n - p_0)\,\mathrm{d}\lambda - (\mathbb{P}_n - \mathbb{P})f\right|.\end{aligned}$$

Recall that

$$-\mathbb{P}D^2L_{(i)}(p_0)(\widehat{p}_n - p_0, g) = \int_\Omega g p_0^{-1}(\widehat{p}_n - p_0)\,\mathrm{d}\lambda$$

holds for $g \in \mathsf{L}^\infty(\Omega)$ by Proposition 3 (and Condition 1). So for $g = (f - \int_\Omega f \mathrm{d}\mathbb{P})p_0 = \Pi(f)$ we obtain

$$\left\|\hat{v}_n - v_n\right\|_{\infty,\mathcal{U}_{t,B}} = \sqrt{n} \sup_{f \in \mathcal{U}_{t,B}} \left|-\mathbb{P}D^2L_{(i)}(p_0)(\widehat{p}_n - p_0, \Pi(f)) - (\mathbb{P}_n - \mathbb{P})f\right|. \quad (35)$$

We now treat the term $-\mathbb{P}D^2L_{(i)}(p_0)(\widehat{p}_n - p_0, \Pi(f))$: by the mean value theorem, we have

$$DL_n(\widehat{p}_n)(\Pi(f)) = DL_n(p_0)(\Pi(f)) + D^2L_n(\bar{p}_n)(\widehat{p}_n - p_0, \Pi(f))$$

where the mean values $\bar{p}_n \equiv \bar{p}_n(f)$ lie, for every $f \in \mathcal{U}_{t,B}$, on the line segment connecting $\widehat{p}_n$ and p_0 which is contained in $\mathcal{P}_t$. This gives

$$\begin{aligned}DL_n(\widehat{p}_n)(\Pi(f)) - \mathbb{P}D^2L_{(i)}(p_0)(\widehat{p}_n - p_0, \Pi(f)) = {} & DL_n(p_0)(\Pi(f)) \\ & + (D^2L_n(\bar{p}_n) - \mathbb{P}D^2L_{(i)}(p_0)) \\ & (\widehat{p}_n - p_0, \Pi(f)).\end{aligned}$$

Note now that the set $\Pi(\mathcal{U}_{t,B}) = \{\Pi(f) = (f - \mathbb{P}f)p_0 : f \in \mathcal{U}_{t,B}\}$ is a $\mathbb{P}$-Donsker class by (31) and Part 5 of Proposition 2. Let k be as in the theorem. W.l.o.g. we may restrict ourselves to the case $k \le t$. Choose a real j, $1/2 < j < k$. Let $\mathcal{U}_{j,1}$ denote the unit ball of $\mathsf{W}_2^j(\Omega, \lambda)$ which is a $\mathbb{P}$-Donsker class by Part 5 of

Proposition 2. Using Lemma 2, Propositions 2, 3, and 6, and (31), we have

$$\begin{aligned}
&\sup_{f\in\mathcal{U}_{t,B}}\left|(D^2L_n(\bar{p}_n)-\mathbb{P}D^2L_{(i)}(p_0))(\widehat{p}_n-p_0,\Pi(f))\right| \\
&\quad\le \sup_{f\in\mathcal{U}_{t,B}}\left|(D^2L_n(\bar{p}_n)-\mathbb{P}D^2L_{(i)}(\bar{p}_n))(\widehat{p}_n-p_0,\Pi(f))\right| \\
&\qquad+\sup_{f\in\mathcal{U}_{t,B}}\left|(\mathbb{P}D^2L_{(i)}(\bar{p}_n)-\mathbb{P}D^2L_{(i)}(p_0))(\widehat{p}_n-p_0,\Pi(f))\right| \\
&\quad\le \sup_{p\in\mathcal{P}_t}\left\|D^2L_n(p)-\mathbb{P}D^2L_{(i)}(p)\right\|_{\infty,\mathcal{U}_{j,1},\Pi(\mathcal{U}_{t,B})}\|\widehat{p}_n-p_0\|_{j,2,\lambda} \\
&\qquad+2\zeta^{-3}C_tD\sup_{f\in\mathcal{U}_{t,B}}\|\Pi(f)\|_\infty\|\widehat{p}_n-p_0\|_{2,\lambda}\left\|\bar{p}_n-p_0\right\|_{2,\lambda} \\
&\quad= O_{\mathbb{P}}(n^{-1/2})O_{\mathbb{P}^*}(n^{-(t-j)/(2t+1)})+O_{\mathbb{P}^*}(n^{-2t/(2t+1)}) \\
&\quad= o_{\mathbb{P}^*}(n^{-1/2-(t-k)/(2t+1)}).
\end{aligned}$$

Here we have used the simple fact that

$$\|\bar{p}_n-p_0\|_{2,\lambda}=\|\xi(f)\widehat{p}_n+(1-\xi(f))p_0-p_0\|_{2,\lambda}=\xi(f)\,\|\widehat{p}_n-p_0\|_{2,\lambda}$$

for some $0\le\xi(f)\le 1$. This together with Lemma 4 gives

$$\sup_{f\in\mathcal{U}_{t,B}}\left|-\mathbb{P}D^2L_{(i)}(p_0)(\widehat{p}_n-p_0,\Pi(f))-DL_n(p_0)(\Pi(f))\right|=o_{\mathbb{P}^*}(n^{-1/2-(t-k)/(2t+1)}).$$

Inserting this is into (35) shows that

$$\begin{aligned}
\left\|\hat{\nu}_n-\nu_n\right\|_{\infty,\mathcal{U}_{t,B}}&\le\sqrt{n}\sup_{f\in\mathcal{U}_{t,B}}|DL_n(p_0)(\Pi(f))-(\mathbb{P}_n-\mathbb{P})f|+o_{\mathbb{P}^*}(n^{-(t-k)/(2t+1)}) \\
&=o_{\mathbb{P}^*}(n^{-(t-k)/(2t+1)})
\end{aligned}$$

upon recalling that $DL_n(p_0)(\Pi(f))=(\mathbb{P}_n-\mathbb{P})f$. The proof of (9) is now complete (observing that $o_{\mathbb{P}^*}$ can be replaced by $o_{\mathbb{P}}$, since $\left\|\hat{\nu}_n-\nu_n\right\|_{\infty,\mathcal{U}_{t,B}}$ is measurable by the results in Sect. 4.2.1). The second claim of the theorem now follows upon observing that $\mathcal{U}_{t,B}$ is a $\mathbb{P}$-Donsker class by Part 5 of Proposition 2. □

4.3.2 Proof of Theorem 2

Since the limiting derivative $\mathbb{P}DL_{(i)}(p_0)(h)$ is zero not only for $h\in\mathcal{T}_t$, but for every element in the 'nonparametric' tangent space $\mathcal{T}_{\mathrm{np}}=\{h:\int_\Omega h\mathrm{d}\lambda=0\}$, one may suspect that (26) above will also hold for 'nice' subsets of $\mathcal{T}_{\mathrm{np}}$ that are *not* contained in $\mathsf{W}_2^t(\Omega,\lambda)$, especially in light of the fact that the rate for the empirical score obtained in Lemma 4 is faster than $n^{-1/2}$. Condition 2 describes

a large class of sets $\mathcal{F}$, for which the projection $\Pi(\mathcal{F})$ on $\mathcal{T}_{\text{np}}$ is indeed a 'nice' set in this sense. For such classes $\mathcal{F}$ one can then readily obtain the corresponding weak convergence result.

Proof of Theorem 2 We start from

$$\left\|\hat{\nu}_n - \nu_n\right\|_{\infty,\mathcal{F}} \leq \sup_{f\in\mathcal{F}} \left|\hat{\nu}_n(f - u_n(f))\right| + \sup_{f\in\mathcal{F}} \left|\hat{\nu}_n(u_n(f)) - \nu_n(u_n(f))\right| + \sup_{f\in\mathcal{F}} \left|\nu_n(u_n(f) - f)\right|.$$

Observe that the first term is equal to

$$n^{1/2} \sup_{f\in\mathcal{F}} \left| \int_\Omega (\widehat{p}_n - p_0)(u_n(f) - f) \mathrm{d}\lambda \right| = n^{1/2} a_n$$

by definition of a_n. The second term equals zero in the trivial case where $u_n(f) = 0$ for every $f \in \mathcal{F}$ and otherwise is equal to

$$\sup_{\substack{f\in\mathcal{F},\\ u_n(f)\neq 0}} \|u_n(f)\|_{t,2,\lambda} \left|(\hat{\nu}_n - \nu_n)(u_n(f)/\|u_n(f)\|_{t,2,\lambda})\right| \leq b_n \left\|\hat{\nu}_n - \nu_n\right\|_{\infty,\mathcal{U}_{t,1}} = b_n o_{\mathbb{P}}(n^{-(t-k)/(2t+1)})$$

for every $k > 1/2$ by Theorem 1 and definition of b_n. Finally, the third term is equal to $n^{1/2} c_n$ by definition of c_n. This establishes (10). The claim (11) then follows by Condition 2 upon choosing k such that $1/2 < k \leq k^*$. The last claim then follows from the second one observing that $\mathcal{F}$ is a $\mathbb{P}$-Donsker class. □

4.3.3 Proof of Proposition 1

Proof of Proposition 1 First consider the case $s < t$. Since $\mathcal{F} \subseteq \mathcal{U}_{s,B}$ for some $0 < B < \infty$, it suffices to prove the proposition for $\mathcal{F}$ equal to the Sobolev ball $\mathcal{U}_{s,B}$. Note that $\mathcal{U}_{s,B} \subseteq \mathcal{L}^1(\Omega, \lambda)$ by Part 3 of Proposition 2. Recall from Part 1 of Proposition 2 that an equivalent norm on $W_2^s(\mathbb{R}, \lambda\,|\mathbb{R})$ is given by $\|\cdot\|_{\wedge,s,2,\lambda|\mathbb{R}}$. By Part 2 of Proposition 2, the restriction of $W_2^s(\mathbb{R}, \lambda\,|\mathbb{R})$ to Ω coincides with $W_2^s(\Omega, \lambda)$ and the restricted norm is equivalent to the intrinsic norm. The restriction operator is a retraction (see Step 3 of the proof of Theorem 1.9.1 in [20]) and hence it follows that there exists a constant $0 < C < \infty$ such that for every $f \in \mathcal{U}_{s,B} \subseteq \mathsf{W}_2^s(\Omega, \lambda)$, there exists a function h with $[h]_\lambda \in W_2^s(\mathbb{R}, \lambda\,|\mathbb{R})$ such that $\|h\|_{\wedge,s,2,\lambda|\mathbb{R}} \leq C\|f\|_{s,2,\lambda|\Omega}$ and $h\,|\Omega = f$ everywhere on Ω hold. Recall that F denotes the Fourier(-Plancherel) transform and define $g(u) := \langle u\rangle^s (Fh)(u)$ where we recall the notation $\langle u\rangle^s = (1 + |u|^2)^{s/2}$. Observe that Part 1 of Proposition 2 and the definition of h imply

$$\|g\|_{2,\lambda|\mathbb{R}} = \left\|\langle u\rangle^s Fh\right\|_{2,\lambda|\mathbb{R}} = \|h\|_{\wedge,s,2,\lambda|\mathbb{R}} \leq C\|f\|_{s,2,\lambda|\Omega} \leq CB < \infty$$

where the constants do not depend on $f \in \mathcal{U}_{s,B}$. For positive real N define now

$$h_N = F^{-1}(\mathbf{1}_{\{|\cdot|\leq N\}} Fh),$$

which is real-valued (since h is real-valued and since $\mathbf{1}_{\{|\cdot|\leq N\}}$ is symmetric). By Part 1 of Proposition 2, Plancherel's Theorem, and the fact that $t > s$ we have

$$\begin{aligned}\|h_N\|^2_{t,2,\lambda|\mathbb{R}} \leq C'_t \left\| \langle\cdot\rangle^t Fh_N \right\|^2_{2,\lambda|\mathbb{R}} &= C'_t \left\| \langle\cdot\rangle^{t-s} \mathbf{1}_{\{|\cdot|\leq N\}} g \right\|^2_{2,\lambda|\mathbb{R}} \\ &\leq C'_t (1+N^2)^{t-s} \|g\|^2_{2,\lambda|\mathbb{R}} < \infty\end{aligned}$$

for a suitable finite constant C'_t. This implies in particular that $[h_N]_\lambda$ is an element of $W^t_2(\mathbb{R}, \lambda|\mathbb{R})$. By a similar reasoning the approximation error can be bounded as follows:

$$\begin{aligned}\|h - h_N\|^2_{k',2,\lambda|\mathbb{R}} \leq C'_{k'} \left\| \langle\cdot\rangle^{k'} F(h-h_N) \right\|^2_{2,\lambda|\mathbb{R}} &= C'_{k'} \left\| \langle\cdot\rangle^{k'} \mathbf{1}_{\{|\cdot|>N\}} Fh \right\|^2_{2,\lambda|\mathbb{R}} \\ = C'_{k'} \left\| \langle\cdot\rangle^{k'-s} \mathbf{1}_{\{|\cdot|>N\}} g \right\|^2_{2,\lambda|\mathbb{R}} &\leq C'_{k'} \|g\|^2_{2,\lambda|\mathbb{R}} (1+N^2)^{k'-s}\end{aligned}$$

for every $k' < s$. By restriction of the equivalence class $[h_N]_\lambda$ to the bounded C^∞-domain Ω, we obtain the function $h_N|\Omega \in \mathsf{W}^t_2(\Omega, \lambda)$ approximating $f \in \mathcal{U}_{s,B}$. Setting $N = n^{1/(2t+1)}$ for $n \in \mathbb{N}$ defines approximating sequence $u_n(f) = h_{n^{1/(2t+1)}}|\Omega$ which satisfies

$$\sup_{f\in\mathcal{U}_{s,B}} \|u_n(f)\|_{t,2,\lambda} = O(n^{(t-s)/(2t+1)}) \tag{36}$$

and

$$\sup_{f\in\mathcal{U}_{s,B}} \|f - u_n(f)\|_{k',2,\lambda} = O(n^{(k'-s)/(2t+1)}) \tag{37}$$

for every $k' < s$ by the above reasoning and since $\|d|\Omega\|_{r,2,\lambda} \leq C'' \|d\|_{r,2,\lambda|\mathbb{R}}$ holds for some $0 < C'' < \infty$, every $r \geq 0$, and $d \in W^s_2(\mathbb{R}, \lambda|\mathbb{R})$ by Part 2 of Proposition 2. Now note that b_n is equal to the left-hand side in (36) which satisfies the growth requirement in Condition 2 for $k^* = s > 1/2$. By applying Cauchy-Schwarz's inequality to the expression defining a_n, and using (24) from Proposition 6 as well as (37) with $k' = 0$, we have

$$a_n \leq \left\|\hat{p}_n - p_0\right\|_{2,\lambda} \sup_{f\in\mathcal{U}_{s,B}} \|f - u_n(f)\|_{0,2,\lambda} = O_{\mathbb{P}}(n^{(-t-s)/(2t+1)}) = o_{\mathbb{P}}(n^{-1/2}),$$

since $s > 1/2$. Finally, for every $1/2 < k' < s$ we obtain by Part 5 of Proposition 2 and by (37)

$$c_n = \sup_{f \in \mathcal{U}_{s,B}} |(\mathbb{P}_n - \mathbb{P})(u_n(f) - f)| \leq \|\mathbb{P}_n - \mathbb{P}\|_{\infty, \mathcal{U}_{k',B}} \sup_{f \in \mathcal{U}_{s,B}} \|u_n(f) - f\|_{k',2,\lambda}$$
$$= O_{\mathbb{P}}(n^{-1/2}) O(n^{(k'-s)/(2t+1)}) = o_{\mathbb{P}}(n^{-1/2}).$$

This shows that the proposition holds in case $s < t$. In case $s \geq t$, the proof of the proposition is trivial by choosing $u_n(f) = f$. □

4.4 Proofs for Sect. 3

Proof of Corollary 1 We first show that $\mathcal{F}_{1,\infty,1}$ is a bounded subset of $\mathsf{W}_2^s(\Omega, \lambda)$ for every $1/2 < s < 1$: note that $\mathcal{F}_{1,\infty,1}$ is a bounded subset of $\mathsf{C}^s(\Omega) = \{f \in \mathcal{L}^0(\Omega) : \|f\|_{s,\infty} < \infty\}$ (cf. (1)) and that $\mathsf{C}^s(\Omega)$ is equal to the Hölder–Zygmund space $\mathsf{B}_{\infty,\infty}^s(\Omega)$ for non-integer s (e.g., [34]), 2.5.7/9 and 3.4.2/2). By Theorem 3.3.1/7 in [34] we have the continuous imbedding $\mathsf{B}_{\infty,\infty}^s(\Omega) \hookrightarrow \mathsf{B}_{2,2}^s(\Omega, \lambda) = \mathsf{W}_2^s(\Omega, \lambda)$. This implies

$$\beta(\hat{\mathbb{P}}_n, \mathbb{P}_n) \leq n^{-1/2} \left\| \hat{\nu}_n - \nu_n \right\|_{\infty, \mathcal{U}_{s,B}}$$

for suitable $0 < B < \infty$. Applying Theorem 3 gives

$$\beta(\hat{\mathbb{P}}_n, \mathbb{P}_n) = o_{\mathbb{P}}(n^{-1/2-(\min(s,t)-k)/(2t+1)})$$

for every $k > 1/2$ and every $1/2 < s < 1$, which is equivalent to (16). Since $\mathcal{F}_{1,\infty,1}$ is a $\mathbb{P}$-Donsker class, the second claim then follows immediately. Measurability of $\beta(\hat{\mathbb{P}}_n, \mathbb{P}_n)$ as well as of $\beta(\hat{\mathbb{P}}_n, \mathbb{P})$ follows from the results in Sect. 4.2.1 and from continuity of β on $\mathfrak{P}(\Omega) \times \mathfrak{P}(\Omega)$. □

Proof of Lemma 1 Since the topology of weak convergence on $\mathfrak{P}(\Omega)$ is metrizable, it is sufficient to show that $\mu_m \to \mu$ weakly if and only if $d_s(\mu_m, \mu) \to 0$ holds as $m \to \infty$ for μ_m, μ in $\mathfrak{P}(\Omega)$. We first show that $d_s(\mu_m, \mu) \to 0$ implies $\mu_m \to \mu$ weakly: by the portmanteau-theorem (e.g. Theorem 11.1.1 in [9]), it is sufficient to show that $\limsup_{m\to\infty} \mu_m(F) \leq \mu(F)$ holds for every $F \subseteq \Omega$ that is closed in Ω. Let such a $F \subseteq \Omega$ be given. Then, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $\mu(F_\delta) - \mu(F) < \varepsilon$ holds where $F_\delta = \{x \in \Omega : |x - F| < \delta\}$, and this remains true if δ is made smaller. Now by Proposition 11.2.3 in [9] there exists a function $f_\delta : \Omega \to [0, 1]$ contained in the space of bounded Lipschitz functions on Ω (that is, in the space $\mathsf{C}^1(\Omega)$ defined in the proof of Corollary 1 above) which satisfies

$$f_\delta(x) = \begin{cases} 1 & \text{for } x \in F \\ 0 & \text{for } x \in \Omega \backslash F_\delta \end{cases}.$$

Since f_δ is bounded Lipschitz it is uniformly continuous on Ω and hence there exists a continuous extension f_δ^{ext} to the closure of Ω, i.e., $f_\delta^{\text{ext}} \in \mathsf{C}(\bar{\Omega})$. By the Stone–Weierstraß theorem, we can find a polynomial g_δ such that $\sup_{x\in\Omega} |g_\delta(x) - f_\delta(x)| \leq \sup_{x\in\bar{\Omega}} \left|g_\delta(x) - f_\delta^{\text{ext}}(x)\right| < \varepsilon$. Obviously, $g_\delta \in \mathsf{W}_2^s(\Omega, \lambda)$ for every s. Now

$$\begin{aligned}
\mu_m(F) &\leq \int_\Omega f_\delta \mathrm{d}\mu_m = \int_\Omega (f_\delta - g_\delta)\mathrm{d}\mu_m + \int_\Omega g_\delta \mathrm{d}\mu_m \\
&\leq \int_\Omega g_\delta \mathrm{d}\mu_m + \varepsilon \underset{m\to\infty}{\to} \int_\Omega g_\delta \mathrm{d}\mu + \varepsilon \\
&\leq \int_\Omega f_\delta \mathrm{d}\mu + 2\varepsilon \leq \mu(F_\delta) + 2\varepsilon \leq \mu(F) + 3\varepsilon
\end{aligned}$$

which proves one direction. To prove the converse, observe that $\mathcal{U}_{s,1}$ $(s > 1/2)$ is uniformly bounded and equicontinuous on Ω by Part 5 of Proposition 2 and thus $\mu_m \to \mu$ weakly implies $d_s(\mu_m, \mu) \to 0$ by Corollary 11.3.4 in [9]. □

Proof of Corollary 2 Follows immediately from Theorem 3 and the fact that $\mathcal{U}_{s,1}$ is a $\mathbb{P}$-Donsker class by Part 5 of Proposition 2. Measurability of $d_s(\hat{\mathbb{P}}_n, \mathbb{P}_n)$ as well as of $d_s(\hat{\mathbb{P}}_n, \mathbb{P})$ follows from the results in Sect. 4.2.1. □

Proof of Corollary 3 The case $r \geq 0$ was proved in Proposition 6 above. The case $r < -1/2$ follows directly from Theorem 3 by Prohorov's theorem. We next consider the case $-1/2 \leq r < 0$. From what has just been established, we have $\left\|\hat{p}_n - p_0\right\|_{0,2,\lambda} = O_{\mathbb{P}}(n^{-t/(2t+1)})$ as well as $\left\|\hat{p}_n - p_0\right\|_{-s,2,\lambda} = O_{\mathbb{P}}(n^{-1/2})$ for $s > 1/2$. From Theorem 1.12.5 (and 1.2.43) in [20], we obtain the interpolation inequality

$$\|f\|_{r,2,\lambda} \leq C \|f\|_{0,2,\lambda}^{1+r/s} \|f\|_{-s,2,\lambda}^{-r/s} \tag{38}$$

for all $f \in \mathcal{L}^2(\Omega, \lambda)$ and $0 > r \geq -1/2$. Choose $s = 1/2 + \varepsilon$ where $\varepsilon > 0$ is arbitrary. Applying (38) to $f = \hat{p}_n - p_0$ gives

$$\left\|\hat{p}_n - p_0\right\|_{r,2,\lambda} = O_{\mathbb{P}}(n^{-(t-r)/(2t+1)} n^{-(r-(r/(1+2\varepsilon)))/(2t+1)}) = O_{\mathbb{P}}(n^{-(t-r)/(2t+1)} n^{\delta})$$

for $0 > r \geq -1/2$. The second exponent δ is positive and can be made arbitrarily close to zero by choosing ε arbitrarily small. □

Proof of Corollary 4 We first prove the claim under the first condition. Observe that the relation $\hat{p}_n \in A$ eventually holds by Part 2 of Proposition 6. Using Corollary 3 with $\delta > 0$ arbitrary, we obtain that

$$\left|\Phi(\hat{p}_n) - \Phi(p_0) - D\Phi(p_0)(\hat{p}_n - p_0)\right| = O_{\mathbb{P}^*}(n^{\omega[-(t-r)/(2t+1)+\delta]}) = o_{\mathbb{P}^*}(n^{-1/2}) \tag{39}$$

holds by the assumption on ω. Since the class $\mathcal{F}$ satisfies Condition 2 by assumption, we can use (39) together with Theorem 2 to obtain that

$$\begin{aligned}\sqrt{n}(\Phi(\hat{p}_n) - \Phi(p_0)) &= \sqrt{n}D\Phi(p_0)(\hat{p}_n - p_0) + o_{\mathbb{P}^*}(1) = \hat{\nu}_n(u_{\Phi,\mathbb{P}}) + o_{\mathbb{P}^*}(1) \\ &= \nu_n(u_{\Phi,\mathbb{P}}) + o_{\mathbb{P}^*}(1)\end{aligned}$$

holds. This proves (19) by the classical central limit theorem, observing that $u_{\Phi,\mathbb{P}} \in \mathcal{L}^2(\Omega, \mathbb{P})$. □

Proof of Corollary 5 We may assume w.l.o.g. that the bounded C^∞-domain Ω equals the open interval (a, b). Denote by $\langle \cdot, \cdot \rangle$ the inner product on $L^2(\Omega, \lambda)$. We first prove the auxiliary result that

$$\left\langle D_w^{2\alpha} p_0, h \right\rangle = (-1)^\alpha \left\langle D_w^\alpha p_0, D_w^\alpha h \right\rangle \tag{40}$$

holds for all h with $[h]_\lambda \in W_2^\alpha(\Omega, \lambda)$. For $\alpha = 0$, this is trivial, hence assume $\alpha \geq 1$. It suffices to show (40) for all $h \in \mathsf{W}_2^\alpha(\Omega, \lambda)$. By Part 2 of Proposition 2 (and [1], Theorem 4.12/III), there exist continuous functions p_0^{ext} and h^{ext}, extending p_0 and h to $\mathbb{R}$, that satisfy $[p_0^{\text{ext}}]_\lambda \in W_2^t(\mathbb{R}, \lambda\,|\mathbb{R})$ and $[h^{\text{ext}}]_\lambda \in W_2^\alpha(\mathbb{R}, \lambda\,|\mathbb{R})$, respectively. Note that both $[D_w^l p_0^{\text{ext}}]_\lambda$ and $[D_w^j h^{\text{ext}}]_\lambda$, $0 \leq l \leq 2\alpha - 1$, $0 \leq j \leq \alpha - 1$, belong to $W_2^1(\mathbb{R}, \lambda\,|\mathbb{R})$ and hence are absolutely continuous on the closed interval $[a, b]$ in the sense that each equivalence class contains an absolutely continuous representative. By the assumption in the corollary we have for these representatives that $D_w^l p_0^{\text{ext}}(a) = D_w^l p_0^{\text{ext}}(b) = 0$ for every $\alpha \leq l \leq 2\alpha - 1$. Applying integration by parts proves (40) since

$$\begin{aligned}\left\langle D_w^{2\alpha} p_0, h \right\rangle &= \int_{(a,b)} D_w^{2\alpha} p_0 h \mathrm{d}\lambda = \int_{[a,b]} D_w^{2\alpha} p_0^{\text{ext}} h^{\text{ext}} \mathrm{d}\lambda = (-1)^\alpha \int_{[a,b]} D_w^\alpha p_0^{\text{ext}} D_w^\alpha h^{\text{ext}} \mathrm{d}\lambda \\ &= (-1)^\alpha \int_{(a,b)} D_w^\alpha p_0 D_w^\alpha h \mathrm{d}\lambda = (-1)^\alpha \left\langle D_w^\alpha p_0, D_w^\alpha h \right\rangle.\end{aligned}$$

We now apply Part 1 of Corollary 4 and verify the respective conditions with $A = W_2^\alpha(\Omega, \lambda)$ and parameters $\omega = 2$ and $r = \alpha$. We first show that the functional $\Phi : W_2^\alpha(\Omega, \lambda) \to \mathbb{R}$ given by $[p]_\lambda \longmapsto \left\| D_w^\alpha p \right\|_{2,\lambda}^2$ is Fréchet-differentiable at $[p_0]_\lambda$, with derivative $D\Phi(p_0)(\cdot) = (-1)^\alpha \left\langle 2D_w^{2\alpha} p_0, \cdot \right\rangle$. Observe that

$$\begin{aligned}&\left| \left\| D_w^\alpha (p_0 + h) \right\|_{2,\lambda}^2 - \left\| D_w^\alpha p_0 \right\|_{2,\lambda}^2 - D\Phi(p_0)(h) \right| \\ &= \left| \left\langle D_w^\alpha (p_0 + h), D_w^\alpha (p_0 + h) \right\rangle - \left\langle D_w^\alpha p_0, D_w^\alpha p_0 \right\rangle - D\Phi(p_0)(h) \right| \\ &= \left| 2 \left\langle D_w^\alpha p_0, D_w^\alpha h \right\rangle + \left\langle D_w^\alpha h, D_w^\alpha h \right\rangle - D\Phi(p_0)(h) \right| \\ &= \left| \left\langle D_w^\alpha h, D_w^\alpha h \right\rangle \right| = O(\|h\|_{\alpha,2,\lambda}^2)\end{aligned}$$

holds for all h with $[h]_\lambda \in W_2^\alpha(\Omega,\lambda)$, because of (40). Fréchet differentiability of Φ on $W_2^\alpha(\Omega,\lambda)$ at the point $[p_0]_\lambda$ follows since $D\Phi(p_0)(\cdot) \in (L^2(\Omega,\lambda))' \subseteq (W_2^\alpha(\Omega,\lambda))'$ in view of the fact that $[D_w^{2\alpha} p_0]_\lambda \in L^2(\Omega,\lambda)$. Now the condition $t - 2\alpha > 1/2$ of this corollary is equivalent to the required inequality for ω in Corollary 4, and also implies that there exists a continuous representative of $[-2D_w^{2\alpha} p_0]_\lambda$, that is, $-2D_w^{2\alpha} p_0 \in \mathsf{W}_2^{t-2\alpha}(\Omega,\lambda)$. Proposition 1 now shows that $\mathcal{F} = \{-2D_w^{2\alpha} p_0\}$ satisfies Condition 2, completing the proof. □

Proof of Corollary 6 Note first that

$$\sqrt{n}(\hat{p}_n * \hat{p}_n - p_0 * p_0) = 2\sqrt{n}(\hat{p}_n - p_0) * p_0 + \sqrt{n}(\hat{p}_n - p_0) * (\hat{p}_n - p_0)$$

holds. The remainder term is asymptotically negligible, since

$$\left\|(\hat{p}_n - p_0) * (\hat{p}_n - p_0)\right\|_\infty \le \left\|\hat{p}_n - p_0\right\|^2_{2,\lambda|(0,1)} = O_{\mathbb{P}}(n^{-2t/2t+1}) = o_{\mathbb{P}}(n^{-1/2})$$

holds by Young's inequality and Proposition 6. It hence remains to show that $2\sqrt{n}(\hat{p}_n - p_0) * p_0$ converges in law in $\mathsf{C}((0,1))$. To do this, we apply Theorem 3 and a result in [22]: denote by $Fp(u)$, $u \in \mathbb{Z}$, the Fourier coefficients of a continuous real-valued function p on $\mathbb{T}$, and define, for $s > 1/2$, the set

$$\mathcal{U}_{\mathbb{T},s,1} = \left\{ f : \mathbb{T} \to \mathbb{R} : f \text{ continuous}, \sum_{u\in\mathbb{Z}} |Ff(u)|^2 (1 + |u|^2)^t \le 1 \right\},$$

let $\mathcal{U}_{(0,1),s,1}$ be the corresponding set of restrictions of elements of $\mathcal{U}_{\mathbb{T},s,1}$ to $(0,1)$, and let $\mathcal{U}_{s,B}$ be the set appearing in Theorem 3 with $\Omega = (0,1)$. It follows (from 3.5.1/13 and 3.5.4/18,19 in [31] and 3.4.2/6 in [34]) that $\mathcal{U}_{(0,1),s,1}$ is contained in $\mathcal{U}_{s,B}$ for some $0 < B < \infty$. Let now h be an element of the space $\mathcal{UC}((0,1))$ of bounded uniformly continuous functions on $(0,1)$, and let $h^{\text{ext}} : \mathbb{T} \to \mathbb{R}$ denote its (periodic) extension obtained by setting $h(0) = \lim_{x\to 1} h(x) = 0$. We then have

$$\begin{aligned}
\left\|h^{\text{ext}}\right\|_{\infty,\mathcal{U}_{\mathbb{T},s,1}} &= \sup_{f\in\mathcal{U}_{\mathbb{T},s,1}} \left| \int_{\mathbb{T}} h^{\text{ext}} f \mathrm{d}\lambda \right| = \sup_{f\in\mathcal{U}_{\mathbb{T},s,1}} \left| \int_{(0,1)} h^{\text{ext}} f \mathrm{d}\lambda \right| \\
&= \sup_{f\in\mathcal{U}_{\mathbb{T},s,1}} \left| \int_{(0,1)} h f \mathrm{d}\lambda \right| \le \sup_{f\in\mathcal{U}_{s,B}} \left| \int_{(0,1)} h f \mathrm{d}\lambda \right| = \|h\|_{\infty,\mathcal{U}_{s,B}} \qquad (41)
\end{aligned}$$

holds for every $h \in \mathcal{UC}((0,1))$. That is, the mapping $h \longmapsto h^{\text{ext}}$ from $\mathcal{UC}((0,1))$ into the space $\mathcal{L}^\infty(\mathbb{T})$ of bounded measurable functions on $\mathbb{T}$ is continuous if $\mathcal{UC}((0,1))$ and $\mathcal{L}^\infty(\mathbb{T})$ are viewed as linear (topological) subspaces of $\ell^\infty(\mathcal{U}_{s,B})$ and $\ell^\infty(\mathcal{U}_{\mathbb{T},s,1})$, respectively. Consequently, since $\sqrt{n}(\hat{p}_n \mathrm{d}\lambda - p_0 \mathrm{d}\lambda)$ converges in law in the metric space $(\mathcal{UC}((0,1)), \|\cdot\|_{\infty,\mathcal{U}_{s,B}})$ for $s > 1/2$ by Theorem 3, we

conclude that the (periodic) extensions $\sqrt{n}(\hat{p}_n \mathrm{d}\lambda - p_0 \mathrm{d}\lambda)$ converge in law in $\ell^\infty(\mathcal{U}_{\mathbb{T},s,1})$ for $s > 1/2$. To complete the proof, note that $\sum_{u\in\mathbb{Z}} |Fp_0(u)|^2 (1 + |u|^2)^t < \infty$ holds by Condition 1, periodicity of p_0, and again 3.5.1/13, 3.5.4/18,19 in [31] and 3.4.2/6 in [34]. Hence, by Part 1 of Theorem 1 in [22], $2\sqrt{n}(\hat{p}_n - p_0) * p_0$ converges in law in the space of continuous functions on $\mathbb{T}$, and hence also in $\mathsf{C}((0,1))$. □

Acknowledgements I am grateful to two anonymous referees whose detailed comments and suggestions helped to improve the paper in every respect. I furthermore wish to thank my advisor and teacher B.M. Pötscher for his constant support during the dissertation phase, for an uncountable number of hours of discussions on the dissertation subject as well as for a meticulous and extensive proof reading of the manuscript that led to numerous important improvements. In particular, he suggested a much more well organized proof of Theorems 1 and 2. I am also indebted to my co-advisor Viktor Losert for helpful remarks on the manuscript. Finally, I wish to thank the participants of the 2005 Conference on High Dimensional Probability—in particular Richard Dudley, Evarist Giné, and Jon Wellner; as well as Hannes Leeb and Sara van de Geer for interesting discussion on the subject of the paper.

References

1. Adams, R.A., Fournier, J.F.: Sobolev spaces, 2nd edn. Academic, New York (2003)
2. Bickel, J.P., Ritov, Y.: Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhya Ser. A **50**, 381–393 (1988)
3. Bickel, J.P., Ritov, Y.: Nonparametric estimators which can be 'plugged-in'. Ann. Stat. **31**, 1033–1053 (2003)
4. Birgé, L., Massart, P.: Rates of convergence of minimum contrast estimators. Probab. Theory Relat. Fields **97**, 113–150 (1993)
5. Birgé, L., Massart, P.: Estimation of integral functionals of a density. Ann. Stat. **23**, 11–29 (1995)
6. Dieudonné, J.: Foundations of Modern Analysis. Academic, New York (1960)
7. Donoho, D.L., Liu R.C.: Geometrizing rates of convergence II, III. Ann. Stat. **19**, 633–667, 668–701 (1991)
8. Dudley, R.M.: Uniform Central Limit Theorems. Cambridge University Press, Cambridge (1999)
9. Dudley, R.M.: Real Analysis and Probability. Cambridge University Press, Cambridge (2002)
10. Dunford, N., Schwartz, J.T.: Linear Operators. Part I: General Theory. Interscience, New York (1966)
11. Frees, E.W.: Estimating densities of functions of observations. J. Am. Stat. Assoc. **89**, 517–525 (1994)
12. Giné, E.: Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev-norms. Ann. Stat. **3**, 1243–1266 (1975)
13. Giné, E., Mason, D.M.: On local U-statistic processes and the estimation of densities of functions of several variables. Ann. Stat. (in press) (2006)
14. Giné, E., Zinn, J.: Empirical processes indexed by Lipschitz functions. Ann. Probab. **14**, 1329–1338 (1986)
15. Hall, P., Marron, J.S.: Estimation of integrated squared density derivatives. Stat. Probab. Lett. **6**, 109–115 (1987)
16. Kiefer, J., Wolfowitz, J.: Asymptotically minimax estimation of concave and convex distribution functions. Z. Wahrscheinlichkeitstheorie verw. Gebiete **34**, 73–85 (1976)
17. Laurent, B.: Efficient estimation of integral functionals of a density. Ann. Stat. **24**, 659–681 (1996)
18. Laurent, B.: Estimation of integral functionals of a density and its derivatives. Bernoulli **3**, 181–211 (1997)
19. Leeb, H., Pötscher, B.M.: Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. Econom. Theory **22**, 69–97 (2006)

20. Lions, J.L., Magenes, E.: Non-Homogeneous Boundary Value Problems and Applications I. Springer, Berlin Heidelberg New York (1972)
21. Nickl, R.: Empirical and Gaussian processes on Besov classes. In: Giné, E., Kolchinskii, V., Li, W., Zinn, J. (eds.) High Dimensional Probability IV, IMS Lecture Notes (in press) (2006a)
22. Nickl, R.: On convergence and convolutions of random signed measures (preprint) (2006b)
23. Nickl, R.: Uniform central limit theorems for density estimators (preprint) (2006c)
24. Nickl, R., Pötscher, B.M.: Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov- and Sobolev-type. J. Theor. Probab. (in press) (2005)
25. Pötscher, B.M., Prucha, I.R.: Dynamic Nonlinear Econometric Models. Asymptotic Theory. Springer, Berlin Heidelberg New York (1997)
26. Radulovic, D., Wegkamp, M.: Weak convergence of smoothed empirical processes. Beyond Donsker classes. In: Giné, E., Mason, D.M., Wellner, J.A. (eds.) High Dimensional Probability II, Progr. Probab. **47**, pp. 89–105 Birkhäuser, Boston (2000)
27. Radulovic, D., Wegkamp, M.: Necessary and sufficient conditions for weak convergence of smoothed empirical processes. Stat. Probab. Lett.**61**, 321–336 (2003)
28. Rost, D.: Limit theorems for smoothed empirical processes. In: Giné, E., Mason, D.M., Wellner, J.A. (eds.) High dimensional probability II, Progr. Probab. **47**, pp. 107–113 Birkhäuser, Boston (2000)
29. Rufibach, K., Dümbgen, L.: Maximum likelihood estimation of a log-concave density. Basic properties and consistency (preprint) (2004)
30. Schick, A., Wefelmeyer, W.: Root n consistent density estimators for sums of independent random variables. J. Nonparametr. Stat. **16**, 925–935 (2004)
31. Schmeisser, H.-J., Triebel, H.: Topics in Fourier Analysis and Function Spaces. Wiley, New York (1987)
32. Stone, C.J.: Optimal rates of convergence for nonparametric estimators. Ann. Stat. **8**, 1348–1360 (1980)
33. Strassen, V., Dudley, R.M.: The central limit theorem and ε-entropy. Probability and information theory. Lect. Notes Math. **1247**, 224–231 (1969)
34. Triebel, H.: Theory of Function Spaces. Birkhäuser, Basel (1983)
35. van de Geer, S.: Hellinger-consistency of certain nonparametric maximum likelihood estimators. Ann. Stat. **21**, 14–44 (1993)
36. van de Geer, S.: Empirical Processes in M-estimation. Cambridge University Press, Cambridge (2000)
37. van der Vaart, A.W.: Weak convergence of smoothed empirical processes. Scand. J. Stat. **21**, 501–504 (1994)
38. van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer, Berlin Heidelberg New York (1996)
39. von Mises, R.: On the asymptotic distribution of differentiable statistical functions. Ann. Math. Stat. **20**, 309–348 (1947)
40. Wong, W.H., Severini, T.A.: On maximum likelihood estimation in infinite dimensional parameter spaces. Ann. Stat. **19**, 603–632 (1991)
41. Wong, W.H., Shen, X.: Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. Ann. Stat. **23**, 339–362 (1995)
42. Yukich, J.E.: Weak convergence of smoothed empirical processes. Scand. J. Stat. **19**, 271–279 (1992)