Uniform Central Limit Theorems for the Grenander Estimator

Jakob Söhl *

University of Cambridge †

November 25, 2014

Abstract

We consider the maximum likelihood estimator for non-increasing densities known as the Grenander estimator. We prove uniform central limit theorems for certain subclasses of bounded variation functions and for Hölder balls of smoothness $s > 1/2$. We do not assume that the density is differentiable or continuous. Since nonparametric maximum likelihood estimators lie on the boundary, the derivative of the likelihood cannot be expected to equal zero as in the parametric case. Nevertheless, our proofs rely on the fact that the derivative of the likelihood can be shown to be small at the estimator.

MSC 2000 subject classification: 60F05, 62G07, 62E20

Key words: UCLT, Grenander estimator, NPMLE, Hölder class

1 Introduction

A fundamental approach to statistical estimation is finding the probability measure which renders the observation most likely. This principle of maximum likelihood estimation has proved very successful in parametric estimation but leads to difficulties in nonparametric problems since the likelihood is typically unbounded so that no maximum is attained. However, in nonparametric problems with shape constraints the maximum likelihood estimator is often well-defined and thus the maximum likelihood approach can be extended to these situations. Examples include non-increasing, convex, concave and log-concave functions.

The classical parametric maximum likelihood theory is based on the estimator $\hat{\theta}_n$ being in the interior of the parameter space and on the resulting fact that the derivative of the likelihood vanishes at $\hat{\theta}_n$. The theory of nonparametric maximum likelihood estimation is quite different from the parametric theory since the estimator lies on the boundary of the parameter space and thus in general the derivative of the likelihood will not be zero. But in some nonparametric situations the derivative of the likelihood can be shown to be sufficiently small and thus enabling a proof strategy parallel to the one in the classical parametric theory. [Nickl (2007)] considered

* I thank Evarist Giné and Richard Nickl for providing me with a preliminary version of their forthcoming book. This note is based on a chapter from their book and extends the results therein. I thank Richard Nickl for helpful discussions on the topic.

† Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, CB3 0WB, Cambridge, UK. Email: j.soehl@statslab.cam.ac.uk
the maximum likelihood estimator for estimating a density in a Sobolev ball and proved uniform central limit theorems using this approach.

We pursue this method of proof in the problem of estimating a non-increasing density $p_0$ on the unit interval. The maximum likelihood estimator $\hat{p}_n$ is called the Grenander estimator in this situation since it was first derived by Grenander (1956). It is well known to be the left-derivative of the least concave majorant of the empirical distribution function. The main results will be uniform central limit theorems for the Grenander estimator that in particular imply for functions $f$

$$\sqrt{n} \int_0^1 (\hat{p}_n(x) - p_0(x))f(x)dx \rightarrow^d N(0, \|f - P_0f\|_2^2),$$

where $P_0$ is the probability measure of the non-increasing density $p_0$ and $P_0f \equiv \int_0^1 f(x)dP_0(x)$.

Our results are uniform for $f$ varying in a class of functions and we cover two different types of classes. The first type is a subclass of the bounded variation functions and for each point of discontinuity $t$ of $p_0$ the indicator function $\mathbb{1}_{[0,t]}$ is contained in such a class. The second type of classes is given by balls in Hölder spaces $C^s$ of order $s > 1/2$.

Under a strict curvature condition and for continuously differentiable $p_0$, Kiefer and Wolfowitz (1976) proved that the difference between the Grenander estimator and the empirical distribution function in supremum norm is with probability one of order $n^{-2/3} \log(n)$. This implies in particular a uniform central limit theorem for the class of all indicator function $\mathbb{1}_{[0,t]}$, $t \in [0,1]$. Results similar to the Kiefer–Wolfowitz theorem hold under other shape constraints as well. Balabdaoui and Wellner (2007) showed such a result in the case where the density is assumed to be convex decreasing and where the maximum likelihood estimator is replaced by the least squares estimator. Dümbgen and Rufibach (2009) derived the rate of estimation for log-concave densities in supremum norm and showed that the difference between empirical and estimated distribution function is $o(n^{-1/2})$. Durot and Lopuhaä (2014) showed a general Kiefer–Wolfowitz type theorem which covers the estimation of monotone regression curves, monotone densities and monotone failure rates.

Jankowski (2014) studied the local convergence rates of the Grenander estimator in situations where it is misspecified and derived the asymptotic distribution of linear functionals under possible misspecification. Our perturbation approach is particularly suitable to domains where a strict curvature condition holds and leads to weaker assumptions on $p_0$ and $f$ in these domains. Beyond asymptotic normality for single functionals our work includes a uniformity in the underlying functional, which is important for the application of the results by Nickl (2009) concerning convolutions of density estimators.

The uniform results can be interpreted in the context of Bickel and Ritov (2003), who consider rate optimal estimators of the density which simultaneously lead to the efficient estimation of functionals with uniform convergence. They coin the expression “plug-in property” for such estimators. Nickl (2007) discusses applications in the context of the uniform central limit theorems for the MLE over a Sobolev ball. Uniform central limit theorems were also shown for kernel density estimators and for wavelet density estimators (Giné and Nickl, 2008, 2009).

This paper is organised as follows. Section 2 states the uniform central limit theorems for a subclass of the bounded variation functions and for Hölder balls. In Section 3 we explain the general approach. In Section 4 we derive upper and lower bounds in probability for the Grenander estimator and recall the $L^2$-convergence rate. In Section 5 the perturbation approach is further developed and the main results are proved.
2 Main results

Let \( X_1, \ldots, X_n \) be i.i.d. on \([0, 1]\) with law \( P_0 \) and distribution function \( F_0(x) = \int_0^x dP_0, x \in [0, 1] \). In order to state the main results we introduce some notation. We define the empirical measure \( P_n = n^{-1} \sum_{i=1}^n \delta_{X_i} \), the empirical cumulative distribution function \( F_n(x) = \int_0^x dP_n, x \in [0, 1] \) and the log-likelihood function

\[
\ell_n(p) = \frac{1}{n} \sum_{i=1}^n \log p(X_i)
\]

(1)

Under the assumption \( E_{P_0} | \log p(X) | < \infty \) for all \( p \in \mathcal{P} \), we can define the limiting log-likelihood function

\[
\ell(p) = \int_0^1 \log p(x) dP_0(x).
\]

(2)

If \( P \) is known to have a monotone decreasing density \( p \) then the associated maximum likelihood estimator \( \hat{p}_n \) maximises the log-likelihood function \( \ell_n(p) \) over

\[
\mathcal{P} = \mathcal{P}^{\text{mon}} = \left\{ p : [0, 1] \rightarrow [0, \infty), \int_0^1 p(x) dx = 1, p \text{ is non-increasing} \right\},
\]

that is,

\[
\max_{p \in \mathcal{P}^{\text{mon}}} \ell_n(p) = \ell_n(\hat{p}_n).
\]

(3)

The maximum likelihood estimator \( \hat{p}_n \) is known to be the left-derivative of the least concave majorant \( \hat{F}_n \) of the empirical distribution function \( F_n \). For a set \( T \) let \( \ell^\infty(T) \) denote the space of bounded real-valued functions with the usual supremum norm \( \| \cdot \|_\infty \). Throughout we will denote by \( \rightarrow^d \) the convergence in distribution as in Chapter 1 in van der Vaart and Wellner (1996). The \( P_0 \)-Brownian bridge \( G_{P_0} \) is defined as the tight Gaussian random variable arising from the centred Gaussian process with covariance

\[
E[G_{P_0}(f)G_{P_0}(g)] = P_0(fg) - P_0fP_0g,
\]

where \( P_0f = \int_0^1 f(x) dP_0(x) \).

The first main result is a uniform central limit theorem for a subclass of the bounded variation functions. We start with the general result in Theorem 1 and consider its consequences in Corollary 1 and Theorem 2. Let \( f, p_0 \in L^1[0, 1] \) and assume that the weak derivatives of \( f \) and \( p_0 \) in the sense of regular Borel signed measures exist and denote them by \( Df \) and \( Dp_0 \), respectively; cf., e.g., p. 42 in Ziemer (1989). We define \( BV[0, 1] := \{ f \in L^1[0, 1] : \| f \|_1 + |Df|(0, 1) < \infty \} \), where \( |Df| \) is total variation of the signed measure \( Df \). In the following theorem it will be important that \( Df \) is absolutely continuous with respect to \( Dp_0 \) since we want to ensure that the perturbations \( p_0 \pm \eta f \) with \( |\eta| \) small (or slightly modified perturbations) are decreasing functions. To this end we denote the Radon–Nikodym derivative by \( Df/Dp_0 \) and assume that its essential supremum with respect to \( Dp_0 \), denoted by \( \| Df/Dp_0 \|_{\infty,Dp_0} \), is bounded.

**Theorem 1.** Suppose \( p_0 \in \mathcal{P} \) is bounded and satisfies \( p_0 \geq \zeta > 0 \) on \([0, 1]\). Then for \( \mathcal{B} = \{ f \in BV[0, 1] : \| f \|_\infty + \| Df/Dp_0 \|_{\infty,Dp_0} \leq B \} \) we have

\[
\sup_{f \in \mathcal{B}} |(\hat{P}_n - P_0)(f)| = O_{P_0}(Bn^{-2/3})
\]

and consequently

\[
\sqrt{n}(\hat{P}_n - P_0) \rightarrow^d G_{P_0} \text{ in } \ell^\infty(\mathcal{B}).
\]
Theorem 2. Suppose on \( F \) we see that the \( s \) in the \( C \) ∥ ∥ of its Kullback–Leibler projection) and assumes on the portion of support where that the Lebesgue measure \( \lambda \) \( p \) allow for discontinuities of \( p \) misspecification, meaning that the true density \( \xi \) > 0.

Theorem 1 and 2 is that they are not for a fixed function \( f \) but they are uniform in \( f \) over classes of functions. A distinct feature of Theorems 1 and 2 is that they are for a fixed function \( f \) but they are uniform in \( f \) over classes of functions. Taking a different perspective Jankowski’s emphasis is on the problem of possible misspecification, meaning that the true density \( p_0 \) does not necessarily has to be non-increasing.

Corollary 1. Suppose \( p \in P \) is bounded and satisfies \( p_0 \geq \zeta > 0 \) on \([0, 1]\). Then for all \( t \in (0, 1) \) where \( p_0 \) is discontinuous we have

\[
|\hat{F}_n(t) - F_0(t)| = O_{P_0}(n^{-2/3})
\]

and consequently

\[
\sqrt{n}(\hat{F}_n(t) - F_0(t)) \rightarrow N(0, F_0(t) - F_0(t)^2).
\]

Under a strict curvature condition the set \( B \) in Theorem 1 contains \( C^1 \)-Hölder balls. By strict curvature condition we have in mind that \( p_0' \) is bounded away from zero, that is \( \inf_{x \in [0, 1]} |p_0'(x)| \geq \xi > 0 \) or equivalently \( \|1/p_0'\|_\infty \leq 1/\xi \). We do not want to assume that \( p_0' \) exists classically. To allow for discontinuities of \( p_0 \) and to stay in the general setting of weak derivatives we assume that the Lebesgue measure \( \lambda \) on \([0, 1]\) is absolutely continuous with respect to \( Dp_0 \) and replace the assumption \( \|1/p_0'\|_\infty \leq 1/\xi \) by the weaker assumption \( \|\lambda/Dp_0\|_\infty \leq 1/\xi \), where \( \lambda/Dp_0 \) is the Radon–Nikodym derivative. We remark that \( \|\lambda/Dp_0\|_\infty, Dp_0 = \|\lambda/Dp_0\|_\infty \). Let \( F \) be a \( C^1 \)-Hölder ball and \( f \in F \). Then

\[
\|Df/Dp_0\|_\infty, Dp_0 = \|Df/\lambda \cdot Dp_0\|_\infty, Dp_0 \leq (1/\xi) \|f'\|_\infty
\]

and we see that the \( C^1 \)-Hölder ball \( F \) is contained in \( B \) for some \( B \). This special case of Theorem 1 with \( F \) instead of \( B \) can be generalized to functions in \( s \)-Hölder spaces \( C^s \) of order \( s > 1/2 \).

Theorem 2. Suppose \( p_0 \in P \) is bounded and satisfies \( p_0 \geq \zeta > 0 \). Let the Lebesgue measure \( \lambda \) on \([0, 1]\) be absolutely continuous with respect to \( Dp_0 \) and let \( \|\lambda/Dp_0\|_\infty < \infty \). Let \( F \) be a ball in the \( s \)-Hölder space of order \( s > 1/2 \). Then

\[
\sup_{F} |(\hat{P}_n - P_n)(f)| = o_{\lambda}(1/\sqrt{n})
\]

as \( n \to \infty \) and thus

\[
\sqrt{n}(\hat{P}_n - P_0) \to^d G_{P_0} \text{ in } \ell^\infty(F).
\]

In particular for any \( f \in C^s \) we have

\[
\sqrt{n} \int_0^1 (\hat{p}_n(x) - p_0(x))f(x)dx \to^d N(0, \|f - P_0 f\|_{L^2(P_0)}^2).
\]

The proof of Theorem 2 will be given at the end of the paper.
is continuously differentiable and that \( |p'_0| \) is bounded, which is used for the application of the Kiefer–Wolfowitz theorem in the proof. The assumption that \( p_0 \) is continuously differentiable is widely used in the literature on the Grenander estimator so it is worthwhile to remark that we do not require \( p_0 \) to be differentiable nor to be continuous. The function \( f \) defining the functional is assumed by Jankowski (2014) to be differentiable on the curved part and to be in \( L^p, p > 2 \), on the flat part. On the curved part Theorem 2 allows for discontinuities of \( f \) at points where \( p_0 \) is discontinuous and Theorem 2 only requires Hölder smoothness of order \( s > 1/2 \). On a possible flat part Theorem 1 assumes \( f \) to be constant which in view of the results by Jankowski (2014) is the natural condition to ensure a Gaussian limit. To summarise the discussion we can say that the perturbation approach has the advantage of providing uniform results under low regularity assumptions on \( p_0 \) and requires stronger assumptions on \( f \) on the flat part while needing weaker assumptions on the curved part.

3 The derivative of the likelihood function

Many classical properties of maximum likelihood estimators \( \hat{\theta}_n \) of regular parameters \( \theta \in \Theta \subset \mathbb{R}^p \), such as asymptotic normality, are derived from the fact that the derivative of the log-likelihood function vanishes at \( \hat{\theta}_n \),

\[
\frac{\partial}{\partial \theta} \ell_n(\theta) |_{\theta = \hat{\theta}_n} = 0.
\]  

This typically relies on the assumption that the true parameter \( \theta_0 \) is interior to \( \Theta \) so that by consistency \( \hat{\theta}_n \) will then eventually also be. In the infinite-dimensional setting, even if one can define an appropriate notion of derivative, this approach is usually not viable since \( \hat{p}_n \) is never an interior point in the parameter space even when \( p_0 \) is.

We now investigate these matters in more detail in the setting where \( \mathcal{P} \) consists of bounded probability densities. In this case we can compute the Fréchet derivatives of the log-likelihood function on the space \( L^\infty = L^\infty(\mathcal{X}) \) equipped with the \( \| \cdot \|_\infty \)-norm. Recall that a real-valued function \( L : U \rightarrow \mathbb{R} \) defined on an open subset \( U \) of a Banach space \( B \) is Fréchet differentiable at \( f \in U \) if

\[
\lim_{||h||_B \to 0} \frac{L(f + h) - L(f) - DL(f)[h]}{||h||_B} = 0
\]  

for some linear continuous map \( DL(f) : B \rightarrow \mathbb{R} \). If \( f' \in U \) is such that the line segment \((1 - t)f + tf', t \in (0, 1)\), joining \( f \) and \( f' \) lies in \( U \) (for instance if \( U \) is convex) then the directional derivative of \( L \) at \( f \) in the direction \( f' \) equals precisely

\[
\lim_{t \to 0^+} \frac{L(f + tf' - f) - L(f)}{t} = DL(f)[f' - f].
\]

The following proposition shows that the log-likelihood function \( \ell_n \) is Fréchet differentiable on the open convex subset of \( L^\infty \) consisting of functions that are positive at the sample points. A similar result holds for \( \ell \) if one restricts to functions that are bounded away from zero. We recall here these results of Proposition 3 in Nickl (2007).

**Proposition 1.** For any finite set of points \( x_1, \ldots, x_n \in \mathcal{X} \) define

\[
\mathcal{U}(x_1, \ldots, x_n) = \left\{ f \in L^\infty : \min_{1 \leq i \leq n} f(x_i) > 0 \right\}
\]

and

\[
\mathcal{U} = \left\{ f \in L^\infty : \inf_{x \in \mathcal{X}} f(x) > 0 \right\}.
\]
Then \( \mathcal{U}(x_1, \ldots, x_n) \) and \( \mathcal{U} \) are open subsets of \( L^\infty(\mathcal{X}) \).

Let \( \ell_n \) be the log-likelihood function from (6) based on \( X_1, \ldots, X_n \) \( \sim \text{i.i.d.} \) \( P_0 \), and denote by \( P_n \) the empirical measure associated with the sample. Let \( \ell \) be as in (3). For \( \alpha \in \mathbb{N} \) and \( f_1, \ldots, f_\alpha \in L^\infty \) the \( \alpha \)-th Fréchet derivatives of \( \ell_n : \mathcal{U}(X_1, \ldots, X_n) \to \mathbb{R} \), \( \ell : \mathcal{U} \to \mathbb{R} \) at a point \( f \in \mathcal{U}(X_1, \ldots, X_n) \), \( f \in \mathcal{U} \), respectively, are given by

\[
D^\alpha \ell_n(f)[f_1, \ldots, f_\alpha] \equiv (-1)^{\alpha-1}(\alpha - 1)! P_n(f^{-\alpha} f_1 \cdots f_\alpha), \quad (6)
\]

\[
D^\alpha \ell(f)[f_1, \ldots, f_\alpha] \equiv (-1)^{\alpha-1}(\alpha - 1)! P_0(f^{-\alpha} f_1 \cdots f_\alpha). \quad (7)
\]

We deduce from the above proposition the intuitive fact that the limiting log-likelihood function has a derivative at the true point \( p_0 > 0 \) that is zero in all ‘tangent space’ directions \( h \) in

\[
\mathcal{H} \equiv \left\{ h : \int_X h = 0 \right\} \quad (8)
\]

since

\[
D\ell(p_0)[h] = \int_X p_0^{-1} h dP_0 = \int_X h = 0. \quad (9)
\]

However, in the infinite-dimensional setting the empirical counterpart of (9),

\[
D\ell_n(\hat{p}_n)[h] = 0 \quad (10)
\]

for \( h \in \mathcal{H} \) and \( \hat{p}_n \), the nonparametric maximum likelihood estimator is not true in general. Even if the set \( \mathcal{P} \) the likelihood was maximised over is contained in \( \mathcal{U}(X_1, \ldots, X_n) \) it will itself in typical nonparametric situations have empty interior in \( L^\infty \), and the maximiser \( \hat{p}_n \) will lie at the boundary of \( \mathcal{P} \). As a consequence we cannot expect that \( \hat{p}_n \) is a zero of \( D\ell_n \).

Following ideas in Nickl (2007) we can circumvent this problem in some situations: if the true value \( p_0 \) lies in the ‘interior’ of \( \mathcal{P} \) in the sense that local \( L^\infty \)-perturbations of \( p_0 \) are contained in \( \mathcal{P} \cap (X_1, \ldots, X_n) \), then we can bound \( D\ell_n \) at \( \hat{p}_n \).

**Lemma 1.** Let \( \hat{p}_n \) be as in (3) and suppose that for some \( h \in L^\infty \), \( \eta > 0 \), the line segment joining \( p_n \) and \( p_0 \pm \eta h \) is contained in \( \mathcal{P} \cap \mathcal{U}(X_1, \ldots, X_n) \). Then

\[
|D\ell_n(\hat{p}_n)[h]| \leq (1/\eta) |D\ell_n(\hat{p}_n)[\hat{p}_n - p_0]|. \quad (11)
\]

**Proof.** Since \( \hat{p}_n \) is a maximiser over \( \mathcal{P} \) we deduce from differentiability of \( \ell_n \) on \( \mathcal{U}(X_1, \ldots, X_n) \) that the derivative at \( \hat{p}_n \) in the direction \( p_0 + \eta h \in \mathcal{P} \cap \mathcal{U}(X_1, \ldots, X_n) \) necessarily has to be nonpositive, that is

\[
D\ell_n(\hat{p}_n)[p_0 + \eta h - \hat{p}_n] = \lim_{t \to 0^+} \frac{\ell_n(\hat{p}_n + t(p_0 + \eta h - \hat{p}_n)) - \ell_n(\hat{p}_n)}{t} \leq 0 \quad (12)
\]

or, by linearity of \( D\ell_n(\hat{p}_n)[\cdot] \),

\[
D\ell_n(\hat{p}_n)[\eta h] \leq D\ell_n(\hat{p}_n)[\hat{p}_n - p_0]. \quad (13)
\]

Applying the same reasoning with \(-\eta\) we see

\[
|D\ell_n(\hat{p}_n)[\eta h]| \leq D\ell_n(\hat{p}_n)[\hat{p}_n - p_0] = |D\ell_n(\hat{p}_n)[\hat{p}_n - p_0]|. \quad (14)
\]

Divide by \( \eta \) to obtain the result. \( \square \)
The above lemma is interesting if we are able to show that

\[ Dl_n(\hat{p}_n)[\hat{p}_n - p_0] = o_P(1/\sqrt{n}), \]
as then the same rate bound carries over to \( Dl_n(\hat{p}_n)[h] \). This can in turn be used to mimic the finite-dimensional asymptotic normality proof of maximum likelihood estimators, which does not require (14) but only that the score is of smaller stochastic order of magnitude than \( 1/\sqrt{n} \). As a consequence we will be able to obtain the asymptotic distribution of linear integral functionals of \( \hat{p}_n \), and more generally, for \( \hat{P}_n \) the probability measure associated with \( \hat{p}_n \), central limit theorems for \( \sqrt{n}(\hat{P}_n - P) \) in ‘empirical process - type’ spaces \( \ell^\infty(F) \). To understand this better we notice that Proposition (13) implies the following relationships: If we define the following projection of \( f \in L^\infty \) onto \( \mathcal{H} \),

\[ \pi_0(f) \equiv (f - P_0f)p_0 \in \mathcal{H}, \quad P_0(f) = \int_X f df, \]

(15) and if we assume \( p_0 > 0 \) then

\[ \int_X (\hat{p}_n - p_0) f dx = \int_X p_0^{-2}(\hat{p}_n - p_0)(f - P_0f)p_0 df = -D^2 \ell(p_0)[\hat{p}_n - p_0, \pi_0(f)] \]

and

\[ Dl_n(p_0)[\pi_0(f)] = (P_n - P_0)f \]

so that:

**Lemma 2.** Suppose \( p_0 > 0 \). Let \( \hat{p}_n \) be as in (13) and let \( \hat{P}_n \) be the random probability measure induced by \( \hat{p}_n \). For any \( f \in L^\infty \) and \( P_n \) the empirical measure we have

\[ |(\hat{P}_n - P_n)(f)| = \left| \int_X f d(\hat{P}_n - P_n) \right| = |Dl_n(p_0)[\pi_0(f)] + D^2 \ell(p_0)[\hat{p}_n - p_0, \pi_0(f)]|. \]

(16) Heuristically the right hand side equals, up to second order

\[ Dl_n(\hat{p}_n)[\pi_0(f)] - D^2 \ell_n(p_0)[\hat{p}_n - p_0, \pi_0(f)] + D^2 \ell(p_0)[\hat{p}_n - p_0, \pi_0(f)] \]

(17) Control of (11) at a rate \( o_P(1/\sqrt{n}) \) combined with stochastic bounds on the second centred log-likelihood derivatives and convergence rates for \( \hat{p}_n - p_0 \to 0 \) thus give some hope that one may be able to prove

\[ (\hat{P}_n - P_0 - P_0)(f) = (\hat{P}_n - P_n)(f) = o_P(1/\sqrt{n}) \]

and that thus, by the central limit theorem for \( (P_n - P_0)f \),

\[ \sqrt{n} \int_X (\hat{p}_n - p_0) f dx \to^d N(0, P_0(f - P_0f)^2) \]
as \( n \to \infty \).

## 4 Bounding the estimator and \( L^2 \)-convergence rate

We establish some first probabilistic properties of \( \hat{p}_n \) that will be useful below: If \( p_0 \) is bounded away from zero then so it \( \hat{p}_n \) on the interval \([0, X_{(n)}]\) where \( X_{(n)} \) is the last order statistics. Similarly if \( p_0 \) is bounded above then so is \( \hat{p}_n \) with high probability.
Lemma 3. a) Suppose the true density \( p_0 \) satisfies \( \inf_{x \in [0,1]} p_0(x) > 0. \) Then, for every \( \epsilon > 0, \) there exists \( \xi > 0 \) and a finite index \( N(\epsilon) \) such that, for all \( n \geq N(\epsilon), \)
\[
P_0^N \left( \inf_{x \in [0,X(n)]} \hat{p}_n(x) < \xi \right) = \Pr \left( \hat{p}_n(X(n)) < \xi \right) < \epsilon
\]
b) Suppose the true density \( p_0(x) \leq K < \infty \) for all \( x \in [0,1]. \) Then, for every \( \epsilon > 0, \) there exists \( 0 < k < \infty \) such that,
\[
P_0^N \left( \sup_{x \in [0,1]} \hat{p}_n(x) > k \right) = \Pr \left( \hat{p}_n(0) > k \right) < \epsilon
\]

**Proof.** a) The first equality is obvious, since \( \hat{p}_n \) is monotone decreasing. Let \( X_1, \ldots, X_n \) denote the order statistic of \( X_1, \ldots, X_n. \) On each interval \( (X_{j-1}, X_j), \) \( f_n \) is the slope of the least concave majorant of \( F_n. \) The least concave majorant connects \( (X_n, 1) \) and at least one other order statistic (possibly \( (X_0, 0) := (0, 0) \), so that
\[
\{\hat{p}_n(X(n)) < \xi\} \subseteq \{X(n) - X_{n-j} > j/\xi \text{ for some } j = 1, \ldots, n\}.
\]
Note next that since \( F_0 \) is strictly monotone we have \( X_i = F_0^{-1} F_0(X_i) \) and
\[
F_0^{-1} F_0(X(n)) - F_0^{-1} F_0(X_{n-j}) \leq \frac{1}{p_0(1)} (F_0(X(n)) - F_0(X_{n-j})) \leq \zeta^{-1} (U_n - U_{n-j})
\]
where the \( U_{(i)}'s \) are distributed as the order statistics of a sample of size \( n \) of a uniform random variable on \( [0,1], \) and where \( U_{(0)} = 0 \) by convention. Hence it suffices to bound
\[
\Pr \left( U_n - U_{n-j} > \frac{\zeta j}{\xi n} \text{ for some } j = 1, \ldots, n \right). \tag{18}
\]
By a standard computation involving order statistics, the joint distribution of \( U_{(i)}, i = 1, \ldots, n, \) is the same as the one of \( Z_i/Z_{n+1} \) where \( Z_n = \sum_{i=1}^n W_i \) and where \( W_i \) are independent standard exponential random variables. Consequently, for \( \delta > 0, \) the probability in (18) is bounded by
\[
\Pr \left( \frac{W_{n-j+1} + \ldots + W_n}{Z_{n+1}} > \frac{\zeta j}{\xi n} \text{ for some } j \right) = \Pr \left( \frac{n}{Z_{n+1}} \frac{W_{n-j+1} + \ldots + W_n}{\frac{Z_{n+1}}{n}} > \frac{\zeta j}{\xi n} \text{ for some } j \right) \leq \Pr(n/Z_{n+1} > 1 + \delta) + \Pr \left( \frac{W_{n-j+1} + \ldots + W_n}{n} > \frac{\zeta j}{\xi n(1 + \delta)} \text{ for some } j \right)
\]
\[
= A + B.
\]
To bound \( A, \) note that it is equal to
\[
\Pr \left( \frac{1}{n+1} \sum_{i=1}^{n+1} (W_i - EW_i) < -\delta - (1 + \delta)/n \frac{n}{1 + \delta} \frac{n}{n+1} \right),
\]
which, since \( \delta > 0, \) is less than \( \epsilon/2 > 0 \) arbitrary, from some \( n \) onwards, by the law of large
numbers. For the term B we have, for \( \xi \) small enough and by Markov's inequality

\[
\Pr \left( W_{n-j+1} + \ldots + W_n > \frac{\xi j}{\xi (1 + \delta)} \text{ for some } j \right)
\]

\[
\leq \sum_{j=1}^{n} \Pr \left( W_{n-j+1} + \ldots + W_n > \frac{\xi j}{\xi (1 + \delta)} \right)
\]

\[
= \sum_{j=1}^{n} \Pr \left( \sum_{l=1}^{j} (W_{n-l+1} - EW_{n-l+1}) > \frac{\xi j}{\xi (1 + \delta)} - j \right)
\]

\[
\leq \sum_{j=1}^{n} \xi^4 E(\sum_{l=1}^{j} (W_{n-l+1} - EW_{n-l+1}))^4
\]

\[
\leq \xi^4 C'(\delta, \zeta) \sum_{j=1}^{n} j^2 \leq \xi^4 C''(\delta, \zeta) < \epsilon/2,
\]

since, for \( Y_i = W_{n-l+1} - EW_{n-l+1} \), by Hoffmann-Jørgensen's inequality (de la Peña and Giné, 1999, Corollary 1.2.7)

\[
\left\| \sum_{l=1}^{j} Y_l \right\|_{4,p} \leq K \left[ \left\| \sum_{l=1}^{j} Y_l \right\|_{2,p} + \left\| \max_l Y_l \right\|_{4,p} \right] \leq K' \left( \sqrt{j} + j^{1/4} \right),
\]

using the fact that \( EW^p_1 = p! \) and \( Var(Y_1) = 1 \).

b) Since \( \hat{p}_n \) is the least concave majorant of the empirical distribution \( F_n \),

\[
\|\hat{p}_n\|_\infty = \hat{p}_n(0) > M \iff F_n(t) > Mt \text{ for some } t.
\]

Since \( F_0 \) is concave and continuous it maps \([0, 1]\) onto \([0, 1]\) and satisfies \( F_0(t) \leq p_0(0)t \leq t\|p_0\|_\infty \) so that we obtain

\[
P_0^{\hat{p}_n}(\|\hat{p}_n\|_\infty > M) \leq P_0^{F_0^U}(\sup_{t>0} F_0(t) > M/\|p_0\|_\infty) = P_0^{F_0^U}(\sup_{t \in [0,1]} \frac{F_0^U(t)}{t} > M/\|p_0\|_\infty)
\]

(19)

where \( F_0^U \) is the empirical distribution function based on a sample of size \( n \) from the uniform distribution. The density of the order statistic of \( n \) uniform \( U(0, 1) \) random variables is \( n! \) on the set of all \( 0 \leq x_1 < \ldots < x_n \leq 1 \). Let \( M \geq \|p_0\|_\infty \) and set \( C \equiv M/\|p_0\|_\infty \) then the complement of the event in (19) has the probability

\[
P_0^{F_0^U}(\frac{F_0^U(t)}{t} \leq C \quad \forall t \in [0,1])
\]

\[
= n! \int_{1/C}^{1} \int_{(n-1)/(nC)}^{x_2} \ldots \int_{(j-1)/(nC)}^{x_{j+1}} dx_1 \ldots dx_{n-1} dx_n
\]

\[
= n! \int_{1/C}^{1} \int_{j/(nC)}^{x_{j+1}} \frac{1}{(j-1)!} x_{j-1} - \frac{1}{nC} \frac{1}{(j-2)!} x_{j-2} dx_j \ldots dx_n
\]

\[
= 1 - \frac{1}{C},
\]

where \( j = 2, \ldots, n-1 \). In particular, the probability in (19) equals \( \|p_0\|_\infty / M \) and can be made small by choosing \( M \) large. \( \square \)
We can now derive the rate of convergence of the maximum likelihood estimator of a monotone density. The rate corresponds to functions that are once differentiable in an $L^1$-sense, which is intuitively correct since a monotone decreasing function has a weak derivative that is a finite signed measure. The following convergence rate in Hellinger distance is given in Example 7.4.2 in van de Geer (2000). It is used here to derive the $L^2$-convergence rate. Kulikov and Lopuha"a (2005) prove a much finer result by deriving the asymptotic distribution of the $L^2$-errors under stronger smoothness assumptions.

**Proposition 2.** Suppose $p_0 \in \mathcal{P}^{mon}$ and that $p_0$ is bounded. Let $\hat{p}_n$ satisfy (22). Then

\[
h(\hat{p}_n, p_0) = O_{p_0}(n^{-1/3})
\]

and also

\[
\|\hat{p}_n - p_0\|_2 = O_{p_0}(n^{-1/3}).
\]

**Proof.** The statement for the Hellinger distance is contained in Example 7.4.2 in van de Geer (2000). $p_0$ is bounded by assumption and $\|\hat{p}_n\|_\infty = O_{p_0}(1)$ by Lemma 3. Then the result in $L^2$-distance follows by the inequality

\[
\|\hat{p}_n - p_0\|_2^2 \leq 2(\|\hat{p}_n\|_\infty + \|p_0\|_\infty) h^2(\hat{p}_n, p_0).
\]

\[
\square
\]

5 Putiing things together

The maximiser $\hat{p}_n$ is in some sense an object that lives on the boundary of $\mathcal{P}$ — it is piecewise constant with step-discontinuities at the observation points, exhausting the possible ‘roughness’ of a monotone function.

We can construct line segments in the parameter space through $p_0$, following the philosophy of Lemma 1. In order to ensure that the perturbed function lies again in $\mathcal{P}$ we will perturb a non-increasing density $p_0 \geq \zeta > 0$ with weak derivative $Dp_0$ by $\eta h$ where $h \in L^\infty$, $\int h = 0$ and $Dh$ is absolutely continuous with respect to $Dp_0$ such that the Radon–Nikodym density satisfies $\|Dh/Dp_0\|_{\infty,Dp_0} < \infty$. Then we have indeed for $\eta$ of absolute value small enough

\[
p_0 + \eta h \geq \zeta - \eta \|h\|_\infty > 0, \quad \int_0^1 (p_0 + \eta h) = 0,
\]

and $D(p_0 + \eta h) = Dp_0 + \eta Dh$ is a negative measure such that $p_0 + \eta h$ is non-increasing where $p_0 + \eta h$ is possibly modified on a nullset to equal the integral of $Dp_0 + \eta Dh$. A similar statement holds if we replace $h$ by $\pi_0(f)$ defined in (15) when $\|f\|_\infty + \|Df/Dp_0\|_{\infty,Dp_0}$ is finite. For simplicity we write $p_0 + \eta \pi_0(f)$ for the integral of its weak derivative.

**Lemma 4.** Let $p_0$ be non-increasing with $K \geq p_0 \geq \zeta > 0$ and $f$ such that $\|f\|_\infty + \|Df/Dp_0\|_{\infty,Dp_0}$ is finite. Then we have $p_0 + \eta \pi_0(f) \in \mathcal{P} \cap \mathcal{U}$ for $|\eta| \leq c(\|f\|_\infty + \|Df/Dp_0\|_{\infty,Dp_0})^{-1}$, where $c > 0$ depends on $K$ and $\zeta$ only.

**Proof.** $\pi_0(f) = (f - P_0 f)p_0$ is bounded by $2K \|f\|_\infty$. The assumption $p_0 \geq \zeta > 0$ yields $p_0 + \eta \pi_0(f) \in \mathcal{U}$ for $|\eta| < \zeta/(2K \|f\|_\infty)$. In addition to $|\eta| < \zeta/(2K \|f\|_\infty)$ we will require

\[
D(p_0 + \eta \pi_0(f)) = (1 - \eta P_0 f + \eta f)Dp_0 + \eta p_0 Df.
\]
to be a negative measure. This is the case if
\[
\frac{K[\eta\|Df/Dp_0\|_\infty,Dp_0]}{1 - 2[\eta\|f\|_\infty]} \leq 1 \iff (K\|Df/Dp_0\|_\infty,Dp_0 + 2\|f\|_\infty)[\eta] \leq 1,
\]
which holds for \( |\eta| \leq (\max(2, K)(\|f\|_\infty + \|Df/Dp_0\|_\infty,Dp_0))^{-1}. \)

For \( p_0 \) and \( f \) as above we thus obtain from Lemma 3 that on events of probability as close to one as desired,
\[
|D\ell_n(\hat{p}_n)[\pi_0(f)]| \leq C(\|f\|_\infty + \|Df/Dp_0\|_\infty,Dp_0)|D\ell_n(\hat{p}_n)[\hat{p}_n - p_0]|
\]
(23)
for some constant \( C \) that depends on \( K \) and \( \zeta \) only. Note that the differential calculus from Proposition 1 applies since \( \hat{p}_n, p_0 \) as well as all points on the line segment \((1 - t)\hat{p}_n + tp_0, t \in (0, 1)\), lie in \( U(X_1, \ldots, X_n) \cap P \), using Lemma 3a).

We next need to derive stochastic bounds of the likelihood derivative at \( \hat{p}_n \) in the direction of \( p_0 \).

**Lemma 5.** Suppose \( p_0 \) is bounded and satisfies \( \inf_{x \in [0,1]} p_0(x) > 0 \). For \( \hat{p}_n \) satisfying (24) we have
\[
|D\ell_n(\hat{p}_n)[\hat{p}_n - p_0]| = O_{P_0}^n(n^{-2/3})
\]
**Proof.** By Lemma 3 we can restrict to an event where
\[
0 < \xi \leq \inf_{x \in [0,1]} \hat{p}_n(x) \leq \sup_{x \in [0,1]} \hat{p}_n(x) \leq k < \infty
\]
and by (21) further to an event where
\[
\|\hat{p}_n - p_0\|_{2,p} \leq \|p_0\|_{\infty}^{1/2}\|\hat{p}_n - p_0\|_2 \leq \|p_0\|_{\infty}^{1/2}Mn^{-1/3}
\]
for some finite constant \( M \). For any \( a_n \to 0 \) with \( na_n \to \infty \) and some \( c > 0 \)
\[
\Pr((1 - X(n)) > a_n) = \Pr((1 - a_n) > X(n)) = (F_0(1 - a_n))^n \leq (1 - ca_n)^n \to 0,
\]
in particular we obtain for \( a_n = \log n/n \) that
\[
1 - X(n) = O_{P_0}^n \left( \frac{\log n}{n} \right).
\]
Let us define the random function \( \hat{p}_n^{-1} \equiv \hat{p}_n^{-1} \) on \([0, X(n)]\) and zero on \((X(n), 1]\). By \( D\ell_n(\hat{p}_n) \) and \( D\ell(\hat{p}_n) \) we denote the corresponding right hand sides in (6) and (7). We observe that
\[
D\ell_n(\hat{p}_n) = D\ell_n(\hat{p}_n). \]
The function \( h \equiv \hat{p}_n^{-1} (\hat{p}_n - p_0) \) on \([0, 1]\) and \( h \equiv 0 \) elsewhere is of bounded variation with norm \( \|h\|_{BV} \equiv \|h\|_1 + \|Dh\|_{\mathbb{R}} \) bounded by a fixed constant \( C \) that depends only on \( k, \xi, \|p_0\|_{\infty} \). We observe that \( D\ell_n(p_0)[\hat{p}_n - p_0] = 0 \) by (7) and obtain
\[
\begin{align*}
|D\ell_n(\hat{p}_n)[\hat{p}_n - p_0]| &= |D\ell_n(\hat{p}_n)[\hat{p}_n - p_0] - D\ell(\hat{p}_n)[\hat{p}_n - p_0] + (D\ell(\hat{p}_n) - D\ell(p_0))[\hat{p}_n - p_0]| \\
&\leq \sup_{h: \|h\|_{BV} \leq C, \|h\|_{2,p} \leq Mn^{-1/3}} \| (P_n - F_0)(h) \| + \|\hat{p}_n - p_0\|_2^2 + \int_{X(n)} |\hat{p}_n - p_0| \\
&= O_{P_0}^n \left( n^{-1/2} + n^{-1/3} + \frac{\log n}{n} \right)
\end{align*}
\]
(25)
where we have used Theorem 3.1 in Giné and Koltchinskii (2006) with
\[ H = \text{id}, \sigma = \tilde{M} n^{-1/3}, F = \text{const} \]
combined with the bracketing entropy bound for monotone functions (van der Vaart and Wellner, 1996, Theorem 2.7.5) and its straightforward generalisation to bounded variation functions to control the supremum of the empirical process, (21) to control the second term, and (23) for the last integral.

**Proposition 3.** Suppose \( p_0 \in \mathcal{P} \) is bounded and satisfies \( p_0 \geq \zeta > 0 \) on \([0,1]\). Let \( f \in L^\infty\) be such that \( \|Df/Dp_0\|_{\infty,Dp_0} \) is finite. Then
\[ |D\ell_n(\hat{p}_0)| = O_{P_0} \left( \|f\|_\infty + \|Df/Dp_0\|_{\infty,Dp_0}^{-2/3} \right). \]

**Proof.** By Lemma 4 we have that \( p_0 + \eta \pi_0(f) \in \mathcal{P} \cap \mathcal{U} \) for \( \eta \) a small multiple of \( \|f\|_\infty + \|Df/Dp_0\|_{\infty,Dp_0} \). The claim of the proposition then follows from (23) and Lemma 5.

We are now ready to prove Theorem 4 and Theorem 2.

**Proof of Theorem 4.** We use Lemma 2, Proposition 1, \( \hat{p}_n, p_0 \in \mathcal{U}(X_1, \ldots, X_n) \) by Lemma 3 and a Taylor expansion up to second order to see
\[ \|(\hat{p}_n - P_0)(f)\| = |D\ell_n(p_0)[\pi_0(f)] + D^2\ell(p_0)[\hat{p}_n - p_0, \pi_0(f)]| \leq |D\ell_n(\hat{p}_n)[\pi_0(f)]| + |D^2\ell_n(p_0) - D^2\ell(p_0)][\hat{p}_n - p_0, \pi_0(f)]| + \frac{1}{2}|D^3\ell(\hat{p}_n)[\hat{p}_n - p_0, \hat{p}_n - p_0, \pi_0(f)]| \]
where \( \hat{p}_n \) equals, on \([0, X_{(n)}]\), some mean values between \( \hat{p}_n \) and \( p_0 \), and \( \hat{p}_n^{-1} \) is zero otherwise by convention. Here again \( D^3\ell_n(\hat{p}_n) \) and \( D^3\ell(\hat{p}_n) \) stand for the corresponding right hand sides in (34) and (37). The first term is bounded using Proposition 8 giving the bound \( Bn^{-2/3} \) in probability. We define \( h \equiv p_0^{-1}(\hat{p}_n - p_0)(f - P_0 f) \) on \([0,1]\) and \( h \equiv 0 \) elsewhere so that the second term equals \( |(P_n - P_0)h| \). With probability arbitrarily close to one we have \( \|f\|_{BV} \lesssim \|f\|_\infty + \|f\|_{BV} \lesssim \|f\|_\infty + \|Df/Dp_0\|_{\infty,Dp_0} \) and \( \|h\|_2 \lesssim \|f\|_\infty n^{-1/3} \). The second term is bounded similarly as in (23) above by
\[ \sup_{h: \|h\|_{BV} \lesssim \tilde{C} Bu, \|h\|_2 \lesssim M Bn^{-1/3}} |(P_n - P_0)(h)| = O_{P_0}(Bn^{-2/3}). \]
The third term is bounded the same way, using \( \|\hat{p}_n - p_0\|_{BV} = O(1) \), and noting that \( \hat{p}_n \) as a convex combination of \( \hat{p}_n,p_0 \) has variation bounded by a fixed constant on \([0, X_{(n)}]\), so that we can estimate the term by the supremum of the empirical process over a fixed BV-ball, and using again Lemma 3 to bound \( \hat{p}_n \) from below on \([0, X_{(n)}]\). Using the last fact the fourth term is also seen to be of order
\[ \|f\|_\infty \|\hat{p}_n - p_0\|_2^2 = O_P(Bn^{-2/3}) \]
in view of (21) completing the proof the first claim. The second claim follows from the fact that \( \mathcal{B} \) is a bounded set in the space of bounded variation functions and thus a Donsker class.

**Proof of Theorem 2.** It is sufficient to prove the result for \( 1/2 < s < 1 \). Let \( \psi_{lk} \) be a boundary corrected Daubechies wavelet basis, where we use the notation \( \psi_{-1,k} = \phi_{0,k} \). We denote Besov spaces by \( B_{pq}^s \) and will use their wavelet characterisation, see for example Giné and Nickl (2009), which also holds for Besov spaces on compact intervals using boundary corrected wavelets.
We decompose the functions $f$ in a ball $\mathcal{F}$ of $C^s$ by using the projection $\pi_{V_j}(f)$ onto the span of the wavelets up to resolution level $j$,

$$
\sup_{f \in \mathcal{F}} |(\hat{P}_n - P_n)(f) - \pi_{V_j}(f)| \leq \sup_{f \in \mathcal{F}} \left| \int_0^1 (\hat{p}_n - p_0)(f - \pi_{V_j}(f)) \right| + \sup_{f \in \mathcal{F}} |(\hat{P}_n - P_n)(\pi_{V_j}(f))| + \sup_{f \in \mathcal{F}} |(P_n - P_0)(f - \pi_{V_j}(f))|.
$$

(26)

Since $C^s = B^s_{\infty \infty}$ for $s \notin \mathbb{N}$ and since the $C^1$-norm is bounded by the $B^1_{\infty \infty}$-norm, we have for the wavelet partial sum $\pi_{V_j}(f)$ of $f \in C^s$

$$
\|\pi_{V_j}(f)\|_{C^1} \leq \sum_{l \leq j} 2^{3l/2} \max_k |\langle f, \psi_{lk} \rangle| \lesssim 2^{(1-s)} \max_{l \leq j} \sup_{l,k} |\langle f, \psi_{lk} \rangle| \leq 2^{(1-s)} \|f\|_{B^s_{\infty \infty}}.
$$

Thus taking $2^j \sim n^{1/3}$ we have by Proposition 4

$$
\sup_{f \in \mathcal{F}} |(\hat{P}_n - P_n)(\pi_{V_j}(f))| = O_P(n^{-2/3} \{1-s\}/3) = o_P(1/\sqrt{n})
$$

since $s > 1/2$. Moreover, by Parseval’s identity

$$
\sup_{f \in \mathcal{F}} \|\pi_{V_j}(f) - f\|_2 = O(2^{-js}).
$$

Also, using (21) and the Cauchy–Schwarz inequality

$$
\sup_{f \in \mathcal{F}} \left| \int_0^1 (\hat{p}_n - p_0)(f - \pi_{V_j}(f)) \right| = O_P(n^{-1/3} s^{-1/\delta}) = o_P(1/\sqrt{n})
$$

and since the class $\{f - \pi_{V_j}(f)\}$ is contained in a fixed $s$-Hölder ball, which is a $P_0$-Donsker class for $s > 1/2$ in view of Corollary 5 in Nickl and Pötscher (2007) and has envelopes that converge to zero we see that the third term in (20) is also $o_P(1/\sqrt{n})$ (since the empirical process is tight and has a degenerate Gaussian limit). The remaining claims follow from the fact that $\mathcal{F}$ is a $P_0$-Donsker class.

References

Balabdaoui, F. and J. A. Wellner (2007). A Kiefer–Wolfowitz theorem for convex densities. In Asymptotics: particles, processes and inverse problems, Volume 55 of IMS Lecture Notes Monogr. Ser., pp. 1–31. Beachwood, OH: Inst. Math. Statist.

Bickel, P. J. and Y. Ritov (2003). Nonparametric estimators which can be “plugged-in”. Ann. Statist. 31(4), 1033–1053.

de la Peña, V. H. and E. Giné (1999). Decoupling. From Dependence to Independence. Probability and its Applications. Springer-Verlag, New York.

Dümbgen, L. and K. Rufibach (2009). Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. Bernoulli 15(1), 40–68.
Durot, C. and H. P. Lopuhaä (2014). A Kiefer–Wolfowitz type of result in a general setting, with an application to smooth monotone estimation. ArXiv: 1308.0417.

Giné, E. and V. Koltchinskii (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* 34(3), 1143–1216.

Giné, E. and R. Nickl (2008). Uniform central limit theorems for kernel density estimators. *Probab. Theory Related Fields* 141(3-4), 333–387.

Giné, E. and R. Nickl (2009). Uniform limit theorems for wavelet density estimators. *Ann. Probab.* 37(4), 1605–1646.

Giné, E. and R. Nickl (2014). *Mathematical Foundations of Infinite-Dimensional Statistical Models.* Cambridge University Press, to appear.

Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* 39, 125–153.

Jankowski, H. (2014). Convergence of linear functionals of the Grenander estimator under mis-specification. *Ann. Statist.* 42(2), 625–653.

Kiefer, J. and J. Wolfowitz (1976). Asymptotically minimax estimation of concave and convex distribution functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 34(1), 73–85.

Kulikov, V. N. and H. P. Lopuhaä (2005). Asymptotic normality of the $L_k$-error of the Grenander estimator. *Ann. Statist.* 33(5), 2228–2255.

Nickl, R. (2007). Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Related Fields* 138(3-4), 411–449.

Nickl, R. (2009). On convergence and convolutions of random signed measures. *J. Theoret. Probab.* 22(1), 38–56.

Nickl, R. and B. M. Pötscher (2007). Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov- and Sobolev-type. *J. Theoret. Probab.* 20(2), 177–199.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press, Cambridge.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer Series in Statistics. Springer-Verlag, New York.

Ziemer, W. P. (1989). *Weakly differentiable functions. Sobolev spaces and functions of bounded variation*, Volume 120 of *Graduate Texts in Mathematics.* Springer-Verlag, New York.