Uniform Central Limit Theorems for the Grenander Estimator

Richard Nickl

University of Connecticut

January 2008

Abstract

The classical result of Kiefer and Wolfowitz (1976) on the asymptotic closeness of the distribution function of the maximum likelihood estimator of a monotone decreasing density and the empirical distribution function is generalized to Donsker classes of functions, paralleling recent results for other density estimators obtained in Nickl (2007) and Giné and Nickl (2007, 2008). These results are then applied to efficiently estimate the entropy functional by the associated plug-in MLE.

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

Key words and phrases: NPMLE, entropy functional, Besov classes.

1 Introduction

Let $X, X_1, ..., X_n$ be i.i.d. on $[0, 1]$ with law P and distribution function F. Define the empirical measure $P_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ and the empirical cdf $F_n(x) = \int_0^x dP_n$. If P is known to have a monotone decreasing density f, it is natural to estimate this density by the (nonparametric) maximum likelihood estimator $f_n(y)$ defined by the solution of the optimization problem $\max_{f\in\mathcal{F}} n^{-1} \sum_{i=1}^n \log f(X_i)$ where

$$\mathcal{F} = \left\{ f : [0, 1] \to [0, \infty), \int_0^1 f(x)dx = 1, f \text{ is monotone decreasing} \right\}.$$

This estimator is known as the 'Grenander' estimator, since it was introduced by Grenander (1956), who also showed that f_n has a simple geometric interpretation: f_n is the left derivative of the least concave majorant $\hat{F}_n$ of the empirical distribution function F_n, see, e.g., Section 24.4 in van der Vaart (1998). Of course $\hat{F}_n(x)$ is, by absolute continuity of $\hat{F}_n$, equal to the distribution function $\int_0^x f_n(y)dy$ of f_n.

The following is a classical result of Kiefer and Wolfowitz (1976): Under several assumptions on f – including strict monotonicity and existence of a continuous derivative of f – Kiefer and Wolfowitz proved that

$$\|\hat{F}_n - F_n\|_\infty = O_{a.s.}\left(\left(\frac{\log n}{n}\right)^{2/3}\right), \tag{1}$$

so that, in particular, $\sqrt{n}(\hat{F}_n - F)$ converges in law to the P-Brownian bridge in the Banach space $\ell^\infty([0,1])$ of bounded functions on $[0,1]$. We also refer to Balabdaoui and Wellner (2007), who recently gave an 'updated' version of the Kiefer and Wolfowitz proof, using results by Kulikov and Lupohaä (2006).

One way to generalize this result is along the lines of recent results on *uniform central limit theorems* (UCLTs) for density estimators, see Nickl (2007), Giné and Nickl (2007, 2008). In these articles, density estimators $\hat{f}_n$ are viewed as random measures $\hat{P}_n$ acting on general classes of functions $\mathcal{H}$ via integration, and quite general central limit theorems were proved for the processes

$$\left(\sqrt{n}(\hat{P}_n - P)\right)_{h\in\mathcal{H}} = \left(\sqrt{n}\int h(\hat{f}_n - f)\right)_{h\in\mathcal{H}},$$

allowing, in particular, for Donsker classes $\mathcal{H}$ of smooth functions (such as bounded variation, Sobolev, Hölder and Besov classes). These results contain (1) as a special case upon choosing $\mathcal{H} = \{1_{[0,x]} : x \in [0,1]\}$. Such more general results have several statistical applications, see Bickel and Ritov (2003) as well as Section 3 in Nickl (2007). In the present article we show how the fact that the rate of convergence in (1) is *faster* than $1/\sqrt{n}$ can be used together with methods from approximation theory to prove such general results also for the Grenander estimator. We then apply these results to efficiently estimate the entropy functional by the plug-in MLE, see Theorem 3. We should remark that results similar to (1) were also proved for estimators of monotone regression functions, see, e.g., Durot and Tocquet (2003) and Wang and Woodroofe (2007). The proof methods of the present article should apply to regression models as well, with the natural modifications.

2 Main Results

For Borel-measurable functions $h : [0,1] \to \mathbb{R}$ and Borel measures μ on $[0,1]$, we set $\mu h := \int_0^1 h d\mu$, and we denote by $L^p([0,1],\mu)$ the usual Lebesgue-spaces of real-valued functions, normed by $\|\cdot\|_{p,\mu}$. If $d\mu(x) = dx$ is Lebesgue measure on $[0,1]$, we set $L^p([0,1]) := L^p([0,1],\mu)$, and we abbreviate the norm by $\|\cdot\|_p$ if $1 \le p < \infty$. We also denote by $L^\infty([0,1])$ the space of bounded measurable functions on $[0,1]$, normed by the supnorm $\|\cdot\|_\infty$, and, for an arbitrary (non-empty) set H, $\ell^\infty(H)$ denotes the space of bounded functions $g : H \to \mathbb{R}$ normed by $\|g\|_H := \sup_{h\in H}|g(h)|$. Throughout, the variables X_i are the coordinate projections of $([0,1]^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, P^{\mathbb{N}})$, and we set $\Pr := P^{\mathbb{N}}$. The empirical process indexed by $\mathcal{H} \subset L^2([0,1], P)$ is given by $f \mapsto \sqrt{n}\,(P_n - P)\,h, h \in \mathcal{H}$. Convergence in law $\to_d$ of random elements in $\ell^\infty(\mathcal{H})$ is defined as, e.g., in Chapter 5 of de la Peña and Giné (1999). The class $\mathcal{H}$ is said to be *P-Donsker* if the centered Gaussian process G_P with covariance $EG_P(f)G_P(g) = P[(f - Pf)(g - Pg)]$ is sample-bounded and sample-continuous w.r.t. the covariance semimetric, and if $\sqrt{n}\,(P_n - P) \to_d G_P$ in $\ell^\infty(\mathcal{H})$.

We start with the classical result of Kiefer and Wolfowitz (1976), that we give, without loss of generality, for densities defined on $[0,1]$. In this case the assumptions simplify essentially to requiring strictly monotone, continuously differentiable f that is bounded away from zero.

Theorem 1 (Kiefer and Wolfowitz (1976)) *Let $X_1, ..., X_n$ be i.i.d. with monotone decreasing density f on $[0,1]$, assume that f' is continuous on $[0,1]$ and that f and f' are bounded away from zero. Then*

$$\sqrt{n} \sup_{t \in [0,1]} |\hat{F}_n(t) - F_n(t)| = O_{a.s.}(n^{-1/6}(\log n)^{2/3}), \tag{2}$$

so that in particular

$$\sqrt{n}(\hat{F}_n - F) \to_d G_P \quad \text{in} \quad \ell^\infty([0,1])$$

Proof. See Kiefer and Wolfowitz (1976) or Balabdaoui and Wellner (2007). ■

Our goal now is to generalize Theorem 1 to general Donsker classes of functions. The following corollary is immediate. $BV([0,1])$ will denote the space of measurable functions $h : [0,1] \mapsto \mathbb{R}$ of bounded variation, equipped with the total variation norm

$$\|h\|_{BV} = \sup \left\{ \sum_{i=1}^{n} |h(x_i) - h(x_{i-1})| : n \in \mathbb{N}, 0 < x_1 < \cdots < x_n < 1 \right\}.$$

Corollary 1 *Let the assumptions of Theorem 1 hold. Define $\hat{P}_n$ by $d\hat{P}_n(y) = f_n(y)dy$. Let $\mathcal{H}_R = \{h \text{ right-continuous} : \|h\|_\infty + \|h\|_{BV} \leq 1\}$. Then*

$$\sqrt{n} \sup_{h \in \mathcal{H}_R} \left| \int_0^1 h d(\hat{P}_n - P_n) \right| = \sqrt{n} \|\hat{P}_n - P_n\|_{\mathcal{H}_R} = O_{a.s.}(n^{-1/6}(\log n)^{2/3}), \tag{3}$$

and, if $\mathcal{H} = \{h : \|h\|_\infty + \|h\|_{BV} \leq 1\}$, then

$$\sqrt{n}(\hat{P}_n - P) \to_d G_P \quad \text{in} \quad \ell^\infty(\mathcal{H}).$$

To obtain a result more general than the previous corollary, we define Besov spaces as follows.

Definition 1 *(Besov spaces) Let $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, $0 < s < \infty$. For a function $h : \mathbb{R} \to \mathbb{R}$, the difference operator Δ_z is defined by $\Delta_z h(\cdot) = h(\cdot + z) - h(\cdot)$ as well as*

$\Delta_z^2 h(\cdot) = \Delta_z(\Delta_z h(\cdot))$, *and, iteratively, also* Δ_z^r. *For* $h : [0,1] \to \mathbb{R}$, *we define* $\Delta_z^r(h)(x)$ *as above if* $x, x+rz \in [0,1]$ *and set it equal to zero otherwise. For* $h \in L^p([0,1])$ *and* $r > s$, *define*

$$\|h\|_{s,p,q}^* := \left(\int_{|z|\le 1} [z^{-s-1/q} \|\Delta_z^r(h)\|_p]^q dz \right)^{1/q}$$

with the modification in case $q = \infty$

$$\|f\|_{s,p,\infty}^* := \sup_{0 \ne |z| \le 1} |z|^{-s} \|\Delta_z^r(f)\|_p.$$

The Besov space is defined as the linear space

$$B_{pq}^s([0,1]) := \{ f \in L^p([0,1]) : \|f\|_{s,p,q}^* < \infty \},$$

normed by $\|f\|_{s,p,q} := \|f\|_p + \|f\|_{s,p,q}^*$.

If $p = q = \infty$, these spaces are the Hölder-Zygmund spaces, which contain the classical Hölder-Lipschitz spaces. Furthermore, Sobolev spaces of order s correspond to $B_{22}^s([0,1])$, so these Besov scales include many classical spaces but also much more, see, e.g., Triebel (1983). In particular, for $p > 1$ and/or $s < 1$, these Besov classes contain functions that are *not* of bounded variation, so that Corollary 1 does not apply.

The following proposition, which can be shown to be (essentially) best possible, follows from results in Nickl and Pötscher (2007).

Proposition 1 *Let* $\mathcal{H}$ *be a bounded subset of* $B_{pq}^s([0,1])$ *where* $1 \le p \le \infty$, $1 \le q \le \infty$, *and let* P *be a probability measure on* $[0,1]$. *Let either* $s > \max(1/p, 1/2)$ *or suppose that* $1 \le p < 2, q = 1, s = 1/p$ *simultaneously hold. Then* $\mathcal{H}$ *is* P*-Donsker.*

Nickl (2007) showed that certain nonparametric MLEs (not including the Grenander estimator) satisfy UCLTs over Besov classes with $p \ge 2$ and $s > 1/2$. More generally, Giné and Nickl (2007, 2008) showed that kernel and wavelet density estimators satisfy UCLTs over *any* Besov class featured in the previous proposition, in fact, they considered Besov spaces over the real line. Using Corollary 1, the rate of convergence of f_n to f in the L^2-norm, and an approximation argument by wavelets, one can prove the main theorem of this article, which shows that f_n satisfies the same UCLTs as kernel and wavelet density estimators, if f satisfies the assumptions from Theorem 1.

Theorem 2 *Let the assumptions of Theorem 1 hold. Define* $\hat{P}_n$ *by* $d\hat{P}_n(y) = f_n(y)dy$. *Let* $\mathcal{H}$ *be a bounded subset of* $B_{pq}^s([0,1])$, *where* s, p, q *satisfy one of the conditions of Proposition 1. Then*

$$\sqrt{n} \|\hat{P}_n - P_n\|_{\mathcal{H}} = o_P(1)$$

so that, in particular,

$$\sqrt{n}(\hat{P}_n - P) \to_d G_P \quad \text{in} \quad \ell^\infty(\mathcal{H}).$$

Corollary 2 *Let*

$$\mathcal{H}^s := \left\{ f \in L^\infty([0,1]) : \|f\|_\infty + \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^s} \leq 1 \right\}$$

for some $1/2 < s \leq 1$. *Then* $\|\hat{P}_n - P_n\|_{\mathcal{H}^s} = o_P(n^{-1/2})$ *and*

$$\sqrt{n}(\hat{P}_n - P) \rightarrow_d G_P \quad \text{in} \quad \ell^\infty(\mathcal{H}^s).$$

A consequence of the above results is, for example, that

$$\beta(\hat{P}_n, P) = \|\hat{P}_n - P\|_{\mathcal{H}_1} = O_P(n^{-1/2}),$$

where β is the usual bounded-Lipschitz metric for weak convergence of probability measures. Note that, under the assumptions of Theorem 2, $\hat{P}_n$ outperforms the empirical measure P_n in that it is not only $\sqrt{n}$-consistent in metrics for the weak topology, but also consistent in the strong total variation norm (see Theorem 24.6 in van der Vaart (1998))

$$\|\hat{P}_n - P\|_{TV} \simeq \|f_n - f\|_1 = O_P(n^{-1/3}).$$

Furthermore, the above results can be applied to construct simple and efficient plug-in estimators of integral functionals. For example, proceeding as in Corollary 5 in Nickl (2007), one shows that $\int_0^1 f_n^2$ is a $\sqrt{n}$-consistent and efficient estimator of the quadratic functional $T(f) = \int_0^1 f^2$. A slightly more involved application is estimation of the entropy functional

$$T(f) = \int_0^1 f(y) \log f(y).$$

If $X_{(n)}$ is the largest order statistic, then f_n is positive on $[0, X_{(n)}]$, but $f_n(y)$ is zero for all $y > X_{(n)}$, so the direct plug in estimator $T(f_n)$ cannot be used. The natural modification is

$$T_n := \int_0^{X_{(n)}} f_n(y) \log f_n(y),$$

which is in fact very easy to compute: As is well known, f_n is a piecewise constant function with jumps only at the observations, so that T_n is a finite sum consisting of terms $c_j \log c_j$, $j = 1, ..., m$, where c_j are some positive numbers and $m \leq n$. Using the results from above together with Fréchet-differentiability of $T(\cdot)$ and control of the probability of the event that $\inf_{x \in [0, X_{(n)}]} f_n(x)$ is small, one can prove the following result. Note that $\sigma^2(f)$ in the following theorem is the efficient Cramér-Rao lower bound for estimation of the parameter $T(f)$, see, e.g., Laurent (1996).

Theorem 3 *Let the assumptions of Theorem 1 hold. Then*

$$\sqrt{n}(T_n - T(f)) \rightarrow_d N(0, \sigma^2(f))$$

where $\sigma^2(f) = \int f \log^2 f - \left(\int f \log f\right)^2$.

3 Proofs

Proof. (Corollary 1) If h is of bounded variation, right-continuous and satisfies $h(0)=0$, then there exists a unique finite signed Borel measure μ_h such that $h(x)=\int 1_{[0,x]}(v)d\mu_h(v)$. Since $(\hat{P}_n - P_n)c = 0$ for c constant, we may assume that the elements in $\mathcal{H}_R$ all satisfy $h(0)=0$. We then have by Fubini, for $h \in \mathcal{H}_R$,

$$\begin{aligned} |(\hat{P}_n - P_n)h| &= \left| \int_0^1 \int_0^1 1_{[0,x]}(v) d\mu_h(v) d(\hat{P}_n - P_n)(x) \right| \\ &= \left| \int_0^1 \int_0^1 1_{[v,1]}(x) d(\hat{P}_n - P_n)(x) d\mu_h(v) \right| \\ &\leq \int_0^1 d|\mu(v)| \|\hat{F}_n - F_n\|_\infty \leq \|\hat{F}_n - F_n\|_\infty. \end{aligned}$$

This already proves the first claim of the corollary by Theorem 1. To prove the second claim, observe that any $h \in \mathcal{H}$ is right-continuous except at most at a countable number of points, in particular there exists a right continuous function $\tilde{h}$ such that $\tilde{h} = h$ almost everywhere. Since $\hat{P}_n, P$ are absolutely continuous measures, we have

$$\sqrt{n}(\hat{P}_n - P)h = \sqrt{n}(\hat{P}_n - P)\tilde{h} = \sqrt{n}(\hat{P}_n - P_n)\tilde{h} + \sqrt{n}(P_n - P)\tilde{h},$$

which proves the second claim by using the first, and since $\mathcal{H}$ is P-Donsker (e.g., Theorem 2.1 in Dudley (1992)). ■

Proof. (Theorem 2)

Step I: We will first approximate $h \in B^s_{pq}([0,1])$ by a suitable sequence of functions $K_j(h) \in B^1_{11}([0,1])$ whose Besov norm $\|K_j(h)\|_{1,1,1}$ does not increase too fast as a function of j. For the approximation argument, we will use wavelets, hence we will imbed our approximation problem into a Besov space $B^s_{pq}(\mathbb{R})$ defined over the real line. In slight abuse of notation, we use $\|\cdot\|_p$ also to denote the norm on the sequence spaces $\ell^p(\mathbb{Z})$, and we denote by $\|\cdot\|_{p,\mathbb{R}}$ the usual L^p-norms on the real line (w.r.t. Lebesgue measure on $\mathbb{R}$), and by $L^p(\mathbb{R})$ the associated function spaces.

Definition 2 *Let $1 \leq p, q \leq \infty$, $0 < s < S$, $s \in \mathbb{R}$, $S \in \mathbb{N}$. Let ϕ, ψ be bounded, compactly supported father and mother wavelet, let ϕ be S-times differentiable, and denote by $\alpha_k(h) = \int_{\mathbb{R}} h(x)\phi(x-k)dx$ and $\beta_{lk}(h) = \int_{\mathbb{R}} 2^{l/2}h(x)\psi(2^l x - k)dx$ the wavelet coefficients of $h \in L^p(\mathbb{R})$. The Besov space $B^s_{pq}(\mathbb{R})$ is defined as the set of all functions*

$$\left\{ h \in L^p(\mathbb{R}) : \|h\|_{s,p,q,\mathbb{R}} := \|\alpha_{(\cdot)}(h)\|_p + \left(\sum_{l=0}^{\infty} \left(2^{l(s+1/2-1/p)} \|\beta_{l(\cdot)}(h)\|_p \right)^q \right)^{1/q} < \infty \right\},$$

with modification in case $q = \infty$

$$\|h\|_{s,p,\infty,\mathbb{R}} := \|\alpha_{(\cdot)}(h)\|_p + \sup_{l \geq 0} 2^{l(s+1/2-1/p)} \|\beta_{l(\cdot)}(h)\|_p.$$

Note that wavelets satisfying the above conditions for any given S exist, e.g., Daubechies' wavelets, see Section 7 in Härdle, Kerkyacharian, Picard and Tsybakov (1998). Given these definitions, the following lemma is not difficult to prove.

Lemma 1 *Let ϕ, ψ be bounded, compactly supported father and mother wavelet and let ϕ be S-times differentiable. Suppose $g \in B^s_{p\infty}(\mathbb{R}) \cap B^s_{1\infty}(\mathbb{R})$ for some $1/2 < s < S$. Then the truncated wavelet series*

$$K_j(g)(y) = \sum_k \alpha_k(g)\phi(y-k) + \sum_{l=0}^{j-1}\sum_k \beta_{lk}(g)2^{l/2}\psi(2^l y - k)$$

(*convergence pointwise and in $\mathcal{L}^p(\mathbb{R})$*) *satisfies, for constants c, c',*

$$\|K_j(g)\|_{s,p,\infty,\mathbb{R}} \leq \|g\|_{s,p,\infty,\mathbb{R}}$$

and

$$\|K_j(g) - g\|_{p,\mathbb{R}} \leq c2^{-js}\|g\|_{s,p,\infty,\mathbb{R}}$$

as well as

$$\|K_j(g)\|_{1,1,1,\mathbb{R}} \leq c'2^{j(1/2-\delta)}\|g\|_{s,1,\infty,\mathbb{R}}.$$

for some $\delta > 0$ as $j \to \infty$. (If $s < 1$, δ can be taken to equal $s - 1/2$.)

Proof. The first claim is obvious from definition of $K_j(g)$ and the Besov norm. For the third claim, we have that $g \in B^s_{1\infty}(\mathbb{R})$ implies, by Definition 2, that

$$\|\beta_{l(\cdot)}(g)\|_1 \leq D2^{-l(s-1/2)}\|g\|_{s,1,\infty,\mathbb{R}}$$

as well as $\|\alpha(g)_{(\cdot)}\|_1 \leq \|g\|_{s,1,\infty,\mathbb{R}}$ so that, since $s > 1/2$

$$\begin{aligned}
\|K_j(g)\|_{1,1,1,\mathbb{R}} &= \|\alpha(g)_{(\cdot)}\|_1 + \sum_{l=0}^{j-1} 2^{l/2}\|\beta_{l(\cdot)}(g)\|_1 \\
&\leq D'\|g\|_{s,1,\infty,\mathbb{R}} \sum_{l=0}^{j-1} 2^{l(1-s)} \\
&\leq c'2^{j(1/2-\delta)}\|g\|_{s,1,\infty,\mathbb{R}}.
\end{aligned}$$

For the second claim, we will need that

$$\sup_u \sum_k |\psi(u-k)| \leq \Psi,$$

where Ψ is some fixed finite constant, which follows from Lemma 8.5 in Härdle et al. (1998). Note also that $g \in B^s_{p\infty}(\mathbb{R})$ implies

$$\|\beta_{l(\cdot)}(g)\|_p \leq C2^{-l(s+1/2-1/p)}\|g\|_{s,p,\infty,\mathbb{R}}$$

by Definition 2. Now, by change of variables and Hölder's inequality (with $1/p + 1/q = 1$), we have

$$
\begin{aligned}
\|K_l(g) - g\|_{p,\mathbb{R}} &= \left\| \sum_{l \ge j} \sum_k \beta_{lk}(g) 2^{l/2} \psi(2^l(\cdot) - k) \right\|_{p,\mathbb{R}} \\
&\le \sum_{l \ge j} 2^{l/2} 2^{-l/p} \left(\int \left(\sum_k |\beta_{lk}(g)| |\psi(u-k)|^{1/p} |\psi(u-k)|^{1/q} \right)^p du \right)^{1/p} \\
&\le \sum_{l \ge j} 2^{l(1/2-1/p)} \left(\int \sum_k |\beta_{lk}(g)|^p |\psi(u-k)| \left(\sum_k |\psi(u-k)| \right)^{p/q} du \right)^{1/p} \\
&\le \sum_{l \ge j} 2^{l(1/2-1/p)} \|\beta_{l(\cdot)(g)}\|_p \Psi^{1/q} \left(\int |\psi(u)| du \right)^{1/p} \\
&\le C' \sum_{l \ge j} 2^{l(1/2-1/p)} \|\beta_{l(\cdot)}(g)\|_p \\
&\le C'' \|g\|_{s,p,\infty,\mathbb{R}} \sum_{l \ge j} 2^{-ls} \le c 2^{-js} \|g\|_{s,p,\infty,\mathbb{R}}
\end{aligned}
$$

which completes the proof. ■

To apply the above lemma to approximate a function $h \in B^s_{pq}([0,1])$, we will use that $B^s_{pq}([0,1])$ is equal (with equivalent norms) to the space of restrictions of elements of $B^s_{pq}(\mathbb{R})$ to $[0,1]$ (equipped with the usual quotient norm). This follows from the fact that the wavelet definition of $B^s_{pq}(\mathbb{R})$ given above coincides (with equivalent norms) with the more classical definitions of Besov spaces on $\mathbb{R}$; in particular it coincides with Definition 2 on p.45 in Triebel (1983) – see, e.g., Theorems 9.1 and 9.6 in Härdle et al. (1998) – so that we can use the extension and restriction theorems in Triebel (1983) in what follows.

To approximate $h \in B^s_{pq}([0,1]) \subset B^s_{p\infty}([0,1])$, observe first that any $h \in B^s_{p\infty}([0,1])$ can be extended to a function $h^{ext} : \mathbb{R} \to \mathbb{R}$ (with $h^{ext} = h$ on $[0,1]$) that is contained in $B^s_{p\infty}(\mathbb{R})$ and satisfies the norm estimate $\|h^{ext}\|_{s,p,\infty,\mathbb{R}} \le c\|h\|_{s,p,\infty}$, see Theorem 3.3.4 and Section 3.4.2 in Triebel (1983). This extension can furthermore be taken to be compactly supported (if it is not, multiply h^{ext} with an infinitely differentiable function that is equal to one on $[0,1]$ and has compact support, and use the fact that $B^s_{pq}(\mathbb{R})$ is a multiplication algebra under the assumptions on s, p, q, see Theorem 2.8.3 in Triebel (1983)). Now, since h^{ext} is compactly supported and contained in $B^s_{p\infty}(\mathbb{R})$, one has also $h^{ext} \in B^s_{1\infty}(\mathbb{R})$ and $\|h^{ext}\|_{s,1,\infty,\mathbb{R}} \le c\|h^{ext}\|_{s,p,\infty,\mathbb{R}}$ (e.g., Theorem 3.3.1 in Triebel (1983)), so that we can apply the previous lemma to approximate h^{ext} by $K_j(h^{ext})$, and use the restriction $K_j(h) := K_j(h^{ext})|[0,1]$ of $K_j(h^{ext})$ to $[0,1]$ to approximate h. The norm $\|\cdot\|_{s,p,q}$ is equivalent to the restricted norm (again Theorem 3.3.4 in Triebel (1983)). Summarizing, we conclude

$$\|K_j(h)\|_{s,p,\infty} \le c'\|h\|_{s,p,\infty}, \tag{4}$$

and

$$\|K_j(h) - h\|_p \le c'' 2^{-js} \|h\|_{s,p,\infty} \tag{5}$$

as well as

$$\|K_j(h)\|_{1,1,1} \le c''' 2^{j(1/2-\delta)} \|h\|_{s,p,\infty} \tag{6}$$

for some $\delta > 0$ as $j \to \infty$, using $\|\cdot\|_{s,1,\infty} \le k \|\cdot\|_{s,p,q}$ on $B^s_{pq}([0,1])$ in the last step.

Step 2: The main idea to take advantage of the approximating sequence from the previous step is similar as in the proof of Theorem 2 in Nickl (2007):

$$\begin{aligned}
\sup_{h \in \mathcal{H}} |(\hat{P}_n - P_n)(h)| &\le \sup_{h \in \mathcal{H}} |(\hat{P}_n - P_n)(K_j(h) - h)| + \sup_{h \in \mathcal{H}} |(\hat{P}_n - P_n)(K_j(h))| \\
&\le \sup_{h \in \mathcal{H}} |(\hat{P}_n - P)(K_j(h) - h)| + \sup_{h \in \mathcal{H}} |(P_n - P)(K_j(h) - h)| \\
&\quad + \sup_{h \in \mathcal{H}} \Big(|(\hat{P}_n - P_n)(K_j(h)/\|K_j(h)\|_{1,1,1})| \|K_j(h)\|_{1,1,1} \Big) \\
&= (i) + (ii) + (iii).
\end{aligned}$$

We first prove $s > \max(1/p, 1/2)$ and comment on the modifications for the limiting case $s = 1/p, q = 1, p < 2$ at the end of the proof. Hence, for $p < 2$ one has $s > 1/p > 1/2$, so it follows that $B^s_{pq}([0,1]) \subset B^r_{2q}([0,1])$ for $r = s - 1/p + 1/2 > 1/2$ (Theorem 3.3.1 in Triebel (1983)), and we may further restrict ourselves to the case $p \ge 2$ and $s > 1/2$. Choose now $j := j_n$ such that $2^j \simeq n^{1/3}$ and recall the continuous injection $B^s_{pq}([0,1]) \subset B^s_{p\infty}([0,1])$.

The term (i) is bounded by

$$\int_0^1 |(f_n - f)(K_j(h) - h)| \le \|f_n - f\|_q \|K_j(h) - h\|_p = O_P(n^{-1/3} 2^{-js}) = o_P(n^{-1/2}),$$

uniformly in $\mathcal{H}$, using Hölder's inequality, (5) and the fact that $\|f_n - f\|_q = O_P(n^{-1/3})$ for all $q \le 2$, e.g., Theorem 24.6 in van der Vaart (1996).

For the second term, note $\sup_{h \in \mathcal{H}} \|K_j(h) - h\|_{s,p,\infty} < \infty$ by (4) and that

$$\sup_{h \in \mathcal{H}} \|K_j(h) - h\|_{2,P} \le c \sup_{h \in \mathcal{H}} \|K_j(h) - h\|_p \to 0$$

as $j \to \infty$ by (5) and boundedness of f. Consequently

$$\sup_{h \in \mathcal{H}} |(P_n - P)(K_j(h) - h)| = o_P(n^{-1/2})$$

has to hold since bounded subsets of $B^s_{p\infty}([0,1])$ are P-Donsker under the maintained conditions on s, p (see Proposition 1).

For the third term (iii), observe that any bounded subset of $B^1_{11}([0,1])$ consists of absolutely continuous functions h (e.g., 2.5.7/10 and 3.4.2 in Triebel (1983)) with uniformly bounded $\|h\|_\infty, \|Dh\|_1$, and hence is bounded in $BV([0,1])$, so that, using Corollary 1 and (6), the term (iii) is bounded by

$$\|\hat{P}_n - P_n\|_{\mathcal{H}_R} \sup_{h \in \mathcal{H}} \|K_j(h)\|_{1,1,1} = O_{a.s.}(n^{-2/3} (\log n)^{2/3} n^{1/6} n^{-\delta/3}) = o_{a.s.}(n^{-1/2}).$$

We finally turn to the limiting case $s = 1/p, q = 1, 1 \leq p < 2$, and consider first $p > 1$: As above, one imbeds $B^{1/p}_{p1}([0,1]) \subset B^{1/2}_{21}([0,1])$, but now one chooses j_n such that $2^{j_n} \simeq (n \log n)^{1/3}$. By the same reasoning as above (with $s = 1/2$), the term (i) is then of order $O_P(n^{-1/3}(n \log n)^{-1/6}) = o_P(n^{-1/2})$. The treatment of the second term is identical to above. For the the third term, note that we can choose $\delta = 1/p - 1/2$ in (6), cf. Lemma 1, so that by Corollary 1 this term is of order

$$O_{a.s.}\left(n^{-2/3}(\log n)^{2/3}(n \log n)^{(1/3)(1-1/p)}\right) = o_{a.s.}(n^{-1/2})$$

since $p < 2$ implies $2/3 - (1/3)(1 - 1/p) > 1/2$. If $p = 1$, then $s = 1$, so the theorem follows from Corollary 1 since, as mentioned above, $B^1_{11}(\mathbb{R}) \subset BV(\mathbb{R})$ with continuous injection. ■

Proof. (Theorem 3) Denote by $X_{(1)}, ..., X_{(n)}$ the order statistics of the sample, and set $X_{(0)} = 0$. Also define $S_n = n^{-1} \sum_{i=1}^n (\log f(X_i) - E \log f(X))$. We will prove that, for all $\epsilon, \delta > 0$ there exists a finite index N such that for $n \geq N$ one has

$$\Pr(\sqrt{n}|T_n - T(f) - S_n| > \epsilon) < \delta,$$

which implies the theorem by the CLT $\sqrt{n} S_n \to_d N(0, \sigma^2(f))$. Define

$$A_1 = \left\{ \inf_{x \in [0, X_{(n)}]} f_n(x) < \xi \right\}.$$

Then by Lemma 2 below, one has

$$\Pr\{\sqrt{n}|T_n - T(f) - S_n| > \epsilon\} \leq \Pr\left(\{\sqrt{n}|T_n - T(f) - S_n| > \epsilon\} \cap A_1^c\right) + \delta/2$$

for some small enough ξ, $0 < \xi < \inf_{x \in [0,1]} f(x)$, and n large enough. So in what follows we can work on the event A_1^c.

The (random) functional

$$g \mapsto \bar{T}_n(g) := \int_0^{X_{(n)}} g \log g$$

from $L^\infty([0, X_{(n)}])$ to $\mathbb{R}$ is Fréchet-differentiable on the open subset

$$\mathcal{V} = \{g \in L^\infty([0, X_{(n)}]) : g(x) > \xi/2 \text{ for all } x \in [0, X_{(n)}]\}$$

of $L^\infty([0, X_{(n)}])$, with first derivative

$$D\bar{T}_n(g)[h] = \int_0^{X_{(n)}} (\log g + 1) h$$

for $g \in \mathcal{V}, h \in L^\infty([0, X_{(n)}])$ and second derivative (for $h_1, h_2 \in L^\infty([0, X_{(n)}])$)

$$D^2\bar{T}_n(g)[h_1, h_2] = \int_0^{X_{(n)}} (1/g) h_1 h_2.$$

To see this, differentiate $f \mapsto f(x)\log f(x) = (w \circ \delta_x)(f)$ from $L^\infty([0, X_{(n)}])$ to $\mathbb{R}$ with $w(t) := t\log t$, using the chain rule on Banach spaces, and then observe that differentiation and integration can be interchanged, using, e.g., Proposition 4 in Nickl (2007).

Now since $f_n, f \in \mathcal{V}$ on A_1^c and since $f_n = 0$ on $(X_{(n)}, 1]$, we can write, by Taylor's expansion and the (pathwise) mean value theorem

$$\begin{aligned}
T_n - T(f) &= \bar{T}_n(f_n) - \bar{T}_n(f) - \int_{X_{(n)}}^1 f\log f \\
&= D\bar{T}_n(f)[f_n - f] + D^2\bar{T}_n(\tilde{f}_n)[f_n - f, f_n - f] - \int_{X_{(n)}}^1 f\log f \\
&= \int_0^{X_n} (\log f + 1)(f_n - f) + \int_0^{X_{(n)}} (1/\tilde{f}_n)(f_n - f)^2 - \int_{X_{(n)}}^1 f\log f \\
&= \int_0^1 (\log f + 1)(f_n - f) + \int_0^{X_{(n)}} (1/\tilde{f}_n)(f_n - f)^2 + \int_{X_{(n)}}^1 f \\
&= I + II + III
\end{aligned}$$

on the event A_1^c. Let now $\varepsilon = \epsilon/3$. Since f has a continuous derivative and is bounded from below, $\log f$ is bounded Lipschitz, so that we know from Corollary 2 that

$$\left|\int_0^1 (\log f + 1)(f_n - f) - S_n\right| = \left|\int_0^1 (\log f)(f_n - f) - S_n\right| = o_P(n^{-1/2}) \tag{7}$$

and hence, $\Pr(\{\sqrt{n}(I - S_n) > \varepsilon\} \cap A_1^c) \le \Pr(\{\sqrt{n}(I - S_n) > \varepsilon\}) \to 0$ as $n \to \infty$.

It therefore remains to prove that also the terms II and III are $o_P(n^{-1/2})$. For the term II, note that on the event A_1^c, the mean values $\tilde{f}_n$ on the line segment between f_n and f are bounded from below by ξ on $[0, X_{(n)}]$, so that II is bounded from above by

$$\xi^{-1}\int_0^{X_{(n)}} (f_n - f)^2 \le \xi^{-1}\|f_n - f\|_2^2$$

which is $O_P(n^{-2/3})$, see, e.g., Theorem 24.6 in van der Vaart (1998), hence $\Pr(\{\sqrt{n}(II) > \varepsilon\} \cap A_1^c) \to 0$.

The quantity III is bounded by $\|f\|_\infty(1 - X_{(n)})$. Using the formula for the density of the n-th order statistic (e.g., Lemma 13.1 in van der Vaart (1998)), we have by the mean value theorem and boundedness from below of $F' = f$ on $[0, 1]$ that, for all $\varepsilon > 0$

$$\begin{aligned}
\Pr(\sqrt{n}(1 - X_{(n)}) > \varepsilon) &= n\int_0^{1-\varepsilon n^{-1/2}} F(x)^{n-1} f(x)dx \\
&= (F(1 - \varepsilon n^{-1/2}))^n = \left(F(1) - F'(\eta)\frac{\varepsilon\sqrt{n}}{n}\right)^n \\
&\le \left(1 - \frac{c\varepsilon\sqrt{n}}{n}\right)^n \to 0
\end{aligned}$$

as $n \to \infty$, which implies that also $\Pr(\{\sqrt{n}(III) > \varepsilon\} \cap A_1^c) \to 0$.

It remains to prove the following lemma, which controls the probability of the event A_1.

Lemma 2 *Suppose the true density f satisfies $f(x) \geq \zeta > 0$ for all $x \in [0,1]$. Then, for every $\delta > 0$, there exists $\xi > 0$ and a finite index $N(\delta)$ such that, for all $n \geq N(\delta)$,*

$$\Pr\left(\inf_{x\in[0,X_{(n)}]} f_n(x) < \xi\right) = \Pr\left(f_n(X_{(n)}) < \xi\right) < \delta/2$$

Proof. The first equality of the lemma is obvious, since f_n is monotone decreasing. On each interval $(X_{(j-1)}, X_{(j)}]$, f_n is the slope of the least concave majorant of F_n (see, e.g., van der Vaart (1998, p.350)). The least concave majorant touches $(X_{(n)}, 1)$ and at least one other order statistic $(X_{(n-j)}, (n-j)/n)$, so that

$$\{f_n(X_{(n)}) < \xi)\} \subseteq \left\{X_{(n)} - X_{(n-j)} > j/(\xi n) \text{ for some } j = 1, ..., n\right\}.$$

Note next that $X_i = F^{-1}F(X_i)$ where F^{-1} is a differentiable function since F is strictly monotone and differentiable. Hence, by the mean value theorem,

$$F^{-1}F(X_{(n)}) - F^{-1}F(X_{(n-j)}) \leq \frac{1}{f(\eta)}\left(F(X_{(n)}) - F(X_{(n-j)})\right) \leq \zeta^{-1}\left(U_{(n)} - U_{(n-j)}\right)$$

where $U_{(i)}$ are the order statistics of a sample of size n of uniform random variables on $[0,1]$, and where $U_{(0)} = 0$ by convention. Hence it suffices to bound

$$\Pr\left(U_{(n)} - U_{(n-j)} > \frac{\zeta j}{\xi n} \text{ for some } j = 1, ..., n\right). \tag{8}$$

By Proposition 13.15 in Breiman (1968), the joint distribution of the order statistics $U_{(i)}$, $i = 1, ..., n$, is the same as the one of Z_i/Z_{n+1} where $Z_n = \sum_{l=1}^{n} W_l$ and where W_l are independent standard exponential random variables. Consequently, for $\gamma > 0$, the probability in (8) is bounded by

$$\begin{aligned}
&\Pr\left(\frac{W_{n-j+1} + ... + W_n}{Z_{n+1}} > \frac{\zeta j}{\xi n} \text{ for some } j\right) \\
&= \Pr\left(\frac{n}{Z_{n+1}}\frac{W_{n-j+1} + ... + W_n}{n} > \frac{\zeta j}{\xi n} \text{ for some } j\right) \\
&\leq \Pr(n/Z_{n+1} > 1+\gamma) + \Pr\left(\frac{W_{n-j+1} + ... + W_n}{n} > \frac{\zeta j}{\xi n(1+\gamma)} \text{ for some } j\right) \\
&= A + B.
\end{aligned}$$

To bound A, note that it is equal to

$$\Pr\left(\frac{1}{n+1}\sum_{l=1}^{n+1}(W_l - EW_l) < \frac{-\gamma - (1+\gamma)/n}{1+\gamma}\frac{n}{n+1}\right),$$

which, since $\gamma > 0$, is less than $\delta/4 > 0$ arbitrary, from some n onwards, by the law of large numbers. For the term B we have, for ξ small enough and by Markov's inequality

$$\begin{aligned}
&\Pr\left(W_{n-j+1}+...+W_n > \frac{\zeta j}{\xi(1+\gamma)} \quad \text{for some} \quad j\right)\\
&\leq \sum_{j=1}^{n} \Pr\left(W_{n-j+1}+...+W_n > \frac{\zeta j}{\xi(1+\gamma)}\right)\\
&= \sum_{j=1}^{n} \Pr\left(\sum_{l=1}^{j}(W_{n-l+1}-EW_{n-l+1}) > \frac{\zeta j}{\xi(1+\gamma)} - j\right)\\
&\leq \sum_{j=1}^{n} \frac{\xi^4 E(\sum_{l=1}^{j}(W_{n-l+1}-EW_{n-l+1}))^4}{j^4 C(\gamma,\zeta,\xi)}\\
&\leq \xi^4 C'(\gamma,\zeta,\xi)\sum_{j=1}^{n} j^{-2} = \xi^4 C''(\gamma,\zeta,\xi) < \delta/4,
\end{aligned}$$

where $C(\gamma,\zeta,\xi) = (1+\gamma)/(\zeta-\xi(1+\delta))$, since, for $Y_l = W_{n-l+1} - EW_{n-l+1}$, by Hoffmann-Jorgensen's inequality (de la Peña and Giné (1999), Theorem 1.5.13)

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

using the fact that $Var(Y_1) = 1$ and $E|Y_1|^p = p!$. ■ ■

Acknowledgement I would like to thank Evarist Giné for several helpful conversations.

References

[1] Balabdaoui, F. and Wellner, J.A. (2007). A Kiefer-Wolfowitz theorem for convex densities. *IMS Lecture notes monograph series* **55** 1-31.

[2] Bickel, J.P. and Ritov, Y. (2003). Nonparametric estimators which can be 'plugged-in'. *Ann. Statist.* **31** 1033-1053.

[3] Breiman, L. (1968). *Probability.* Addison-Wesley.

[4] de la Peña, V. and Giné, E. (1999). *Decoupling. From dependence to independence.* Springer, New York.

[5] Dudley, R.M. (1992). Fréchet differentiability, p-variation and uniform Donsker classes. *Ann. Prob.* **20** 1968-1982.

[6] Durot, C. and Tocquet, A.S. (2003). On the distance between the empirical process and its concave majorant in a monotone regression framework. *Ann. Inst. H. Poincaré Probab. Statist.* **39** 217-240.

[7] GINÉ, E. and NICKL, R. (2007). Uniform limit theorems for wavelet density estimators. *preprint.*

[8] GINÉ, E. and NICKL, R. (2008). Uniform central limit theorems for kernel density estimators. *Probab. Theory Related Fields*, forthcoming.

[9] GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Akt. Tid.* **39** 125-153.

[10] HÄRDLE, W.; KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, approximation, and statistical applications.* Lecture Notes in Statistics **129**. Springer, New York.

[11] KIEFER, J. and WOLFOWITZ, J. (1976). Asymptotically minimax estimation of concave and convex distribution functions. *Z. Wahr. verw. Gebiete* **34** 73-85.

[12] KULIKOV, V.N. and LUPOHAÄ (2006). The limit process of the difference between the empirical distribution function and its concave majorant. *Statist. Probab. Lett.* **76** 1781-1786.

[13] LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659-681.

[14] NICKL, R. (2007). Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Related Fields* **138** 411-449. Correction, ibid., forthcoming.

[15] NICKL, R. and PÖTSCHER, B.M. (2007). Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov- and Sobolev-type. *J. Theoret. Probab.* **20** 177-199.

[16] TRIEBEL, H. (1983). *Theory of function spaces.* Birkhäuser, Basel.

[17] VAN DER VAART, A.W. (1998). *Asymptotic statistics.* Cambridge University Press; Cambridge England.

[18] WANG, X. and WOODROOFE, M. (2007). A Kiefer-Wolfowitz comparison theorem for Wicksell's problem. *Ann. Statist.* **35** 1559-1575.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CONNECTICUT
STORRS, CT 06269-3009, USA
E-MAIL: nickl@math.uconn.edu