Submitted to the Annals of Statistics

CONVERGENCE OF LINEAR FUNCTIONALS OF THE GRENANDER ESTIMATOR UNDER MISSPECIFICATION

By Hanna Jankowski*

York University

Under the assumption that the true density is decreasing, it is well known that the Grenander estimator converges at rate $n^{1/3}$ if the true density is curved (Prakasa Rao, 1969) and at rate $n^{1/2}$ if the density is flat (Groeneboom and Pyke, 1983; Carolan and Dykstra, 1999). In the case that the true density is misspecified, the results of Patilea (2001) tell us that the global convergence rate is of order $n^{1/3}$ in Hellinger distance. Here, we show that the local convergence rate is $n^{1/2}$ at a point where the density is misspecified. This is not in contradiction with the results of Patilea (2001): the global convergence rate simply comes from locally curved *well-specified* regions. Furthermore, we study global convergence under misspecification by considering linear functionals. The rate of convergence is $n^{1/2}$ and we show that the limit is made up of two independent terms: a mean-zero Gaussian term and a second term (with non-zero mean) which is present only if the density has well-specified locally flat regions.

1. Introduction. Shape-constrained nonparametric maximum likelihood estimators provide an intriguing alternative to kernel-based density estimators. For example, one can compare the standard histogram with the Grenander estimator for a decreasing density. Rules exist to pick the bandwidth (or bin width) for the histogram to attain optimal convergence rates, cf. Wasserman (2006). On the other hand, the Grenander estimator gives a piecewise constant density, or histogram, but the bin widths are now chosen completely automatically by the estimator. Furthermore, the bin widths selected by the Grenander estimator are naturally locally adaptive (Birgé, 1987; Cator, 2011). Similar comparisons can also be made between the log-concave nonparametric MLE and the kernel density estimator.

The Grenander estimator was first introduced in Grenander (1956) and has been considered extensively in the literature since then. A recent review of the history of the problem appears in Durot et al. (2012). The latter paper establishes that the Grenander estimator converges to a true strictly

*Supported in part by an NSERC Discovery Grant

AMS 2000 subject classifications: Primary 62E20, 62G20, 62G07

Keywords and phrases: Grenander estimator, monotone density, misspecification, linear functional, nonparametric maximum likelihood

decreasing density at a rate of $(n/\log n)^{1/3}$ in the L_∞ norm. Other rates have also been derived over the years, most notably, convergence at a point at a rate of $n^{1/3}$ if the true density is locally strictly decreasing (Prakasa Rao, 1969; Groeneboom, 1985) and at a rate of $n^{1/2}$ if the true density is locally flat (Groeneboom, 1983; Carolan and Dykstra, 1999).

As noted in Cule et al. (2010); Dümbgen et al. (2011) the "success story" of maximum likelihood estimators is their robustness. Namely, let $\mathcal{F}$ denote the space of decreasing densities on $\mathbb{R}_+$. Next, let f_0 denote the true density and $\widehat{f_0}$ denote the density closest to f_0 in the Kullback-Leibler sense. That is,

$$\widehat{f_0} = \operatorname{argmin}_{g\in\mathcal{F}} \int_0^\infty f_0(x) \log \frac{f_0(x)}{g(x)} dx. \tag{1.1}$$

We will call the density $\widehat{f_0}$ the KL projection density of f_0, or the KL projection for short. Note that if $f_0 \in \mathcal{F}$ then $\widehat{f_0} = f_0$. Patilea (2001) showed that the density $\widehat{f_0}$ exists, and that the Grenander estimator converges to $\widehat{f_0}$ when the observed samples come from the true density f_0, regardless if $f_0 \in \mathcal{F}$. Similar results were proved for the log-concave maximum likelihood estimator in Cule and Samworth (2010); Cule et al. (2010); Dümbgen et al. (2011); Balabdaoui et al. (2013).

In order to understand the local behaviour of the Grenander estimator when $f_0 \notin \mathcal{F}$, we first need to define regions where f_0 is considered to be miss– and well–specified. Let $\widehat{F_0}$ denote the cumulative distribution function of $\widehat{f_0}$ defined in (1.1). The regions where $\widehat{F_0} \neq F_0$ are then the regions where f_0 is misspecified, and f_0 is considered to be well-specified otherwise. Note that, if f_0 is misspecified in a region, it may still be decreasing on some portion of this region, see e.g. Figure 1.

Let $\widehat{f_n}$ denote the Grenander estimator of a decreasing density. We show here that at a point where the density is misspecified the rate of convergence of $\widehat{f_n}$ to $\widehat{f_0}$ is $n^{1/2}$, and we also identify the limiting distribution. This is not in contradiction with the results of Patilea (2001): the slower $n^{1/3}$ global convergence rate simply comes from locally curved well-specified regions. To be more specific, if the density f_0 is misspecified at a point, then $\widehat{F_0}$ must be linear (and $\widehat{f_0}$ is flat), and in regions where $\widehat{f_0}$ is flat the rate of convergence is $n^{1/2}$. In fact, the $n^{1/2}$ rate holds at all flat regions of $\widehat{f_0}$, irrespective of whether these are miss– or well–specified. The complete result is given in Section 2, where some properties of $\widehat{f_0}$ are also discussed.

Next, we consider convergence of linear functionals. Let

$$\widehat{\mu}_0(g) = \int_0^\infty g(x)\widehat{f_0}(x)dx \quad \text{and} \quad \widehat{\mu}_n(g) = \int_0^\infty g(x)\widehat{f_n}(x)dx. \tag{1.2}$$

In Section 3 we show that $n^{1/2}(\widehat{\mu}_n(g)-\widehat{\mu}_0(g)) = O_p(1)$, and we again identify the limiting distribution. Notably, the limit is made up of two *independent* terms: a mean-zero Gaussian term and a second term with non-zero mean. Furthermore, the second term is present only if the density has well-specified locally flat regions. Our results apply to a wide range of KL projections with both strictly curved and flat regions. The work in the strictly curved case follows from the rates of convergence of $\widehat{F}_n(y) = \int_0^y \widehat{f}_n(y)dy$ to the empirical distribution function established in Kiefer and Wolfowitz (1976). However, as mentioned above, this is only for the well-specified regions of f_0. A related work here is that of Kulikov and Lopuhaä (2008), who consider functionals in the strictly curved case but at the distribution function level.

In Section 4 we go beyond the linear setting, and consider convergence of the entropy functional in the misspecified case. The limit in this case is Gaussian, irrespective of the local properties of $\widehat{f}_0$. Most proofs appear in Section 6 and some technical details are left to the Supplementary Material. Throughout, our results are illustrated by reproducible simulations. Code for these is available online at `www.math.yorku.ca/∼hkj/`.

To our best knowledge, previous work on rates of convergence under misspecification in the shape-constrained context is limited to the rates established in van de Geer (2000) and Patilea (2001), as well as the more recent results of Balabdaoui et al. (2013). In Balabdaoui et al. (2013), the pointwise asymptotic distribution under misspecification was derived for the log-concave probability mass function.

The implications of the new results obtained here are as follows. First, we now understand that $\widehat{f}_0$ will be made up of local well-specified and misspecified regions, and that the rate of convergence in the misspecified regions is always $n^{1/2}$. We conjecture that this type of behaviour will be seen in other situations, such as the log-concave setting for $d = 1$. That is, the rate of convergence in misspecified regions will be $n^{1/2}$ whereas in well-specified regions the rate of convergence will depend on whether locally the density lies on the boundary or the interior of the underlying space. In the log-concave $d = 1$ case, this "interior" rate is known to be $n^{2/5}$ (Balabdaoui et al., 2009). The interesting case of $d > 1$ is more mysterious though, as the relationship between the slower boundary points and faster interior points is harder to identify.

Secondly, we show that linear functionals (as well as the non-linear entropy functional) converge at rate $n^{1/2}$, and we also conjecture that this behaviour will continue to hold for other shape constraints. Let $\mu_0(g) = \int_0^\infty g(x)f_0(x)dx$. Our results show that

$$\sqrt{n}\,(\widehat{\mu}_n(g) - \mu_0(g)) = O_p(1) + \sqrt{n}\,(\widehat{\mu}_0(g) - \mu_0(g))\,. \tag{1.3}$$

Therefore, global rates of *divergence* are $n^{1/2}$ for linear functionals in the misspecified case. A similar statement also holds for the entropy functional, and here the random $O_p(1)$ term is always Gaussian. Such results are well-understood in parametric settings, and are key in power calculations. The exact conditions necessary for (1.3) to hold are given in Section 3 for $\mu_0(g)$ and in Section 4 for the entropy. Our work can also be easily extended to locally misspecified settings such as those studied in Le Cam (1960).

Lastly, the fact that the limiting distribution of the linear functional $\widehat{\mu}_n(g)$ depends on properties of $\widehat{f_0}$, whereas the limiting distribution of the entropy functional is always Gaussian, makes the entropy functional potentially more appealing in terms of testing procedures. Hypothesis testing based on functionals was considered, for example, in Cule et al. (2010) and Chen and Samworth (2013). The latter reference develops the "trace test" which depends on a nearly linear functional, the variance. Both, however, are developed in the context of log-concavity, and it would be of great interest to extend the results presented here to that setting, particularly for higher dimensions.

2. The Kullback-Leibler projection and pointwise convergence under misspecification. Properties of the KL projection onto the space of log-concave densities were studied in Dümbgen et al. (2011). When projecting onto the space of decreasing densities, the behaviour is a little easier to characterize.

THEOREM 2.1. *(Patilea, 1997, 2001) Let f_0 be a density with support on $[0,\infty)$ with $F_0(x) = \int_0^x f_0(u)du$. Let $\widehat{F}_0$ denote the least concave majorant of F_0. Then the left derivative of $\widehat{F}_0$, $\widehat{f}_0$, satisfies the inequality $\int \log \frac{\widehat{f}_0}{f} dF_0 \geq 0$, for all decreasing densities f.*

REMARK 2.2. *The density $\widehat{f_0}$ satisfying $\int \log \frac{\widehat{f}_0}{f} dF_0 \geq 0$ for all $f \in \mathcal{F}$ is called the "pseudo-true" density by Patilea (2001). If we additionally assume that $\sup_{f\in\mathcal{F}} \int \log f \, dF_0$ and $\int \log f_0 \, dF_0$ are both finite, then this $\widehat{f_0}$ is also the unique minimizer of the Kullback-Leibler divergence*

$$\widehat{f_0} = \operatorname{argmin}_{f\in\mathcal{F}} \int \log \tfrac{f_0}{f} dF_0.$$

See Patilea (2001, page 95) for more details. In what follows we continue to refer to $\widehat{f_0}$ as defined in Theorem 2.1, as the KL projection, even if it comes from the slightly more general definition of Patilea (2001).

Thus, in our setting, we have a complete graphical representation of the distribution function $\widehat{F}_0$ of the KL projection. This representation makes it

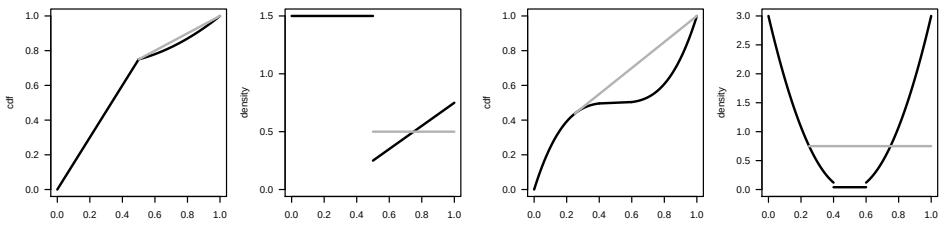

FIG 1. *Two examples of f_0 and $\widehat{f}_0 = \text{gren}(F_0)$. The two left panels show the cdf and density for example* (2.1) *while the two right panels show the cdf and density for example* (2.2). *F_0 (resp. f_0) is shown in black, and $\widehat{F}_0$ (resp. $\widehat{f}_0$) is shown in gray, but only if different from the truth (namely F_0 and f_0 respectively).*

possible to calculate $\widehat{f}_0$ in many cases. It also allows us to easily visualize the various F_0 which yield the same $\widehat{f}_0$. Moreover, the representation is key in understanding the behaviour of the estimator, both on the finite sample and asymptotic levels. Therefore, for a function g we define the operator $\text{gren}(g)$ to denote the (left) derivative of the least concave majorant of g. When the least concave majorant is restricted to a set $[a, b]$, we will write $\text{gren}_{[a,b]}(g)$.

Let $\mathcal{S}_0$ denote the support of f_0. We write $\mathcal{S}_0 = \mathcal{M} \cup \mathcal{W}$, where $\mathcal{M} = \{x \geq 0 : \widehat{F}_0(x) > F_0(x)\}$ and $\mathcal{W} = \{x \geq 0 : \widehat{F}_0(x) = F_0(x)\}$. Since f_0 is a density, it follows that F_0 is continuous, as is $\widehat{F}_0$, and therefore $\mathcal{W}$ is a closed set and $\mathcal{M}$ is open. For a fixed point $x_0 \in \mathcal{M}$, we thus know that x_0 lies in some open interval. Indeed, let $a_0 = \sup\{x < x_0 : \widehat{F}_0(x) = F_0(x)\}$ and $b_0 = \inf\{x > x_0 : \widehat{F}_0(x) = F_0(x)\}$. Then $x_0 \in (a_0, b_0)$ with $(a_0, b_0) \subset \mathcal{M}$.

Two examples are given in Figure 1. For the first example we have

$$f_0(x) = \begin{cases} 1.5 & x \in [0, 0.5] \\ x - 0.25 & x \in (0.5, 1]. \end{cases} \tag{2.1}$$

Here $\mathcal{M} = (0.5, 1)$ and $\mathcal{W} = [0, 0.5] \cup \{1\}$. For the second example we have

$$f_0(x) = \begin{cases} 12(x - 0.5)^2 & x \in [0, 0.4] \cup [0.6, 1] \\ 0.04 & x \in (0.4, 0.6). \end{cases} \tag{2.2}$$

Here $\mathcal{M} = (0.25, 1)$ and $\mathcal{W} = [0, 0.25] \cup \{1\}$.

The next proposition gives some additional properties of the KL projection.

PROPOSITION 2.3. *The density, $\widehat{f}_0$, satisfies the following:*

1. *Fix $x_0 \in \mathcal{M}$ and define a_0, b_0 as above. Then $b_0 < \infty$, and $\widehat{f}_0$ is constant on $(a_0, b_0]$ and satisfies the mean-value property*

$$\widehat{f}_0(x_0) \;=\; \frac{1}{b_0 - a_0} \int_{a_0}^{b_0} f_0(x)dx.$$

2. *Suppose that $\int_0^\infty f_0^2(x)dx < \infty$. Then $\widehat{f}_0 = \text{argmin}_{g\in\mathcal{F}} \int_0^\infty (g(x) - f_0(x))^2 dx$.*
3. *For any increasing function $h(x)$, $\int_0^\infty h(x)\widehat{f}_0(x)dx \le \int_0^\infty h(x)f_0(x)dx$.*
4. *Let $g_0 \in \mathcal{F}$ and let $G_0(y) = \int_0^y g_0(x)dx$. Then*

$$\sup_{x\ge 0} |\widehat{F}_0(x) - G_0(x)| \;\le\; \sup_{x\ge 0} |F_0(x) - G_0(x)|.$$

Point (3) above tells us that if g is increasing then $\mu_0(g) \ge \widehat{\mu}_0(g)$. Point (4) is Marshall's Lemma (Marshall, 1970). The proof of Proposition 2.3 appears in the Supplementary Material.

Suppose that $X_1, \ldots X_n$ are independent and identically distributed with density f_0 on $\mathbb{R}_+ = [0, \infty)$. Let $\widehat{f}_n$ denote the Grenander estimator of a decreasing density

$$\widehat{f}_n = \text{argmax}_{g\in\mathcal{F}} \int \log g(x)\, d\mathbb{F}_n(x),$$

where $\mathcal{F}$ denotes the class of decreasing densities on $\mathbb{R}_+$, and $\mathbb{F}_n(x) = n^{-1}\sum_{i=1}^n 1_{(-\infty,x]}(X_i)$ denotes the empirical distribution function. The next theorem is our first main result.

THEOREM 2.4. *Fix a point $x_0 \in \mathcal{M}$, and let $[a, b]$ denote the largest interval I containing x_0 such that $\widehat{F}_0(x)$ is linear on I. Let $\mathbb{U}$ denote a standard Brownian bridge process on $[0, 1]$, and let $\mathbb{U}_{F_0}(x) = \mathbb{U}(F_0(x))$ for $x \in \mathcal{S}_0$. Then*

$$\sqrt{n}(\widehat{f}_n(x_0) - \widehat{f}_0(x_0)) \;\Rightarrow\; \text{gren}_{[a,b]}\left(\mathbb{U}_{F_0}^{mod}\right)(x_0),$$

where

$$\mathbb{U}_{F_0}^{mod}(u) \;=\; \begin{cases} \mathbb{U}_{F_0}(u) & u \in [a, b] \cap \mathcal{W}, \\ -\infty & u \in [a, b] \cap \mathcal{M}. \end{cases}$$

If it happens that $[a, b] \cap \mathcal{W} = \{a, b\}$, then

$$\sqrt{n}(\widehat{f}_n(x_0) - \widehat{f}_0(x_0)) \;\Rightarrow\; \sigma Z,$$

where Z is a standard normal random variable and

$$\sigma^2 \;=\; \widehat{f}_0(x_0)\left[\frac{1}{b-a} - \widehat{f}_0(x_0)\right].$$

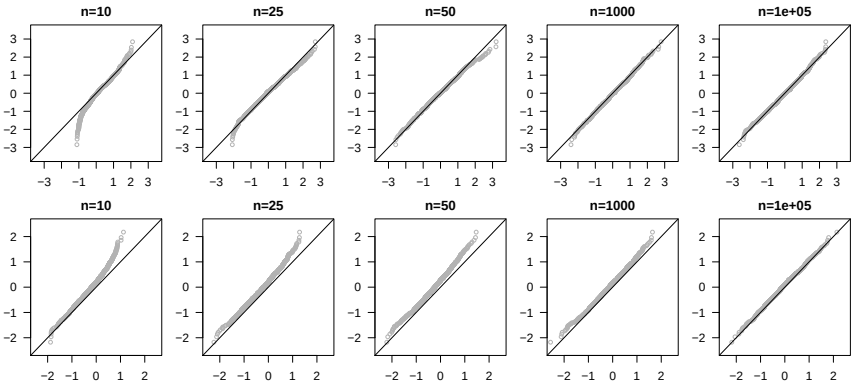

FIG 2. *Empirical quantiles of* $\sqrt{n}(\widehat{f}_n(x_0) - \widehat{f}_0(x_0))$ *vs. the true quantiles of the limiting* $N(0, \sigma^2)$ *distributions at the point* $x_0 = 0.75$ *for* f_0 *given by* (2.1) *in the top row* ($\sigma^2 = 3/4$) *and* (2.2) *in the bottom row* ($\sigma^2 = 7/16$). *The sample size varies from* $n = 10$ *to* $n = 100\,000$. *The straight line goes through the origin and has slope one. Each plot is based on* $B = 1000$ *samples.*

Recall that Patilea (2001, Corollary 5.6) shows that the rate of convergence (in Hellinger distance) of $\widehat{f}_n$ to $\widehat{f}_0$ is $n^{1/3}$. The above theorem shows that the *local* rate of convergence will be $\sqrt{n}$ where the KL projection is flat. When the KL density is curved, the KL density and true density are actually equal, and hence the convergence rate from the correctly specified case applies. The next formulation of the limiting process is similar to that of Carolan and Dykstra (1999) for a density with a flat region on $[a, b]$.

REMARK 2.5. *Let* $p_0 = F_0(b) - F_0(a) = \widehat{F}_0(b) - \widehat{F}_0(a)$. *Since* $\widehat{F}_0$ *is linear on* $[a, b]$ *the limiting distribution may also be expressed as*

$$\operatorname{gren}_{[a,b]}\left(\mathbb{U}_{F_0}^{mod}\right)(x_0) \;=\; \frac{1}{b-a}\left\{Z + \sqrt{p_0}\,\operatorname{gren}(\mathbb{U}^{mod})\left(\frac{x_0 - a}{b-a}\right)\right\},$$

where Z *is a mean zero normal random variable with variance* $p_0(1 - p_0)$, $\mathbb{U}$ *is an independent standard Brownian bridge, and*

$$\mathbb{U}^{mod}(u) \;=\; \begin{cases} \mathbb{U}(u) & u \in ([a, b] \cap \mathcal{W} - a)/(b - a), \\ -\infty & u \in ([a, b] \cap \mathcal{M} - a)/(b - a). \end{cases}$$

Notably, if $[a, b] \cap \mathcal{W} = \{a, b\}$, *then* $\operatorname{gren}(\mathbb{U}^{mod})(u) = 0$.

Figure 2 illustrates the theory. The convergence is surprisingly fast, although it appears to be a little slower in the second example (2.2). We conjecture that this difference is caused by the presence/absence of the strictly curved region of f_0.

PROOF OF THEOREM 2.4. By the switching relation (Balabdaoui et al., 2011), we have

$$\begin{aligned} & P\left(\sqrt{n}(\widehat{f}_n(x_0) - \widehat{f}_0(x_0)) < t\right) \\ &= P\left(\text{argmax}_{z\geq 0}\left\{\mathbb{F}_n(z) - (\widehat{f}_0(x_0) + n^{-1/2}t)z\right\} < x\right) \\ &= P\left(\text{argmax}_{z\geq 0}\left\{\sqrt{n}\left(\mathbb{F}_n(z) - \mathbb{F}_n(a) - (F_0(z) - F_0(a))\right)\right.\right. \\ &\qquad \left.\left. + \sqrt{n}\left(F_0(z) - F_0(a) - \widehat{f}_0(x_0)(z-a)\right) - tz\right\} < x\right). \end{aligned}$$

We now look more closely at the "second" term. That is,

$$\begin{aligned} & F_0(z) - F_0(a) - \widehat{f}_0(x_0)(z-a) \\ &\qquad = -\left\{\widehat{F}_0(z) - F_0(z)\right\} + \left\{\widehat{F}_0(z) - \widehat{F}_0(a) - \widehat{f}_0(x_0)(z-a)\right\}, \end{aligned}$$

noting that $\widehat{F}_0(a) = F_0(a)$, since $a \in [a,b] \cap \mathcal{W}$. On the other hand, for all $z \in [a,b] \cap \mathcal{M}$, we have $\widehat{F}_0(z) > F_0(z)$. Furthermore, $\widehat{F}_0$ is concave with derivative $\widehat{f}_0(x_0)$ (at any point $z \in (a,b)$), and hence

$$\widehat{F}_0(z) - \widehat{F}_0(a) - \widehat{f}_0(x_0)(z-a) \leq 0$$

for all $z \geq 0$. For $z \in [a,b] \cap \mathcal{W}$ this is an equality, and a strict inequality otherwise. Therefore, the weak limit of

$$\sqrt{n}\left\{\mathbb{F}_n(z) - \mathbb{F}_n(a) - (F_0(z) - F_0(a))\right\} - \sqrt{n}\left(F_0(z) - F_0(a) - \widehat{f}_0(x_0)(z-a)\right)$$

is $\mathbb{U}_{F_0}^{mod}(z) - \mathbb{U}_{F_0}^{mod}(a) = \mathbb{U}_{F_0}^{mod}(z) - \mathbb{U}_{F_0}(a)$, for all $z \in [a,b]$. For $z \notin [a,b] \cap \mathcal{W}$, the limit of this process is always $-\infty$, and therefore the maximum must occur inside of $[a,b]$. By the argmax continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 3.2.2, page 287),

$$\begin{aligned} P\left(\sqrt{n}(\widehat{f}_n(x_0) - \widehat{f}_0(x_0)) < t\right) &\to P\left(\text{argmax}_{z\in[a,b]}\left\{\mathbb{U}_{F_0}^{mod}(z) - tz\right\} < x\right) \\ &= P\left(\text{gren}_{[a,b]}(\mathbb{U}_{F_0}^{mod})(x_0) < t\right), \end{aligned}$$

by switching again. When $[a,b] \cap \mathcal{W} = \{a,b\}$, then the least concave majorant is simply the line joining $\mathbb{U}_{F_0}(a)$ and $\mathbb{U}_{F_0}(b)$, with slope equal to

$$\frac{\mathbb{U}_{F_0}(b) - \mathbb{U}_{F_0}(a)}{b-a},$$

a Gaussian random variable with mean zero and variance

$$\frac{1}{(b-a)^2}(F_0(b) - F_0(a))\left[1 - (F_0(b) - F_0(a))\right] = \widehat{f}_0(x_0)\left[\frac{1}{b-a} - \widehat{f}_0(x_0)\right].$$

□

PROOF OF REMARK 2.5. Recall that $\widehat{F}_0$ is linear on $[a,b]$. Therefore, for $x \in [a,b]$, we can write $\mathbb{U}(\widehat{F}_0(x)) - \mathbb{U}(\widehat{F}_0(a)) = \frac{x-a}{b-a}\,\mathbb{W} + \mathbb{V}(x)$, where

$$\begin{aligned}
\mathbb{W} &= \mathbb{U}(\widehat{F}_0(b)) - \mathbb{U}(\widehat{F}_0(a)), \\
\mathbb{V}(x) &= \mathbb{U}(\widehat{F}_0(x)) - \mathbb{U}(\widehat{F}_0(a)) - \frac{\widehat{F}_0(x) - \widehat{F}_0(a)}{\widehat{F}_0(b) - \widehat{F}_0(a)}\,\mathbb{W} \\
&= \mathbb{U}(\widehat{F}_0(x)) - \mathbb{U}(\widehat{F}_0(a)) - \frac{x-a}{b-a}\,\mathbb{W}.
\end{aligned}$$

Since all variables are jointly Gaussian, a careful calculation of the covariances reveals that $\mathbb{W}$ and $\mathbb{V}(x)$ are independent (also as processes), and $\mathbb{W}$ is mean-zero Gaussian with variance $p_0(1-p_0)$. Furthermore,

$$\mathbb{V}(s) \stackrel{d}{=} \sqrt{p_0}\,\mathbb{U}\left(\frac{s-a}{b-a}\right).$$

This decomposition is similar to that of Shorack and Wellner (1986, Exercise 2.2.11, page 32). Now, note that the Grenander operator satisfies $\mathrm{gren}_{[a,b]}(g)(x) = \beta + \frac{\gamma}{b-a}\,\mathrm{gren}_{[0,1]}(h)\left(\frac{t-a}{b-a}\right)$ if $g(t) = \alpha + \beta t + \gamma h\left(\frac{t-a}{b-a}\right)$. It follows that

$$\mathrm{gren}_{[a,b]}\left(\mathbb{U}_{\widehat{F}_0}\right)(x_0) \;=\; \frac{1}{b-a}\,Z + \frac{\sqrt{p_0}}{b-a}\,\mathrm{gren}(\mathbb{U})\left(\frac{x_0-a}{b-a}\right),$$

with $Z, \mathbb{U}$ defined as in the Remark. The full result follows since, $\mathbb{U}^{mod}_{F_0}(x) = \mathbb{U}^{mod}_{\widehat{F}_0}(x) = \frac{x-a}{b-a}\mathbb{W} + \mathbb{V}^{mod}(x)$. □

3. $\sqrt{n}$-convergence of linear functionals. Consider a density f_0 with support $\mathcal{S}_0$ and let $\widehat{f}_0$ denote its KL projection. We write $\mathcal{S}_0 = \mathcal{S}_c \cup \mathcal{S}_f$, where $\mathcal{S}_c$ denotes the portion of the support where $\widehat{f}_0$ is curved and $\mathcal{S}_f$ denotes the portion of the support where $\widehat{f}_0$ is flat. By definition of S_f as well as Proposition 2.3, the KL projection can be written as

$$\widehat{f}_0(x) = \textstyle\sum_{j=1}^{J} \widehat{q}_j\, 1_{I_j}(x) \tag{3.1}$$

on S_f, where the intervals are disjoint and each is of the form $I_j = (a_j, b_j]$. Our results for linear functionals hold under the following assumptions.

(S). The support, $\mathcal{S}_0$, of f_0 is bounded.
(C). When the KL projection is curved, $\sup_{x\in\mathcal{S}_c} |f_0'(x)| < +\infty$.
(P). The true density is strictly positive: $\inf_{x\in S_0} f_0(x) > 0$.
(F). When the KL projection is flat, J is finite in (3.1).

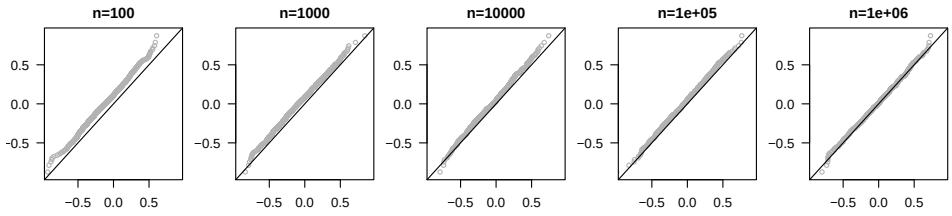

FIG 3. *Empirical quantiles of* $\sqrt{n}(\widehat{\mu}_n(g) - \widehat{\mu}_0(g))$ *with* $g(x) = x$ *vs. the true quantiles of the limiting* $N(0, \sigma^2)$ *distribution for* f_0 *given in* (2.2)*, with* $\sigma^2 \approx 0.07032$.

Let $g : \mathcal{S}_0 \mapsto \mathbb{R}$ and define $\widehat{\mu}_n(g)$ by (1.2). Then we require that g satisfy the following conditions.

(G1). $\int_{\mathcal{S}_c} |g'(x)|dx < \infty$.
(G2). $g \in L_\beta(\mathcal{S}_f)$ for some $\beta > 2$.

In order to state our main result for linear functionals we need to define the following functions,

$$\begin{aligned} g_j(u) &= g\left((b_j - a_j)u + a_j\right) \quad u \in [0,1], \\ \overline{g}_j &= (b_j - a_j)^{-1} \textstyle\int_{a_j}^{b_j} g(x)dx. \end{aligned} \tag{3.2}$$

$$\overline{g}(x) = \begin{cases} g(x) & x \in \mathcal{S}_c, \\ \overline{g}_j & x \in I_j,\ j = 1, \ldots, J. \end{cases} \tag{3.3}$$

Thus, $\overline{g}_1, \ldots, \overline{g}_J$ are the local averages of the function g, and each $g_j(u)$ is a localized version of g.

THEOREM 3.1. *Suppose that the density* f_0 *satisfies conditions (S), (C), (P), and (F). Consider a function* $g : \mathcal{S}_0 \mapsto \mathbb{R}$ *which satisfies conditions (G1) and (G2). Let* $\mathbb{U}, \mathbb{U}_1, \ldots, \mathbb{U}_J$ *denote independent Brownian bridges,* $\mathbb{U}_{F_0}(x) = \mathbb{U}(F_0(x))$*, and define* $\mathbb{U}_j^{mod}$ *as in Theorem 2.4. Then*

$$\begin{aligned} \sqrt{n}\left(\widehat{\mu}_n(g) - \widehat{\mu}_0(g)\right) \;\Rightarrow\; & \int_{\mathcal{S}_0} \overline{g}(x)d\mathbb{U}_{F_0}(x) \\ & + \sum_{j=1}^{J} \sqrt{p_j} \int_0^1 g_j(u)\, \mathrm{gren}(\mathbb{U}_j^{mod})(u)du, \end{aligned}$$

where $p_j = F_0(b_j) - F_0(a_j) = \widehat{F}_0(b_j) - \widehat{F}_0(a_j)$*. Furthermore,* $\int_{\mathcal{S}_0} \overline{g}(x)d\mathbb{U}_{F_0}(x) = \int_{\mathcal{S}_0} \overline{g}(x)d\mathbb{U}_{\widehat{F}_0}(x)$*. Also, if* $I_j \cap \mathcal{W} = \{a_j, b_j\}$*, then* $\mathrm{gren}(\mathbb{U}_j^{mod}) \equiv 0$.

It follows that $\sqrt{n}(\widehat{\mu}_n(g)-\widehat{\mu}_0(g))$ will converge to a Gaussian limit for true density (2.2) but not for (2.1), as the latter has well-specified flat regions. A simulation for (2.2) is shown in Figure 3. The proof of Theorem 3.1 is given in Section 6. The simulations show that there appears a systematic bias prior to convergence (the empirical quantiles appear on the x-axis in Figure 3, the negative bias translates to a left-shift in the plot). The proof of Proposition 6.1 shows that one source of the bias is the term $\sqrt{n}\int x d(\widehat{F}_n - \mathbb{F}_n) \approx -\sqrt{n}\int(\widehat{F}_n - \mathbb{F}_n) \leq 0$. When $\mathcal{S}_0 = \mathcal{S}_c$, this term is the only source of bias, and from Kiefer and Wolfowitz (1976), it converges to zero at a rate of at least $n^{1/6}(\log n)^{-2/3}$. Since (3) of Proposition 2.3 also holds at the empirical level, similar behaviour will be seen for all increasing functions g.

The results of Theorem 3.1 also show that $\sqrt{n}(\widehat{\mu}_n(g) - \widehat{\mu}_0(g))$ is asymptotically normal with variance $\mathrm{var}_{f_0}(\overline{g}(X)) = \mathrm{var}_{\widehat{f}_0}(\overline{g}(X))$ if $\mathcal{S}_0$ has no *well-specified flat* regions. Additionally, if $\mathcal{S}_0 = \mathcal{S}_c$, then $\overline{g}(x) = g(x)$ and the model is well-specified. In this case, $\widehat{\mu}_n(g)$ has the same asymptotic distribution as the empirical estimator $n^{-1}\sum_{i=1}^n g(X_i)$ (see also Proposition 6.1). This shows that the maximum likelihood estimator is asymptotically efficient, as in the strictly curved case the family of decreasing densities is complete, and hence the "naive estimator" $n^{-1}\sum_{i=1}^n g(X_i)$ is asymptotically efficient (van de Geer, 2003, Example 4.7).

Finally, we make a few comments on the assumptions required for Theorem 3.1 to hold. The assumptions which we use on S_c are (S), (P), and (C). These are quite standard assumptions in the literature for the strictly curved setting, see for example Kiefer and Wolfowitz (1976); Durot et al. (2012); Kulikov and Lopuhaä (2008); Groeneboom et al. (1999); Durot and Lopuhaä (2013). In the misspecified region, the required assumptions are (P), and (F). Note also that by Remark 3.2, the assumption (G2) is required in the result. Additional discussions of these assumptions, including directions for future research, are provided in the Supplementary Material.

To further illustrate these assumptions, as well as Theorem 3.1, we consider the examples (2.1) and (2.2). In example (2.1), we have that

$$\widehat{f}_0(x) \quad = \quad 1.5\,1_{[0,0.5]}(x) + 0.5\,1_{(0.5,1]}(x). \tag{3.4}$$

The conditions (S) and (P) are clearly satisfied, as is (C) since $\mathcal{S}_0 = \mathcal{S}_f$. Lastly, (F) holds with $J = 2, \widehat{q}_1 = 1.5, \widehat{q}_2 = 0.5, I_1 = (0, 0.5], I_2 = (0.5, 1]$. Applying Theorem 3.1 for $g(x) = x$, we find that $\overline{g}(x) = 0.75\,1_{[0,0.5]}(x) +$

$0.25\,1_{(0.5,1]}(x)$, and $I_2 \cap \mathcal{W} = \{a_2, b_2\}$ (hence $\mathrm{gren}(\mathbb{U}_2^{mod}) = 0$). Therefore,

$$\begin{aligned}\sqrt{n}(\widehat{\mu}_n(g) - \widehat{\mu}_0(g)) \quad &\Rightarrow \quad \int_0^1 \overline{g}(x) d\mathbb{U}_{F_0}(x) + \sqrt{\frac{3}{4}} \int_0^1 \frac{u}{2}\,\mathrm{gren}(\mathbb{U}_1^{mod})(u) du \\ &= \quad -\frac{1}{2}\mathbb{U}_{F_0}(0.5) + \sqrt{\frac{3}{16}} \int_0^1 u\,\mathrm{gren}(\mathbb{U}_1)(u) du, \end{aligned} \tag{3.5}$$

where $\mathbb{U}_{F_0}, \mathbb{U}_1$ are independent Brownian bridges as defined in Theorem 3.1. Notably, the limit has a non-Gaussian component.

Example (2.2) can be analysed similarly. Here,

$$\widehat{f_0}(x) \quad = \quad 12(x - 0.5)^2\, 1_{[0,0.25]}(x) + 0.75\, 1_{(0.25,1]}(x).$$

Again, the conditions (S) and (P) clearly hold. On $S_c = [0, 0.25]$, we have $\sup_{x \in S_c} |f_0'(x)| = 12$, and therefore condition (C) holds. On $\mathcal{S}_f = (0.25, 1]$ we have $J = 1$ and hence (F) also holds. Applying Theorem 3.1 for $g(x) = x$, we find that $\overline{g}(x) = x\, 1_{[0,0.25]}(x) + (5/8) 1_{(0.25,1]}(x)$, and $I_1 \cap \mathcal{W} = \{a_1, b_1\}$ (hence $\mathrm{gren}(\mathbb{U}_1^{mod}) = 0$). Therefore,

$$\sqrt{n}(\widehat{\mu}_n(g) - \widehat{\mu}_0(g)) \quad \Rightarrow \quad \int_0^1 \overline{g}(x) d\mathbb{U}_{F_0}(x),$$

That is, the limit is zero-mean Gaussian with variance $\sigma^2 \approx 0.07032$.

Remark 3.2. *Marginal properties of the process* $\mathrm{gren}(\mathbb{U})$ *were studied in Carolan and Dykstra (2001). The results include marginal densities and moments, including* $\mathrm{E}[(\mathrm{gren}(\mathbb{U})(x))^2] = 0.5(x^2/(1-x) + (1-x)^2/x)$. *It follows that* $\mathrm{E}[\int_0^1 (\mathrm{gren}(\mathbb{U})(x))^2 dx] = \int_0^1 (1-x)^2/x\, dx = \infty$, *and hence the limiting process*

$$< g, \mathrm{gren}(\mathbb{U}) > \quad = \quad \int_0^1 g(x)\,\mathrm{gren}(\mathbb{U})(x) dx,$$

exists only for $g \in L_\beta(\mathcal{S}_f)$ *for* $\beta > 2$. *We would therefore not expect convergence of* $\widehat{\mu}_n(g)$ *for* $g \in L_\beta(\mathcal{S}_f)$ *with* $\beta \in [1, 2]$.

4. Beyond linear functionals: a special case. Entropy measures the amount of disorder or uncertainty in a system and is closely related to the Kullback-Leibler divergence. Let $T(f) = \int_0^\infty f(x) \log f(x) dx$ denote the entropy functional. A review of testing and other applications of entropy appears, for example, in Beirlant et al. (1997).

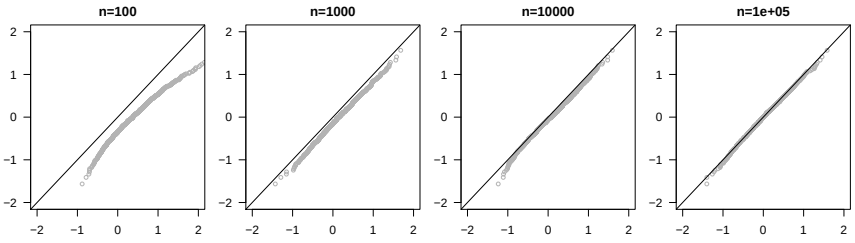

FIG 4. *Empirical quantiles of* $\sqrt{n}(T(\widehat{f}_n) - T(\widehat{f}_0))$ *vs. the true quantiles of the limiting* $N(0, \sigma^2)$ *distribution for* f_0 *given in* (2.1)*, with* $\sigma^2 \approx 0.2263$.

THEOREM 4.1. *Suppose that* $\widehat{f}_0$ *is bounded, the support of* f_0 *is also bounded, and that* $f_0/\widehat{f}_0 \leq c_0^2 < \infty$. *Then*

$$\sqrt{n}\left(T(\widehat{f}_n) - T(\widehat{f}_0)\right) \quad \Rightarrow \quad \sigma Z,$$

where Z *is a standard normal random variable and*

$$\sigma^2 = \mathrm{var}_{f_0}(\log(\widehat{f}_0(X))) = \mathrm{var}_{\widehat{f}_0}(\log(\widehat{f}_0(X))).$$

The proof is made up of two key pieces: (1) tight bounds on the likelihood ratio from Lemma 4.2 and (2) specialized equalities which hold for the Grenander estimator.

LEMMA 4.2. *Suppose that* $\widehat{f}_0$ *is bounded, the support of* f_0 *is also bounded, and that* $f_0/\widehat{f}_0 \leq c_0^2 < \infty$. *Then*

$$\int \log \frac{\widehat{f}_n}{\widehat{f}_0} d\mathbb{F}_n \quad = \quad O_p(n^{-2/3}).$$

We note that the conditions we require here are stronger than those of Patilea (2001, Corollary 5.6). However, under those conditions Patilea (2001) establishes convergence rates on $\int \log \frac{2\widehat{f}_n}{\widehat{f}_n + \widehat{f}_0} d\mathbb{F}_n$, which is not sufficient for our purposes. The condition that $f_0/\widehat{f}_0$ is bounded above was also used in the study of misspecification in van de Geer (2000, Section 10.4). The condition that the support of f_0 is bounded is the strongest, whereas the condition that $\widehat{f}_0$ is bounded may be relaxed somewhat. We discuss this further in the Supplementary Material.

PROOF. We first show that $\int_0^\infty \varphi(\widehat{f}_n) d(\widehat{F}_n - \mathbb{F}_n) = 0$ for any function φ. This follows since $\widehat{F}_n(x) \geq \mathbb{F}_n(x)$ with equality at finitely many touch points,

and also $\widehat{f}_n$ is constant between all touch points. Thus, letting $\tau_1, \tau_2, \ldots, \tau_m$ enumerate the (random) points of touch, we have

$$\int_0^\infty \varphi(\widehat{f}_n) d(\widehat{F}_n - \mathbb{F}_n) \;=\; \sum_{i=1}^{m} \varphi(\widehat{f}_n(\tau_i)) \left((\widehat{F}_n - \mathbb{F}_n)(\tau_i) - (\widehat{F}_n - \mathbb{F}_n)(\tau_{i-1}) \right) \;=\; 0,$$

with $\tau_0 = 0$ and $\tau_m = X_{(n)}$. A similar argument also establishes that

$$\int_0^\infty \varphi(\widehat{f}_0) d(\widehat{F}_0 - F_0) = \int_{\mathcal{M}} \varphi(\widehat{f}_0) d(\widehat{F}_0 - F_0) = 0. \tag{4.1}$$

For $\varphi(v) = \log v$, it follows that

$$\begin{aligned} \sqrt{n} \left(T(\widehat{f}_n) - T(\widehat{f}_0) \right) &= \sqrt{n} \left(\int \log \widehat{f}_n \, d\widehat{F}_n - \int \log \widehat{f}_0 \, d\widehat{F}_0 \right) \\ &= \sqrt{n} \int \log \left(\frac{\widehat{f}_n}{\widehat{f}_0} \right) d\mathbb{F}_n + \sqrt{n} \int \log \widehat{f}_0 \, d(\mathbb{F}_n - F_0). \end{aligned}$$

The first term is $O_p(n^{-1/6})$ by Lemma 4.2. The second term has a Gaussian limit with variance $\mathrm{var}_{f_0}(\log \widehat{f}_0(X))$. By (4.1) (with $\varphi(v) = \log^2 v, \log v$) this is equal to $\mathrm{var}_{\widehat{f}_0}(\log \widehat{f}_0(X))$. □

A simulation of this result is shown in Figure 4 based on the true density (2.1). The KL projection of (2.1) is given in (3.4). One can easily check that the conditions of Theorem 4.1 are satisfied in this case. Note that this density has well-specified flat regions, and therefore linear functionals that do not ignore $\mathcal{S}_f \cap \mathcal{W}$ should have non-Gaussian terms in their limit; see, for example, (3.5) for the case when $g(x) = x$. On the other hand, the entropy functional will *always* result in a Gaussian limit. The simulations exhibit a systematic positive bias. The proof shown above reveals the cause: The term $\int \log(\widehat{f}_n/\widehat{f}_0) d\mathbb{F}_n \geq 0$ since $\widehat{f}_n$ is the MLE. In the plots the quantiles of $\sqrt{n}(T(\widehat{f}_n) - T(\widehat{f}_0))$ are shown on the x–axis, and these quantiles appear to be shifted to the right – that is, they are larger than the quantiles of the limiting Gaussian distribution.

5. Conclusion. We anticipate that extensions of this work to other one-dimensional shape-constrained models, such as the log-concave and convex decreasing constraints, are within reach, although certain technical difficulties will need to be overcome. In particular, the results of Patilea (2001) for convex models should yield some results for convex decreasing densities under misspecification. The Grenander estimator has a particular simplicity

of form, which we have exploited here. Some progress for the log-concave setting has already been made in Balabdaoui et al. (2013), albeit for the discrete (i.e. probability mass function) setting. We conjecture that statements such as (1.3) will continue to hold for other shape-constraints in $d = 1$ for linear functionals. Similar results for higher dimensional shape-constrained models seem premature in view of the current lack of rate of convergence results even when the model is correctly specified.

6. Proofs for Section 3. We now present the proof for Theorem 3.1. We proceed by proving convergence results for the different types of behaviours of the density separately (curved, flat, misspecified), and combine the results together at the end. We believe that the intermediate results are of independent interest to the reader, and we also hope that this approach makes the proof more accessible.

6.1. *Strictly curved well-specified density.* We first suppose that the true density f_0 satisfies the conditions introduced in Kiefer and Wolfowitz (1976).

PROPOSITION 6.1. *Suppose that f_0 satisfies conditions (S) and (C), and that g satisfies condition (G1). Then*

$$\sqrt{n}(\widehat{\mu}_n(g) - \mu_0(g)) \quad \Rightarrow \quad \sigma Z,$$

where Z is a standard normal random variable and $\sigma^2 = \mathrm{var}(g(X)) < \infty$.

We note that this result is similar to that in Kulikov and Lopuhaä (2008).

PROOF. Without loss of generality, we assume that $\mathcal{S}_0 = \mathcal{S}_c = [0, 1]$. Let $\overline{\mu}_n(g) = n^{-1} \sum_{i=1}^n g(X_i)$ denote the empirical estimator of $\mu_0(g)$. Using Fubini, we have

$$\begin{aligned} |\widehat{\mu}_n(g) - \overline{\mu}_n(g)| &= \left| \int_0^1 g'(x) \left[\widehat{F}_n(x) - \mathbb{F}_n(x) \right] dx \right| \\ &\leq \left\{ \int_{\mathcal{S}_c} |g'(x)| dx \right\} \sup_{x \in \mathcal{S}_c} |\widehat{F}_n(x) - \mathbb{F}_n(x)|. \end{aligned}$$

From the results of Kiefer and Wolfowitz (1976) (see also Durot and Lopuhaä (2013, Corollary 2.2)), we have that $\sup_{x \in [0,1]} \sqrt{n} |\widehat{F}_n(x) - \mathbb{F}_n(x)| = o_p(n^{-1/6} \log^{2/3} n)$. Therefore,

$$\sqrt{n}(\widehat{\mu}_n(g) - \mu_0(g)) = \sqrt{n}(\overline{\mu}_n(g) - \mu_0(g)) + o_p(n^{-1/6} \log^{2/3} n),$$

from which the result follows. □

6.2. *Piecewise constant well-specified density.* Suppose next that $\mathcal{S}_0 = \mathcal{S}_f = \mathcal{W} \cap \mathcal{S}_0$. That is, the true density is piecewise constant decreasing and can be expressed as

$$f_0(x) = \sum_{j=1}^{J} q_j 1_{(a_j, b_j]}(x) \tag{6.1}$$

where $q_1 > q_2 > \cdots > q_J > 0$, J is finite, and $\cup I_j = \mathcal{S}_0$ where the sets $I_j = (a_j, b_j]$ are disjoint. Indeed, we have $b_j = a_{j+1}$ for $j = 1, \ldots, J-1$. Note that $p_j = q_j(b_j - a_j)$. Also, let $\mathbb{U}_1, \ldots, \mathbb{U}_J$ denote independent standard Brownian bridge processes (each defined on $[0,1]$), and let $\{Z_1, \ldots, Z_J\}$ be an independent multivariate normal with mean zero and covariance $\mathrm{diag}(p) - pp^T$ for $p = (p_1, \ldots, p_J)^T$.

PROPOSITION 6.2. *Suppose that f_0 is as in (6.1). Then $\sqrt{n}(\widehat{f}_n(x) - f_0(x))$ converges weakly to $\mathbb{S}(x)$ in $L_\alpha(\mathcal{S}_f) = L_\alpha(\mathcal{S}_0)$ for any $\alpha \in [1,2)$, where*

$$\mathbb{S}(x) \quad = \quad \sum_{j=1}^{J} \frac{1}{b_j - a_j} \left\{ Z_j + \sqrt{p_j}\, \mathrm{gren}(\mathbb{U}_j)\left(\frac{x - a_j}{b_j - a_j}\right) \right\} 1_{I_j}(x).$$

A pointwise version of Proposition 6.2 was originally proved in Carolan and Dykstra (1999). Here, we extend these results to convergence in L_α, which is a much stronger statement, requiring tight bounds on the tail behaviour at a point of the kind proved in Groeneboom et al. (1999, Theorem 2.1). In the case of the decreasing probability mass function, ℓ_k, $k \geq 1$ convergence has been established in Jankowski and Wellner (2009). An immediate corollary of this work is convergence of the linear functionals $\widehat{\mu}_n(g)$: see Corollary 6.3 below.

Groeneboom (1986, Theorem 4.1) shows that for f_0 equal to the uniform density on $[0,1]$ we have

$$\sqrt{n} \int_0^1 |\hat{f}_n(x) - f_0(x)| dx \Rightarrow \int_0^1 |\,\mathrm{gren}(\mathbb{U})(x)| dx = 2 \sup_{0 \leq x \leq 1} \mathbb{U}(x),$$

where $\mathbb{U}$ is again a standard Brownian bridge process on $[0,1]$. This is an immediate corollary of Proposition 6.2 with $J = 1$. On the other hand, Groeneboom (1983) (see also Groeneboom and Pyke (1983)) shows that

$$\frac{\int_0^1 (\sqrt{n}(\widehat{f}_n(x) - f_0(x)))^2 dx - \log n}{\sqrt{3 \log n}} \Rightarrow Z \sim N(0,1)$$

and hence convergence of $\sqrt{n}(\widehat{f}_n(x) - f_0(x))$ to $\text{gren}(\mathbb{U})(x)$ in $L_2([0,1])$ fails. See also Remark 3.2.

COROLLARY 6.3. *Suppose that f_0 takes the form* (6.1) *with bounded support $\mathcal{S}_0 = \mathcal{S}_f \cap \mathcal{W}$ and with J finite. Suppose further that g satisfies condition (G2). Then $\sqrt{n}(\widehat{\mu}_n(g) - \mu_0(g)) \Rightarrow Y_J$, where*

$$Y_J \;=\; \sum_{j=1}^{J} \left\{ \overline{g}_j \, Z_j + \sqrt{p_j} \int_0^1 g_j(u) \, \text{gren}(\mathbb{U}_j)(u) du \right\},$$

with $\overline{g}_j$ and g_j defined in (3.2).

In what follows, unless stated otherwise, we assume that $\mathcal{S}_0 = [0,1]$.

LEMMA 6.4. *Suppose that f_0 is as in* (6.1) *with a discontinuity at a point $x_0 \neq 0$. Then, for all $c > 0$,*

$$\sup_{0 \leq x \leq c/n} \left| \widehat{f}_n(x_0 + x) - f_0(x_0 + x) \right| = O_p(1).$$

PROOF. It was shown in Anevski and Hössjer (2002, Theorem 2) that

$$\widehat{f}_n(x_0 + t/n) - \frac{f_0(x_0-) + f_0(x_0+)}{2} \Rightarrow h(t), \tag{6.2}$$

where $h(t)$ is the left derivative of the least concave majorant (over $\mathbb{R}$) of the process

$$\mathbb{N}(\lambda(s)) - \lambda(s) - \left\{ \frac{f_0(x_0-) - f_0(x_0+)}{2} \right\} |s|,$$

where the rate function is equal to

$$\lambda(s) \;=\; \begin{cases} f_0(x_0+)s, & s > 0, \\ f_0(x_0-)s, & s < 0. \end{cases}$$

Here, $\mathbb{N}$ denotes a standard two-sided Poisson process. The result in Anevski and Hössjer (2002, Theorem 2) is established by a "switching" argument similar to that in the proof of Theorem 2.4. The switching argument can also be extended to this situation even if $f_0(x_0-) = f_0(x_0+)$. A similar argument may also be used to show convergence in finite dimensional distributions as well. We next show convergence of the supremum norm

$$\sup_{0 \leq x \leq c/n} \left| \widehat{f}_n(x_0 + x) - f_0(x_0 + x) \right|$$
$$\Rightarrow \quad \sup_{0 \leq x \leq c} \left| h(x) + \frac{f_0(x_0-) - f_0(x_0+)}{2} \right| = O_p(1).$$

This is done by (1) showing that the convergence in (6.2) also holds in $D[0,\infty)$, and (2) showing that this implies convergence of the supremum norm (as above). Both of these steps follow exactly the same argument as the proof of Theorem 1.1 in Balabdaoui et al. (2011), and we therefore omit the details. □

LEMMA 6.5. *Suppose that f_0 is decreasing on $\mathbb{R}$ and flat on $(a,b]$ and fix $x \in (a,b)$. Then, for any $t_0 > 0$ and $k_0 > 0$, there exists a constant $c_0 = t_0/(f_0(b) + t_0/k_0)$ such that*

$$P\left(\widehat{f}_n(x) > f_0(x) + n^{-1/2}t\right) \;\leq\; \exp\left\{-c_0\frac{t(x-a)}{2}\right\} \quad \textit{for all } t \geq t_0,$$

for all $n \geq (k_0/3)^2$. Also,

$$P\left(\widehat{f}_n(x) < f_0(x) - n^{-1/2}t\right) \;\leq\; \exp\left\{-\frac{t^2(b-x)}{2f_0(b)}\right\} \textit{ for all } t \in [0, \sqrt{n}f_0(x)],$$

and otherwise the probability is equal to zero.

PROOF. Let $\mathbb{F}_n(a,s) = \mathbb{F}_n(s) - \mathbb{F}_n(a)$, and we write $f_0(a+) = \lim_{x\to a+} f_0(x)$. By the switching relation,

$$\begin{aligned}
&P\left(\widehat{f}_n(x) > f_0(x) + n^{-1/2}t\right) \\
&\quad= P\left(\operatorname{argmax}_{s\in[0,1]}\{\mathbb{F}_n(s) - (f_0(x) + n^{-1/2}t)s\} > x\right) \\
&\quad= P\left(\operatorname{argmax}_{s\in[0,1]}\{\mathbb{F}_n(a,s) - (f_0(a+) + n^{-1/2}t)(s-a)\} > x\right) \\
&\quad\leq P\left(\mathbb{F}_n(a,s) \geq (f_0(a+) + n^{-1/2}t)(s-a)\} \text{ for some } s \in (x,1]\right) \\
&\quad= P\left(\frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq \frac{(f_0(a+) + n^{-1/2}t)(s-a)}{F_0(a,s)} \text{ for some } s \in (x,1]\right) \\
&\quad= P\left(\frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq 1 + \frac{n^{-1/2}t}{f_0(a+)} \text{ for some } s \in (x,1]\right) \\
&\quad\leq P\left(\sup_{s\in(x,1]}\frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq 1 + \frac{n^{-1/2}t}{f_0(a+)}\right)
\end{aligned}$$

Since $\mathbb{F}_n(a,s)$ is a binomial random variable, we can bound the above using Shorack and Wellner (1986, Inequality 10.3.2, page 416), with $h(v) =$

$v(\log v - 1) + 1$ and $\psi(v) = 2h(1+v)/v^2 \geq (1+v/3)^{-1}$. It therefore follows that

$$\begin{aligned} & P\left(\sup_{s\in(x,1]} \frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq 1 + \frac{n^{-1/2}t}{f_0(a+)}\right) \\ & \quad \leq \exp\left\{-nF_0(a,x)h\left(1+\frac{n^{-1/2}t}{f_0(a+)}\right)\right\} \\ & \quad = \exp\left\{-\frac{t^2(x-a)}{2f_0(a+)}\psi\left(\frac{n^{-1/2}t}{f_0(a+)}\right)\right\} \\ & \quad \leq \exp\left\{-\frac{t(x-a)}{2}\ \frac{t/f_0(a+)}{1+(t/f_0(a+))/(3\sqrt{n})}\right\}. \end{aligned}$$

Write $u = t/f_0(a_+)$ and note that for all $n \geq (k_0/3)^2$ we have

$$\frac{u}{1+u/(3\sqrt{n})} \ \geq \ \frac{u}{1+u/k_0},$$

which is a increasing function of u. Fix $t_0 > 0$ and let $u_0 = t_0/f_0(a+)$. Then, with $c_0 = u_0/(1+u_0/k_0) = t_0/(f_0(a+) + t_0/k_0)$ we have that

$$P\left(\widehat{f}_n(x) > f_0(x) + n^{-1/2}t\right) \ \leq \ \exp\left\{-c_0\ t\ \left(\frac{x-a}{2}\right)\right\}.$$

We handle the other side in a similar manner.

$$\begin{aligned} & P\left(\widehat{f}_n(x) < f_0(x) - n^{-1/2}t\right) \\ & \quad = P\left(\text{argmax}_{s\in[0,1]}\{\mathbb{F}_n(s,b) - (f_0(b) - n^{-1/2}t)(x-b)\} < x\right) \\ & \quad \leq P\left(\frac{\mathbb{F}_n(s,b)}{F_0(s,b)} \leq 1 - \frac{n^{-1/2}t}{f_0(b)} \text{ for some } s \in [0,x)\right) \\ & \quad \leq P\left(\inf_{s\in[0,x)} \frac{\mathbb{F}_n(s,b)}{F_0(s,b)} \leq 1 - \frac{n^{-1/2}t}{f_0(b)}\right). \end{aligned}$$

We now bound this using the martingale inequality from Groeneboom et al. (1999, Lemma 2.3).

$$\begin{aligned} P\left(\inf_{s\in[0,x)} \frac{\mathbb{F}_n(s,b)}{F_0(s,b)} \leq 1 - \frac{n^{-1/2}t}{f_0(b)}\right) & \leq \exp\left\{-nF(x,b)h\left(1-\frac{n^{-1/2}t}{f_0(b)}\right)\right\} \\ & = \exp\left\{-\frac{t^2(b-x)}{2f_0(b)}\ \psi\left(-\frac{n^{-1/2}t}{f_0(b)}\right)\right\}. \end{aligned}$$

Now, note that since $\widehat{f}_n$ is a density, we only consider $t \leq \sqrt{n} f_0(b) = \sqrt{n} f_0(x)$. Therefore, we bound only $\psi(-v)$ for $v \in [0,1]$, for which we have that $\psi(-v) \geq 1$. Thus it follows that

$$P\left(\widehat{f}_n(x) < f_0(x) - n^{-1/2}t\right) \leq \exp\left\{-\frac{t^2(b-x)}{2f_0(b)}\right\}.$$

□

Let $(x)_+ = \max(x,0)$ and $(x)_- = \min(x,0)$.

LEMMA 6.6. *Suppose that f_0 is flat on $(a,b]$ and fix $x \in (a,b)$, and fix $\alpha > 0$. Then, there exists a constant C such that*

$$\begin{aligned} E\left[\left|\sqrt{n}(\widehat{f}_n(x) - f_0(x))_-\right|^\alpha\right] &\leq C(b-x)^{-\alpha/2}, \\ E\left[\left|\sqrt{n}(\widehat{f}_n(x) - f_0(x))_+\right|^\alpha\right] &\leq C(x-a)^{-\alpha/2}, \end{aligned}$$

with the second bound valid only for $(x-a) \geq \widetilde{c}_0/n$, for some $\widetilde{c}_0 > 0$.

PROOF. Using the bounds obtained in Lemma 6.5, we find that

$$\begin{aligned} E\left[\left|\sqrt{n}(\widehat{f}_n(x) - f_0(x))_-\right|^\alpha\right] &= \int_0^\infty \alpha t^{\alpha-1} P(\sqrt{n}(\widehat{f}_n(x) - f_0(x))_- > t)dt \\ &= \int_0^{n^{1/2}f_0(x)} \alpha t^{\alpha-1} P(\widehat{f}_n(x) < f_0(x) - n^{-1/2}t)dt \\ &\leq \int_0^\infty \alpha t^{\alpha-1} \exp\left\{-\frac{t^2(b-x)}{2f_0(b)}\right\} dt \\ &= \Gamma(1+\alpha/2)\left(\frac{2f_0(b)}{b-x}\right)^{\alpha/2}. \end{aligned}$$

For the second inequality, we first fix $t_0 > 0$. We then have

$$\begin{aligned} E\left[\left|\sqrt{n}(\widehat{f}_n(x) - f_0(x))_+\right|^\alpha\right] &= \int_0^\infty \alpha t^{\alpha-1} P(\sqrt{n}(\widehat{f}_n(x) - f_0(x))_+ > t)dt \\ &= \int_0^{t_0} \alpha t^{\alpha-1} P(\widehat{f}_n(x) > f_0(x) + n^{-1/2}t)dt \\ &\quad + \int_{t_0}^\infty \alpha t^{\alpha-1} P(\widehat{f}_n(x) > f_0(x) + n^{-1/2}t)dt \\ &\leq t_0^\alpha + \int_{t_0}^\infty \alpha t^{\alpha-1} \exp\left\{-c_0\frac{t(x-a)}{2}\right\} dt \\ &\leq t_0^\alpha + \Gamma(\alpha+1)\left(\frac{2}{c_0(x-a)}\right)^\alpha. \end{aligned}$$

Now, recall that c_0 takes the form $t_0/(f_0(b) + t_0/k_0)$. Therefore, we obtain the bounds

$$\begin{aligned}\Gamma(\alpha+1)\left(\frac{2}{c_0(x-a)}\right)^{\alpha} &\leq 2^{\alpha}\Gamma(\alpha+1)(x-a)^{-\alpha}\left(\frac{f_0(b)+t_0/k_0}{t_0}\right)^{\alpha} \\ &\leq C_{\alpha}(1+K)^{\alpha}\left(\frac{f_0(b)}{x-a}\right)^{\alpha} t_0^{-\alpha},\end{aligned}$$

as long as $t_0/k_0 \leq K f_0(b)$ for some choice of K. We optimize the entire quantity in t_0 to find that

$$E\left[\left|\sqrt{n}(\widehat{f}_n(x) - f_0(x))_+\right|^{\alpha}\right] \leq A_{\alpha}\left(\frac{f_0(b)}{x-a}\right)^{\alpha/2},$$

for some new constant A_α. Now, in order for this optimized bound to hold, we need $t_0 \leq K f_0(b) k_0$, and

$$K^2 f_0(b)^2 k_0^2 \geq C_{\alpha}(1+K)^{\alpha}\left(\frac{f_0(b)}{x-a}\right)^{\alpha}.$$

The latter translates to $(x-a) \geq \widetilde{c}_0 n^{-1}$ by using $k_0^2 \leq 9n$. □

Proof of Proposition 6.2. The outline of the proof is as follows. We first require pointwise convergence, which follows from Carolan and Dykstra (1999, Theorem 6.4). One can also easily extend this to convergence in finite dimensional distributions. The particular form of the limit follows from the following decomposition of a (time-transformed) Brownian bridge, which is a generalization of Shorack and Wellner (1986, Exercise 2.2.11, page 32). Let F denote any distribution function with compact support, which, without loss of generality, we assume to be $[0,1]$. Let $0 = a_1 < b_1 = a_2 < \ldots < b_{J-1} = a_J < b_J = 1$. Let $\mathbb{V}, \mathbb{U}_1, \ldots, \mathbb{U}_J$ denote independent Brownian bridges. Then

$$\begin{aligned}\mathbb{V}(F(t)) \equiv \sum_{i=1}^{J}\Bigg\{&\sum_{j=1}^{i-1}\Delta\mathbb{V}(F(a_i)) + \Delta\mathbb{V}(F(a_i))\frac{F(t)-F(a_i)}{F(b_i)-F(a_i)} \\ &+\sqrt{F(b_i)-F(a_i)}\,\mathbb{U}_i\left(\frac{F(t)-F(a_i)}{F(b_i)-F(a_i)}\right)\Bigg\} 1_{(a_i,b_i]}(t) \qquad (6.3)\end{aligned}$$

where $\Delta\mathbb{V}(F(a_j)) = \mathbb{V}(F(b_j)) - \mathbb{V}(F(a_j))$.

Recall that the Grenander operator satisfies $\text{gren}(a + bt + ch(t)) = b + c\,\text{gren}(h(t))$. Also note that F_0 is linear on $(a_i, b_i]$ by assumption. Therefore,

from Carolan and Dykstra (1999), the limit of $\sqrt{n}(\widehat{f}_n(x) - f_0(x))$ at a point $x \in I_i = (a_i, b_i]$ can be written as

$$\begin{aligned}\operatorname{gren}_{(a_i,b_i]}(\mathbb{V}(F_0(t))) &= \Delta\mathbb{V}(F(a_i))\frac{1}{b_i - a_i} + \sqrt{p_i}\operatorname{gren}\left(\mathbb{U}_i\left(\frac{t-a_i}{b_i-a_i}\right)\right) \\ &= \frac{1}{b_i - a_i}\left\{\Delta\mathbb{V}(F(a_i)) + \sqrt{p_i}\operatorname{gren}(\mathbb{U}_i)\left(\frac{t-a_i}{b_i-a_i}\right)\right\},\end{aligned}$$

from the above characterization. Finally $\{\Delta\mathbb{V}(F(a_1)), \ldots, \Delta\mathbb{V}(F(a_J))\} \stackrel{d}{=} \{Z_1, \ldots, Z_J\}$ as in Proposition 6.2.

The second step is to show that the process $\mathbb{S}_n(x) = \sqrt{n}(\widehat{f}_n(x) - f_0(x))$ is tight in $L_\alpha(\mathcal{S})$. For this, we first need a characterization of compact sets in $L_\alpha(\mathcal{S})$ for $\alpha \geq 1$. These appear, for example in Dunford and Schwartz (1958, page 298) (see also Simon (1987)). For $\mathcal{S}$ bounded, a set $\mathcal{K} \subset L_\alpha(\mathcal{S})$ is relatively compact if for all $f \in \mathcal{K}$

1. $\sup_{f\in\mathcal{K}} \int_{\mathcal{S}} |f(x)|^\alpha dx < \infty$,
2. $\lim_{\delta\to 0} \sup_{f\in\mathcal{K}} \int_{\mathcal{S}} |f(x+\delta) - f(x)|^\alpha dx \to 0$.

We want to show that for each $\epsilon > 0$ we can find a compact subset $\mathcal{K} = \mathcal{K}_\epsilon$ of $L_\alpha(\mathcal{S})$ such that $\limsup_n P(\mathbb{S}_n \in \mathcal{K}^c) < \epsilon$. Thus we want to show that

$$\limsup_n P\left(\int_0^1 |\mathbb{S}_n(x)|^\alpha dx > M\right) \to 0 \text{ as } M \to \infty, \text{ and} \tag{6.4}$$

$$\limsup_n P\left(\int_0^1 |\mathbb{S}_n(x+\delta) - \mathbb{S}_n(x)|^\alpha dx > \epsilon\right) \to 0 \tag{6.5}$$

as $\delta \to 0$, for every $\epsilon > 0$.

To show the first of these we proceed as follows: for f_0 as in (6.1),

$$\int_0^1 |\mathbb{S}_n(x)|^\alpha dx = \sum_{j=1}^J \int_{(a_j,b_j]} |\mathbb{S}_n(x)|^\alpha dx,$$

and hence we have

$$P\left(\int_0^1 |\mathbb{S}_n(x)|^\alpha dx > M\right) \leq \sum_{j=1}^J P\left(\int_{(a_j,b_j]} |\mathbb{S}_n(x)|^\alpha dx > M/J\right). \tag{6.6}$$

Thus, it suffices to show that

$$\limsup_n P\left(\int_{(a,b]} |\mathbb{S}_n(x)|^\alpha dx > M\right) \to 0$$

as $M \to \infty$ for each fixed $(a, b]$ with f_0 flat on $(a, b]$. Now,

$$\begin{aligned}P\left(\int_{(a,b]} |\mathbb{S}_n(x)|^\alpha dx > M\right) &\leq P\left(\int_{(a,a+\tilde{c}_0/n]} |\mathbb{S}_n(x)|^\alpha dx > M/2\right) \\ &\quad + P\left(\int_{(a+\tilde{c}_0/n,b]} |\mathbb{S}_n(x)|^\alpha dx > M/2\right),\end{aligned} \tag{6.7}$$

and we handle each term separately. From Lemma 6.4, it follows that

$$\begin{aligned}\int_{(a,a+\tilde{c}_0/n]} |\mathbb{S}_n(x)|^\alpha dx &= \int_{(a,a+\tilde{c}_0/n]} n^{\alpha/2} |\widehat{f}_n(x) - f_0(x)|^\alpha dx \\ &= n^{\alpha/2}\tilde{c}_0 n^{-1} O_p(1) = o_p(1) \qquad (6.8)\end{aligned}$$

for $\alpha < 2$. For the second term, we use Markov's inequality, Lemma 6.6, and Fubini's theorem to get

$$\begin{aligned}&P\left(\int_{(a+\tilde{c}_0/n,b]} |\mathbb{S}_n(x)|^\alpha dx > M/2\right) \\ &\quad\le \frac{2}{M} 2^{\alpha-1} \left\{ \int_{(a+\tilde{c}_0/n,b]} E|\mathbb{S}_n(x)_+|^\alpha \, dx + \int_{(a+\tilde{c}_0/n,b]} E|\mathbb{S}_n(x)_-|^\alpha \, dx \right\} \\ &\quad\le \frac{2^\alpha C}{M} \left\{ \int_{(a,b]} (x-a)^{-\alpha/2} dx + \int_{(a,b]} (b-x)^{-\alpha/2} dx \right\} \le \widetilde{C}/M, \qquad (6.9)\end{aligned}$$

for some new, finite, constant $\widetilde{C}$ depending on a, b, α, noting that $\alpha < 2$. Combining (6.8) and (6.9) yields (6.4) for our choice of f_0.

Now, to prove (6.5). Since $f_0(x)$ is constant for $x \in (a_j, b_j]$ for each j, the processes $\mathbb{S}_n(x) = \sqrt{n}(\widehat{f}_n(x) - f_0(x))$, are piecewise monotone, and hence the convergence in $L_\alpha((a_j, b_j])$ for $\alpha \in [1, 2)$ and each $j \le J$ follows as in Huang and Zhang (1994, Corollary 2, page 1260). We conclude that (6.5) holds, and hence $\mathbb{S}_n$ is tight in $L_\alpha(\mathcal{S})$ when $\alpha < 2$. □

Proof of Corollary 6.3. Convergence follows immediately by continuity of the linear functional $\int g(x)\mathbb{S}_n(x)dx$ by Hölder's inequality. We need only check the final form, that is, $\int g(x)\mathbb{S}(x)dx$ is equal to

$$\begin{aligned}&\sum_{i=1}^J \left\{ Z_i \frac{\int_{a_i}^{b_i} g(x)dx}{b_i - a_i} + \sqrt{p_i} \int_{a_i}^{b_i} \frac{g(x)}{b_i - a_i} \operatorname{gren}(\mathbb{U}_i)\left(\frac{x - a_i}{b_i - a_i}\right) dx \right\} \\ &\quad= \sum_{j=1}^J \left\{ Z_i \frac{\int_{a_i}^{b_i} g(x)dx}{b_i - a_i} + \sqrt{p_i} \int_0^1 g((b_i - a_i)u + a_i) \operatorname{gren}(\mathbb{U}_j)(u)\, du \right\}.\end{aligned}$$

□

6.3. *Piecewise constant KL density.* We next consider the case that $\widehat{f}_0(x)$ can be written in the form (3.1) with condition (F). Let $\mathbb{U}_1, \ldots, \mathbb{U}_J$, denote independent standard Brownian bridge processes (each defined on $[0, 1]$),

and for each j define $\mathbb{U}_j^{mod}$ as in Remark 2.5 with $I_j = [a_j, b_j]$ replacing $[a, b]$. Also, let $\{Z_1, \ldots, Z_J\}$ be an independent multivariate normal with mean zero and covariance $\mathrm{diag}(p) - pp^T$ for $p = (p_1, \ldots, p_J)^T$, where $p_j = F_0(b_j) - F_0(a_j) = \widehat{F}_0(b_j) - \widehat{F}_0(a_j)$.

PROPOSITION 6.7. *Suppose that $\widehat{f}_0$ satisfies conditions (P) and (F) with $\mathcal{S}_0 = \mathcal{S}_f$ and that g satisfies condition (G2). Then $\sqrt{n}(\widehat{f}_n(x) - f_0(x))$ converges weakly to $\mathbb{S}^{mod}(x)$ in $L_\alpha(\mathcal{S})$ for $\alpha \in (0, 2)$, where*

$$\mathbb{S}^{mod}(x) \;=\; \sum_{j=1}^{J} \frac{1}{b_j - a_j} \left\{ Z_j + \sqrt{p_j}\, \mathrm{gren}(\mathbb{U}_j^{mod}) \left(\frac{x - a_j}{b_j - a_j} \right) \right\} 1_{(a_j, b_j]}(x).$$

COROLLARY 6.8. *Suppose that $\widehat{f}_0$ satisfies conditions (P) and (F) with $\mathcal{S}_0 = \mathcal{S}_f$ and that g satisfies condition (G2). Then $\sqrt{n}(\widehat{\mu}_n(g) - \widehat{\mu}_0(g)) \Rightarrow Y_J$, where*

$$Y_J^{mod} \;=\; \sum_{j=1}^{J} \left\{ \overline{g}_j\, Z_j + \sqrt{p_j} \int_0^1 g_j(u)\, \mathrm{gren}(\mathbb{U}_j^{mod})(u) du \right\}.$$

with $g_j(u)$ and $\overline{g}_j$ defined as in (3.2).

The proof of these results is very close to that of Proposition 6.2, and we omit any details which are the same. The difference lies in the following modifications to Lemmas 6.4 and 6.5. Note that we add the additional requirement that f_0 be bounded below (P).

LEMMA 6.9. *Fix a point $x \in \mathcal{S}_0$ and let $[a, b]$ denote the largest interval I such that $x \in I$ and $\widehat{f}_0$ is constant on I. Then, for all $c > 0$,*

$$\sup_{0 \le u \le c/n} \left| \widehat{f}_n(a + u) - \widehat{f}_0(a + u) \right| = O_p(1).$$

PROOF. By the switching relation, it follows that

$$\begin{aligned} & P\left(\widehat{f}_n(a + u/n) - \widehat{f}_0(a + u/n) < t \right) \\ &\quad = P\left(\widehat{f}_n(a + u/n) - \widehat{f}_0(b) < t \right) \\ &\quad = P\left(\mathrm{argmax}_{z \in [0,1]} \left\{ \mathbb{F}_n(z) - (\widehat{f}_0(b) + t) z \right\} < a + u/n \right) \\ &\quad = P\left(n(\mathrm{argmax}_{z \in [0,1]} \left\{ \mathbb{F}_n(z) - (\widehat{f}_0(b) + t) z \right\} - a) < u \right), \end{aligned}$$

and the inner process

$$\begin{aligned} & n\left(\text{argmax}_{z\in[0,1]}\left\{\mathbb{F}_n(z)-(\widehat{f}_0(b)+t)z\right\}-a\right) \\ &= \text{argmax}_{h\geq -na}\left\{\mathbb{F}_n(a+h/n)-(t+\widehat{f}_0(b))(a+h/n)\right\} \\ &= \text{argmax}_{h\geq -na}\left\{\mathbb{V}_n(h)\right\}, \end{aligned}$$

where $\mathbb{V}_n(h) = A_n(h) + B_n(h) - th$, with $A_n(h) = n(\mathbb{F}_n(a+h/n) - \mathbb{F}_n(a)) - n(F_0(a+h/n) - F_0(a))$ and $B_n(h) = n(F_0(a+h/n) - F_0(a)) - \widehat{f}_0(b)h$.

Now, the term $\mathbb{N}_n(h) = n(\mathbb{F}_n(a+h/n) - \mathbb{F}_n(a))$ is binomial with mean $n(F_0(a+h/n) - F_0(a)) \to f_0(a)h$. Therefore, $A_n(h)$ converges to a centered Poisson random variable with mean $f_0(a)$. A similar argument may be used to show convergence as a process of $A_n(h) \Rightarrow \mathbb{N}(h) - f_0(a)h$, where $\mathbb{N}(\cdot)$ is a Poisson process with rate $\lambda(h) = f_0(a)$. The second piece, $B_n(h)$ satisfies

$$n^{-1}B_n(h) \qquad \begin{cases} = 0 & h \in n\{[a,b]\cap\mathcal{W}-a\} \\ < 0 & h \in n\{[a,b]\cap\mathcal{M}-a\} \end{cases}.$$

Thus, if for all $\delta > 0$ $[a,\delta)\cap\mathcal{W} = \{a\}$ then the limit of $B_n(h)$ is 0 if $h = 0$ and is equal to $-\infty$ otherwise (we will call this setting case (A)). If the above assumption is not true (we will call this setting case (B)), then $\lim_{n\to\infty} B_n(h) = 0$ for all $h \geq 0$. In case (A), it follows that the limit of $\mathbb{V}_n(h)$ is equal to 0 at $h = 0$ and is equal to $-\infty$ otherwise. Therefore, $\text{argmax}_{h\geq 0}\{\mathbb{V}_n(h)\} = 0$ here. In case (B) the limit of $\mathbb{V}_n(h)$ is a centered (a.k.a. compensated) Poisson process with rate $f_0(a)$. We therefore have that, in case (A),

$$\begin{aligned} & P\left(\widehat{f}_n(a+u/n) - \widehat{f}_0(a+u/n) < t\right) \\ &\qquad = P\left(\text{argmax}_{h\geq -na}\left\{\mathbb{V}_n(h)\right\} < u\right) \to 1, \end{aligned}$$

and in case (B),

$$\begin{aligned} & P\left(\widehat{f}_n(a+u/n) - \widehat{f}_0(a+u/n) < t\right) = P\left(\text{argmax}_{h\geq -na}\left\{\mathbb{V}_n(h)\right\} < u\right) \\ & \to P\left(\text{argmax}_{h\geq 0}\left\{\mathbb{N}(h) - f_0(a)h - th\right\} < u\right), \end{aligned}$$

which gives us pointwise convergence in distribution in both cases.

Lastly, note that $\widehat{f}_0(a+u) = \widehat{f}_0(b)$ is a constant, and $\widehat{f}_n(a+u)$ is decreasing in u by definition. Therefore, $\sup_{0\leq u\leq c/n}|\widehat{f}_n(a+u) - \widehat{f}_0(a+u)| = |\widehat{f}_n(a) - \widehat{f}_0(b)|$, which converges as described above. □

LEMMA 6.10. *Suppose that $\widehat{f}_0$ is flat on $(a,b]$ and fix $x \in (a,b)$. Assume also that $\inf_{x\in(a,b]} f_0(x) = \alpha_0 > 0$, and let $\widehat{c}_0 = \alpha_0/\widehat{f}_0(b)$. Then, for any $t_0 > 0$ and $k_0 > 0$, there exists a constant $c_0 = t_0/(\widehat{f}_0(b) + t_0/k_0)$ such that*

$$P\left(\widehat{f}_n(x) > \widehat{f}_0(x) + n^{-1/2}t\right) \;\leq\; \exp\left\{-\widehat{c}_0\; c_0 \frac{t(x-x_0)}{2}\right\} \quad \textit{for all } t \geq t_0,$$

for all $n \geq (k_0/3)^2$. Also, for all $t \in [0, \sqrt{n}\widehat{f}_0(x)]$,

$$P\left(\widehat{f}_n(x) < \widehat{f}_0(x) - n^{-1/2}t\right) \;\leq\; \exp\left\{-\widehat{c}_0 \frac{t^2(b-x)}{2\widehat{f}_0(x)}\right\},$$

and otherwise the probability is equal to zero.

PROOF. Let $\mathbb{F}_n(a,s) = \mathbb{F}_n(s) - \mathbb{F}_n(a)$, and we write $\widehat{\theta} = \widehat{f}_0(x)$. Repeating the argument for the proof of Lemma 6.5 we obtain that

$$\begin{aligned}
&P\left(\widehat{f}_n(x) > \widehat{f}_0(x) + n^{-1/2}t\right) \\
&\quad\leq\; P\left(\frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq \frac{(\widehat{\theta} + n^{-1/2}t)(s-a)}{F_0(a,s)} \text{ for some } s \in (x,1]\right) \\
&\quad\leq\; P\left(\sup_{s\in(x,1]} \frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq \inf_{s\in(x,1]} \frac{(\widehat{\theta} + n^{-1/2}t)(s-a)}{F_0(a,s)}\right) \\
&\quad\leq\; P\left(\sup_{s\in(x,1]} \frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq \inf_{s\in(x,1]} \frac{(\widehat{\theta} + n^{-1/2}t)(s-a)}{\widehat{F}_0(a,s)}\right) \\
&\quad=\; P\left(\sup_{s\in(x,1]} \frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq 1 + \frac{n^{-1/2}t}{\widehat{\theta}}\right)
\end{aligned}$$

since $\widehat{F}_0(s) > F_0(s)$ with equality at $s = a, b$. Applying the exponential bounds for binomial variables as before, we find that

$$\begin{aligned}
&P\left(\sup_{s\in(x,1]} \frac{\mathbb{F}_n(a,s)}{F_0(a,s)} \geq 1 + \frac{n^{-1/2}t}{\widehat{\theta}}\right) \\
&\quad\leq\; \exp\left\{-nF_0(a,x)h\left(1 + \frac{n^{-1/2}t}{\widehat{\theta}}\right)\right\} \\
&\quad\leq\; \exp\left\{-\left[\inf_{x\in[a,b]} \frac{f_0(x)}{\widehat{\theta}}\right] \frac{t(x-a)}{2} \frac{t/\widehat{\theta}}{1 + (t/\widehat{\theta})/(3\sqrt{n})}\right\}.
\end{aligned}$$

Therefore, assuming that $\inf_{x\in[a,b]} \frac{f_0(x)}{\widehat{\theta}} = \widehat{c}_0 > 0$, we can repeat the same argument as for Lemma (6.5).

We handle the other side in a similar manner.

$$\begin{aligned} & P\left(\widehat{f}_n(x) < \widehat{f}_0(x) - n^{-1/2}t\right) \\ & \quad \leq \quad P\left(\inf_{s\in[0,x)} \frac{\mathbb{F}_n(s,b)}{F_0(s,b)} \leq \sup_{s\in[0,x)} \frac{(\widehat{\theta} - n^{-1/2}t)(s-b)}{F_0(s,b)}\right) \\ & \quad \leq \quad P\left(\inf_{s\in[0,x)} \frac{\mathbb{F}_n(s,b)}{F_0(s,b)} \leq 1 - \frac{n^{-1/2}t}{\widehat{\theta}}\right) \end{aligned}$$

We again bound this using the martingale inequality from Groeneboom et al. (1999, Lemma 2.3).

$$\begin{aligned} & P\left(\inf_{s\in[0,x)} \frac{\mathbb{F}_n(s,b)}{F_0(s,b)} \leq 1 - \frac{n^{-1/2}t}{\widehat{\theta}}\right) \\ & \quad \leq \quad \exp\left\{-nF(x,b)h\left(1 - \frac{n^{-1/2}t}{\widehat{\theta}}\right)\right\} \\ & \quad = \quad \exp\left\{-\left[\inf_{x\in[a,b]} \frac{f_0(x)}{\widehat{\theta}}\right] \frac{t^2(b-x)}{2\widehat{\theta}} \, \psi\left(-\frac{n^{-1/2}t}{\widehat{\theta}}\right)\right\}. \end{aligned}$$

□

6.4. *Putting it all together.*

Proof of Theorem 3.1. To illustrate the method of proof, we consider a simplified case. Since $g \in L_\beta(\mathcal{S}_f)$ for some $\beta > 2$ and $\mathbb{S}_n$ converges in $L_\alpha(\mathcal{S})$ the proof easily extends to a general setting. Suppose then that $\mathcal{S}_c = [0, a]$ and $\mathcal{S}_f = [a, b]$, so that the support is $\mathcal{S}_0 = [0, b]$. Furthermore, we assume that on $\mathcal{S}_f$ we have $J = 1$. Let $\mathbb{U}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F_0(x))$. Then

$$\begin{aligned} & \sqrt{n}(\widehat{\mu}_n(g) - \widehat{\mu}_0(g)) \\ & \quad = \quad \int_0^a g(x)d\mathbb{U}_n(x) + \int_a^b g(x)\sqrt{n}\left(\text{gren}(\mathbb{F}_n)(x) - \widehat{f}_0(x)\right)dx + \varepsilon_n, \end{aligned}$$

where $\varepsilon_n = \sqrt{n}\int_0^a g(x)d(\widehat{F}_n - \mathbb{F}_n)(x)$. From assumptions (C) and (G1) it follows that $\varepsilon_n = o_p(1)$ as in Proposition 6.1.

Next, let $\mathbb{W}_n = \mathbb{U}_n(b) - \mathbb{U}_n(a)$, and let $\mathbb{V}_n(x) = \mathbb{U}_n(x) - \mathbb{U}_n(a) - \frac{x-a}{b-a}\mathbb{W}_n$ for $x \in [a,b]$. Lastly, let $\ell(x) = F_0(a) + \widehat{f}_0(b)(x-a)$. Then for $x \in (a,b]$,

$$\begin{aligned}\sqrt{n}\left(\operatorname{gren}(\mathbb{F}_n)(x) - \widehat{f}_0(x)\right) &= \sqrt{n}\left(\operatorname{gren}(\mathbb{F}_n)(x) - \widehat{f}_0(b)\right) \\ &= \operatorname{gren}(\mathbb{U}_n + \sqrt{n}(F_0 - \ell)) \\ &= \frac{1}{b-a}\mathbb{W}_n + \operatorname{gren}(\mathbb{V}_n + \sqrt{n}(F_0 - \ell)),\end{aligned}$$

and we also define $\mathbb{V}_n^{mod} = \mathbb{V}_n + \sqrt{n}(F_0 - \ell)$. Therefore, $\sqrt{n}(\widehat{\mu}_n(g) - \widehat{\mu}_0(g))$ is equal to

$$\begin{aligned}&\int_0^a g(x)d\mathbb{U}_n(x) + \mathbb{W}_n\frac{1}{b-a}\int_a^b g(x) + \int_a^b g(x)\operatorname{gren}(\mathbb{V}_n^{mod})(x)dx + o_p(1) \\ &= \int_0^b \overline{g}(x)d\mathbb{U}_n(x) + \int_a^b g(x)\operatorname{gren}(\mathbb{V}_n^{mod})(x)dx + o_p(1),\end{aligned}$$

from the definition of $\mathbb{W}_n$ and of $\overline{g}$. The weak limit of $\mathbb{V}_n^{mod}$ can be established similarly as in Theorem 2.4 and Remark 2.5. The outline of the rest of the proof proceeds as follows:

1. Joint weak convergence of $\{\int_0^b \bar{g}d\mathbb{U}_n, \mathbb{V}_n^{mod}(x_1), \ldots, \mathbb{V}_n^{mod}(x_k)\}$ to a Gaussian limit.
2. Joint weak convergence of $\{\int_0^b \bar{g}d\mathbb{U}_n, \operatorname{gren}(\mathbb{V}_n^{mod})(x_1), \ldots, \operatorname{gren}(\mathbb{V}_n^{mod})(x_k)\}$ via the switching relation.
3. We have that
$$\operatorname{gren}(\mathbb{V}_n^{mod})(x) = \sqrt{n}\left(\widehat{f}_n(x) - \widehat{f}_0(x)\right) - \frac{1}{b-a}\mathbb{W}_n,$$
where in Proposition 6.7 we showed that the first term on the right hand side is tight in $L_\alpha(a,b)$. The second term on the right hand side is a tight constant, and therefore $\operatorname{gren}(\mathbb{V}_n^{mod})(x)$ is also tight in $L_\alpha(a,b)$.
4. From (1) and (3) we obtain marginal tightness of the terms $\int_0^b \bar{g}d\mathbb{U}_n$ in $\mathbb{R}$ and $\operatorname{gren}(\mathbb{V}_n^{mod})(\cdot)$ in $L_\alpha(a,b)$, which implies joint tightness in $\mathbb{R}\times L_\alpha$. The full result now follows by the continuous mapping theorem.

Lastly, we note that since $F_0(z) = \widehat{F}_0(z)$ at $z = a, b$ and $\overline{g}$ is constant on $[a,b]$ then $\int_{\mathcal{S}_0} \overline{g}(x)\, d\mathbb{U}_{F_0}(x) = \int_{\mathcal{S}_0} \overline{g}(x)\, d\mathbb{U}_{\widehat{F}_0}(x)$. □

Acknowledgements. The author thanks Valentin Patilea for sharing a copy of his thesis, Takumi Saegusa for pointing out a small error in one of the proofs, and the referees for a number of helpful suggestions. Parts of this work were completed while the author was visiting the University of

Washington and the University of Heidelberg, and the author thanks both institutions for their hospitality and financial travel support. The author also thanks Jon Wellner for generous contributions to this work.

References.

Anevski, D. and Hössjer, O. (2002). Monotone regression and density function estimation at a point of discontinuity. *J. Nonparametr. Stat.* **14** 279–294.

Balabdaoui, F., Jankowski, H., Pavlides, M., Seregin, A. and Wellner, J. A. (2011). On the Grenander estimator at zero. *Statistica Sinica* **21** 873–899.

Balabdaoui, F., Jankowski, H., Rufibach, K. and Pavlides, M. (2013). Maximum likelihood estimation and confidence bands for a discrete log-concave distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75** 769–790.

Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.* **37** 1299–1331.

Beirlant, J., Dudewicz, E. J., Györfi, L. and van der Meulen, E. C. (1997). Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.* **6** 17–39.

Birgé, L. (1987). On the risk of histograms for estimating decreasing densities. *Ann. Statist.* **15** 1013–1022.

Carolan, C. and Dykstra, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *Canad. J. Statist.* **27** 557–566.

Carolan, C. and Dykstra, R. (2001). Marginal densities of the least concave majorant of Brownian motion. *Ann. Statist.* **29** 1732–1750.

Cator, E. (2011). Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli* **17** 714–735.

Chen, Y. and Samworth, R. J. (2013). Smoothed log-concave maximum likelihood estimation with applications. *Statistica Sinica* **23** 1373–1398.

Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.* **4** 254–270.

Cule, M., Samworth, R. and Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 545–607.

Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730.

Dunford, N. and Schwartz, J. T. (1958). *Linear Operators. I. General Theory.* With the assistance of W. G. Bade and R. G. Bartle. Pure and Applied Mathematics, Vol. 7, Interscience Publishers, Inc., New York.

Durot, C., Kulikov, V. and Lopuhaä, H. (2012). The limit distribution of the L_∞-error of Grenander-type estimators. *Annals of Statistics* **40** 1578–1608.

Durot, C. and Lopuhaä, H. (2013). A Kiefer-Wolfowitz type of result in a general setting, with an application to smooth monotone estimation .

Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153 (1957).

Groeneboom, P. (1983). The concave majorant of Brownian motion. *Ann. Probab.* **11** 1016–1027.

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*. Wadsworth Statist./Probab. Ser., Wadsworth, Belmont, CA.

Groeneboom, P. (1986). Some current developments in density estimation. In *Mathemat-*

ics and Computer Science (Amsterdam, 1983), vol. 1 of *CWI Monogr.* North-Holland, Amsterdam, 163–192.

GROENEBOOM, P., HOOGHIEMSTRA, G. and LOPUHAÄ, H. P. (1999). Asymptotic normality of the L_1 error of the Grenander estimator. *Ann. Statist.* **27** 1316–1347.

GROENEBOOM, P. and PYKE, R. (1983). Asymptotic normality of statistics based on the convex minorants of empirical distribution functions. *Ann. Probab.* **11** 328–345.

HUANG, Y. and ZHANG, C.-H. (1994). Estimating a monotone density from censored observations. *Ann. Statist.* **22** 1256–1274.

JANKOWSKI, H. K. and WELLNER, J. A. (2009). Estimation of a discrete monotone distribution. *Electron. J. Stat.* **3** 1567–1605.

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

KIEFER, J. and WOLFOWITZ, J. (1977). Asymptotically minimax estimation of concave and convex distribution functions. II. In *Statistical Decision Theory and Related Topics. II (Proc. Sympos., Purdue Univ., Lafayette, Ind., 1976)*. Academic Press, New York, 193–211.

KULIKOV, V. N. and LOPUHAÄ, H. P. (2008). Distribution of global measures of deviation between the empirical distribution function and its concave majorant. *J. Theoret. Probab.* **21** 356–377.

LE CAM, L. (1960). Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. *Univ. California Publ. Statist.* **3** 37–98.

MARSHALL, A. W. (1970). Discussion on Barlow and van Zwet's paper. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*. Cambridge Univ. Press, London, 174–176.

PATILEA, V. (1997). *Convex models, NPMLE and misspecification.* Ph.D. thesis, Univ. catholique de Louvain.

PATILEA, V. (2001). Convex models, MLE and misspecification. *Ann. Statist.* **29** 94–123.

PRAKASA RAO, B. L. S. (1969). Estimation of a unimodal density. *Sankhya Series A* **31** 23–36.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York.

SIMON, J. (1987). Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl. (4)* **146** 65–96.

VAN DE GEER, S. (2003). Asymptotic theory for maximum likelihood in nonparametric mixture models. *Comput. Statist. Data Anal.* **41** 453–464.

VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer Series in Statistics, Springer-Verlag, New York.

WASSERMAN, L. (2006). *All of Nonparametric Statistics.* Springer Texts in Statistics, Springer, New York.

DEPARTMENT OF MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ON, CANADA
E-MAIL: hkj@yorku.ca

Supplementary Material for "Convergence of linear functionals of the Grenander estimator under misspecification"

Hanna Jankowski*

Department of Mathematics and Statistics
York University
Toronto, ON, Canada
e-mail: hkj@yorku.ca

Abstract: In this supplement we present some additional proofs, and discuss further the assumptions of the main manuscript.

Additional Proofs

Proof of Proposition 2.3. **(1).** Let ℓ denote a general linear function. Then $\widehat{F}_0(x_0) = \min\{\ell(x_0) : \ell \geq F_0\}$. Now consider, $\ell(x)$ the (specific) linear function which passes through the points $(a_0, F_0(a_0)), (b_0, F_0(b_0))$. Then clearly, there is no linear function which always lies above F_0 and strictly below ℓ on (a_0, b_0). It follows that $\widehat{F}_0(x) = \ell(x)$ on (a_0, b_0).

Next, suppose that $F_0(a_0) < 1$, with $\widehat{F}_0(a_0) = F_0(a_0)$ but with $\widehat{F}_0(x) > F_0(x)$ for all $x > a_0$. Since F_0 is increasing, it follows that any straight line $\ell(x)$ such that $\ell(a_0) = F_0(a_0)$ and $\ell(x) \geq F_0(x)$ has the form $\ell(x) = F_0(a_0) + m(x - a_0)$ with $m \geq \delta$ for some $\delta > 0$. But then $\widehat{F}_0(x) \geq F_0(x) + \delta(x - a_0)$, which implies that $\lim_{x\to\infty} \widehat{F}_0(x) = \infty > 1$, a contradiction to the fact that $\widehat{f}_0$ is a density.

The last statement follows immediately since $\widehat{F}_0(b_0) - \widehat{F}_0(a_0) = F_0(b_0) - F_0(a_0)$ and $\widehat{F}_0$ is linear on $(a_0, b_0]$.

(2). Define $\varphi(g) = \int_0^\infty (g(x) - f_0(x))^2 dx$. Recall that any (left-continuous) decreasing function may be written as $g(x) = \int_0^\infty 1_{[0,y]}(x) d\mu_g(y)$, where μ_g is a positive measure. We will use this "decomposition" throughout this proof.

To prove the result, we show that a function $\widehat{f}_0$ minimizes $\varphi(g)$ over $\mathcal{D}$, the space of decreasing nonnegative functions on $\mathbb{R}_+$, if and only if (M) holds, where

$$\text{(M)} \quad = \quad \begin{cases} \widehat{F}_0(y) = \int_0^y \widehat{f}_0(x)dx \geq F_0(y) \text{ for all } y \geq 0, \\ \text{and } \int_0^\infty (\widehat{F}_0 - F_0)(x) d\mu_{\widehat{f}_0}(x) = 0. \end{cases}$$

We first show that a function $\widehat{f}_0$ which minimizes φ must satisfy (M). To do this, we calculate the directional derivative of $\varphi(g)$, $\nabla_b \varphi(g) = \int_0^\infty b(x)(g -$

*Supported in part by an NSERC Discovery Grant

$f_0)(x)dx$. Then, if $\varphi(g) \geq \varphi(\widehat{f}_0)$ for all g, it follows that $\nabla_b \varphi(\widehat{f}_0) \geq 0$ for all b such that $\widehat{f}_0 + \varepsilon b \in \mathcal{D}$ for sufficiently small $\varepsilon > 0$. Similarly, if $\widehat{f}_0 \pm \varepsilon b \in \mathcal{D}$ for sufficiently small $\varepsilon > 0$, then $\nabla_b \varphi(\widehat{f}_0) = 0$. Choosing $b(x) = 1_{[0,y]}(x)$ yields $\widehat{F}_0(y) - F_0(y) \geq 0$, and $b = \widehat{f}_0$ yields $\int_0^\infty (\widehat{F}_0 - F_0)(x) d\mu_{\widehat{f}_0}(x) = 0$.

We will next show that a function which satisfies (M) is the least concave majorant of F_0. That $\widehat{F}_0$ satisfying (M) is a concave majorant is immediate from the definition. Note also that both $\widehat{F}_0$ and F_0 are continuous, and therefore $\{x : \widehat{F}_0(x) > F_0(x)\}$ is open. Therefore, if y is such that $\widehat{F}_0(x) > F_0(x)$ at $x = y$ then this must also be true for all $x \in (y - \delta, y + \delta)$, some $\delta > 0$. It follows then from (M) that $\mu_{\widehat{f}_0}((y - \delta, y + \delta)) = 0$. Let $a_0 = \sup\{x < x_0 : \widehat{F}_0(x) = F_0(x)\}$ and $b_0 = \inf\{x > x_0 : \widehat{F}_0(x) = F_0(x)\}$. Then $\widehat{F}_0 > F_0$ on (a_0, b_0) and, from our previous arguments it follows that $\mu_{\widehat{f}_0}((a_0, b_0)) = 0$. Note that $b_0 < \infty$, since otherwise $\mu_{\widehat{f}_0}((a_0, \infty)) = 0$ which implies that $\widehat{F}_0(x) = \widehat{F}_0(a_0)$ for all $x \geq a_0$ and hence $\widehat{F}_0(x) = F_0(x)$ for all $x \geq a_0$. From the characterization, this implies that $\widehat{f}_0$ is constant on (a_0, b_0) and hence $\widehat{F}_0$ is linear on $[a_0, b_0]$. Since $\widehat{F}_0(x) = F_0(x)$ at $x = a_0, b_0$, there exists no smaller concave function than $\widehat{F}_0$ on $[a_0, b_0]$ which is also greater than F_0. Therefore, $\widehat{F}_0$ is the least concave majorant of F_0. Note that this implies that $\lim_{x\to\infty} \widehat{F}_0(x) = 1$, that $\widehat{f}_0$ is a density, and also that $\widehat{f}_0(x) \leq 1/x$. In particular, we learn that $\int_\varepsilon^\infty \widehat{f}_0^2 dx < \infty$, for all $\varepsilon > 0$.

To complete the proof, we need to show that if $\widehat{f}_0$ satisfies (M), then it must minimize φ over $\mathcal{D}$. Suppose then that there exists $\delta > 0$ such that $\widehat{F}_0(x) > F_0(x)$ for all $x \in (0, \delta]$. Then $\widehat{F}_0$ is linear on $(0, \delta]$, which implies that $\widehat{f}_0(0+) < \infty$. In this case, $\int_0^\infty \widehat{f}_0^2 dx < \infty$. If there is no such δ, then there exists a sequence $\varepsilon_n \to 0$, such that $\widehat{F}_0(\varepsilon_n) = F_0(\varepsilon_n)$. In the former case, the proof below is simplified greatly (i.e take $\varepsilon_n = 0$), and therefore we consider only the latter.

Now, assume that $\int_{\varepsilon_n}^\infty (g - f_0)^2 dx < \infty$, as if it is not, we automatically have that $\int_{\varepsilon_n}^\infty (\widehat{f}_0 - f_0)^2 dx < \int_{\varepsilon_n}^\infty (g - f_0)^2 dx \leq \varphi(g)$ for all ε_n. We may thus also assume that $g(x) < \infty$ for all $x \geq \varepsilon_n$. Then,

$$\begin{aligned}
&\int_{\varepsilon_n}^\infty (g - f_0)^2 dx - \int_{\varepsilon_n}^\infty (\widehat{f}_0 - f_0)^2 dx \\
&\qquad = \int_{\varepsilon_n}^\infty \left\{ (g - \widehat{f}_0)^2 + 2(g - \widehat{f}_0)(\widehat{f}_0 - f_0) \right\} dx \\
&\qquad \geq 2 \int_{\varepsilon_n}^\infty (g - \widehat{f}_0)(\widehat{f}_0 - f_0) dx.
\end{aligned}$$

Note that for $x \geq \varepsilon_n$, $g(x) = \int_0^\infty 1_{[0,y]}(x) d\mu_g(y) = \int_{\varepsilon_n}^\infty 1_{[0,y]}(x) d\mu_g(y)$. Writing g and $\widehat{f}_0$ both in terms of this revised decomposition, we find that the second term is equal to

$$2 \int_{\varepsilon_n}^\infty \int_{\varepsilon_n}^\infty 1_{[0,y]}(x)(\widehat{f}_0 - f_0)(x) d(\mu_g - \mu_{\widehat{f}_0})(y) dx$$

$$= \quad 2\int_{\varepsilon_n}^{\infty} (\widehat{F}_0 - F_0)(y) d(\mu_g - \mu_{\widehat{f}_0})(y),$$

since $\widehat{F_0}(\varepsilon_n) = F_0(\varepsilon_n)$. Note also that $\int_{\varepsilon_n}^{\infty} \widehat{F}_0(y) d\mu_g(y) \leq \int_{\varepsilon_n}^{\infty} d\mu_g(y) = g(\varepsilon_n) < \infty$. Similar arguments may be used to prove finiteness of all terms in the above expression. Since $\widehat{F}_0$ satisfies (M), the right hand side is greater than or equal to zero. Finally, we rearrange to obtain that

$$\int_{\varepsilon_n}^{\infty} (g - f_0)^2 dx \;\geq\; \int_{\varepsilon_n}^{\infty} (\widehat{f}_0 - f_0)^2 dx.$$

The result is proved by letting $\varepsilon_n \to 0$.

(3). This follows from the argument in **(2)**, since h is increasing and therefore $\nabla_{(-h)}\varphi(\widehat{f}_0) \geq 0$.

(4). Let $a = \sup_{x \geq 0} |F_0(x) - G_0(x)|$. Then, by definition, $\widehat{F}_0(x) + a \geq G_0(x)$. Also, since $G_0(x) + a$ is concave and always greater than $F_0(x)$, we have that $G_0(x) + a \geq \widehat{F_0}(x)$, and the result follows. □

Proof of Lemma 4.2

We note that within this section we make reference to the "bracketing entropy", in the sense of empirical process theory (van der Vaart and Wellner, 1996), which is different than the entropy functional discussed in Section 4. Define the Bernstein "norm" to be $d_B^2(f,g) = 2\int (e^{|f-g|} - 1 - |f-g|) dF_0$, and let $h^2(f,g) = 1/2 \int (\sqrt{f} - \sqrt{g})^2 dx$ denote the square of the Hellinger distance.

Lemma 0.1. *Suppose that $f_0/\widehat{f}_0 \leq c_0^2$. For $m_f = \log((f + \widehat{f}_0)/2\widehat{f}_0)$, we have*

1. $d_B^2(m_f, m_{\widehat{f}_0}) \leq 24 c_0^2 h^2(f, \widehat{f}_0)$,
2. $d_B^2(m_f, m_g) \leq 48 c_0^2 h^2(f,g)$.

Proof. For $x \geq -2$, we have $e^{|x|} - 1 - |x| \leq 12(e^{x/2} - 1)^2$. Since $m_f \geq -\log 2$, it follows

$$\begin{aligned} d_B^2(m_f, m_{\widehat{f}_0}) \;&\leq\; 24 \int (e^{m_f/2} - 1)^2 dF_0 \;=\; 24 \int \left(\sqrt{\frac{f + \widehat{f}_0}{2\widehat{f}_0}} - 1 \right)^2 dF_0 \\ &\leq\; 48 c_0^2 h^2((f + \widehat{f}_0)/2, \widehat{f}_0) \;\leq\; 24 c_0^2 h^2(f, \widehat{f}_0), \end{aligned}$$

applying van de Geer (2000, Lemma 4.2, page 48) in the last inequality. Now, suppose that $m_f - m_g \geq 0$. Then using a similar argument, we have that $d_B^2(m_f, m_g)$ is bounded above by

$$\begin{aligned} 24 \int \left(\sqrt{\frac{f + \widehat{f}_0}{g + \widehat{f}_0}} - 1 \right)^2 dF_0 \;&=\; 48 \int \left(\sqrt{\frac{f + \widehat{f}_0}{2}} - \sqrt{\frac{g + \widehat{f}_0}{2}} \right)^2 \frac{f_0}{g + \widehat{f}_0} dx \\ &\leq\; 96 c_0^2 h^2 \left(\frac{f + \widehat{f}_0}{2}, \frac{g + \widehat{f}_0}{2} \right) \;\leq\; 48 c_0^2 h^2(f,g). \end{aligned}$$

If $m_f - m_g \leq 0$, we use instead the bound $e^{|x|} - 1 - |x| \leq 12(e^{|x|/2} - 1)^2$, and obtain $d_B^2(m_f, m_g) \leq 48c_0^2 h^2(g, f) = 48c_0^2 h^2(f, g)$. □

Proof of Lemma 4.2. Let $h_0^2(f, \widehat{f}_0) = \frac{1}{2}\int \left(\sqrt{\frac{f}{\widehat{f}_0}} - 1\right)^2 dF_0$ denote the modified Hellinger distance used in Patilea (2001). Since $\widehat{f}_n$ is the MLE, $h(x) = \log(x)$ is convex, and $\log x \leq x - 1$,

$$\begin{aligned} 0 \leq \frac{1}{4}\int \log \frac{\widehat{f}_n}{\widehat{f}_0} d\mathbb{F}_n &\leq \frac{1}{2}\int \log \frac{\widehat{f}_n + \widehat{f}_0}{2\widehat{f}_0} d\mathbb{F}_n \\ &\leq \frac{1}{2}\int m_{\widehat{f}_n} d(\mathbb{F}_n - F_0) - \int \left(1 - \sqrt{\frac{\widehat{f}_n + \widehat{f}_0}{2\widehat{f}_0}}\right) dF_0, \end{aligned}$$

where $m_f = \log\left(\frac{f+\widehat{f}_0}{2\widehat{f}_0}\right)$. Now, from Patilea (2001, Lemma 2.2, (2.2)), we have $\int \left(1 - \sqrt{\frac{f+\widehat{f}_0}{2\widehat{f}_0}}\right) dF_0 \geq 0$. Hence

$$0 \leq \frac{1}{2}\int \log \frac{\widehat{f}_n}{\widehat{f}_0} d\mathbb{F}_n \leq \int m_{\widehat{f}_n} d(\mathbb{F}_n - F_0),$$

and therefore to prove the lemma it is sufficient to prove that the term on the right hand side is also of order $O_p(n^{-2/3})$. To this end, define $\mathbb{G}_n(m_f) = \sqrt{n}\int m_f(x) d(\mathbb{F}_n - F_0)$, and $||\mathbb{G}_n||_{\mathcal{H}} = \sup_{h\in\mathcal{H}} |\mathbb{G}_n(h)|$, and let $\mathcal{M}_{\delta,K} = \{m_f, h_0(f, \widehat{f}_0) \leq \delta, f \in \mathcal{F}_K\}$, with $\mathcal{F}_K$ denoting the class of positive decreasing functions with bounded support and bounded above by K.

Now, by Markov's inequality

$$\begin{aligned} &P(n^{-1/2}\mathbb{G}_n(m_{\widehat{f}_n}) \geq \delta_n^2) \\ &\quad\leq P(n^{-1/2}\mathbb{G}_n(m_{\widehat{f}_n}) \geq \delta_n^2, m_{\widehat{f}_n} \in \mathcal{M}_{\delta_n,K}) + P(h_0(\widehat{f}_n, \widehat{f}_0) > \delta_n) \\ &\qquad\qquad + P(\widehat{f}_n \notin \mathcal{F}_K) \\ &\quad\leq \frac{\mathrm{E}\left[||\mathbb{G}_n||_{\mathcal{M}_{\delta_n,K}}\right]}{n^{1/2}\delta_n^2} + P(h_0(\widehat{f}_n, \widehat{f}_0) > \delta_n) + P(\widehat{f}_n \notin \mathcal{F}_K). \end{aligned} \tag{S-1}$$

To finish the proof we will use the results of Patilea (2001) and empirical process theory as in Gao and Wellner (2009).

To obtain the appropriate bracketing entropy bounds on $\mathrm{E}[||\mathbb{G}_n||_{\mathcal{M}_{\delta_n}}]$, we will use van der Vaart and Wellner (1996, Lemma 3.4.3, page 324) (we can do this by the first inequality in Lemma 0.1 from which it follows that if $h_0(\widehat{f}_n, \widehat{f}_0) \leq \delta$ then $d_B(\widehat{f}_n, \widehat{f}_0) \leq 5\delta$). Next, the results of Lemma 0.1 further show that an L_2 bracket $[\sqrt{f}, \sqrt{g}]$ of densities of size δ leads to a bracket $[m_f, m_g]$ of Bernstein norm size a multiple of δ. Therefore, the bracketing integral of $\mathcal{M}_{\delta,K}$ under the Bernstein norm can be bounded by the bracketing integral of $\widetilde{\mathcal{M}}_{\delta,K}$, where $\widetilde{\mathcal{M}}_{\delta,K}$ is equal to

$$\{f, f \text{ decreasing with } f(0+) \leq K \text{ and support on } [0, A] \text{ and } ||f||_2 \leq \delta\},$$

under the L_2 norm $||f||_2^2 = \int f^2(x)dx$. By Gao and Wellner (2009, Theorem 4), we have that the bracketing integral $J_{[\cdot]}(\delta, \widetilde{\mathcal{M}}_\delta, ||\cdot||_2) \leq C(\log K)^{1/4}\delta^{1/2}$, for some constant C and we assume without loss of generality that $A = 1$. Applying van der Vaart and Wellner (1996, Lemma 3.4.3, page 324) it follows that

$$\mathrm{E}[||\mathbb{G}_n||_{\mathcal{M}_{\delta_n,K}}] \quad \leq \quad C(\log K)^{1/4}\delta_n^{1/2}\left(1 + (\log K)^{1/4}\frac{\delta_n^{1/2}}{\delta_n^2\sqrt{n}}\right).$$

We now choose $\delta_n = Mn^{-1/3}$ for $M > 1$ and plug this into (S-1). We obtain

$$\begin{aligned} &P(n^{-1/2}\mathbb{G}_n(m_{\widehat{f}_n}) \geq Mn^{-2/3}) \\ &\qquad \leq \quad C\frac{(\log K)^{1/4}}{M^{3/2}}\left(1 + \frac{(\log K)^{1/4}}{M^{3/2}}\right) \\ &\qquad\qquad + \quad P(h_0(\widehat{f}_n, \widehat{f}_0) > Mn^{-1/3}) + P(\widehat{f}_n(0+) \geq K). \end{aligned} \tag{S-2}$$

Now, by definition

$$\widehat{f}_n(0+) \quad = \quad \sup_{t\geq 0}\frac{\mathbb{F}_n(t)}{t} \quad = \quad \sup_{t>0}\frac{\mathbb{F}_n(t)}{F_0(t)}\frac{F_0(t)}{t} \quad \leq \quad \sup_{t>0}\frac{F_0(t)}{t}\sup_{t>0}\frac{\mathbb{F}_n(t)}{F_0(t)},$$

and the term $\sup_{t>0}\mathbb{F}_n(t)/F_0(t) = O_p(1)$ (Shorack and Wellner, 1986, Theorem 2, page 345), while $\sup_{t>0}\frac{F_0(t)}{t} = \widehat{f}_0(0+)$ which is bounded by assumption. Therefore, $\lim_{K\to\infty}\limsup_{n\to\infty}P(\widehat{f}_n(0+) \geq K) = 0$, and we can choose $K = M$, say. Using Patilea (2001, Corollary 5.6 with $\varepsilon = 1$ and $\alpha = 0$) to handle the middle term, we obtain

$$\lim_{M\to\infty}\limsup_{n\to\infty}P(n^{-1/2}\mathbb{G}_n(m_{\widehat{f}_n}) \geq Mn^{-2/3}) = 0,$$

as required. □

On the conditions of Theorem 4.1

We conclude this section by commenting on how the conditions of Theorem 4.1 could be relaxed. The conditions that (A) $f_0/\widehat{f}_0 < \infty$ and (B) f_0 has bounded support are necessary in our method of proof. However, one could relax the condition that $\widehat{f}_0$ is bounded above and replace it with both

(C) for some $\varepsilon \in (0,1)$, $\int_{\widehat{f}_0\geq 1}\widehat{f}_0^{\varepsilon}dF_0(x) < \infty$, and

(D) $\limsup_n P(\sup_{t>0}\mathbb{F}_n(t)/t > K_n) = 0$, for $K_n = n^{1/2}\delta_n^{3/2}$ and $\delta_n = n^{\alpha}$ for $\alpha \in (-1/3, -1/4)$.

The key inequality in (S-2) provides a bound made of up of three terms. The second of these is handled by Patilea (2001, Corollary 5.6), and this continues to hold if (A), (B), and (C) are true. We then have to pick $K = K_n$ such that both the first and third terms go to zero, which is exactly condition (D). Condition

(C) is discussed in Patilea (2001). Some ideas on how to achieve (D) can be gained from the results in Balabdaoui et al. (2011). Note that this approach will change the result of Lemma 4.2 to $\int \log(\widehat{f}_n/\widehat{f}_0)d\mathbb{F}_n = o(n^{-1/2})$, which is still sufficient for Theorem 4.1.

On the conditions of Theorem 3.1

The aim of our work has been to examine misspecification in the Grenander estimator. We therefore focus here on the conditions required for the flat part of the KL projection, namely, (P), and (F). It is not at all clear how to remove condition (P) based on our method of proof. We briefly sketch below some ideas on how one could weaken condition (F) on $\mathcal{S}_f$. Based on our general approach, relaxing these conditions would involve additional assumptions again, as discussed below. Future directions for research would involve developing new methodology, so that potentially weaker assumptions could be obtained.

The key lies in generalizing the results so that $J = \infty$ is possible in the decomposition (3.1), and reproving tightness of $\int |\mathbb{S}_n(x)|^\alpha dx$ under these conditions. Let $\delta_j = b_j - a_j$. Then, (6.6) may be replaced with

$$P\left(\int |\mathbb{S}_n(x)|^\alpha dx > M\right) \leq \sum_{j=1}^J P\left(\int_{(a_j,b_j]} |\mathbb{S}_n(x)|^\alpha dx > M\delta_j\right).$$

To provide the required bounds here, there are two key steps:

1. For each term in the sum above, we first work on the integrand from $a_j + \widetilde{c}_0/n$ to b_j, and apply the calculations from (6.9). We find an upper bound of
$$\frac{C\frac{2^{\alpha+1}}{2-\alpha}\delta_j^{-\alpha/2}}{M}$$
Thus we would need to assume that $\sum_{j=1}^J \delta_j^{-\alpha/2} < \infty$.
2. Secondly, we would require that
$$\sup_j \sup_{x\in[a_j,a_j+\widetilde{c}_0/n)} n^{\alpha/2-1}|\widehat{f}_n(x) - \widehat{f}_0(x)|^\alpha \quad = \quad O_p(1). \qquad \text{(S-3)}$$
Now, repeating the arguments in Balabdaoui et al. (2011) (see also Woodroofe and Sun (1993)), one has that
$$\sup_{x\in[a_j,a_j+\widetilde{c}_0/n)} |\widehat{f}_n(x) - \widehat{f}_0(x)| \quad \Rightarrow \quad \widehat{f}_0(a_j)\left(\frac{1}{U_j} - 1\right),$$
where U_j is a uniform random variable. Moreover, it is not difficult to see that these variables are independent across j. In order to show that (S-3) holds, one would require at least that $\sup_j \widehat{f}_0(a_j)(U_j^{-1} - 1)$ is finite a.s..

To see that the latter holds, note that

$$\begin{aligned} P\left(\sup_j f_0^\alpha(a_j)(U_j^{-1}-1)^\alpha > M\right) &= 1-\prod_j\left(1-\frac{\widehat{f}_0(a_j)}{\widehat{f}_0(a_j)+M^{1/\alpha}}\right) \\ &\leq 1-\exp\left\{-\frac{\sum_j \widehat{f}_0(a_j)}{M^{1/\alpha}}\right\}, \end{aligned}$$

since $1-x > e^{-x/(1-x)}$, for $x<1$. Assuming then that $\sum_j \widehat{f}_0(a_j) < \infty$, yields that $\sup_j \widehat{f}_0(a_j)(U_j^{-1}-1)$ is finite a.s..

References

Balabdaoui, F., Jankowski, H., Pavlides, M., Seregin, A. and Wellner, J. A. (2011). On the Grenander estimator at zero. *Statistica Sinica* **21** 873–899.

Gao, F. and Wellner, J. A. (2009). On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Sci. China Ser. A* **52** 1525–1538.

Patilea, V. (2001). Convex models, MLE and misspecification. *Ann. Statist.* **29** 94–123.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York.

van de Geer, S. A. (2000). *Applications of Empirical Process Theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

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

Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is nonincreasing. *Statistica Sinica* **3** 501–515.