Submitted to the Annals of Statistics

APPROXIMATION BY LOG-CONCAVE DISTRIBUTIONS WITH APPLICATIONS TO REGRESSION

By Lutz Dümbgen*

University of Bern

and

By Richard Samworth

University of Cambridge

and

By Dominic Schuhmacher*

University of Bern

We study the approximation of arbitrary distributions P on d-dimensional space by distributions with log-concave density. Approximation means minimizing a Kullback–Leibler type functional. We show that such an approximation exists if, and only if, P has finite first moments and is not supported by some hyperplane. Furthermore we show that this approximation depends continuously on P with respect to Mallows' distance $D_1(\cdot,\cdot)$. This result implies consistency of the maximum likelihood estimator of a log-concave density under fairly general conditions. It also allows us to prove existence and consistency of estimators in regression models with a response $Y = \mu(X) + \epsilon$, where X and ϵ are independent, $\mu(\cdot)$ belongs to a certain class of regression functions while ϵ is a random error with log-concave density.

1. Introduction. Log-concave distributions, i.e. distributions with a Lebesgue density the logarithm of which is concave, are an interesting non-parametric model comprising many parametric families of distributions. Bagnoli and Bergstrom (2005) give an overview of many interesting properties and applications in econometrics. Indeed, these distributions have received a lot of attention among statisticians recently as described in the review by Walther (2009). The nonparametric maximum likelihood estimator was studied in the univariate setting by Pal *et al.* (2007), Rufibach (2006), Dümbgen *et al.* (2007), Balabdaoui *et al.* (2009) and Dümbgen and Ru-

*Work supported by Swiss National Science Foundation

AMS 2000 subject classifications: 62E17, 62G05, 62G07, 62G08, 62G35, 62H12

Keywords and phrases: convex support, isotonic regression, linear regression, Mallows' distance, projection, weak semi-continuity

imsart-aos ver. 2009/08/13 file: LogConcRegr_AoS.tex date: February 25, 2010

fibach (2009). These references contain chracterizations of the estimators, consistency results and explicit algorithms. Extensions of one or more of these aspects to the multivariate setting are presented by Cule *et al.* (2010), Cule and Samworth (2010), Koenker and Mizera (2010), Seregin and Wellner (2009) and Schuhmacher and Dümbgen (2010). Both Cule and Samworth (2010) and Schuhmacher *et al.* (2009) show that multivariate log-concave distributions are a very well-behaved nonparametric class. For instance, moments of arbitrary order are continuous statistical functionals with respect to weak convergence.

The first aim of the present paper is a deeper understanding of the approximation scheme underlying the maximum-likelihood estimator of a log-concave density. Precisely, let $\hat{P}_n$ be the empirical distribution of independent random vectors $X_1, X_2, \ldots, X_n$ with distribution P. Suppose that P has a density f such that $\log f : \mathbb{R}^d \to [-\infty, \infty)$ is concave. If we denote the set of all such densities by $\mathcal{F}$, the maximum likelihood estimator of f may be written as

$$\hat{f}_n \ = \ \operatorname*{arg\,max}_{\tilde{f} \in \mathcal{F}} \int \log(\tilde{f}) \, d\hat{P}_n.$$

But even if P fails to have a log-concave density, one may view $\hat{f}_n$ as an estimator of the approximating density

$$f(\cdot \,|\, P) \ := \ \operatorname*{arg\,max}_{\tilde{f} \in \mathcal{F}} \int \log(\tilde{f}) \, dP,$$

provided that this exists and is unique. In fact, if P has a density $f \notin \mathcal{F}$ such that the integral $\int f(x) \log f(x) \, dx$ exists in $\mathbb{R}$, one may rewrite $f(\cdot \,|\, P)$ as the minimizer of the Kullback–Leibler divergence

$$D_{KL}(\tilde{f}, f) \ = \ \int \log(f/\tilde{f})(x) f(x) \, dx$$

over all $\tilde{f} \in \mathcal{F}$. This point of view is common in statistics, see for instance Doksum *et al.* (2007), and was adopted by Cule and Samworth (2010).

In Section 2 of the present paper we will show that $f(\cdot \,|\, P)$ exists and is unique in $L^1(\mathbb{R}^d)$ if, and only if,

$$\int \|x\| \, P(dx) \ < \ \infty$$

and

$$P(H) \ < \ 1 \quad \text{for any hyperplane } H \subset \mathbb{R}^d.$$

Some additional properties of $f(\cdot \,|\, P)$ will be established as well. We show that the mapping $P \mapsto f(\cdot \,|\, P)$ is continuous with respect to Mallows' distance D_1 (Mallows, 1972), also known as a Wasserstein, Monge–Kantorovich or Earth Mover's distance. Precisely, let P satisfy the properties just mentioned, and let $(P_n)_n$ be a sequence of probability distributions that converge in D_1; in other words,

$$P_n \;\to_w\; P \quad \text{and} \quad \int \|x\| \, P_n(dx) \;\to\; \int \|x\| \, P(dx)$$

as $n \to \infty$. Then $f(\cdot \,|\, P_n)$ is well-defined for sufficiently large n, and

$$\lim_{n\to\infty} \int \bigl| f(x \,|\, P_n) - f(x \,|\, P) \bigr| \, dx \;=\; 0.$$

In addition we show that $P \mapsto \max_{\tilde{f}\in\mathcal{F}} \int \log(\tilde{f}) \, dP$ is upper semicontinuous with respect to weak convergence. These findings entail consistency of the maximum likelihood estimator $\hat{f}_n$, because $(\hat{P}_n)_n$ converges to P in probability with respect to Mallows' distance $D_1(\cdot,\cdot)$, and because if P has a log-concave Lebesgue density f on $\mathbb{R}^d$, then $f(\cdot \,|\, P) = f$.

In Section 3 we apply these results to the following type of regression problem: Suppose that we observe independent real random variables Y_1, Y_2, ..., Y_n such that

$$Y_i \;=\; \mu(x_i) + \epsilon_i$$

for given fixed design points $x_1, x_2, \ldots, x_n$ in some set $\mathcal{X}$, some unknown regression function $\mu : \mathcal{X} \to \mathbb{R}$ and independent, identically distributed random errors ϵ_i with unknown log-concave density f and mean zero. We will show that a maximum likelihood estimator of (μ, f) exists and is consistent under certain regularity conditions in the following two cases: (i) $\mathcal{X} = \mathbb{R}^q$ and μ is affine (i.e. affine linear); (ii) $\mathcal{X} = \mathbb{R}$ and μ is non-decreasing.

The main proofs are deferred to Section 4. For some technical details and auxiliary results we refer to an accompanying technical report by Dümbgen *et al.* (2010).

2. Log-Concave Approximations. For a fixed dimension $d \in \mathbb{N}$, let $\mathcal{G}$ be the family of concave functions $g : \mathbb{R}^d \to [-\infty, \infty)$ which are upper semicontinuous and coercive in the sense that

$$g(x) \;\to\; -\infty \quad \text{as } \|x\| \to \infty.$$

In particular, for any $g \in \mathcal{G}$ there exist constants a and $b > 0$ such that $g(x) \le a - b\|x\|$, so $\int e^{g(x)} \, dx$ is finite. Further let $\mathcal{P} = \mathcal{P}(d)$ be the family of all probability distributions P on $\mathbb{R}^d$. Then we define a log-likelihood type functional

$$L(g, P) \ := \ \int g \, dP - \int e^{g(x)} \, dx + 1$$

and a profile log-likelihood

$$L(P) \ := \ \sup_{g \in \mathcal{G}} L(g, P).$$

If, for fixed P, there exists a function $\psi \in \mathcal{G}$ such that $L(\psi, P) = L(P) \in \mathbb{R}$, then it will automatically satisfy

$$\int e^{\psi(x)} \, dx \ = \ 1.$$

To verify this, note that $g + c \in \mathcal{G}$ for any function $g \in \mathcal{G}$ and constant $c \in \mathbb{R}$, and

$$\frac{\partial}{\partial c} L(g + c, P) \ = \ 1 - e^c \int e^{g(x)} \, dx$$

if $L(g, P) > -\infty$. Then $c \mapsto L(g + c, P)$ has a unique minimum at $c = -\log \int e^{g(x)} \, dx$.

2.1. *Existence, uniqueness and basic properties.* The next theorem provides a complete characterization of all distributions $P \in \mathcal{P}$ such that $L(P) \in \mathbb{R}$. To state the result we first need to define the convex support of a distribution $P \in \mathcal{P}$ and establish some of its properties:

LEMMA 2.1. *For any $P \in \mathcal{P}$, the set*

$$\mathrm{csupp}(P) \ := \ \bigcap \{C : C \subset \mathbb{R}^d \ \textit{closed and convex}, P(C) = 1\}$$

is itself closed and convex with $P(\mathrm{csupp}(P)) = 1$. The following three properties of P are equivalent:
(a) $\mathrm{csupp}(P)$ *has non-empty interior.*
(b) *$P(H) < 1$ for any hyperplane $H \subset \mathbb{R}^d$.*
(c) *With* Leb *denoting Lebesgue measure on $\mathbb{R}^d$,*

$$\lim_{\delta \downarrow 0} \sup\{P(C) : C \subset \mathbb{R}^d \ \textit{closed and convex}, \mathrm{Leb}(C) \le \delta\} \ < \ 1.$$

THEOREM 2.2. *For any $P \in \mathcal{P}$, the value of $L(P)$ is real if, and only if,*

$$\int \|x\| P(dx) \ < \ \infty \quad \textit{and} \quad \mathrm{interior}(\mathrm{csupp}(P)) \ \neq \ \emptyset.$$

In that case, there exists a unique function

$$\psi = \psi(\cdot \,|\, P) \;\in\; \operatorname*{arg\,max}_{g \in \mathcal{G}} L(g, P).$$

This function ψ satisfies $\int e^{\psi(x)}\,dx = 1$ and

$$\mathrm{interior}(\mathrm{csupp}(P)) \;\subset\; \mathrm{dom}(\psi) := \{x \in \mathbb{R}^d : \psi(x) > -\infty\} \;\subset\; \mathrm{csupp}(P).$$

Remark 2.3 (Moment (in)equalities). Let $P \in \mathcal{P}$ satisfy the properties stated in Theorem 2.2. Then the log-density $\psi = \psi(\cdot \,|\, P)$ satisfies the following requirements: $\int \psi\, dP > -\infty$, and for any function $\Delta : \mathbb{R}^d \to \mathbb{R}$,

$$\int \Delta\, dP \;\le\; \int \Delta(x) e^{\psi(x)}\, dx \quad \text{if } \psi + t\Delta \in \mathcal{G} \text{ for some } t > 0. \tag{1}$$

This follows from

$$\lim_{t \downarrow 0} t^{-1}\big(L(\psi + t\Delta, P) - L(\psi, P)\big) \;=\; \int \Delta\, dP - \int \Delta(x) e^{\psi(x)}\, dx.$$

Let P^* be the approximating probability measure with $P^*(dx) = e^{\psi(x)}\,dx$. It satisfies the following (in)equalities:

$$\int h\, dP^* \;\le\; \int h\, dP \quad \text{for any convex } h : \mathbb{R}^d \to (-\infty, \infty], \tag{2}$$

$$\int x\, P^*(dx) \;=\; \int x\, P(dx). \tag{3}$$

For let $v \in \mathbb{R}^d$ be a subgradient of h at 0, i.e. $h(x) \ge h(0) + v^\top x$ for all $x \in \mathbb{R}^d$. Since $\psi(x) \le a - b\|x\|$ for arbitrary $x \in \mathbb{R}^d$ and suitable constants a and $b > 0$, the function $\Delta := -h$ satisfies the requirement that $\psi + t\Delta \in \mathcal{G}$ whenever $0 < t < b/\|v\|$. Hence the asserted inequality follows from (1). The equality for the first moments follows by setting $h(x) := \pm v^\top x$ for arbitrary $v \in \mathbb{R}^d$.

In what follows let

$$\begin{aligned} \mathcal{P}^1 = \mathcal{P}^1(d) \;&:=\; \Big\{P \in \mathcal{P} : \int \|x\|\, P(dx) < \infty\Big\}, \\ \mathcal{P}_o = \mathcal{P}_o(d) \;&:=\; \{P \in \mathcal{P} : \mathrm{interior}(\mathrm{csupp}(P)) \neq \emptyset\}. \end{aligned}$$

Thus $L(P) \in \mathbb{R}$ if, and only if, $P \in \mathcal{P}_o \cap \mathcal{P}^1$. Moreover, the proof of Theorem 2.2 shows that

$$L(P) \;=\; \begin{cases} -\infty & \text{for } P \in \mathcal{P} \setminus \mathcal{P}^1, \\ +\infty & \text{for } P \in \mathcal{P}^1 \setminus \mathcal{P}_o. \end{cases}$$

REMARK 2.4 (Affine equivariance). Suppose that $P \in \mathcal{P}_o \cap \mathcal{P}^1$. For arbitrary vectors $a \in \mathbb{R}^d$ and nonsingular, real $d \times d$ matrices B define $P_{a,B}$ to be the distribution of $a + BX$ when X has distribution P .

Then $P_{a,B} \in \mathcal{P}_o \cap \mathcal{P}^1$, too, and elementary considerations reveal that

$$\begin{aligned} L(P_{a,B}) &= L(P) - \log|\det B| \quad \text{and} \\ \psi(x \,|\, P_{a,B}) &= \psi(B^{-1}(x-a) \,|\, P) - \log|\det B| \quad \text{for } x \in \mathbb{R}^d. \end{aligned}$$

2.2. *The one-dimensional case.* For the case of $d = 1$ one can generalize Theorem 2.4 of Dümbgen and Rufibach (2009) as follows: For a function $g \in \mathcal{G}$ let

$$\mathcal{S}(g) \ := \ \Big\{ x \in \mathrm{dom}(g) : g(x) > 2^{-1}\big(g(x-\delta) + g(x+\delta)\big) \text{ for all } \delta > 0 \Big\}.$$

The log-concave approximation of a distribution on $\mathbb{R}$ can be characterized in terms of distribution functions only:

THEOREM 2.5. *Let P be a non-degenerate distribution on $\mathbb{R}$ with finite first moment and distribution function F. Let $\tilde{P}$ be a distribution on $\mathbb{R}$ with distribution function $\tilde{F}$ and log-density $g \in \mathcal{G}$. Then $g = \psi(\cdot \,|\, P)$ if, and only if, $\int_{-\infty}^{\infty} (\tilde{F}(t) - F(t))\, dt = 0$ and*

$$\int_{-\infty}^{x} (\tilde{F}(t) - F(t))\, dt \begin{cases} \leq 0 & \textit{for all } x \in \mathbb{R}, \\ = 0 & \textit{for all } x \in \mathcal{S}(g). \end{cases}$$

REMARK 2.6. One consequence of this theorem is that the distribution function F^* of P^* follows the distribution function F quite closely in that

$$F(x\,-) \ \leq \ F^*(x) \ \leq \ F(x) \quad \text{for arbitrary } x \in \mathcal{S}(\psi(\cdot \,|\, P)).$$

For it follows from the inequalities in Theorem 2.5 that for $x \in \mathcal{S}(\psi(\cdot \,|\, P))$ and $\delta > 0$,

$$\begin{aligned} 0 \ &\leq \ \frac{1}{\delta} \int_{x-\delta}^{x} (F^*(t) - F(t))\, dt \ \to \ F^*(x) - F(x\,-) \quad \text{and} \\ 0 \ &\geq \ \frac{1}{\delta} \int_{x}^{x+\delta} (F^*(t) - F(t))\, dt \ \to \ F^*(x) - F(x) \quad \text{as } \delta \downarrow 0. \end{aligned}$$

EXAMPLE 2.7. Let P be a rescaled version of student's distribution t_2 with density and distribution function

$$f(x) \ = \ 2^{-1}(1+x^2)^{-3/2} \quad \text{and} \quad F(x) \ = \ 2^{-1}(1 + (1+x^2)^{-1/2}x),$$

respectively. The best approximating log-concave distribution is the Laplace distribution with density and distribution function

$$f^*(x) \;=\; 2^{-1}e^{-|x|} \quad \text{and} \quad F^*(x) \;=\; \begin{cases} f^*(x) & \text{for } x \le 0, \\ 1 - f^*(x) & \text{for } x \ge 0, \end{cases}$$

respectively. To verify this claim, note that by symmetry it suffices to show that

$$\int_{-\infty}^{x} \bigl(F^*(t) - F(t)\bigr)\, dt \; \begin{cases} \le \; 0 & \text{for } x \le 0, \\ = \; 0 & \text{for } x = 0. \end{cases}$$

Indeed the integral on the left hand side equals

$$2^{-1}\bigl(\exp(x) - x - (1 + x^2)^{1/2}\bigr)$$

for all $x \le 0$. Clearly this expression is zero for $x = 0$, and elementary considerations show that it is non-positive all $x \le 0$. Figure 1 shows f and F (green/dotted lines) together with f^* and F^* (blue lines). The maximum of $|F^* - F|$ is less than 0.04.

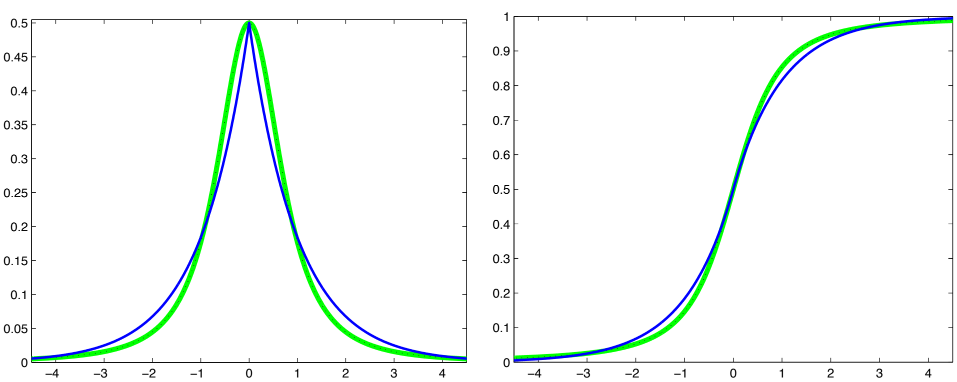

FIG 1. *Densities f, f^* (left panel) and distribution functions F, F^* (right panel) for Example 2.7.*

REMARK 2.8. Suppose that $P(a, b) = 0$ for some interval (a, b) contained in $\mathrm{csupp}(P)$. Then $\psi = \psi(\cdot \,|\, P)$ is linear on (a, b). For otherwise we could replace ψ on $[a, b]$ by a linear function $\tilde{\psi}$ with $\tilde{\psi}(a) = \psi(a)$ and $\tilde{\psi}(b) = \psi(b)$ without changing $\int \psi \, dP$ while decreasing $\int e^{\psi(x)}\, dx$ strictly. Note that $\psi(a) > -\infty$ and $\psi(b) > -\infty$, because otherwise $\int_{(-\infty,a]} \psi \, dP = -\infty$ or $\int_{[b,\infty)} \psi \, dP = -\infty$ and hence $L(\psi, P) = -\infty$.

REMARK 2.9. Suppose that P has a continuous but not log-concave density f. Nevertheless one can say the following about the approximating log-density $\psi = \psi(\cdot \,|\, P)$; see Dümbgen *et al.* (2010):

(i) Suppose that $\log f$ is concave on an interval $(-\infty, a]$ with $f(a) > 0$ and $\psi(a) \le \log f(a)$. Then there exists a point $a' \in [-\infty, a]$ such that ψ is linear on $(a', a]$, and $\psi = \log f$ on $(-\infty, a']$.

(ii) Suppose that $\log f$ is differentiable everywhere, convex on a bounded interval $[a, b]$ and concave on both $(-\infty, a]$ and $[b, \infty)$. Then there exist points $a' \in (-\infty, a]$ and $b' \in [b, \infty)$ such that ψ is linear on $[a', b']$ while $\psi = \log f$ on $(-\infty, a'] \cup [b', \infty)$.

(iii) Suppose that $\log f$ is convex on an interval $(-\infty, a]$ such that $-\infty < \log f(a) \le \psi(a)$. Then ψ is linear on $(-\infty, a]$.

EXAMPLE 2.10. Let us illustrate part (ii) of Remark 2.9 with a numerical example. Figure 2 shows the bimodal density f of the Gaussian mixture $0.3\mathcal{N}(-1.5, 1) + 0.7\mathcal{N}(1.5, 1)$ together with its log-concave approximation f^*. As predicted, there exists an interval $[a', b']$ such that $f^* = f$ on $\mathbb{R} \setminus (a', b')$ and $\log f^*$ is linear on $[a', b']$.

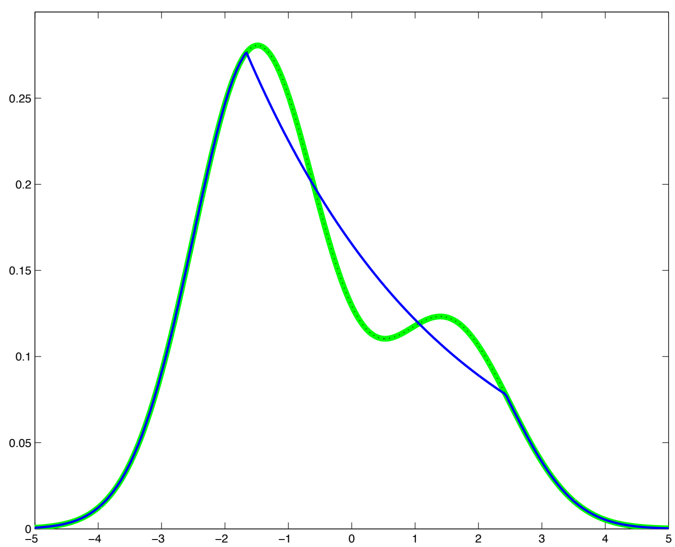

FIG 2. *Density of a Gaussian mixture and its log-concave approximation.*

2.3. *Continuity in P.* For the applications to regression problems to follow we need to understand the properties of both $P \mapsto L(P)$ and $P \mapsto$

$\psi(\cdot \mid P)$ on $\mathcal{P}^1 \cap \mathcal{P}_o$. Our first hope was that both mappings would be continuous with respect to the weak topology. It turned out, however, that we need a somewhat stronger notion of convergence, namely, convergence with respect to Mallows' distance D_1 which is defined as follows: For two probability distributions $P, Q \in \mathcal{P}^1$,

$$D_1(P,Q) \ := \ \inf_{(X,Y)} \mathbb{E}\|X - Y\|,$$

where the infimum is taken over all pairs (X, Y) of random vectors $X \sim P$ and $Y \sim Q$ on a common probability space. It is well-known that the infimum in $D_1(P,Q)$ is a minimum. Moreover, for a sequence $(P_n)_n$ in $\mathcal{P}^1$, the following two statements are known to be equivalent (Mallows, 1972; Bickel and Freedman, 1981):

$$D_1(P_n, P) \ \to \ 0 \quad (n \to \infty),$$

$$P_n \ \to \ P \text{ weakly} \quad \text{and} \quad \int \|x\| \, P_n(dx) \ \to \ \int \|x\| \, P(dx) \quad (n \to \infty).$$

In case of $d = 1$, the optimal coupling is given by the quantile transformation: If F and G denote the distribution function of P and Q, respectively, then

$$D_1(P,Q) \ = \ \int_0^1 \bigl|F^{-1}(u) - G^{-1}(u)\bigr| \, du \ = \ \int_{-\infty}^{\infty} \bigl|F(x) - G(x)\bigr| \, dx.$$

Before presenting the main results of this section we mention two useful facts about the convex support of distributions.

LEMMA 2.11. *Given a distribution $P \in \mathcal{P}$, a point $x \in \mathbb{R}^d$ belongs to* interior(csupp(P)) *if, and only if,*

$$h(P,x) := \sup\bigl\{P(C) : C \subset \mathbb{R}^d \textit{ closed and convex}, \ x \notin \operatorname{interior}(C)\bigr\} \ < \ 1.$$

Moreover, if $(P_n)_n$ is a sequence in $\mathcal{P}$ converging weakly to P, then

$$\limsup_{n \to \infty} h(P_n, x) \ \le \ h(P,x) \quad \textit{for any } x \in \mathbb{R}^d.$$

This lemma implies that the set $\mathcal{P}_o$ is an open subset of $\mathcal{P}$ with respect to the topology of weak convergence. The supremum $h(P,x)$ will be seen to be a maximum over closed halfspaces and is related to Tukey's halfspace depth; see also Section 6 of Donoho and Gasko (1992). Now we are ready for stating the main results of this section.

THEOREM 2.12 (Weak upper semicontinuity). *Let $(P_n)_n$ be a sequence of distributions in $\mathcal{P}_o$ converging weakly to some $P \in \mathcal{P}_o$. Then*

$$\limsup_{n\to\infty} L(P_n) \ \le \ L(P).$$

Moreover,

$$\liminf_{n\to\infty} L(P_n) \ < \ L(P)$$

if, and only if,

$$\limsup_{n\to\infty} \int \|x\| \, P_n(dx) \ > \ \int \|x\| \, P(dx).$$

This result already entails continuity of $L(\cdot)$ on $\mathcal{P}_o \cap \mathcal{P}^1$ with respect to Mallows' distance D_1. The next theorem extends this result to $L : \mathcal{P}^1 \to (-\infty, \infty]$:

THEOREM 2.13 (Continuity with respect to Mallows' distance D_1). *Let $(P_n)_n$ be a sequence of distributions in $\mathcal{P}^1$ such that $\lim_{n\to\infty} D_1(P, P_n) = 0$ for some $P \in \mathcal{P}^1$. Then*

$$\lim_{n\to\infty} L(P_n) \ = \ L(P).$$

In case of $P \in \mathcal{P}_o \cap \mathcal{P}^1$, the probability densities $f := \exp\circ\psi(\cdot \,|\, P)$ and $f_n := \exp\circ\psi(\cdot \,|\, P_n)$ are well-defined for sufficiently large n and satisfy

$$\begin{aligned}
\lim_{n\to\infty,\ x\to y} f_n(x) \ &= \ f(y) \quad \textit{for all } y \in \mathbb{R}^d \setminus \partial\{f > 0\}, \\
\limsup_{n\to\infty,\ x\to y} f_n(x) \ &\le \ f(y) \quad \textit{for all } y \in \partial\{f > 0\}, \\
\lim_{n\to\infty} \int \bigl| f_n(x) - f(x) \bigr| \, dx \ &= \ 0.
\end{aligned}$$

REMARK 2.14 (Stronger modes of convergence). The convergence of $(f_n)_n$ to f in total variation distance may be strengthened considerably. It follows from recent results of Cule and Samworth (2010) or Schuhmacher *et al.* (2009) that $(f_n)_n \to f$ uniformly on arbitrary closed subsets of $\mathbb{R} \setminus D(f)$, where $D(f)$ is the set of discontinuity points of f. The latter set is contained in the boundary of the convex set $\{f > 0\}$, hence a null set with respect to Lebesgue measure. Moreover, there exists a number $\epsilon(f) > 0$ such that

$$\lim_{n\to\infty} \int e^{\epsilon(f)\|x\|} \bigl| f_n(x) - f(x) \bigr| \, dx \ = \ 0.$$

More generally,

$$\lim_{n\to\infty} \int e^{A(x)} \big| f_n(x) - f(x) \big| \, dx \ = \ 0$$

for any sublinear function $A : \mathbb{R}^d \to \mathbb{R}$ such that $\lim_{\|x\|\to\infty} e^{A(x)} f(x) = 0$.

Example 2.15 (Discontinuity w.r.t. weak convergence). Here is a simple example illustrating that convergence with respect to D_1 in Theorem 2.13 cannot be replaced with weak convergence: Let $P := \mathrm{Unif}\{-1, 1\}$, and let $P_n := (1 - n^{-1})P + n^{-1}\mathrm{Unif}\{-(n+1), n+1\}$. Clearly the sequence $(P_n)_n$ converges weakly to P. But

$$\int |x| \, P_n(dx) \ = \ 2 \ = \ \int |x| \, P(dx) + 1$$

for all $n \in \mathbb{N}$. One can easily verify that $\psi := \psi(\cdot \,|\, P)$ is given by

$$\psi(x) \ = \ \begin{cases} -\log(2) & \text{if } |x| \le 1, \\ -\infty & \text{otherwise.} \end{cases}$$

In particular, $L(P) = -\log(2)$. Moreover, $\psi_n = \psi(\cdot \,|\, P_n)$ has to be an even function, linear on the three intervals $[-(n+1), -1]$, $[-1, 1]$ and $[1, n+1]$, while $\psi_n(x) = -\infty$ for $|x| > n+1$. It is shown by Dümbgen *et al.* (2010) that as $n \to \infty$

$$\begin{aligned} \psi_n(x) \ &\to \ -\log 2 - \log(1 + \kappa^{-1}) - \kappa(|x| - 1)^+, \\ L(P_n) = L(\psi_n, P_n) \ &\to \ -\log(2) - \kappa - \log(1 + \kappa^{-1}), \\ \int \big| f_n(x) - f(x) \big| \, dx \ &\to \ 2/(\kappa + 1), \end{aligned}$$

where $\kappa := (5^{1/2} - 1)/2$.

3. Regression problems.

Now we consider the regression setting described in the introduction with observations $Y_i = \mu(x_i) + \epsilon_i$, $1 \le i \le n$, where the $x_i \in \mathcal{X}$ are given fixed design points, $\mu : \mathcal{X} \to \mathbb{R}$ is an unknown regression function, and the ϵ_i are independent random errors with mean zero and unknown distribution P on $\mathbb{R}$ such that $\psi = \psi(\cdot \,|\, P)$ is well-defined. The regression function μ is assumed to belong to a given family $\mathcal{M}$ with the property that

$$m + c \in \mathcal{M} \quad \text{for arbitrary } m \in \mathcal{M}, c \in \mathbb{R}.$$

3.1. *Maximum likelihood estimation.* We propose to estimate (ψ, μ) by a maximizer of

$$\hat{\Lambda}(g, m) \ := \ \frac{1}{n}\sum_{i=1}^{n} g(Y_i - m(x_i)) - \int e^{g(x)}\, dx + 1$$

over all $(g, m) \in \mathcal{G} \times \mathcal{M}$. Note that $\hat{\Lambda}(g, m)$ remains unchanged if we replace (g, m) with $(g(\cdot + c), m + c)$ for an arbitrary $c \in \mathbb{R}$. For fixed m the maximizer $\hat{g} = \hat{g}_m$ of $\hat{\Lambda}(\cdot, m)$ over $\mathcal{G}$ will automatically satisfy $\int \exp(\hat{g}(x))\, dx = 1$ and $\int x \exp(\hat{g}(x))\, dx = n^{-1}\sum_{i=1}^{n}(Y_i - m(x_i))$. Thus if $(\hat{g}, \hat{m})$ maximizes $\hat{\Lambda}(g, m)$ over $\mathcal{G} \times \mathcal{M}$, then

$$(\hat{\psi}, \hat{\mu}) \ := \ (\hat{g}(\cdot + c), \hat{m} + c) \quad \text{with} \quad c \ := \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}(x_i))$$

maximizes $\hat{\Lambda}(g, m)$ over all $(g, m) \in \mathcal{G} \times \mathcal{M}$ satisfying the additional constraint that $\exp \circ g$ defines a probability density with mean zero.

Define $\boldsymbol{x} := (x_i)_{i=1}^{n}$ and $m(\boldsymbol{x}) := (m(x_i))_{i=1}^{n}$. Then we may write

$$\hat{\Lambda}(g, m) \ = \ L(g, \hat{P}_{m(\boldsymbol{x})})$$

with the empirical distributions

$$\hat{P}_{\boldsymbol{v}} \ := \ \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i - v_i}$$

for $\boldsymbol{v} = (v_i)_{i=1}^{n} \in \mathbb{R}^n$. Thus our procedure aims to find

$$(\hat{g}, \hat{m}) \ \in \ \mathop{\arg\max}_{(g,m) \in \mathcal{G}\times\mathcal{M}} L(g, \hat{P}_{m(\boldsymbol{x})}),$$

and this representation is our key to proving the existence of $(\hat{\psi}, \hat{\mu})$. Before doing so we state a simple inequality of independent interest.

LEMMA 3.1. *For any distribution $P \in \mathcal{P}^1(1)$,*

$$L(P) \ \le \ -\log\Bigl(2\int |x - \mathrm{Med}(P)|\, P(dx)\Bigr) \ \le \ -\log\Bigl(\int |x - \mu(P)|\, P(dx)\Bigr),$$

where $\mathrm{Med}(P)$ is a median of P while $\mu(P)$ denotes its mean $\int x\, P(dx)$.

THEOREM 3.2 (Existence of regression estimator). *Suppose that the set $\mathcal{M}(\boldsymbol{x}) := \{m(\boldsymbol{x}) : m \in \mathcal{M}\} \subset \mathbb{R}^n$ is closed and does not contain $\boldsymbol{Y} := (Y_i)_{i=1}^{n}$. Then there exists a maximizer $(\hat{g}, \hat{m})$ of $\hat{\Lambda}(g, m)$ over all $(g, m) \in \mathcal{G} \times \mathcal{M}$.*

The constraint $\boldsymbol{Y} \notin \mathcal{M}(\boldsymbol{x})$ prevents situations with perfect fit. In that case, the Dirac measure δ_0 would be the most plausible error distribution.

EXAMPLE 3.3 (Linear Regression). Let $\mathcal{X} = \mathbb{R}^q$, and let $\mathcal{M}$ consist of all affine functions on $\mathbb{R}^q$. Then $\mathcal{M}(\boldsymbol{x})$ is the column space of the design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^{n\times(q+1)},$$

hence a linear subspace of $\mathbb{R}^n$. Consequently there exists a maximizer $(\hat{g}, \hat{m})$ of $\hat{\Lambda}$ over $\mathcal{G} \times \mathcal{M}$, unless $\boldsymbol{Y} \in \mathcal{M}(\boldsymbol{x})$.

EXAMPLE 3.4 (Isotonic Regression). Let $\mathcal{X}$ be some interval on the real line, and let $\mathcal{M}$ consist of all isotonic functions $m : \mathcal{X} \to \mathbb{R}$. Then the set $\mathcal{M}(\boldsymbol{x})$ is a closed convex cone in $\mathbb{R}^n$. Here the condition that $\boldsymbol{Y} \notin \mathcal{M}(\boldsymbol{x})$ is equivalent to the existence of two indices $i, j \in \{1, 2, \ldots, n\}$ such that $x_i \le x_j$ but $Y_i > Y_j$.

Fisher consistency. So far we were not able to prove uniqueness of the maximum likelihood estimator $(\hat{\psi}, \hat{\mu})$. Nevertheless we will prove it to be consistent under certain regularity conditions. A key point here is Fisher consistency in the following sense: Note that the expectation measure of the empirical distribution $\hat{P}_{\boldsymbol{v}}$ equals

$$\mathbb{E}\hat{P}_{m(\boldsymbol{x})} \;=\; \frac{1}{n}\sum_{i=1}^n P \star \delta_{\mu(x_i)-m(x_i)} \;=\; P \star Q_{(\mu-m)(\boldsymbol{x})}$$

with

$$Q_{\boldsymbol{v}} \;:=\; \frac{1}{n}\sum_{i=1}^n \delta_{v_i}.$$

But

$$L(P \star Q_{(\mu-m)(\boldsymbol{x})}) \;\le\; L(P)$$

with equality if, and only if, $\mu - m$ is constant on $\{x_1, x_2, \ldots, x_n\}$. This follows from a more general inequality which is somewhat reminiscent of Anderson's Lemma (Anderson, 1955):

THEOREM 3.5. *Let $P \in \mathcal{P}_o(d) \cap \mathcal{P}^1(d)$ and $Q \in \mathcal{P}^1(d)$. Then $P \star Q \in \mathcal{P}_o \cap \mathcal{P}^1$ and*

$$L(P \star Q) \;\le\; L(P).$$

Equality holds if, and only if, $Q = \delta_a$ for some $a \in \mathbb{R}^d$.

3.2. *Consistency.* In this subsection we consider a triangular scheme of independent observations (x_{ni}, Y_{ni}), $1 \le i \le n$, with fixed design points $x_{ni} \in \mathcal{X}_n$ and

$$Y_{ni} \ = \ \mu_n(x_{ni}) + \epsilon_{ni},$$

where μ_n is an unknown regression function in $\mathcal{M}_n$ and $\epsilon_{n1}, \epsilon_{n2}, \ldots, \epsilon_{nn}$ are unobserved independent random errors with mean zero and unknown distribution $P_n \in \mathcal{P}_o(1) \cap \mathcal{P}^1(1)$. Two basic assumptions are:

(A.1) $\mathcal{M}_n(\boldsymbol{x}_n)$ is a closed subset of $\mathbb{R}^n$ for every $n \in \mathbb{N}$;
(A.2) $D_1(P_n, P) \to 0$ for some distribution $P \in \mathcal{P}_o(1) \cap \mathcal{P}^1(1)$.

We write $(\hat{\psi}_n, \hat{\mu}_n)$ for a maximizer of $L(g, \hat{P}_{n,m})$ over all pairs $(g, m) \in \mathcal{G} \times \mathcal{M}_n$ such that $\int e^{g(x)}\,dx = 1$ and $\int x e^{g(x)}\,dx = 0$, where $\hat{P}_{n,m}$ stands for the empirical distribution of the residuals $Y_{ni} - m(x_{ni})$, $1 \le i \le n$. We also need to consider its expectation measure

$$P_{n,m} := \mathbb{E}\hat{P}_{n,m} \ = \ P_n \star Q_{(\mu_n - m)(\boldsymbol{x}_n)}.$$

Furthermore we write

$$\|\boldsymbol{v}\|_n \ := \ \frac{1}{n}\sum_{i=1}^n |v_i| \quad \text{for } \boldsymbol{v} = (v_i)_{i=1}^n \in \mathbb{R}^n.$$

It is also convenient to metrize weak convergence. In Theorem 3.6 below we utilize the bounded Lipschitz distance: For probability distributions P, Q on the real line let

$$D_{BL}(P, Q) \ := \ \sup_{h \in \mathcal{H}_{BL}} \Big| \int h\,d(Q - P) \Big|,$$

where $\mathcal{H}_{BL}$ is the family of all functions $h : \mathbb{R} \to [-1, 1]$ such that $\big|h(x) - h(y)\big| \le |x - y|$ for all $x, y \in \mathbb{R}$.

THEOREM 3.6 (Consistency of regression estimator). *Let Assumptions* (A.1-2) *be satisfied. Suppose further that*
(A.3) *for arbitrary fixed $c > 0$,*

$$\sup_{m \in \mathcal{M}_n \,:\, \|(m - \mu_n)(\boldsymbol{x}_n)\|_n \le c} D_{BL}(\hat{P}_{n,m}, P_{n,m}) \ \to_p \ 0.$$

Then, with $f_n := \exp \circ \psi(\cdot \,|\, P_n)$ and $\hat{f}_n := \exp \circ \hat{\psi}_n$, the maximum likelihood estimator $(\hat{f}_n, \hat{\mu}_n)$ of (f_n, μ_n) exists with asymptotic probability one and satisfies

$$\int \big|\hat{f}_n(x) - f_n(x)\big|\,dx \ \to_p \ 0, \quad \big\|(\hat{\mu}_n - \mu_n)(\boldsymbol{x}_n)\big\|_n \ \to_p \ 0.$$

We know already that Assumption (A.1) is satisfied for multiple linear regression and isotonic regression. Assumption (A.2) is a generalization of assuming a fixed error distribution for all sample sizes. The crucial point, of course, is Assumption (A.3). In our two examples it is satisfied under mild conditions:

THEOREM 3.7 (Linear Regression). *Let $\mathcal{M}_n$ be the family of all affine functions on $\mathcal{X}_n := \mathbb{R}^{q(n)}$. If Assumption (A.2) is satisfied, then (A.3) follows from*

$$\lim_{n\to\infty} \frac{q(n)}{n} \;=\; 0.$$

THEOREM 3.8 (Isotonic Regression). *Let $\mathcal{M}_n$ be the set of all non-decreasing functions on an interval $\mathcal{X}_n \subset \mathbb{R}$. If Assumption (A.2) holds true, then (A.3) follows from*

$$\big\|\mu_n(\boldsymbol{x}_n)\big\|_n \;=\; O(1).$$

3.3. *Numerical examples.* We have conducted some basic simulation experiments for the proposed estimator in the context of linear and isotonic regression. In summary, these experiments suggest that the maximum likelihood estimator gives a rather accurate estimate $\hat{\psi}$ of the true error density even if n is only moderately large. Also, at least for various skewed error distributions, we obtain better estimates $\hat{\mu}$ of the regression functions than with the least squares method.

Below we present more detailed results for the special case of a centered gamma error distribution, with density $f(x) \sim (x + \sqrt{r}\sigma)^{r-1} e^{-(\sqrt{r}x/\sigma + r)}$ for $x \geq -\sqrt{r}\sigma$, where σ denotes the standard deviation and r the usual shape parameter of the gamma distribution. For the latter we consider the values $r = 1$ and 3.

An alternating algorithm. In order to maximize

$$\hat{\Lambda}(g, m) = \frac{1}{n}\sum_{i=1}^{n} g(Y_i - m(x_i)) - \int e^{g(x)}\,dx + 1$$

over all $(g, m) \in \mathcal{G} \times \mathcal{M}$, we propose an algorithm that alternates between maximizing this expression in g and in m.

The details are as follows. The algorithm is initialized with some starting function $m^{(0)} \in \mathcal{M}$ that satisfies $\sum_{i=1}^{n}(Y_i - m^{(0)}(x_i)) = 0$. In our implementation, $m^{(0)}$ is the least squares estimate of the regression function, but one could also use $m^{(0)} \equiv \bar{Y}$. Compute then the log-concave density estimate $g^{(0)}$ for the residuals $Y_i - m^{(0)}(x_i)$, $i = 1, 2, \ldots, n$ (see also Step C below). Then the following three steps are iterated for $k \geq 1$ until convergence:

(A) Determine

$$\tilde{m}^{(k)} \in \arg\max_{m \in \mathcal{M}} \sum_{i=1}^{n} g^{(k-1)}(Y_i - m(x_i)).$$

(B) Shift the function $\tilde{m}^{(k)}$ so that the sum of the residuals is zero, i.e. set

$$m^{(k)}(x) := \tilde{m}^{(k)}(x) + c^{(k)}$$

for all $x \in \mathcal{X}$, where $c^{(k)} := n^{-1} \sum_{i=1}^{n} (Y_i - \tilde{m}^{(k)}(x_i))$.

(C) Determine

$$g^{(k)} := \arg\max_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} g(Y_i - m^{(k)}(x_i)) - \int e^{g(x)}\, dx.$$

The algorithm is terminated if either $\sum_{i=1}^{n} |m^{(k)}(x_i) - m^{(k-1)}(x_i)| < \eta$ or if the maximal number $k_{\max}$ of iterations has been reached. Here $\eta > 0$ and $k_{\max} \in \mathbb{N}$ are constants that are specified in advance. We chose $\eta = 10^{-6}$ and $k_{\max} = 30$ in the examples below.

The maximization problem in Step C can be solved efficiently by an active set algorithm, which was described in Dümbgen *et al.* (2007). We use the implementation in the contributed package `logcondens` by Rufibach and Dümbgen (2006) in R (R Development Core Team, 2009). The maximization in Step A is carried out in a problem-specific way, depending on the set $\mathcal{M}$; see the paragraphs on linear and isotonic regression below.

It is currently not clear under what conditions our alternating algorithm converges to an overall maximizer of $\hat{\Lambda}$. However, it is easily seen that every time the cycle (A)–(C) is completed, the likelihood is at least as large as before. For at the end of Step A we obviously have $L(g^{(k-1)}, \tilde{m}^{(k)}) \geq L(g^{(k-1)}, m^{(k-1)})$. While in Step B the likelihood typically decreases, we can be sure, that after Step C is complete, it is at least as high as after Step A, because

$$L(g^{(k)}, m^{(k)}) \geq L(g^{(k-1)}(\,\cdot + c^{(k)}), m^{(k)}) = L(g^{(k-1)}, \tilde{m}^{(k)}).$$

Our experience with this algorithm shows that it typically converges rather quickly to some pair $(\hat{g}_{\mathrm{algo}}, \hat{m}_{\mathrm{algo}}) \in \mathcal{G} \times \mathcal{M}$. However, it may happen that this pair is not in $\arg\max_{(g,m)} \hat{\Lambda}(g,m)$, but that simply $\hat{g}_{\mathrm{algo}}$ maximizes $\hat{\Lambda}(g, \hat{m}_{\mathrm{algo}})$ in g and $\hat{m}_{\mathrm{algo}}$ maximizes $\hat{\Lambda}(\hat{g}_{\mathrm{algo}}, m)$ in m. See Figure 3 to observe how this behavior affects the performance of the estimator in the linear regression examples. A more reliable algorithm that can be shown always to converge to an element of $\arg\max_{(g,m)} \hat{\Lambda}(g,m)$ under reasonable conditions would be very desirable.

Linear Regression. In the context of linear regression the maximization over $\mathcal{M}$ in Step A of the alternating algorithm given a particular function $g \in \mathcal{G}$ can be rephrased as maximization of the concave function $\mathbb{R}^{q+1} \to \mathbb{R}$, $\theta \mapsto \sum_{i=1}^n g(Y_i - (1, x_i^\top)\theta)$. We use the general-purpose optimizer function `optim` in R that is based on an algorithm by Nelder and Mead (1965). Although more problem-specific methods may be faster, we found this function to be sufficiently fast and reliable for our first investigations.

For our simulation experiment we generate i.i.d. design points x_i from the uniform distribution on $[0, 3]$, and values y_i from the simple linear regression model

$$y_i = m_{a,b}(x_i) + \epsilon_i, \quad 1 \le i \le n,$$

where $m_{a,b}(x) := a + bx$ and the ϵ_i are i.i.d. (independent also of the x_i) and follow a centered gamma distribution with standard deviation σ and shape parameter r. As concrete values we chose $n = 200$, $a = 5$, $b = 2$, $\sigma = 1$, and $r \in \{1, 3\}$.

In Dümbgen *et al.* (2010) typical realizations are shown, together with the corresponding non-parametric maximum likelihood estimates of the regression line and the error density, which are quite close to the true curves (see also the realizations in Figure 4 for the isotonic regression case).

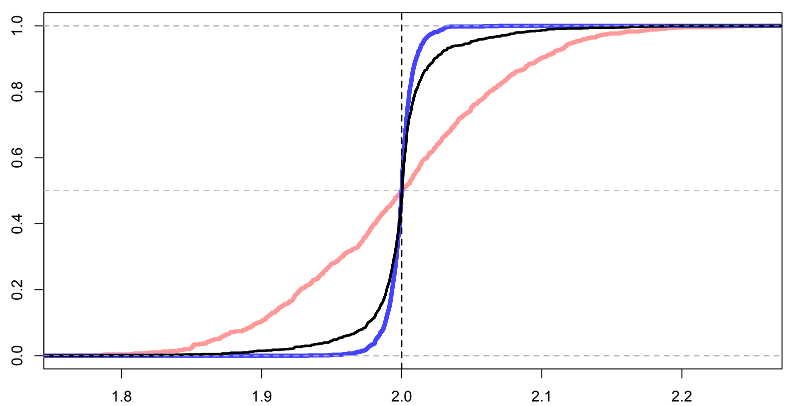

FIG 3. *Empirical cumulative distribution functions (ecdfs) of the computed estimators of the slope b based on 1000 simulations with centered exponentially distributed errors. Thick faint red: least squares estimator $\hat{b}_{LSQ}$; thick dark blue: exact maximum likelihood estimator $\hat{b}_{ML}$; thinner black: algorithmic "approximation" $\hat{b}_{algo}$ to maximum likelihood estimator.*

Figure 3 compares the ecdfs for three different estimators of the slope b in 1000 simulations of the above model with shape parameter $r = 1$. The corresponding picture for $r = 3$ is shown in Dümbgen *et al.* (2010) and has the same general flavor although the difference between the curves is much less pronounced. Table 1 gives estimates of certain distributional characteristics of the estimators for $r = 1$ and $r = 3$. In addition to the ordinary least squares estimator $\hat{b}_{\mathrm{LSQ}}$ (thick faint red), and the estimator $\hat{b}_{\mathrm{algo}}$ (thin black) that results from our algorithm, we have computed the true (semi-parametric) maximum likelihood estimator $\hat{b}_{\mathrm{ML}}$ (thick dark blue) in a rather laborious way. The function $\Phi : \mathbb{R} \to \mathbb{R}, b \mapsto \max_{g \in \mathcal{G}} \hat{\Lambda}(g, m_{a,b})$, which is independent of $a \in \mathbb{R}$, was evaluated (compare Step C of the alternating algorithm) at the points of a fine grid and then maximized in neighborhoods of the detected modes. Unfortunately, Φ turned out to be not unimodal in general (from the proof of Theorem 3.2 it can be seen to be continuous and coercive, however), so that there is no easy way of maximizing Φ in a less laborious way.

Figure 3 and Table 1 suggest that in the present examples $\hat{b}_{\mathrm{algo}}$ and $\hat{b}_{\mathrm{ML}}$ both perform clearly better than $\hat{b}_{\mathrm{LSQ}}$, especially in the extreme case $r = 1$. There is, however, a noticeable gap between the performance of $\hat{b}_{\mathrm{algo}}$ and $\hat{b}_{\mathrm{ML}}$. This gap stems from the fact that the alternating algorithm sometimes gets stuck in a "bi-conditional" maximum as mentioned earlier.

$r = 1$	$\hat{b}_{\mathrm{ML}}$	$\hat{b}_{\mathrm{algo}}$	$\hat{b}_{\mathrm{LSQ}}$
mean	1.9995	2.0001	1.9982
RMSE	0.0106	0.0327	0.0778
$q_{0.025}$	1.9762	1.9239	1.8506
$q_{0.975}$	2.0205	2.0769	2.1464

$r = 3$	$\hat{b}_{\mathrm{ML}}$	$\hat{b}_{\mathrm{algo}}$	$\hat{b}_{\mathrm{LSQ}}$
mean	1.9991	1.9997	2.0002
RMSE	0.0648	0.0713	0.0827
$q_{0.025}$	1.8747	1.8656	1.8315
$q_{0.975}$	2.1263	2.1357	2.1602

TABLE 1

Estimated characteristics for the three estimators. RMSE denotes the root mean squared error, q_α the α-quantile of the estimator.

Isotonic Regression. The maximization over $\mathcal{M}$ in Step A is possible within finitely many steps by means of a rather general form of the pool adjacent violators algorithm by Ayer *et al.* (1955); see Dümbgen *et al.* (2010) for more details.

For the simulation experiment we generate design points x_i independently from the uniform distribution on $[0, 3]$ again. Then we simulate values y_i from the isotonic regression model

$$y_i = \mu(x_i) + \epsilon_i, \quad 1 \le i \le n,$$

where $\mu(x) = \lfloor x + 0.5 \rfloor$, $n = 300$, and the ϵ_i are i.i.d. and follow a centered gamma distribution with standard deviation $1/2$ and shape parameter $r \in \{1, 3\}$.

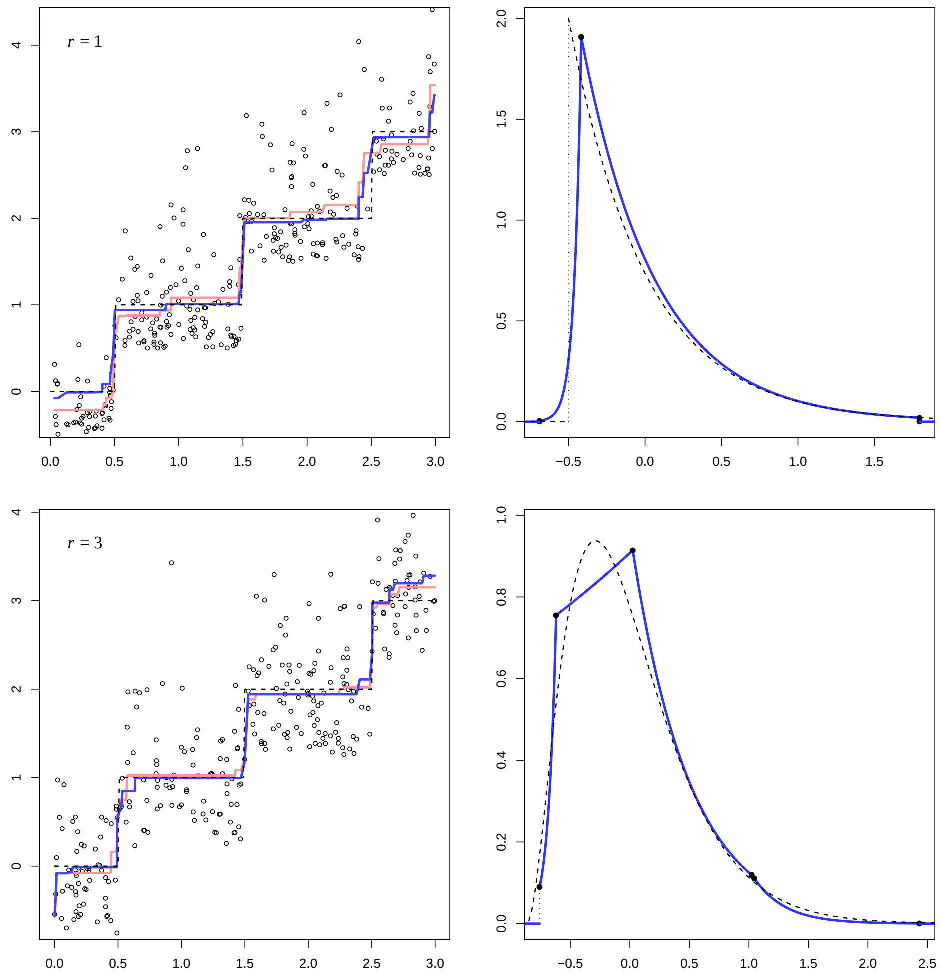

Fig 4. *Regression lines (left panel) and error densities (right panel) for isotonic regression with centered gamma distributed errors with shape parameters $r = 1$ (above) and $r = 3$ (below). Black dashed: true underlying curves; thick faint red: estimate obtained by least squares; thick dark blue: estimates obtained by the alternating algorithm.*

n

Figure 4 shows typical realizations for $r = 1$ and $r = 3$, respectively, each with the corresponding least squares estimate of the regression line (faint red) and our algorithmic estimate of the regression line and the error density (dark blue). We can see that the global structure in the function μ is detected rather better by the algorithmic estimator $\hat{m}_{\rm algo}$ than by $\hat{m}_{\rm LSQ}$.

Table 2 gives estimates of certain distributional characteristics of the mean absolute deviation $\|\hat{m}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{1,n} = \frac{1}{n}\sum_{i=1}^n |\hat{m}(x_i) - \mu(x_i)|$ and the root mean squared deviation $\|\hat{m}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{2,n} = \bigl(\frac{1}{n}\sum_{i=1}^n (\hat{m}(x_i) - \mu(x_i))^2\bigr)^{1/2}$ of the estimators $\hat{m}_{\rm algo}(\boldsymbol{x})$ and $\hat{m}_{\rm LSQ}(\boldsymbol{x})$ from the true values $\mu(\boldsymbol{x})$ based on 1000 simulations of the above models. Again the algorithmic estimator performs better than the least squares estimator, especially if $r = 1$. It should be noted that it is not to be expected that the algorithmic estimator realizes the true likelihood maximum in general; compare the relevant discussion at the end of the linear regression example.

$\boldsymbol{r = 1}$	$\|\hat{m}_{\rm algo}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{1,n}$	$\|\hat{m}_{\rm LSQ}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{1,n}$	$\|\hat{m}_{\rm algo}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{2,n}$	$\|\hat{m}_{\rm LSQ}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{2,n}$
mean	0.04713	0.08187	0.09146	0.12007
sd	0.01707	0.01860	0.02714	0.02472
$q_{0.8}$	0.06162	0.09748	0.11155	0.13958
$q_{0.95}$	0.07867	0.11367	0.13691	0.16248

$\boldsymbol{r = 3}$	$\|\hat{m}_{\rm algo}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{1,n}$	$\|\hat{m}_{\rm LSQ}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{1,n}$	$\|\hat{m}_{\rm algo}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{2,n}$	$\|\hat{m}_{\rm LSQ}(\boldsymbol{x}) - \mu(\boldsymbol{x})\|_{2,n}$
mean	0.07635	0.08154	0.11900	0.12034
sd	0.01849	0.01890	0.02560	0.02341
$q_{0.8}$	0.09221	0.09841	0.13935	0.14129
$q_{0.95}$	0.10881	0.11152	0.16352	0.15721

TABLE 2

Estimated characteristics for the mean absolute deviation and root mean squared deviation from the true values $\mu(\boldsymbol{x})$.

4. Proofs.

4.1. *Proofs for Section 2.*

PROOF OF LEMMA 2.1. It is well-known that

$$\begin{aligned}\operatorname{supp}(P) &:= \{x \in \mathbb{R}^d : P(U) > 0 \text{ for all open } U \subset \mathbb{R}^d \text{ containing } x\} \\ &= \bigcap\{S : S \subset \mathbb{R}^d \text{ closed}, P(S) = 1\}\end{aligned}$$

is itself a closed set with $P(\operatorname{supp}(P)) = 1$. The set $\operatorname{csupp}(P)$, being the intersection of all closed and convex sets $C \subset \mathbb{R}^d$ with $P(C) = 1$, is certainly closed and convex, too, and contains $\mathrm{supp}(P)$. Thus $P(\mathrm{csupp}(P)) = 1$. The equivalence of statements (a) to (c) can be seen as follows.

(a) $\Longrightarrow$ (b): Since hyperplanes are closed and convex sets, $P(H) = 1$ for some hyperplane H implies that $\mathrm{csupp}(P) \subset H$ and consequently that $\mathrm{interior}(\mathrm{csupp}(P)) = \emptyset$.

(b) $\Longrightarrow$ (c): $P(H) < 1$ for any hyperplane $H \subset \mathbb{R}^d$ implies that $\mathrm{supp}(P)$ cannot be contained in a hyperplane. Therefore $\mathrm{supp}(P)$ contains $d+1$ points $x_0, x_1, \ldots, x_d$ such that the simplex $\Delta := \mathrm{conv}(x_0, x_1, \ldots, x_d)$ has non-empty interior, where $\mathrm{conv}(\cdot)$ stands for convex hulls. Fix $z \in \mathrm{interior}(\Delta)$ and $\alpha \in (0, 1/2)$. Writing $\tilde{x}_i := (1-\alpha)x_i + \alpha z$, we define the contracted simplex

$$\tilde{\Delta} = \tilde{\Delta}(z, \alpha) \;:=\; \mathrm{conv}\big(\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_d\big)$$

which still has non-empty interior and hence positive Lebesgue measure. Defining the i-th corner simplex of $\tilde{\Delta}$ as

$$\Delta_i \;:=\; \{2\tilde{x}_i - x : x \in \tilde{\Delta}\},$$

it is easily seen that

$$2\tilde{x}_i - x_i \;=\; \frac{1-2\alpha}{1-\alpha}\,\tilde{x}_i + \frac{\alpha}{1-\alpha}\,z \;\in\; \mathrm{interior}(\tilde{\Delta})$$

and hence $x_i \in \mathrm{interior}(\Delta_i)$. Consequently, since all x_i belong to $\mathrm{supp}(P)$, the number $\eta := \min_{0 \le i \le d} P(\Delta_i)$ is strictly positive. Any convex closed set C with $P(C) > 1 - \eta$ must satisfy $C \cap \Delta_i \neq \emptyset$ for every i. Therefore it contains a simplex $\Delta^* := \mathrm{conv}\big(x_0^*, x_1^*, \ldots, x_d^*\big)$ with $x_i^* \in \Delta_i$. As shown by Schuhmacher *et al.* (2009) in the proof of their Lemma 4.1, this simplex Δ^* must contain $\tilde{\Delta}$. Hence $\mathrm{Leb}(C) \ge \mathrm{Leb}(\Delta^*) \ge \mathrm{Leb}(\tilde{\Delta}) > 0$. In total, we obtain that any convex closed set C with $\mathrm{Leb}(C) < \mathrm{Leb}(\tilde{\Delta})$ satisfies $P(C) \le 1 - \eta$, which yields statement (c).

(c) $\Longrightarrow$ (a): Statement (c) implies that $\mathrm{Leb}(\mathrm{csupp}(P)) > 0$. But the boundary of the convex set $\mathrm{csupp}(P)$ has Lebesgue measure zero (Lang, 1986). Therefore $\mathrm{interior}(\mathrm{csupp}(P)) \neq \emptyset$. $\square$

For the proof of Theorem 2.2 we need an elementary bound for the Lebesgue measure of level sets of log-concave distributions.

LEMMA 4.1. *Let $g \in \mathcal{G}$ be such that $\int e^{g(x)}\,dx = 1$. For real t define the level set $D_t := \{x \in \mathbb{R}^d : g(x) \ge t\}$. Then for $r < M \le \max_{x \in \mathbb{R}^d} g(x)$,*

$$\mathrm{Leb}(D_r) \;\le\; (M - r)^d e^{-M} \Big/ \int_0^{M-r} t^d e^{-t}\,dt.$$

PROOF OF LEMMA 4.1. Concavity of g entails that for $s \in [r, M]$,

$$\mathrm{Leb}(D_s) \ \geq \ \Bigl(\frac{M-s}{M-r}\Bigr)^d \mathrm{Leb}(D_r).$$

Thus by Fubini's theorem,

$$\begin{aligned} 1 \ \geq \ \int_{D_r} (e^{g(x)} - e^r)\, dx \ &\geq \ \int_{D_r} \int_r^M 1_{\{g(x) \geq s\}} e^s \, ds \, dx \\ &= \ \int_r^M e^s \, \mathrm{Leb}(D_s) \, ds \\ &\geq \ \mathrm{Leb}(D_r) \int_r^M \Bigl(\frac{M-s}{M-r}\Bigr)^d e^{M-(M-s)} \, ds \\ &= \ \mathrm{Leb}(D_r) \frac{e^M}{(M-r)^d} \int_0^{M-r} t^d e^{-t} \, dt. \end{aligned}$$

□

Another key ingredient for the proofs of Theorems 2.2 and 2.13 is a lemma on pointwise limits of sequences in $\mathcal{G}$. A complete proof of this auxiliary result is given by Dümbgen *et al.* (2010):

LEMMA 4.2. *Let $\bar{g}$ and $g_1, g_2, g_3, \ldots$ be functions in $\mathcal{G}$ such that $g_n \leq \bar{g}$ for all $n \in \mathbb{N}$. Further suppose that the set*

$$C \ := \ \Bigl\{ x \in \mathbb{R}^d : \liminf_{n \to \infty} \, g_n(x) > -\infty \Bigr\}$$

is non-empty. Then there exist a subsequence $(g_{n(k)})_k$ of $(g_n)_n$ and a function $g \in \mathcal{G}$ such that $C \subset \mathrm{dom}(g) = \{g > -\infty\}$ and

$$\begin{aligned} \lim_{k \to \infty, \ x \to y} g_{n(k)}(x) \ &= \ g(y) \quad \textit{for all } y \in \mathrm{interior}(\mathrm{dom}(g)), \\ \limsup_{k \to \infty, \ x \to y} g_{n(k)}(x) \ &\leq \ g(y) \ \leq \ \bar{g}(y) \quad \textit{for all } y \in \mathbb{R}^d. \end{aligned}$$

PROOF OF THEOREM 2.2. Suppose first that $\int \|x\| \, P(dx) = \infty$. Since any $g \in \mathcal{G}$ is majorized by $x \mapsto a - b\|x\|$ for suitable constants a and $b > 0$, this entails that $L(P) = -\infty$.

Secondly, suppose that $\int \|x\| \, P(dx) < \infty$ but $\mathrm{interior}(\mathrm{csupp}(P)) = \emptyset$. According to Lemma 2.1, the latter fact is equivalent to $P(H) = 1$ for some hyperplane $H \subset \mathbb{R}^d$. for $c \in \mathbb{R}$ define a function $g_c \in \mathcal{G}$ via $g_c(x) := c - \|x\|$ for $x \in H$ and $g_c(x) := -\infty$ for $x \notin H$. Then $L(g_c, P) = c - \int \|x\| \, P(dx) + 1 \to \infty$ as $c \to \infty$.

For the remainder of this proof suppose that $\int \|x\| P(dx)$ is finite and $\mathrm{csupp}(P)$ has non-empty interior. Since the concave function $h(x) = -\|x\|$

satisfies $\int h\,dP > -\infty$, we have $L(P) > -\infty$. When maximising $L(g,P)$ over all $g \in \mathcal{G}$ we may and do restrict our attention to functions $g \in \mathcal{G}$ such that $\int e^{g(x)}\,dx = 1$ and $\mathrm{dom}(g) = \{g > -\infty\} \subset \mathrm{csupp}(P)$. For if $0 < \int e^{g(x)}\,dx \neq 1$, then $L(g+c,P) > L(g,P)$ with $c := -\log\int e^{g(x)}\,dx$. If $\mathrm{dom}(g) \not\subset \mathrm{csupp}(P)$, replacing $g(x)$ with $-\infty$ for all $x \notin \mathrm{csupp}(P)$ would also increase $L(g,P)$ strictly. Let $\mathcal{G}(P)$ be the family of all $g \in \mathcal{G}$ with these properties.

Now we show that $L(P) < \infty$. Suppose that $g \in \mathcal{G}(P)$ is such that $M := \max_{x\in\mathbb{R}^d} g(x) > 0$. With $D_t := \{g \geq t\}$ and for $c > 0$ we get the bound

$$\begin{aligned} L(g,P) = \int g\,dP &\leq -cMP(\mathbb{R}^d \setminus D_{-cM}) + MP(D_{-cM}) \\ &= -(c+1)M\Big(\frac{c}{c+1} - P(D_{-cM})\Big). \end{aligned}$$

According to Lemma 4.1,

$$\begin{aligned} \mathrm{Leb}(D_{-cM}) &\leq (1+c)^d M^d e^{-M} \Big/ \int_0^{(1+c)M} t^d e^{-t}\,dt \\ &= (1+c)^d M^d e^{-M}/(d! + o(1)) \;\to\; 0 \quad (M \to \infty) \end{aligned}$$

for any fixed $c > 0$. But Lemma 2.1 entails that for sufficiently large c and sufficiently small $\delta > 0$,

$$\sup\{P(C) : C \subset \mathbb{R}^d \text{ closed and convex}, \mathrm{Leb}(C) \leq \delta\} \;<\; \frac{c}{c+1},$$

whence

$$L(g,P) \;\to\; -\infty \quad \text{as } \max_{x\in\mathbb{R}^d} g(x) \to \infty.$$

Note also that for any $g \in \mathcal{G}(P)$,

$$L(g,P) \;\leq\; \max_{x\in\mathbb{R}^d} g(x).$$

These considerations show that $L(P)$ is finite and, for suitable constants $M_o < M_*$, equals the supremum of $L(g,P)$ over all $g \in \mathcal{G}(P)$ such that $M_o \leq \max_x g(x) \leq M_*$.

Next we show the existence of a maximizer $g \in \mathcal{G}(P)$ of $L(\cdot,P)$. Let $(g_n)_n$ be a sequence of functions in $\mathcal{G}(P)$ such that $-\infty < L(g_n,P) \uparrow L(P)$ as $n \to \infty$, where $M_n := \max_{x\in\mathbb{R}^d} g_n(x) \in [M_o, M_*]$ for all $n \geq 1$. Now we show that

$$\inf_{n\geq 1} g_n(x_o) \;>\; -\infty \quad \text{for any } x_o \in \mathrm{interior}(\mathrm{csupp}(P)). \tag{4}$$

If $g_n(x_o) < M_n$, then x_o is not an interior point of the closed, convex set $\{g_n \ge g_n(x_o)\}$. Hence

$$\begin{aligned}\int g_n\,dP &\le g_n(x_o) + (M_n - g_n(x_o))P\{g_n \ge g_n(x_o)\} \\ &\le g_n(x_o) + (M_n - g_n(x_o))h(P, x_o) \\ &\le g_n(x_o)(1 - h(P, x_o)) + \max(M_n, 0)\end{aligned}$$

with $h(P, x_o) < 1$ defined in Lemma 2.11. In the case of $g_n(x_o) = M_n$ these inequalities are true as well. Thus

$$g_n(x_o) \ \ge\ -\frac{\max(M_n, 0) - L(g_n, P)}{1 - h(P, x_o)} \ \ge\ -\frac{\max(M_*, 0) - L(g_1, P)}{1 - h(P, x_o)},$$

which establishes (4). Combining (4) with $g_n \le M_*$, we may deduce from Lemma 3.3 of Schuhmacher *et al.* (2009) that there exist constants a and $b > 0$ such that

$$g_n(x) \ \le\ a - b\|x\| \quad \text{for all } n \in \mathbb{N}, x \in \mathbb{R}^d. \tag{5}$$

The inequalities (4-5) and Lemma 4.2 with $C \supset \mathrm{interior}(\mathrm{csupp}(P))$ and $\bar{g}(x) := a - b\|x\|$ imply existence of a function $\psi \in \mathcal{G}$ and a subsequence $(g_{n(k)})_k$ of $(g_n)_n$ such that $\psi = -\infty$ on $\mathbb{R}^d \setminus \mathrm{csupp}(P)$ and

$$\begin{aligned}\limsup_{k\to\infty} g_{n(k)}(x) &\le \psi(x) \ \le\ a - b\|x\| \quad \text{for all } x \in \mathbb{R}^d, \\ \lim_{k\to\infty} g_{n(k)}(x) &= \psi(x) \ >\ -\infty \quad \text{for all } x \in \mathrm{interior}(\mathrm{csupp}(P)).\end{aligned}$$

Since the boundary of csupp(P) has Lebesgue measure zero, it follows from dominated convergence that $\int e^{\psi(x)}\,dx = 1$. Moreover, applying Fatou's lemma to the nonnegative functions $x \mapsto a - b\|x\| - g_{n(k)}(x)$ yields

$$\limsup_{k\to\infty} \int g_{n(k)}\,dP \ \le\ \int \psi\,dP.$$

Hence

$$L(P) \ \ge\ L(\psi, P) \ \ge\ \limsup_{k\to\infty} L(g_{n(k)}, P) \ =\ L(P)$$

and thus $L(\psi, P) = L(P)$.

Uniqueness of the maximizer ψ follows essentially from strict convexity of the exponential function: If $\tilde{\psi} \in \mathcal{G}(P)$ with $L(\tilde{\psi}, P) > -\infty$, then $L\big((1-t)\psi + t\tilde{\psi}, P\big)$ is strictly concave in $t \in [0, 1]$, unless $\mathrm{Leb}\{\psi \ne \tilde{\psi}\} = 0$. But for $\psi, \tilde{\psi} \in \mathcal{G}(P)$, the latter requirement is equivalent to $\psi = \tilde{\psi}$ everywhere. $\square$

In our proofs of Theorems 2.5 and 2.13 we utilize a special approximation scheme for functions in $\mathcal{G}$; see Dümbgen *et al.* (2010) for a proof:

LEMMA 4.3. *For any function $g \in \mathcal{G}$ with non-empty domain and any parameter $\epsilon > 0$ set*

$$g^{(\epsilon)}(x) \ := \ \inf_{(v,c)} (v^\top x + c)$$

with the infimum taken over all $(v, c) \in \mathbb{R}^d \times \mathbb{R}$ such that $\|v\| \le \epsilon^{-1}$ and $g(y) \le v^\top y + c$ for all $y \in \mathbb{R}^d$. This defines a function $g^{(\epsilon)} \in \mathcal{G}$ which is real-valued and Lipschitz-continuous with constant ϵ^{-1}. Moreover, it satisfies $g^{(\epsilon)} \ge g$ with equality if, and only if, g is real-valued and Lipschitz-continuous with constant ϵ^{-1}. In general, $g^{(\epsilon)} \downarrow g$ pointwise as $\epsilon \downarrow 0$.

PROOF OF THEOREM 2.5. Suppose first that $g = \psi(\cdot \,|\, P)$. Then it follows from (3) and Fubini's theorem that

$$\begin{aligned}
0 \ &= \ \int_{\mathbb{R}} x \, (P - \tilde{P})(dx) \\
&= \ \int_{\mathbb{R}} \int_{\mathbb{R}} (1_{\{0<t<x\}} - 1_{\{x \le t \le 0\}}) \, dt \, (P - \tilde{P})(dx) \\
&= \ \int_{\mathbb{R}} (1\{0 < t\}(\tilde{F} - F)(t) - 1_{\{t \le 0\}}(F - \tilde{F})(t)) \, dt \\
&= \ \int_{\mathbb{R}} (\tilde{F} - F)(t) \, dt.
\end{aligned}$$

Moreover, for any $x \in \mathbb{R}$, the function $s \mapsto (s - x)^+$ is convex, so that (2) yields

$$\begin{aligned}
0 \ &\le \ \int_{\mathbb{R}} (s - x)^+ \, (P - \tilde{P})(ds) \\
&= \ \int_{\mathbb{R}} \int_{\mathbb{R}} 1_{\{x \le t < s\}} \, dt \, (P - \tilde{P})(ds) \\
&= \ \int_{\mathbb{R}} 1\{t \ge x\}(F - \tilde{F})(t) \, dt \\
&= \ -\int_{-\infty}^{x} (\tilde{F} - F)(t) \, dt.
\end{aligned}$$

It remains to be shown that $\int_{-\infty}^{x} (\tilde{F} - F)(t) \, dt \ge 0$ for $x \in \mathcal{S}(g)$. Suppose first that $x \in \mathrm{interior}(\mathrm{dom}(g))$. Note that $g' := g'(\cdot\, +)$ is non-increasing on the interior of $\mathrm{dom}(g)$ with

$$g(x_2) - g(x_1) \ = \ \int_{x_1}^{x_2} g'(u) \, du \quad \text{for } x_1, x_2 \in \mathrm{interior}(\mathrm{dom}(g)) \text{ with } x_1 < x_2.$$

Moreover, $x \in \mathcal{S}(g)$ implies that $g'(x-\delta) > g'(x+\delta)$ for all $\delta > 0$ satisfying $x \pm \delta \in \mathrm{interior}(\mathrm{dom}(g))$. For such $\delta > 0$ we define

$$H_\delta(s) \ := \ \int_{-\infty}^s H_\delta'(u)\, du$$

with

$$H_\delta'(u) \ := \ \begin{cases} 0 & \text{for } u \le x - \delta, \\ \dfrac{g'(x-\delta) - g'(u)}{g'(x-\delta) - g'(x+\delta)} & \text{for } x - \delta < u \le x + \delta, \\ 1 & \text{for } u \ge x + \delta. \end{cases}$$

One can easily verify that $g + tH_\delta$ is upper semicontinuous and concave whenever $0 < t \le g'(x-\delta) - g'(x+\delta)$. In case of $t < -\inf_{u\in\mathbb{R}} g'(u)$ it is also coercive. Thus it follows from (1) that

$$\begin{aligned} 0 \ &\le \ \int_{\mathbb{R}} H_\delta(s)\, (\tilde{P} - P)(ds) \\ &\to \ \int_{\mathbb{R}} (s-x)^+\, (\tilde{P} - P)(ds) \quad (\delta \downarrow 0) \\ &= \ \int_{-\infty}^x (\tilde{F} - F)(t)\, dt. \end{aligned}$$

When $x \in \mathcal{S}(g)$ is the left or right endpoint of $\mathrm{dom}(g)$, we define $\Delta(s) := (s-x)^+$ and conclude analogously that $\int_{-\infty}^x (\tilde{F} - F)(t)\, dt \ge 0$.

Now suppose that the distribution function $\tilde{F}$ with log-density $g \in \mathcal{G}$ satisfies the integral (in)equalities stated in Theorem 2.5. Let $\Delta : \mathbb{R} \to \mathbb{R}$ be Lipschitz-continuous with constant L, so that for arbitrary $x, y \in \mathbb{R}$ with $x < y$,

$$\Delta(y) - \Delta(x) \ = \ \int_x^y \Delta'(t)\, dt$$

for some measurable function $\Delta' : \mathbb{R} \to [-L, L]$. Then

$$\begin{aligned} \int \Delta\, d(P - \tilde{P}) \ &= \ \int \big(\Delta(x) - \Delta(0)\big)\, (P - \tilde{P})(dx) \\ &= \ \int_{\mathbb{R}} \int_{\mathbb{R}} \Delta'(t) \big(1_{\{0 \le t < x\}} - 1_{\{x \le t < 0\}}\big)\, dt\, (P - \tilde{P})(dx) \\ &= \ \int_{\mathbb{R}} \Delta'(t) \big(1_{\{0 \le t\}} (\tilde{F} - F)(t) - 1_{\{t < 0\}} (F - \tilde{F})(t)\big)\, dt \\ &= \ \int_{\mathbb{R}} \Delta'(t) (\tilde{F} - F)(t)\, dt. \end{aligned}$$

Since $\int (\tilde{F} - F)(t)\,dt = 0$, we may continue with

$$\begin{aligned}\int \Delta\,d(P - \tilde{P}) &= \int_{\mathbb{R}} (\Delta'(t) + L)(\tilde{F} - F)(t)\,dt \\ &= \int_{\mathbb{R}} \int_{-L}^{L} 1_{\{s<\Delta'(t)\}}\,ds\,(\tilde{F} - F)(t)\,dt \\ &= \int_{-L}^{L} \int_{A(\Delta',s)} (\tilde{F} - F)(t)\,dt\,ds,\end{aligned}$$

with $A(\Delta', s) := \{t \in \mathbb{R} : \Delta'(t) > s\}$. Now we apply this representation to the function $\Delta := g^{(\epsilon)}$ for some $\epsilon > 0$, i.e. $L = \epsilon^{-1}$. Here one can show that $A(\Delta', s)$ equals either $\emptyset$ or $\mathbb{R}$ or a half-line with right endpoint $a(g, s) = \min\{t \in \mathbb{R} : g'(t+) \le s\}$. But this entails that $a(g, s) \in \mathcal{S}(g)$, whence $\int_{A(\Delta',s)} (\tilde{F} - F)(t)\,dt = 0$ for all $s \in (-L, L)$. Consequently,

$$\int g^{(\epsilon)}\,d(P - \tilde{P}) \;=\; 0.$$

If we consider $\Delta := \psi^{(\epsilon)}$ with $\psi := \psi(\cdot\,|\,P)$, the sets $A(\Delta', s)$ are still half-lines with right endpoint or empty or equal to $\mathbb{R}$. Thus $\int_{A(\Delta',s)} (\tilde{F} - F)(t)\,dt \le 0$ for all $s \in (-L, L)$, whence

$$\int \psi^{(\epsilon)}\,d(P - \tilde{P}) \;\le\; 0.$$

Since $g^{(\epsilon)} \downarrow g$ and $\psi^{(\epsilon)} \downarrow \psi$ as $\epsilon \downarrow 0$, and since $\int g\,d\tilde{P}$ and $\int \psi\,dP$ exist in $\mathbb{R}$, we can deduce from monotone convergence that

$$\int g\,d(P - \tilde{P}) \;=\; 0 \;\ge\; \int \psi\,d(P - \tilde{P}).$$

Since $\int e^{g(x)}\,dx = \int e^{\psi(x)}\,dx = 1$, this entails that

$$L(g, P) \;=\; L(g, \tilde{P}) \;\ge\; L(\psi, \tilde{P}) \;\ge\; L(\psi, P),$$

where the first displayed inequality follows from log-concavity of $\tilde{P}$ with log-density g. Thus $g = \psi$. $\square$

Proof of Lemma 2.11. If C is closed and convex with $x \notin \mathrm{interior}(C)$, then there exists a unit vector $u \in \mathbb{R}^d$ such that C is contained in the closed halfspace

$$H(x, u) \;:=\; \{y \in \mathbb{R}^d : u^\top y \le u^\top x\}.$$

Since $x \in \partial H(x,u)$, this shows that $h(P,x)$ is the supremum of $P(H(x,u))$ over all unit vectors $u \in \mathbb{R}^d$. Indeed,

$$h(P,x) \ = \ \max_{u \in \mathbb{R}^d : \|u\|=1} P(H(x,u)).$$

For if $(u_n)_n$ is a sequence of unit vectors such that $\lim_{n\to\infty} P(H(x,u_n))$ equals $h(P,x)$, we may replace it with a subsequence, if necessary, such that $u := \lim_{n\to\infty} u_n$ exists. Then it follows from Fatou's lemma that $h(P,x)$ equals

$$1 - \lim_{n\to\infty} \int 1_{\{u_n^\top y > u_n^\top x\}} \, P(dy) \ \le \ 1 - \int 1_{\{u^\top y > u^\top x\}} \, P(dy) \ = \ P(H(x,u)).$$

Now suppose that $x \notin \mathrm{interior}(\mathrm{csupp}(P))$. Then $h(P,x) \ge P(\mathrm{csupp}(P)) = 1$ by definition of $h(P,x)$ and $\mathrm{csupp}(P)$. On the other hand, if $h(P,x) = 1$, then $P(H(x,u)) = 1$ for a suitable unit vector u, so $\mathrm{csupp}(P) \subset H(x,u)$ and $x \notin \mathrm{interior}(H(x,u)) \supset \mathrm{interior}(\mathrm{csupp}(P))$.

Now let $(P_n)_n$ be a sequence in $\mathcal{P}$ converging weakly to P, and let $x \in \mathbb{R}^d$ be such that $h(P,x) < 1$. Then there exist unit vectors $u_n \in \mathbb{R}^d$ such that $h(P_n,x) = P_n(H(x,u_n))$ for all $n \in \mathbb{N}$. For a suitable subsequence, we may assume that $(u_{n(k)})_k$ converges to a unit vector u and $\lim_{k\to\infty} h(P_{n(k)},x) = \limsup_{n\to\infty} h(P_n,x)$. Then for any fixed $\delta > 0$,

$$\begin{aligned}
&P\{y \in \mathbb{R}^d : u^\top(y-x) \le \delta\} \\
&\quad \ge \ \limsup_{k\to\infty} \ P_{n(k)}\{y \in \mathbb{R}^d : u^\top(y-x) \le \delta\} \\
&\quad \ge \ \limsup_{k\to\infty} \ P_{n(k)}\{y \in \mathbb{R}^d : u^\top(y-x) \le \delta, u_{n(k)}^\top(y-x) \le 0\} \\
&\quad \ge \ \lim_{k\to\infty} \ P_{n(k)}\{y \in \mathbb{R}^d : u_{n(k)}^\top(y-x) \le 0\} \\
&\qquad\quad - \ \limsup_{k\to\infty} P_{n(k)}\{y \in \mathbb{R}^d : (u - u_{n(k)})^\top(y-x) > \delta\} \\
&\quad \ge \ \limsup_{n\to\infty} \ h(P_n,x) \\
&\qquad\quad - \ \limsup_{k\to\infty} P_{n(k)}\{y \in \mathbb{R}^d : \|y-x\| > \delta/\|u - u_{n(k)}\|\} \\
&\quad = \ \limsup_{n\to\infty} \ h(P_n,x),
\end{aligned}$$

because the sequence $(P_n)_n$ is tight and $\delta/\|u - u_{n(k)}\| \to \infty$ as $k \to \infty$. As $\delta \downarrow 0$, the left hand side tends to $P(H(x,u)) \le h(P,x)$, whence $h(P,x) \ge \limsup_{n\to\infty} h(P_n,x)$. $\square$

Theorem 2.12 and the second part of Theorem 2.13 are a consequence of the following result:

THEOREM 4.4. *Let $(P_n)_n$ be a sequence of distributions in $\mathcal{P}_o$ converging weakly to some $P \in \mathcal{P}_o$ such that*

$$L(P_n) \ \to \ \lambda \in [-\infty, \infty] \quad \text{and} \quad \int \|x\| \, P_n(dx) \ \to \ \gamma \in [0, \infty]$$

as $n \to \infty$. Then $\gamma \geq \int \|x\| \, P(dx)$, and $\lambda > -\infty$ if and only if $\gamma < \infty$. Moreover,

$$\lambda \begin{cases} < \ L(P) & \text{if } \gamma > \int \|x\| \, P(dx), \\ = \ L(P) \in \mathbb{R} & \text{if } \gamma = \int \|x\| \, P(dx) < \infty. \end{cases}$$

In the latter case, the densities $f := \exp \circ \psi(\cdot \,|\, P)$ and $f_n := \exp \circ \psi(\cdot \,|\, P_n)$ are well-defined for sufficiently large n and satisfy

$$\begin{aligned} \lim_{n \to \infty, \ x \to y} f_n(x) \ &= \ f(y) \quad \text{for all } y \in \mathbb{R}^d \setminus \partial\{f > 0\}, \\ \limsup_{n \to \infty, \ x \to y} f_n(x) \ &\leq \ f(y) \quad \text{for } y \in \partial\{f > 0\}, \\ \lim_{n \to \infty} \int \bigl| f_n(x) - f(x) \bigr| \, dx \ &= \ 0. \end{aligned}$$

Before presenting the proof of this result, let us recall two elementary facts about weak convergence and unbounded functions:

LEMMA 4.5. *Suppose that $(P_n)_n$ is a sequence in $\mathcal{P}$ converging weakly to some distribution P. If h is a nonnegative and continuous function on $\mathbb{R}^d$, then*

$$\liminf_{n \to \infty} \int h \, dP_n \ \geq \ \int h \, dP.$$

If the stronger statement $\lim_{n \to \infty} \int h \, dP_n = \int h \, dP < \infty$ holds, then

$$\lim_{n \to \infty} \int f \, dP_n \ = \ \int f \, dP$$

for any continuous function f on $\mathbb{R}^d$ such that $|f|/(1 + h)$ is bounded.

PROOF OF THEOREM 4.4. The asserted inequality $\gamma \geq \int \|x\| \, P(dx)$ follows from the first part of Lemma 4.5 with $h(x) := \|x\|$.

Suppose that $\gamma < \infty$. Then with $g(x) := -\|x\|$,

$$\lambda \ \geq \ \lim_{n \to \infty} L(g, P_n) \ = \ -\gamma - \int e^{-\|x\|} \, dx + 1 \ > \ -\infty.$$

In other words, $\lambda = -\infty$ entails that $\gamma = \infty$.

From now on suppose that $\lambda > -\infty$, and without loss of generality let $L(P_n) > -\infty$ for all $n \in \mathbb{N}$. We have to show that $\gamma < \infty$ and that $\lambda \leq L(P)$

with equality if, and only if, $\gamma = \int \|x\| \, P(dx)$. To this end we analyze the functions $\psi_n := \psi(\cdot \,|\, P_n)$ and their maxima $M_n := \max_{x \in \mathbb{R}^d} \psi_n(x)$. First of all,

$$(M_n)_n \text{ is bounded.} \tag{6}$$

This can be verified as follows: Since $L(P_n) = \int \psi_n \, dP_n \le M_n$, the sequence $(M_n)_n$ satisfies $\liminf_{n\to\infty} M_n \ge \lambda$. With similar arguments as in the proof of Theorem 2.2 one can deduce that $(M_n)_n$ is bounded from above, provided that

$$\limsup_{n\to\infty} \, P_n(C_n) \;<\; 1$$

for any sequence of closed and convex sets $C_n \subset \mathbb{R}^d$ with $\lim_n \mathrm{Leb}(C_n) = 0$. To this end recall our proof of Lemma 2.1: There exist a simplex $\tilde{\Delta} = \mathrm{conv}(\tilde{x}_0, \ldots, \tilde{x}_d)$ with positive Lebesgue measure and open sets U_0, U_1, ..., U_d with $P(U_j) \ge \eta > 0$ for $0 \le j \le d$, such that $\tilde{\Delta} \subset C$ for any convex set C with $C \cap U_j \ne \emptyset$ for $0 \le j \le d$. But $\liminf_n P_n(U_j) \ge P(U_j) \ge \eta$ for all j. Hence $\mathrm{Leb}(C_n) < \mathrm{Leb}(\tilde{\Delta})$ entails that $P_n(C_n) \le 1 - \min_{0 \le j \le d} P_n(U_j) \le 1 - \eta + o(1)$ as $n \to \infty$.

Another key property of the functions ψ_n is that

$$\liminf_{n\to\infty} \, \psi_n(x_o) \;>\; -\infty \quad \text{for any } x_o \in \mathrm{interior}(\mathrm{csupp}(P)). \tag{7}$$

For

$$L(P_n) = \int \psi_n \, dP_n \;\le\; \psi_n(x_o) + (M_n - \psi_n(x_o)) h(P_n, x_o),$$

whence as $n \to \infty$,

$$\psi_n(x_o) \;\ge\; -\frac{\max(M_n, 0) - L(P_n)}{1 - h(P_n, x_o)} \;\ge\; -\frac{\limsup_{\ell\to\infty} \max(M_\ell, 0) - \lambda}{1 - h(P, x_o)} + o(1)$$

by virtue of Lemma 2.11. Combining (4) with (6) we may again deduce that there exist constants a and $b > 0$ such that

$$\psi_n(x) \;\le\; a - b\|x\| \quad \text{for all } n \in \mathbb{N}, x \in \mathbb{R}^d. \tag{8}$$

As in the proof of Theorem 2.2 we can replace $(P_n)_n$ with a subsequence such that for suitable constants a, $b > 0$ and a function $\tilde{\psi} \in \mathcal{G}$ the following conditions are met: $\mathrm{interior}(\mathrm{csupp}(P)) \subset \mathrm{dom}(\tilde{\psi})$ and

$$\begin{aligned}
\psi_n(y), \tilde{\psi}(y) \;&\le\; a - b\|y\| \quad \text{for all } y \in \mathbb{R}^d, n \in \mathbb{N}, \\
\lim_{n\to\infty,\; x\to y} \psi_n(x) \;&=\; \tilde{\psi}(y) \quad \text{for all } y \in \mathrm{interior}(\mathrm{dom}(\tilde{\psi})) \\
\limsup_{n\to\infty,\; x\to y} \psi_n(x) \;&\le\; \tilde{\psi}(y) \quad \text{for all } y \in \mathbb{R}^d.
\end{aligned}$$

In particular,

$$\lambda = \lim_{n\to\infty} \int \psi_n \, dP_n \ \le \ \lim_{n\to\infty} \int (a - b\|x\|) \, P_n(dx) \ = \ a - b\gamma,$$

whence

$$\gamma \ < \ \infty.$$

Moreover, $\int \exp(\tilde{\psi}(x)) \, dx = \lim_{n\to\infty} \int \exp(\psi_n(x)) \, dx = 1$ by dominated convergence.

By Skorohod's theorem, there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with random variables $X_n \sim P_n$ and $X \sim P$ such that $\lim_{n\to\infty} X_n = X$ almost surely. Hence Fatou's lemma, applied to the random variables $H_n := a - b\|X_n\| - \psi_n(X_n)$, yields

$$\begin{aligned}
\lambda &= \lim_{n\to\infty} \int \psi_n \, dP_n \\
&= \lim_{n\to\infty} \Big(\int (a - b\|x\|) \, dP_n - \mathbb{E}(H_n) \Big) \\
&= a - b\gamma - \liminf_{n\to\infty} \mathbb{E}(H_n) \\
&\le a - b\gamma - \mathbb{E}\Big(\liminf_{n\to\infty} H_n \Big) \\
&\le a - b\gamma - \mathbb{E}\big(a - b\|X\| - \tilde{\psi}(X)\big) \\
&= b\Big(\int \|x\| \, P(dx) - \gamma \Big) + \int \tilde{\psi}(x) \, P(dx) \\
&\le b\Big(\int \|x\| \, P(dx) - \gamma \Big) + L(P).
\end{aligned}$$

Thus $\lambda < L(P)$ if $\gamma > \int \|x\| \, P(dx)$.

It remains to analyze the case $\gamma = \int \|x\| \, P(dx) < \infty$. In this case, $\lambda \le L(\tilde{\psi}, P) \le L(P)$, and it remains to show that $\lambda \ge L(P)$ which would entail that $\tilde{\psi}$ equals the unique maximizer $\psi := \psi(\cdot \,|\, P)$ of $L(\cdot, P)$ over $\mathcal{G}$. With the approximations $\psi^{(1)} \ge \psi^{(\epsilon)} \ge \psi$, $0 < \epsilon \le 1$, introduced in Lemma 4.3, it follows from their Lipschitz-continuity and Lemma 4.5 that

$$\begin{aligned}
\lambda = \lim_{n\to\infty} L(\psi_n, P_n) \ &\ge \ \lim_{n\to\infty} L(\psi^{(\epsilon)}, P_n) \\
&= \lim_{n\to\infty} \int \psi^{(\epsilon)} \, dP_n - \int \exp(\psi^{(\epsilon)}(x)) \, dx + 1 \\
&= \int \psi^{(\epsilon)} \, dP - \int \exp(\psi^{(\epsilon)}(x)) \, dx + 1.
\end{aligned}$$

By monotone convergence, applied to the functions $\psi^{(1)} - \psi^{(\epsilon)}$, and dominated convergence, applied to $\exp \circ \psi^{(\epsilon)}$,

$$\lim_{\epsilon \downarrow 0} L(\psi^{(\epsilon)}, P) \ = \ L(\psi, P),$$

whence $\lambda \geq L(P)$.

Note that the probability densities $f = \exp \circ \psi$ and $f_n = \exp \circ \psi_n$ obviously satisfy

$$\begin{aligned} \lim_{n\to\infty,\ x\to y} f_n(x) &= f(y) \quad \text{for all } y \in \mathbb{R}^d \setminus \partial\{f > 0\} \\ \limsup_{n\to\infty,\ x\to y} f_n(x) &\leq f(y) \quad \text{for all } y \in \partial\{f > 0\}. \end{aligned}$$

In particular, $(f_n)_n$ converges to f almost everywhere w.r.t. Lebesgue measure, whence $\int |f_n(x) - f(x)|\, dx \to 0$.

The only problem is that we established these properties only for a *subsequence* of the original sequence $(P_n)_n$. But suppose that any of the assertions about the densisties f_n would be false. Then one could replace the initial sequence $(P_n)_n$ from the start with a subsequence such that one of the following three conditions is satisfied:

(i) $\lim_{n\to\infty} f_n(x_n) > f(y)$ for some sequence $(x_n)_n$ in $\mathbb{R}^d$ converging to some point y;

(ii) $\lim_{n\to\infty} f_n(x_n) < f(y)$ for some sequence $(x_n)_n$ in $\mathbb{R}^d$ converging to some point $y \in \operatorname{interior}\{f > 0\}$;

(iii) $\lim_{n\to\infty} \int |f_n(x) - f(x)|\, dx > 0$.

Since any of these properties (i-iii) are transmitted to subsequences of $(P_n)_n$, we would obtain a contradiction. □

PROOF OF THEOREM 2.13. The assertions of this theorem are essentially covered by Theorem 4.4 as long as $P \in \mathcal{P}_o \cap \mathcal{P}^1$. It only remains to show that $L(P_n) \to \infty$ if $D_1(P_n, P) \to 0$ for some $P \in \mathcal{P}^1 \setminus \mathcal{P}_o$. Thus $\int \|x\|\, P(dx) < \infty$ and $P(H) = 1$ for a hyperplane $H = \{x \in \mathbb{R}^d : u^\top x = r\}$ with a unit vector $u \in \mathbb{R}^d$ and some $r \in \mathbb{R}$. For $k \geq 1$ we define $g_k \in \mathcal{G}$ via

$$g_k(x) \ := \ -\|a_k + B_k x\| + \log(k),$$

where $B_k := I - uu^\top + kuu^\top$ is a real, $d \times d$ matrix and $a_k := -kru$. Note that $\det(B_k) = k$ and $g_k(x) = \log(k) - \|x\|$ for $x \in H$. Thus

$$\begin{aligned} L(g_k, P_n) \ \to \ L(g_k, P) &= \log(k) - \int \|x\|\, P(dx) + \int \exp(g_k(x))\, dx \\ &= \log(k) - \int \|x\|\, P(dx) + \int \exp(-\|x\|)\, dx. \end{aligned}$$

Clearly the right hand side tends to infinity as $k \to \infty$, whence $L(P_n) \to \infty$ as $n \to \infty$. □

4.2. *Proofs for Section 3.*

PROOF OF LEMMA 3.1. We may assume that $P \in \mathcal{P}_o$ because otherwise the assertion is trivial (with the convention that $\log(0) := -\infty$). By definition of a median of P, there exist two probability measures P_l on $(-\infty, \mathrm{Med}(P)]$ and P_r on $[\mathrm{Med}(P), \infty)$ such that $P = 2^{-1}(P_l + P_r)$. With

$$\mu_l := \int x\, P_l(dx) \ \le \ \mathrm{Med}(P) \ \le \ \int x\, P_r(dx) =: \mu_r,$$

it follows from $P \in \mathcal{P}_o \cap \mathcal{P}^1$ that

$$0 \ < \ \mu_r - \mu_l \ = \ 2\int |x - \mathrm{Med}(P)|\, P(dx) \ < \ \infty.$$

For any $g \in \mathcal{G}$ it follows from Jensen's inequality that

$$\int g\, dP \ = \ 2^{-1}\Bigl(\int g\, dP_l + \int g\, dP_r\Bigr) \ \le \ 2^{-1}\bigl(g(\mu_l) + g(\mu_r)\bigr) \ = \ \int g\, dP_o$$

with the discrete disribution $P_o := \mathrm{Unif}\{\mu_l, \mu_r\}$. Thus $L(P) \le L(P_o) = -\log(\mu_r - \mu_l)$, because one easily verifies that $\psi(x \,|\, P_o)$ equals $(\mu_r - \mu_l)^{-1}$ for $x \in [\mu_l, \mu_r]$ and $-\infty$ otherwise.

The second asserted inequality is just a consequence of

$$\begin{aligned}\int |x - \mu(P)|\, P(dx) \ &\le \ \int |x - \mathrm{Med}(P)|\, P(dx) + |\mu(P) - \mathrm{Med}(P)| \\ &\le \ 2\int |x - \mathrm{Med}(P)|\, P(dx).\end{aligned}$$

□

PROOF OF THEOREM 3.2. Note that $\boldsymbol{v} \mapsto \hat{P}_{\boldsymbol{v}}$ defines a continuous mapping from $\mathbb{R}^n$ into the space of probability distributions on $\mathbb{R}$ with finite first moment, equipped with Mallows' distance D_1. Moreover, by our assumption that $\boldsymbol{Y} \notin \mathcal{M}(\boldsymbol{x})$, none of the distributions $\hat{P}_{m(\boldsymbol{x})}$, $m \in \mathcal{M}$, degenerates to a Dirac measure. According to Theorem 2.13, the mapping $\boldsymbol{v} \mapsto L(\hat{P}_{\boldsymbol{v}})$ is thus continuous from $\mathcal{M}(\boldsymbol{x})$ into $\mathbb{R}$.

When proving existence of a maximizer, as explained in Subsection 3.1, we may restrict our attention to the closed subset

$$\mathcal{M}(\boldsymbol{x}, \bar{Y}) \ := \ \Bigl\{\boldsymbol{v} \in \mathcal{M}(\boldsymbol{x}) : \bar{v} = \bar{Y}\Bigr\}$$

of $\mathcal{M}(\boldsymbol{x})$, where generally $\bar{w}$ denotes the arithmetic mean $n^{-1}\sum_{i=1}^n w_i$ for a vector $\boldsymbol{w} \in \mathbb{R}^n$. But for $\boldsymbol{v} \in \mathcal{M}(\boldsymbol{x}, \bar{Y})$,

$$\int \bigl|x - \mu(\hat{P}_{\boldsymbol{v}})\bigr|\, \hat{P}_{\boldsymbol{v}}(dx) \ = \ \frac{1}{n}\sum_{i=1}^n |Y_i - v_i| \ \ge \ \frac{1}{n}\sum_{i=1}^n |v_i| - \frac{1}{n}\sum_{i=1}^n |Y_i|,$$

and the right hand side tends to infinity as $\|\boldsymbol{v}\| \to \infty$. Thus it follows from Lemma 3.1 that

$$L(\hat{P}_{\boldsymbol{v}}) \ \to \ -\infty \quad \text{as } \|\boldsymbol{v}\| \to \infty, \boldsymbol{v} \in \mathcal{M}(\boldsymbol{x}, \bar{Y}),$$

and this coercivity, combined with continuity of $\boldsymbol{v} \mapsto L(\hat{P}_{\boldsymbol{v}})$ and $\mathcal{M}(\boldsymbol{x}, \bar{Y})$ being closed, yields the existence of a maximizer. □

Proof of Theorem 3.5. For any hyperplane $H \subset \mathbb{R}^d$,

$$P \star Q(H) \ = \ \int P(H - y)\, Q(dy) \ < \ 1,$$

and

$$\int \|z\|\, (P \star Q)(dz) \ \le \ \int \|x\|\, P(dx) + \int \|y\|\, Q(dy) \ < \ \infty.$$

Thus $P \star Q$ belongs to $\mathcal{P}_o \cap \mathcal{P}^1$. Moreover, by affine equivariance (Remark 2.4),

$$L(P \star Q) \ = \ L((P \star Q) \star \delta_{-a}) \ = \ L(P \star (Q \star \delta_{-a}))$$

for any $a \in \mathbb{R}^d$. Setting $a := \int y\, Q(dy)$ we may and do assume that $a = 0$.

Now let $\psi := \psi(\cdot \,|\, P)$ and $\tilde{\psi} := \psi(\cdot \,|\, P \star Q)$. Then

$$L(P \star Q) \ = \ \int \int \tilde{\psi}(x + y)\, P(dx) Q(dy) \ = \ \int \tilde{\psi}_Q \, dP,$$

where

$$\tilde{\psi}_Q(x) \ := \ \int \tilde{\psi}(x + y)\, Q(dy) \ \le \ \tilde{\psi}(x)$$

by Jensen's inequality. Hence

$$L(P \star Q) \ \le \ \int \tilde{\psi}\, dP \ = \ L(\tilde{\psi}, P) \ \le \ L(P).$$

Now suppose that $L(P \star Q) = L(P)$, so in particular, $\tilde{\psi} = \psi$. It follows from $\tilde{\psi}_Q \le \tilde{\psi} \in \mathcal{G}$ and Fatou's lemma that $\tilde{\psi}_Q \in \mathcal{G}$ with $\int \exp(\tilde{\psi}_Q(x))\, dx \le \int \exp(\tilde{\psi}(x))\, dx = 1$. Thus

$$L(P) \ = \ L(P \star Q) \ \le \ L(\tilde{\psi}_Q, P) \ \le \ L(P),$$

i.e. $\tilde{\psi}_Q = \psi = \tilde{\psi}$, and

$$\psi(x) \ = \ \int \psi(x + y)\, Q(dy) \quad \text{for all } x \in \mathbb{R}^d. \tag{9}$$

It remains to be shown that (9) entails $Q = \delta_0$. Note that $K := \{x \in \mathbb{R}^d : \psi(x) = M_o\}$ with $M_o := \max_{y \in \mathbb{R}^d} \psi(y)$ defines a compact set. Hence for any unit vector $u \in \mathbb{R}^d$ there exists a vector $x(u) \in K$ such that $u^\top x(u) \ge u^\top x$ for all $x \in K$. But then $\psi(x(u)+y) < M_o$ for all $y \in \mathbb{R}^d$ with $u^\top y > 0$. Hence

$$M_o \;=\; \psi(x(u)) \;=\; \int \psi(x(u)+y)\, Q(dy)$$

implies that $Q\{y : u^\top y > 0\} = 0$. Since u is an arbitrary unit vector, this entails that $\mathrm{csupp}(Q) = \{0\}$, i.e. $Q = \delta_0$. ☐

Proof of Theorem 3.6. One can easily deduce from (A.2-3) that the empirical distribution $\hat{P}_n := \hat{P}_{n,\mu_n}$ of the errors ϵ_{ni} themselves satisfies $D_1(\hat{P}_n, P) \to_p 0$.

To verify the assertions of the theorem it suffices to consider a sequence of *fixed* vectors $\boldsymbol{\epsilon}_n = (\epsilon_{ni})_{i=1}^n \in \mathbb{R}^n$ such that for a constant $c > 0$ to be specified later,

$$D_1(\hat{P}_n, P) + \sup_{m \in \mathcal{M} \,:\, \|(m-\mu_n)(\boldsymbol{x}_n)\|_n \le c} D_{BL}(\hat{P}_{n,m}, P_{n,m}) \;\to\; 0. \tag{10}$$

Our goal is to show that $(\hat{f}_n, \hat{\mu}_n)$, viewed as a function of $\boldsymbol{\epsilon}_n$ and thus fixed, too, is well-defined for sufficiently large n with

$$\int \bigl|\hat{f}_n(x) - f(x)\bigr|\, dx \;\to\; 0 \quad \text{and} \quad \bigl\|(\hat{\mu}_n - \mu_n)(\boldsymbol{x}_n)\bigr\|_n \;\to\; 0. \tag{11}$$

Note that we replaced f_n with $f = \exp \circ \psi(\cdot \,|\, P)$ because the integral of $|f_n - f|$ tends to zero by (A.2).

We know already that we have to restrict our attention to the set $\hat{\mathcal{M}}_n$ of all $m \in \mathcal{M}_n$ such that $\int t\, \hat{P}_{n,m}(dt) = 0$, i.e. $\int t\, Q_{(\mu_n - m)(\boldsymbol{x}_n)}(dt) = -\int t\, \hat{P}_n(dt) = o(1)$. Since $\{m(\boldsymbol{x}_n) : m \in \hat{\mathcal{M}}_n\}$ is a closed subset of $\mathbb{R}^n$ by (A.1), we may argue as in the proof of Theorem 3.2 that a maximizer $\hat{\mu}_n$ of $L(\hat{P}_{n,m})$ over all $m \in \hat{\mathcal{M}}_n$ does exist. It is possible that $L(\hat{P}_{n,\hat{\mu}_n}) = \infty$, but if we can show that $D_1(\hat{P}_{n,\hat{\mu}_n}, P) \to 0$, then $\hat{f}_n$ exists for sufficiently large n, too. Thus we may rephrase (11) as

$$D_1(\hat{M}_n, P) \to 0 \quad \text{and} \quad \int |t|\, \hat{Q}_n(dt) \;\to\; 0, \tag{12}$$

where $\hat{M}_n := \hat{P}_{n,\hat{\mu}_n}$ and $\hat{Q}_n := Q_{(\mu_n - \hat{\mu}_n)(\boldsymbol{x}_n)}$.

Note first that $\check{\mu}_n := \mu_n + \int t\, \hat{P}_n(dt)$ belongs to $\hat{\mathcal{M}}_n$, whence

$$L(\hat{M}_n) \;\ge\; L(\hat{P}_{n,\check{\mu}_n}) = L(\hat{P}_n) \;\to\; L(P) \tag{13}$$

by Theorem 2.13. On the other hand

$$\int |t| \, \hat{M}_n(dt) \; = \; \frac{1}{n} \sum_{i=1}^{n} \bigl| \epsilon_{ni} + (\mu_n - \hat{\mu}_n)(x_{ni}) \bigr| \; \geq \; \int |t| \, \hat{Q}_n(dt) - \int |t| \, \hat{P}_n(dt).$$

Thus by Lemma 3.1, $\hat{\mu}_n$ satisfies $\int |t| \, \hat{Q}_n(dt) \leq c$ for sufficiently large $n \in \mathbb{N}$, provided that

$$c \; > \; \int |t| \, P(dt) + \exp(-L(P)).$$

In particular,

$$D_{BL}(\hat{M}_n, P_n \star \hat{Q}_n) \; = \; D_{BL}(\hat{M}_n, P_{n,\hat{\mu}_n}) \; \to \; 0.$$

Since $D_{BL}(P_n \star \hat{Q}_n, P \star \hat{Q}_n) \leq D_{BL}(P_n, P) \to 0$, we know that even

$$D_{BL}(\hat{M}_n, P \star \hat{Q}_n) \; \to \; 0.$$

Since $(\hat{Q}_n)_n$ is a tight sequence, to verify (12) we may consider a subsequence $(\hat{Q}_{n(k)})_k$ that converges weakly to some distribution Q as $k \to \infty$. Then $\hat{M}_{n(k)} \to_w P \star Q$, so

$$\limsup_{k \to \infty} L(\hat{M}_{n(k)}) \; \leq \; L(P \star Q) \; \leq \; L(P)$$

by Theorems 2.12 and 3.5. By (13) we even know that $\lim_{k \to \infty} L(\hat{M}_{n(k)}) = L(P \star Q) = L(P)$. Consequently, we may deduce from Theorems 2.12 and 3.5 that

$$\lim_{k \to \infty} D_1(\hat{M}_{n(k)}, P \star Q) \; = \; 0 \quad \text{and} \quad Q \; = \; \delta_a \text{ for some } a \in \mathbb{R}.$$

It remains to be shown that $a = 0$ and $\lim_{k \to \infty} \int |t| \, \hat{Q}_{n(k)}(dt) = 0$. To this end we write $\boldsymbol{v}_n := (\mu_n - \hat{\mu}_n)(\boldsymbol{x}_n)$. We obtain for abitrary $r > 0$ and $n \in \mathbb{N}$ that

$$\begin{aligned}
\int |t| \, \hat{M}_n(dt) \; &= \; \int \min(|t|, r) \, \hat{M}_n(dt) + \frac{1}{n} \sum_{i=1}^{n} (|\epsilon_{ni} + v_{ni}| - r)^+ \\
&\geq \; \int \min(|t|, r) \, \hat{M}_n(dt) + \frac{1}{n} \sum_{i=1}^{n} (|v_{ni}| - |\epsilon_{ni}| - r)^+ \\
&\geq \; \int \min(|t|, r) \, \hat{M}_n(dt) + \int (|t| - 2r)^+ \, \hat{Q}_n(dt) \\
&\quad - \int (|t| - r)^+ \, \hat{P}_n(dt) \\
&= \; \int \min(|t|, r) \, \hat{M}_n(dt) + \int |t| \, \hat{Q}_n(dt) \\
&\quad - \int \min(|t|, 2r) \, \hat{Q}_n(dt) - \int (|t| - r)^+ \, \hat{P}_n(dt).
\end{aligned}$$

Hence $\int |t| \, \hat{Q}_{n(k)}(dt)$ is not greater than

$$\begin{aligned} &\int \min(|t|, 2r) \, \hat{Q}_{n(k)}(dt) + \int (|t| - r)^+ \, \hat{M}_{n(k)}(dt) + \int (|t| - r)^+ \, \hat{P}_{n(k)}(dt) \\ &\quad \to \int \min(|t|, 2r) \, Q(dt) + \int (|t| - r)^+ \, P \star Q(dt) + \int (|t| - r)^+ \, P(dt) \end{aligned}$$

as $k \to \infty$. As $r \uparrow \infty$, the limit on the right hand side converges to $\int |t| \, Q(dt) = |a|$. Consequently, $\lim_{k\to\infty} D_1(\hat{Q}_{n(k)}, Q) = 0$. But then $0 = \lim_{k\to\infty} \int t \, \hat{Q}_{n(k)}(dt)$ coincides with $\int t \, Q(dt) = a$. $\square$

In our proofs of Theorems 3.7 and 3.8 we utilize a simple inequality for the bounded Lipschitz distance in terms of the Kolmogorov–Smirnov distance

$$D_{KS}(P, Q) \ := \ \sup_{t\in\mathbb{R}} \bigl| (Q - P)((-\infty, t]) \bigr|$$

of two distributions $P, Q \in \mathcal{P}(1)$; see Dümbgen *et al.* (2010):

LEMMA 4.6. *Let P and Q be distributions on the real line. Then for arbitrary $r > 0$,*

$$D_{BL}(P, Q) \ \le \ 4P(\mathbb{R} \setminus (-r, r]) + 4(r + 1) D_{KS}(P, Q).$$

PROOF OF THEOREM 3.7. A key insight is that the empirical distributions $\hat{P}_{n,m}$ are close to their expectations $P_{n,m}$ with respect to Kolmogorov–Smirnov distance, uniformly over all $m \in \mathcal{M}_n$. Namely,

$$\begin{aligned} &\sup_{m\in\mathcal{M}_n, r\in\mathbb{R}} \bigl| (\hat{P}_{n,m} - P_{n,m})((-\infty, r]) \bigr| \\ &= \sup_{b\in\mathbb{R}^{q(n)}, c\in\mathbb{R}} \Bigl| \frac{1}{n} \sum_{i=1}^n \bigl(1\{Y_{ni} - b^\top x_{ni} \le c\} - \Pr(Y_{ni} - b^\top x_{ni} \le c) \bigr) \Bigr| \\ &\le \sup_{H\in\mathcal{H}_n} \bigl| (\hat{M}_n - M_n)(H) \bigr|, \end{aligned}$$

where $\mathcal{H}_n$ denotes the family of all closed half-spaces in $\mathbb{R}^{q(n)+1}$ while

$$\hat{M}_n \ := \ \frac{1}{n} \sum_{i=1}^n \delta_{(Y_{ni}, x_{ni}^\top)^\top}$$

and $M_n := \mathbb{E}\hat{M}_n$ are probability measures on $\mathbb{R}^{q(n)+1}$. Now we utilize well-known results from empirical process theory: $\mathcal{H}_n$ is a Vapnik–Červonenkis class with VC-dimension $q(n)+3$, and $\hat{M}_n$ is the arithmetic mean of n independent random probability measures. Thus

$$\mathbb{E} \sup_{m \in \mathcal{M}_n} D_{KS}(\hat{P}_{n,m}, P_{n,m}) \ \le \ C \sqrt{\frac{q(n)+3}{n}}$$

for some universal constant C; see Theorems 2.2 and 3.5 of Pollard (1990) and Theorem 2.6.4 and Lemma 2.6.16 of van der Vaart and Wellner (1996).

Since for fixed $c > 0$ the family $\{P_{n,m} : n \in \mathbb{N}, m \in \mathcal{M}_n \text{ with } \|(m - \mu_n)(\boldsymbol{x}_n)\|_n \le c\}$ is tight, the previous finding, combined with Lemma 4.6, implies that

$$\lim_{n\to\infty} \mathbb{E} \sup_{m \in \mathcal{M}_n : \|(m-\mu_n)(\boldsymbol{x}_n)\|_n \le c} D_{BL}(\hat{P}_{n,m}, P_{n,m}) \ = \ 0. \qquad \square$$

Proof of Theorem 3.8. Since $\big\|\mu_n(\boldsymbol{x}_n)\big\|_n = O(1)$ by assumption, it suffices to show that for any fixed $c > 0$,

$$\sup_{m \in \mathcal{M}_n : \|m(\boldsymbol{x}_n)\|_n \le c} D_{BL}(\hat{P}_{n,m}, P_{n,m}) \ \to_p \ 0. \tag{14}$$

One can easily show that for arbitrary $m, \tilde{m} \in \mathcal{M}_n$ and $h \in \mathcal{H}_{BL}$,

$$\left.\begin{array}{l} \Big|\displaystyle\int h\, d(P_{n,m} - P_{n,\tilde{m}})\Big| \\ \Big|\displaystyle\int h\, d(\hat{P}_{n,m} - \hat{P}_{n,\tilde{m}})\Big| \end{array}\right\} \ \le \ d_n\big((m - \tilde{m})(\boldsymbol{x}_n)\big)$$

with $d_n(\boldsymbol{v}) := n^{-1} \sum_{i=1}^n \min(|v_i|, 2)$, i.e. both distances $D_{BL}(P_{n,m}, P_{n,\tilde{m}})$ and $D_{BL}(\hat{P}_{n,m}, \hat{P}_{n,\tilde{m}})$ are not greater than $d_n\big((m - \tilde{m})(\boldsymbol{x}_n)\big)$. Suppose that the following assumption is satisfied:

(A.4) For each n there exist a linear subspace $\mathbb{V}_n$ of $\mathbb{R}^n$ with dimension $q(n)$ and a constant $\delta(n) > 0$ such that $q(n)/n \to 0$, $\delta(n) \to 0$, and for each $m \in \mathcal{M}_n$ there exists a $\tilde{\boldsymbol{v}} \in \mathbb{V}_n$ with

$$\|\tilde{\boldsymbol{v}}\|_n \ \le \ \|m(\boldsymbol{x}_n)\|_n \quad \text{and} \quad d_n\big(\tilde{\boldsymbol{v}} - m(\boldsymbol{x}_n)\big) \ \le \ \delta(n)\big(\|m(\boldsymbol{x}_n)\|_n + 1\big).$$

This assumption implies that

$$\begin{aligned} &\sup_{m \in \mathcal{M}_n : \|m(\boldsymbol{x}_n)\|_n \le c} D_{BL}(\hat{P}_{n,m}, P_{n,m}) \\ &\qquad \le \ 2\delta(n)(c+1) + \sup_{\boldsymbol{v} \in \mathbb{V}_n : \|\boldsymbol{v}\|_n \le c} D_{BL}(\hat{P}_{\boldsymbol{v}}, P_{\boldsymbol{v}}), \end{aligned}$$

where $\hat{P}_{\boldsymbol{v}} := n^{-1} \sum_{i=1}^n \delta_{Y_{ni}-v_i}$ and $P_{\boldsymbol{v}} = \mathbb{E}\hat{P}_{\boldsymbol{v}}$. Since

$$\int |y|\, P_{\boldsymbol{v}}(dy) \;=\; \frac{1}{n}\sum_{i=1}^n \mathbb{E}|Y_{ni} - v_i| \;\le\; \int |y|\, P_n(dy) + \|\mu_n(\boldsymbol{x}_n)\|_n + \|\boldsymbol{v}\|_n,$$

Lemma 4.6 entails that (14) is a consequence of

$$\mathbb{E} \sup_{\boldsymbol{v}\in\mathbb{V}_n} D_{KS}(\hat{P}_{\boldsymbol{v}}, P_{\boldsymbol{v}}) \;\to\; 0. \tag{15}$$

But the latter claim follows again from empirical process theory: Let $\boldsymbol{B}_n = (b_{nij})_{i,j}$ be an $n \times q(n)$ matrix whose columns form a basis of $\mathbb{V}_n$, and let $\boldsymbol{v} = \boldsymbol{B}_n\lambda$ for some $\lambda \in \mathbb{R}^{q(n)}$. Then

$$\hat{P}_{\boldsymbol{v}}((-\infty, r]) \;=\; \frac{1}{n}\sum_{i=1}^n 1\Big\{Y_{ni} - \sum_{j=1}^{q(n)} b_{nij}\lambda_j \le r\Big\} \;=\; \hat{M}_n(H_{n,\lambda,r}),$$

where $\hat{M}_n$ is the empirical distribution of the points $(Y_{ni}, b_{ni1}, \ldots, b_{niq(n)})^\top \in \mathbb{R}^{q(n)+1}$ and $H_{n,\lambda,r} \subset \mathbb{R}^{q(n)+1}$ is a suitable closed halfspace. Consequently, the expected value in (15) is bounded by a universal constant times $\big((q(n)+3)/n\big)^{1/2}$.

It remains to verify assumption (A.4) in case of isotonic regression. Without loss of generality let $x_{n1} \le x_{n2} \le \cdots \le x_{nn}$. Then $\mathcal{M}_n(\boldsymbol{x}_n)$ is contained in the cone $\mathbb{K}_n$ of all vectors in $\mathbb{R}^n$ with non-decreasing components. For $n \ge 3$ let $I_{n1}, I_{n2}, \ldots, I_{nq(n)}$ be a partition of $\{1, 2, \ldots, n\}$ into index intervals $I_{nj} = \{a_{nj}, a_{n,j}+1, \ldots, b_{nj}\}$ with $b_{nj} < a_{n,j+1}$ if $1 \le j < q(n)$ such that

$$\#I_{n2} = \#I_{n3} = \cdots = \#I_{n,q(n)-1} \;=\; k(n),$$
$$k(n) \;\to\; \infty, \quad \frac{k(n)}{\#I_{n1}} + \frac{k(n)}{\#I_{nq(n)}} \;\to\; 0, \quad \frac{\#I_{n1} + \#I_{nq(n)}}{n} \;\to\; 0.$$

Since $n \ge q(n)k(n)$ for sufficiently large n, these conditions entail that $q(n)/n \to 0$. For $\boldsymbol{v} \in \mathbb{K}_n$ we define $\tilde{\boldsymbol{v}}$ as follows: For $i \in I_{nj}$ let

$$\tilde{v}_i \;:=\; \begin{cases} v_{a_{nj}} & \text{if } v_{a_{nj}} > 0, \\ v_{b_{nj}} & \text{if } v_{b_{nj}} < 0, \\ 0 & \text{if } v_{a_{nj}} \le 0 \le v_{b_{nj}}. \end{cases}$$

Obviously this vector $\tilde{\boldsymbol{v}}$ belongs to the $q(n)$-dimensional vector space $\mathbb{V}_n$ of all vectors whose components are constant on each index interval I_{nj},

$1 \le j \le q(n)$. Since $|\tilde{v}_i| \le |v_i|$ for all indices i, we have $\|\tilde{\boldsymbol{v}}\|_n \le \|\boldsymbol{v}\|_n$. Moreover,

$$d_n(\tilde{\boldsymbol{v}} - \boldsymbol{v}) \;\le\; \frac{2(\#I_{n1} + \#I_{nq(n)})}{n} + \frac{1}{n} \sum_{j=2}^{q(n)-1} \sum_{i \in I_{nj}} |\tilde{v}_i - v_i|.$$

By construction of $\tilde{\boldsymbol{v}}$, for $2 \le j < q(n)$,

$$\begin{aligned} \sum_{i \in I_{nj}} |\tilde{v}_i - v_i| \;&\le\; \sum_{i \in I_{nj}} v_i^+ - 1_{\{j>2\}} \sum_{i \in I_{n,j-1}} v_i^+ \\ &\quad + \sum_{i \in I_{nj}} v_i^- - 1_{\{j<q(n)-1\}} \sum_{i \in I_{n,j+1}} v_i^- . \end{aligned}$$

For instance, if $v_{a_{nj}} > 0$, then

$$\sum_{i \in I_{nj}} |\tilde{v}_i - v_i| \;=\; \sum_{i \in I_{nj}} v_i^+ - k(n) v_{a_{nj}}^+ \;\le\; \sum_{i \in I_{nj}} v_i^+ - 1_{\{j>2\}} \sum_{i \in I_{n,j-1}} v_i^+ ,$$

while $v_i^- = 0$ for $i \in I_{nj} \cup I_{n,j+1}$. The cases $v_{b_{nj}} < 0$ and $v_{a_{nj}} \le 0 \le v_{b_{nj}}$ may be treated similarly. Consequently,

$$\begin{aligned} \sum_{j=2}^{q(n)-1} \sum_{i \in I_{nj}} |\tilde{v}_i - v_i| \;&\le\; \sum_{i \in I_{n2}} v_i^- + \sum_{i \in I_{n,q(n)-1}} v_i^+ \\ &\le\; \frac{k(n)}{\#I_{n1}} \sum_{i \in I_{n1}} v_i^- + \frac{k(n)}{\#I_{nq(n)}} \sum_{i \in I_{nq(n)}} v_i^+ \\ &\le\; n \|\boldsymbol{v}\|_n \max\Bigl(\frac{k(n)}{\#I_{n1}}, \frac{k(n)}{\#I_{nq(n)}} \Bigr). \end{aligned}$$

It follows that $d_n(\tilde{\boldsymbol{v}} - \boldsymbol{v}) \le \delta(n)(\|\boldsymbol{v}\|_n + 1)$ with $\delta(n)$ being the maximum of the three numbers $2(\#I_{n1} + \#I_{nq(n)})/n$, $k(n)/\#I_{n1}$ and $k(n)/\#I_{nq(n)}$. $\square$

References.

Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* **6** 170-176.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26** 641-647.

Bagnoli, M. and Bergstrom, T. (2005). Log-concave probability and its applications. *Econometric Theory* **26** 445-469.

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.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some Asymptotic Theory for the Bootstrap. *Annals of Statistics* **9** 1196-1217.

CULE, M. L. and SAMWORTH, R. J. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic J. Statist.* **4** 254-270. Available at http://arxiv.org/abs/0908.4400

CULE, M. L., SAMWORTH, R. J. and STEWART, M. I. (2010). Maximum likelihood estimation of a multidimensional log-concave density (with discussion). *J. Roy. Statist. Soc., Ser. B* **72** to appear. Available at http://arxiv.org/abs/0804.3989

DOKSUM, K., OZEKI, A., KIM, J. and NETO, E. C. (2007). Thinking outside the box: statistical inference based on Kullback-Leibler empirical projections. *Statistics and Probability Letters* **77** 1201-1213.

DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics* **20** 1803-1827.

DÜMBGEN, L., HÜSLER, A. and RUFIBACH, K. (2007). Active set and EM algorithms for log-concave densities based on complete and censored Data Technical report report No. 61, IMSV, University of Bern. Available at http://arxiv.org/abs/0707.4643

DÜMBGEN, L. and RUFIBACH, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli* **15** 40-68. Available at http://arxiv.org/abs/0709.0334.

DÜMBGEN, L., SAMWORTH, R. and SCHUHMACHER, D. (2010). Approximation by log-concave distributions with applications to regression Technical Report No. 75, IMSV, University of Bern. Available at http://arxiv.org/abs/1002.3448

KOENKER, R. and MIZERA, I. (2010). Quasi-convex density estimation. *Ann. Statist.* **38** to appear.

LANG, R. (1986). A note on the measurability of convex sets. *Arch. Math.* **47** 90-92. MR855142 (87j:28002)

MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Annals of Mathematical Statistics* **43** 508-515.

NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *Computer Journal* **7** 308-313.

PAL, J., WOODROOFE, M. and MEYER, M. (2007). Estimating a Polya frequency function$_2$. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, (R. Liu, W. Strawderman and C.-H. Zhang, eds.). *IMS Lecture Notes and Monograph Series* **54** 239-249. Institute of Mathematical Statistics.

POLLARD, D. (1990). *Empirical processes: theory and applications. NSF-CBMS Regional Conference Series in Probability and Statistics, 2.* Institute of Mathematical Statistics, Hayward, CA.

R DEVELOPMENT CORE TEAM, (2009). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org

RUFIBACH, K. (2006). Log-concave density estimation and bump hunting for i.i.d. observations PhD thesis, University of Bern, Department of Mathematics and Statistics.

RUFIBACH, K. and DÜMBGEN, L. (2006). logcondens: Estimate a Log-Concave Probability Density from i.i.d. Observations R package version 1.3.5.

SCHUHMACHER, D. and DÜMBGEN, L. (2010). Consistency of multivariate log-concave density estimators. *Statistics and Probability Letters* **80** 376-380.

SCHUHMACHER, D., HÜSLER, A. and DÜMBGEN, L. (2009). Multivariate log-concave distributions as a nearly parametric model Technical Report No. 74, IMSV, University of Bern. Available at http://arxiv.org/abs/0907.0250

SEREGIN, A. and WELLNER, J. A. (2009). Nonparametric estimation of multivariate convex-transformed densities. Preprint. Available at http://arxiv.org/abs/0911.4151

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

Walther, G. (2009). Inference and modeling with log-concave distributions. *Statistical Science* **24** to appear.

Institute of Math. Statistics and Actuarial Science
Alpeneggstrasse 22
CH-3012 Bern
Switzerland
E-mail: duembgen@stat.unibe.ch

Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge, CB3 0WB
United Kingdom
E-mail: r.samworth@statslab.cam.ac.uk

Institute of Math. Statistics and Actuarial Science
Alpeneggstrasse 22
CH-3012 Bern
Switzerland
E-mail: dominic.schuhmacher@stat.unibe.ch