The Annals of Statistics
2011, Vol. 39, No. 2, 702–730
DOI: 10.1214/10-AOS853

APPROXIMATION BY LOG-CONCAVE DISTRIBUTIONS, WITH APPLICATIONS TO REGRESSION

BY LUTZ DÜMBGEN[1], RICHARD SAMWORTH AND DOMINIC SCHUHMACHER[1]

University of Bern, University of Cambridge and 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 $Y = \mu(X) + \varepsilon$, 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 and mean zero.

1. Introduction. Log-concave distributions, that is, distributions with a Lebesgue density the logarithm of which is concave, are an interesting nonparametric 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, Woodroofe and Meyer (2007), Rufibach (2006), Dümbgen, Hüsler and Rufibach (2007), Balabdaoui, Rufibach and Wellner (2009) and Dümbgen and Rufibach (2009). These references contain characterizations of the estimators, consistency results and explicit algorithms. Extensions of one or more of these aspects to the multivariate setting are presented by Cule, Samworth and Stewart (2010), Cule and Samworth (2010), Koenker and Mizera (2010), Seregin and Wellner (2010) and Schuhmacher and Dümbgen (2010). Both Cule and Samworth (2010) and Schuhmacher, Hüsler and Dümbgen (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.

Received February 2010; revised August 2010.
[1]Supported by the Swiss National Science Foundation.
MSC2010 subject classifications. 62E17, 62G05, 62G07, 62G08, 62G35, 62H12.
Key words and phrases. Convex support, isotonic regression, linear regression, Mallows distance, projection, weak semicontinuity.

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. Let us put this into a somewhat broader context: let $\hat{Q}_n$ be the empirical distribution of independent random vectors $X_1, X_2, \dots, X_n$ with distribution Q on a given open set $\mathcal{X} \subseteq \mathbb{R}^d$. Suppose that Q has a density f belonging to a given class $\mathcal{F}$ of probability densities on $\mathcal{X}$. The maximum likelihood estimator of f may be written as

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

(provided this exists and is unique). Even if Q fails to have a density within $\mathcal{F}$, one may view $\hat{f}_n$ as an estimator of the approximating density

$$f(\cdot|Q) := \operatorname*{arg\,max}_{f \in \mathcal{F}} \int \log(f)\, dQ.$$

In fact, if Q has a density $g \notin \mathcal{F}$ on $\mathcal{X}$ such that the integral $\int g(x) \log g(x)\, dx$ exists in $\mathbb{R}$, one may rewrite $f(\cdot|Q)$ as the minimizer of the Kullback–Leibler divergence,

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

over all $f \in \mathcal{F}$. Note the well-known fact that $D_{\mathrm{KL}}(f, g) > 0$ unless $f = g$ almost everywhere. Viewing a maximum likelihood estimator as an estimator of an approximation within a given model is common in statistics [see, e.g., Pfanzagl (1990), Patilea (2001), Doksum et al. (2007) and Cule and Samworth (2010)]. Pfanzagl (1990) and Patilea (2001) show that under suitable regularity conditions on Q and $\mathcal{F}$, the estimator $\hat{f}_n$ is consistent with certain large deviation bounds or rates of convergence, even in the case of misspecified models. To the best of our knowledge, their results are not directly applicable in the setting of log-concave densities, which is treated by Cule and Samworth (2010). Our ambition is to identify the *largest* possible class of distributions Q such that $f(\cdot|Q)$ is well defined and unique. Moreover, we want to show that the mapping $Q \mapsto f(\cdot|Q)$ is continuous on that class with respect to a *coarse* topology, ideally the topology of weak convergence.

With these goals in mind, let us tell a short success story about Grenander's estimator [Grenander (1956)], also a key example of Patilea (2001): let $\mathcal{F}_{\mathrm{mon}}$ be the class of all nonincreasing and left-continuous probability densities on $\mathcal{X} = (0, \infty)$. Then for *any* distribution Q on $(0, \infty)$, the maximizer

$$f_{\mathrm{mon}}(\cdot|Q) := \operatorname*{arg\,max}_{f \in \mathcal{F}_{\mathrm{mon}}} \int_{(0,\infty)} \log f(x) Q(dx)$$

is well defined and unique. Namely, if G denotes the distribution function of Q, then $f_{\mathrm{mon}}(\cdot|Q)$ is the left-sided derivative of the smallest concave majorant of G

on $(0, \infty)$ [see Barlow et al. (1972)]. With this characterization one can show that for any sequence of distributions Q_n on $(0, \infty)$ converging weakly to Q,

$$\int_{(0,\infty)} \left| f_{\mathrm{mon}}(x|Q_n) - f_{\mathrm{mon}}(x|Q) \right| dx \to 0 \qquad (n \to \infty).$$

Since the sequence of empirical measures $\hat{Q}_n$ converges weakly to Q almost surely, this entails strong consistency of the Grenander estimator $f(\cdot|\hat{Q}_n)$ in total variation distance.

In the remainder of the present paper we consider the class $\mathcal{F}$ of log-concave probability densities on $\mathcal{X} = \mathbb{R}^d$. We will show in Section 2 that $f(\cdot|Q)$ exists and is unique in $L^1(\mathbb{R}^d)$ if and only if

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

and

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

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

$$Q_n \to_w Q \quad \text{and} \quad \int \|x\| \, Q_n(dx) \to \int \|x\| \, Q(dx) \tag{1}$$

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

$$\lim_{n\to\infty} \int \left| f(x|Q_n) - f(x|Q) \right| dx = 0.$$

This entails strong consistency of the maximum likelihood estimator $\hat{f}_n$, because $(\hat{Q}_n)_n$ converges almost surely to Q with respect to Mallows distance $D_1(\cdot, \cdot)$. In addition we show that $Q \mapsto \max_{f \in \mathcal{F}} \int \log(f) \, dQ$ is convex and upper semicontinuous with respect to weak convergence.

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

$$Y_i = \mu(x_i) + \varepsilon_i$$

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 nondecreasing. These methods are illustrated with a real data set.

Many proofs and technical arguments are deferred to Section 4. A longer and more detailed version of this paper is the technical report by Dümbgen, Samworth and Schuhmacher (2010), referred to as [DSS 2010] hereafter. It contains all proofs, additional examples and plots, a detailed description of our algorithms and extensive simulation studies. There we also indicate potential applications to change-point analyses.

2. Log-concave approximations. For a fixed dimension $d \in \mathbb{N}$, let $\Phi = \Phi(d)$ be the family of concave functions $\phi : \mathbb{R}^d \to [-\infty, \infty)$ which are upper semicontinuous and coercive in the sense that

$$\phi(x) \to -\infty \qquad \text{as } \|x\| \to \infty.$$

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

$$L(\phi, Q) := \int \phi\,dQ - \int e^{\phi(x)}\,dx + 1$$

and a profile log-likelihood

$$L(Q) := \sup_{\phi \in \Phi} L(\phi, Q).$$

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

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

To verify this, note that $\phi + c \in \Phi$ for any fixed function $\phi \in \Phi$ and arbitrary $c \in \mathbb{R}$, and

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

if $L(\phi, Q) \in \mathbb{R}$. Thus $L(\phi + c, Q)$ is minimal for $c = -\log \int e^{\phi(x)}\,dx$.

2.1. *Existence, uniqueness and basic properties.* The next theorem provides a complete characterization of all distributions $Q \in \mathcal{Q}$ with real profile log-likelihood $L(Q)$. To state the result we first define the convex support of a distribution $Q \in \mathcal{Q}$ and collect some of its properties.

LEMMA 2.1 (DSS 2010). *For any $Q \in \mathcal{Q}$, the set*

$$\mathrm{csupp}(Q) := \bigcap \{C : C \subseteq \mathbb{R}^d \text{ closed and convex}, Q(C) = 1\}$$

is itself closed and convex with $Q(\mathrm{csupp}(Q)) = 1$. *The following three properties of Q are equivalent*:

(a) $\mathrm{csupp}(Q)$ *has nonempty interior*;
(b) $Q(H) < 1$ *for any hyperplane* $H \subset \mathbb{R}^d$;
(c) *with* Leb *denoting Lebesgue measure on* $\mathbb{R}^d$,

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

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

$$\int \|x\| Q(dx) < \infty \quad \text{and} \quad \mathrm{interior}(\mathrm{csupp}(Q)) \neq \varnothing.$$

In that case, there exists a unique function

$$\psi = \psi(\cdot|Q) \in \operatorname*{arg\,max}_{\phi \in \Phi} L(\phi, Q).$$

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

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

REMARK 2.3 [Moment (in)equalities]. Let $Q \in \mathcal{Q}$ satisfy the properties stated in Theorem 2.2. Then the log-density $\psi = \psi(\cdot|Q)$ satisfies the following requirements: $L(\psi, Q) = \int \psi\,dQ \in \mathbb{R}$, and for any function $\Delta : \mathbb{R}^d \to \mathbb{R}$,

$$\int \Delta\,dQ \le \int \Delta(x) e^{\psi(x)}\,dx \qquad \text{if } \psi + t\Delta \in \Phi \qquad \text{for some } t > 0. \tag{2}$$

This follows from

$$\lim_{t \downarrow 0} t^{-1}\bigl(L(\psi + t\Delta, Q) - L(\psi, Q)\bigr) = \int \Delta\,dQ - \int \Delta(x) e^{\psi(x)}\,dx.$$

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

$$\int h\,dP \le \int h\,dQ \qquad \text{for any convex } h : \mathbb{R}^d \to (-\infty, \infty], \tag{3}$$

$$\int x P(dx) = \int x Q(dx). \tag{4}$$

To verify (3), let $v \in \mathbb{R}^d$ be a subgradient of h at 0, that is, $h(x) \ge h(0) + v^\top x$ 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 $\psi + t\Delta \in \Phi$ whenever $0 < t < b/\|v\|$. Hence the asserted inequality follows from (2). 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

$$\mathcal{Q}^1 = \mathcal{Q}^1(d) := \left\{ Q \in \mathcal{Q} : \int \|x\| \, Q(dx) < \infty \right\},$$
$$\mathcal{Q}_o = \mathcal{Q}_o(d) := \{ Q \in \mathcal{Q} : \text{interior}(\text{csupp}(Q)) \neq \varnothing \}.$$

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

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

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

$$L(Q_{a,B}) = L(Q) - \log|\det B|$$

and

$$\psi(x | Q_{a,B}) = \psi\big(B^{-1}(x - a) | Q\big) - \log|\det B| \qquad \text{for } x \in \mathbb{R}^d.$$

REMARK 2.5 (Convexity, DSS 2010). The profile log-likelihood L is convex on $\mathcal{Q}^1$. Precisely, for arbitrary $Q_0, Q_1 \in \mathcal{Q}^1$ and $0 < t < 1$,

$$L\big((1-t)Q_0 + tQ_1\big) \leq (1-t)L(Q_0) + tL(Q_1).$$

The two sides are equal and real if and only if $Q_0, Q_1 \in \mathcal{Q}_o \cap \mathcal{Q}^1$ with $\psi(\cdot|Q_0) = \psi(\cdot|Q_1)$.

REMARK 2.6 (Concave majorants, DSS 2010). Let $\psi = \psi(\cdot|Q)$ for a distribution $Q \in \mathcal{Q}_o \cap \mathcal{Q}^1$. For any open set $U \subset \mathbb{R}^d$ there exists a (pointwise) minimal function $\psi_U \in \Phi$ such that $\psi_U \geq \psi$ on $\mathbb{R}^d \setminus U$. In particular, $\psi_U \leq \psi$ with equality on $\mathbb{R}^d \setminus U$. One can also show that ψ_U is the pointwise infimum of all affine functions ϕ such that $\phi \geq \psi$ on $\mathbb{R}^d \setminus U$. If $\text{supp}(Q)$ denotes the smallest closed set $A \subseteq \mathbb{R}^d$ with $Q(A) = 1$, then

$$\psi = \psi_{\mathbb{R}^d \setminus \text{supp}(Q)}.$$

Furthermore, suppose that Q has a density g on an open set U such that $\psi > \log g$ on this set. Then

$$\psi = \psi_U.$$

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 $\phi \in \Phi(1)$ let

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

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

THEOREM 2.7. *Let Q be a nondegenerate distribution on $\mathbb{R}$ with finite first moment and distribution function G. Let F be a distribution function with log-density $\phi \in \Phi$. Then $\phi = \psi(\cdot|Q)$ if and only if*

$$\int_{-\infty}^{\infty} (F(t) - G(t))\,dt = 0$$

and

$$\int_{-\infty}^{x} (F(t) - G(t))\,dt \begin{cases} \le 0, & \text{for all } x \in \mathbb{R}, \\ = 0, & \text{for all } x \in \mathcal{S}(\phi). \end{cases}$$

REMARK 2.8 (DSS 2010). One consequence of this theorem is that the c.d.f. F of $\psi(\cdot|Q)$ follows the c.d.f. G of Q quite closely in that

$$G(x-) \le F(x) \le G(x) \qquad \text{for arbitrary } x \in \mathcal{S}(\psi(\cdot|Q)).$$

EXAMPLE 2.9. Let Q be a rescaled version of Student's distribution t_2 with density and distribution function

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

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

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

Indeed the integral on the left-hand side equals

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

for all $x \le 0$. Clearly this expression is zero for $x = 0$, and elementary considerations show that it is nonpositive for all $x \le 0$. Numerical calculations reveal that $|F - G|$ is smaller than 0.04 everywhere.

REMARK 2.10. Let $Q \in \mathcal{Q}_o \cap \mathcal{Q}^1$ such that $Q(a, b) = 0$ for some bounded interval $(a, b) \subset \mathrm{csupp}(Q)$. Then $\psi = \psi(\cdot|Q)$ is linear on $[a, b]$. This follows from Remark 2.6, applied to $U = (a, b)$. Note that $\psi(a) > -\infty$ and $\psi(b) > -\infty$, because otherwise $\psi \equiv -\infty$ on $(-\infty, a]$ or on $[b, \infty)$. But this would be incompatible with $\int \psi \, dQ \in \mathbb{R}$, because both $Q((-\infty, a])$ and $Q([b, \infty))$ are positive.

REMARK 2.11 (DSS 2010). Suppose that Q has a continuous but not log-concave density g. Nevertheless one can say the following about the approximating log-density $\psi = \psi(\cdot|Q)$:

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

(ii) Suppose that $\log g$ 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 $\psi = \log g$ on $(-\infty, a'] \cup [b', \infty)$.

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

EXAMPLE 2.12. Let us illustrate part (ii) of Remark 2.11 with a numerical example. Figure 1 shows the bimodal density g (green/dotted line) of the Gaussian mixture $Q = 0.7 \cdot \mathcal{N}(-1.5, 1) + 0.3 \cdot \mathcal{N}(1.5, 1)$ together with its log-concave approximation $f = f(\cdot|Q)$ (blue line). As predicted, there exists an interval $[a', b']$ such that $f = g$ on $\mathbb{R} \setminus (a', b')$ and $\log f$ is linear on $[a', b']$.

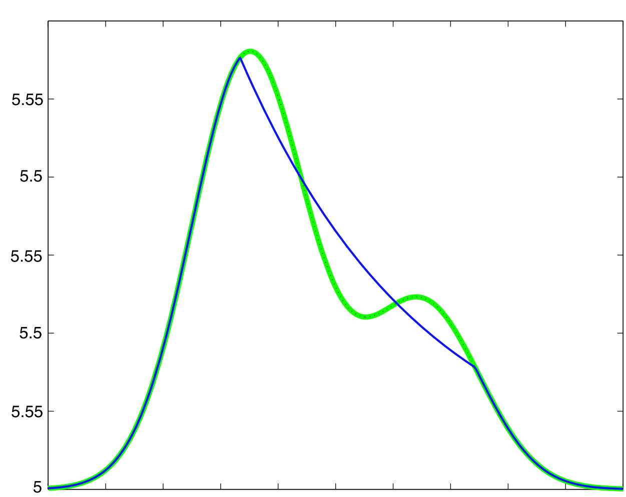

FIG. 1. *Density of a Gaussian mixture and its log-concave approximation.*

2.3. *Continuity in Q.* For the applications to regression problems to follow we need to understand the properties of both $Q \mapsto L(Q)$ and $Q \mapsto \psi(\cdot|Q)$ on $\mathcal{Q}^1 \cap \mathcal{Q}_o$. 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 $Q, Q' \in \mathcal{Q}^1$,

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

where the infimum is taken over all pairs (X, X') of random vectors $X \sim Q$ and $X' \sim Q'$ on a common probability space. It is well known that the infimum in $D_1(Q, Q')$ is a minimum. The distance D_1 is also known as Wasserstein, Monge–Kantorovich or Earth Mover's distance. An alternative representation due to Kantorovič and Rubinšteĭn (1958) is

$$D_1(Q, Q') = \sup_{h \in \mathcal{H}_L} \left| \int h \, dQ - \int h \, dQ' \right|,$$

where $\mathcal{H}_L$ consists of all $h : \mathbb{R}^d \to \mathbb{R}$ such that $|h(x) - h(y)| \le \|x - y\|$ for all $x, y \in \mathbb{R}^d$. Moreover, for a sequence $(Q_n)_n$ in $\mathcal{Q}^1$, it is known that (1) is equivalent to $D_1(Q_n, Q) \to 0$ as $n \to \infty$ [Mallows (1972), Bickel and Freedman (1981)]. In case of $d = 1$, the optimal coupling of Q and Q' is given by the quantile transformation: if G and G' denote the respective distribution functions, then

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

A good starting point for more detailed information on Mallows distance is Chapter 7 of Villani (2003).

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

LEMMA 2.13. *Given a distribution $Q \in \mathcal{Q}$, a point $x \in \mathbb{R}^d$ is an interior point of* csupp(Q) *if and only if*

$$h(Q, x) := \sup\{Q(C) : C \subset \mathbb{R}^d \text{ closed and convex}, x \notin \mathrm{interior}(C)\} < 1.$$

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

$$\limsup_{n \to \infty} h(Q_n, x) \le h(Q, x) \qquad \text{for any } x \in \mathbb{R}^d.$$

This lemma implies that the set $\mathcal{Q}_o$ is an open subset of $\mathcal{Q}$ with respect to the topology of weak convergence. The supremum $h(Q, x)$ is a maximum over closed halfspaces and is related to Tukey's halfspace depth [Donoho and Gasko (1992), Section 6]. For a proof of Lemma 2.13 we refer to [DSS 2010]. Now we are ready to state the main results of this section.

THEOREM 2.14 (Weak upper semicontinuity). *Let $(Q_n)_n$ be a sequence of distributions in $\mathcal{Q}_o$ converging weakly to some $Q \in \mathcal{Q}_o$. Then*

$$\limsup_{n\to\infty} L(Q_n) \le L(Q).$$

Moreover, $\liminf_{n\to\infty} L(Q_n) < L(Q)$ *if and only if*

$$\limsup_{n\to\infty} \int \|x\| Q_n(dx) > \int \|x\| Q(dx).$$

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

THEOREM 2.15 (Continuity with respect to Mallows distance D_1). *Let $(Q_n)_n$ be a sequence of distributions in $\mathcal{Q}^1$ such that* $\lim_{n\to\infty} D_1(Q_n, Q) = 0$ *for some $Q \in \mathcal{Q}^1$. Then*

$$\lim_{n\to\infty} L(Q_n) = L(Q).$$

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

$$\lim_{n\to\infty, x\to y} f_n(x) = f(y) \qquad \text{for all } y \in \mathbb{R}^d \setminus \partial\{f > 0\},$$

$$\limsup_{n\to\infty, x\to y} f_n(x) \le f(y) \qquad \text{for all } y \in \partial\{f > 0\},$$

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

REMARK 2.16 (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, Hüsler and Dümbgen (2009) that $(f_n)_n \to f$ uniformly on arbitrary closed subsets of $\mathbb{R}^d \setminus \mathrm{disc}(f)$, where $\mathrm{disc}(f)$ 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 $\varepsilon(f) > 0$ such that

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

More generally,

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

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

3. Applications to regression problems. Now we consider the regression setting described in the Introduction with observations $Y_i = \mu(x_i) + \varepsilon_i$, $1 \le i \le n$, 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 Q on $\mathbb{R}$ such that $\psi = \psi(\cdot|Q)$ is well defined. The regression function μ is assumed to belong to a given family $\mathcal{M}$ with the property that

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

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

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

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

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

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

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

$$\hat{\Lambda}(\phi, m) = L\big(\phi, \hat{Q}_{m(\mathbf{x})}\big)$$

with the empirical distributions

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

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

$$(\hat{\phi}, \hat{m}) \in \operatorname*{arg\,max}_{(\phi, m) \in \Phi \times \mathcal{M}} L\big(\phi, \hat{Q}_{m(\mathbf{x})}\big),$$

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, which follows from Jensen's inequality and elementary considerations:

LEMMA 3.1 (DSS 2010). *For any distribution $Q \in \mathcal{Q}^1(1)$,*

$$L(Q) \le -\log\Big(2\int |x - \operatorname{Med}(Q)|\,Q(dx)\Big) \le -\log\Big(\int |x - \mu(Q)|\,Q(dx)\Big),$$

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

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

The constraint $\mathbf{Y} \notin \mathcal{M}(\mathbf{x})$ excludes 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 $\mathcal{M}(\mathbf{x})$ is the column space of the design matrix

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

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

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 $\mathcal{M}(\mathbf{x})$ is a closed convex cone in $\mathbb{R}^n$. Here the condition that $\mathbf{Y} \notin \mathcal{M}(\mathbf{x})$ 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. The maximum likelihood estimator $(\hat{\psi}, \hat{\mu})$ need not be unique in general. 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 $\hat{Q}_{m(\mathbf{x})}$ equals

$$\mathbb{E}\hat{Q}_{m(\mathbf{x})} = \frac{1}{n}\sum_{i=1}^n Q \star \delta_{\mu(x_i)-m(x_i)} = Q \star R_{(\mu-m)(\mathbf{x})}$$

with

$$R_{\mathbf{v}} := \frac{1}{n}\sum_{i=1}^n \delta_{v_i}.$$

But

$$L\big(Q \star R_{(\mu-m)(\mathbf{x})}\big) \le L(Q)$$

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* $Q \in \mathcal{Q}_o(d) \cap \mathcal{Q}^1(d)$ *and* $R \in \mathcal{Q}^1(d)$. *Then* $Q \star R \in \mathcal{Q}_o \cap \mathcal{Q}^1$ *and*

$$L(Q \star R) \le L(Q).$$

Equality holds if and only if $R = \delta_a$ *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

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

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

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

We write $(\hat{\psi}_n, \hat{\mu}_n)$ for a maximizer of $L(\phi, \hat{Q}_{n,m})$ over all pairs $(\phi, m) \in \Phi \times \mathcal{M}_n$ such that $\int e^{\phi(x)}\, dx = 1$ and $\int x e^{\phi(x)}\, dx = 0$, where $\hat{Q}_{n,m}$ 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

$$Q_{n,m} := \mathbb{E}\hat{Q}_{n,m} = Q_n \star R_{(\mu_n - m)(\mathbf{x}_n)}.$$

Furthermore we write

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

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

$$D_{\mathrm{BL}}(Q, Q') := \sup_{h \in \mathcal{H}_{\mathrm{BL}}} \left| \int h\, d(Q - Q') \right|,$$

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

THEOREM 3.6 (Consistency in regression). *Let assumptions* (A.1) *and* (A.2) *be satisfied. Suppose further that*:

(A.3) *for arbitrary fixed* $c > 0$,

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

Then, with $f_n := \exp \circ \psi(\cdot | Q_n)$ *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*

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

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*

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

THEOREM 3.8 (Isotonic regression). *Let $\mathcal{M}_n$ be the set of all nondecreasing functions on an interval $\mathcal{X}_n \subseteq \mathbb{R}$. If assumption* (A.2) *holds true, then* (A.3) *follows from*

$$\|\mu_n(\mathbf{x}_n)\|_n = O(1).$$

The proof of Theorem 3.7 is given in Section 4. For the proof of Theorem 3.8, which uses similar ideas and an additional approximation argument, we refer to [DSS 2010].

3.3. *Algorithms and numerical results.* Computing the maximum likelihood estimator $(\hat{\psi}, \hat{\mu})$ from Section 3.1 turns out to be a rather difficult task, because the function $\hat{\Lambda}$ can have multiple local maxima. In [DSS 2010] we discuss strengths and weaknesses of three different algorithms, including an alternating and a stochastic search algorithm. The third procedure, which is highly successful in the case of linear regression, is global maximization of the profile log-likelihood $\hat{\Lambda}(\theta) := \max_{\phi\in\Phi} \hat{\Lambda}(\phi, m_\theta)$, where $m_\theta(x) = \theta^\top x$ for every $x \in \mathbb{R}^q$, by means of differential evolution [Price, Storn and Lampinen (2005)].

Extensive simulation studies in [DSS 2010] suggest that $(\hat{\psi}, \hat{\mu})$ provides rather accurate estimates even if n is only moderately large. For various skewed error distributions, $\hat{\mu}$ may be considerably better than the corresponding least squares estimator. As an example consider the simple linear regression model with observations

$$Y_i = c + \theta X_i + \varepsilon_i, \qquad 1 \le i \le n := 100,$$

where $X_1, \ldots, X_n$ are independent design points from the Unif[0, 3] distribution and $\varepsilon_1, \ldots, \varepsilon_n$ are independent errors from a centered gamma distribution with shape parameter r and variance 1. Note that the distribution of $(\hat{\psi}, \hat{\theta} - \theta)$ does not depend on c or θ. Monte Carlo estimation of the root mean squared error based on 1000 simulations of this model gives 0.023 for the estimator $\hat{\theta}$ versus 0.118 for the least squares estimator of θ if $r = 1$, and 0.095 versus 0.113 for the same comparison if $r = 3$.

3.4. *A data example.* A familiar task in econometrics is to model expenditure (Y) of households as a function of their income (X). Not only the mean curve (Engel curve) but also quantile curves play an important role. A related application are growth charts in which, for instance, X is the age of a newborn or infant and Y is its height or weight.

We applied our methods to a survey of $n = 7125$ households in the United Kingdom in 1973 (data courtesy of W. Härdle, HU Berlin). The two variables we considered were annual income (X_{raw}) and annual expenditure for food (Y_{raw}). Figure 2 shows scatter plots of the raw and log-transformed data. To enhance visibility we only show a random subsample of size $n' = 1000$. In addition, isotonic quantile curves $x \mapsto \hat{q}_\beta(x)$ are added for $\beta = 0.1, 0.25, 0.5, 0.75, 0.9$ (based on all observations). These pictures show clearly that the raw data are heteroscedastic, whereas for the log-transformed data, $(X_i, Y_i) = (\log_{10} X_{\mathrm{raw},i}, \log_{10} Y_{\mathrm{raw},i})$, an additive model seems appropriate.

Interestingly, neither linear nor quadratic nor cubic regression yield convincing fits to these data. Polynomial regression of degree four or cubic splines with knot points at, say, 4.1, 4.3, 4.5, 4.7, 4.9 seem to fit the data quite well. Moreover, exact Monte Carlo goodness-of-fits test, assuming the regression function to be a cubic spline and based on a Kolmogorov–Smirnov statistic applied to studentized residuals, revealed the regression errors ε_i to be definitely non-Gaussian.

Figure 3 shows the data and estimated β-quantile curves for $\beta = 0.1, 0.25, 0.5, 0.75, 0.9$, based on our additive regression model. Note that the estimated β-quantile curve is simply the estimated mean curve plus the β-quantile of the estimated error distribution. On the left-hand side, we only assumed μ to be non-decreasing, on the right-hand side we fitted the aforementioned spline model. In both cases the fitted quantile curves are similar to the quantile curves in Figure 2

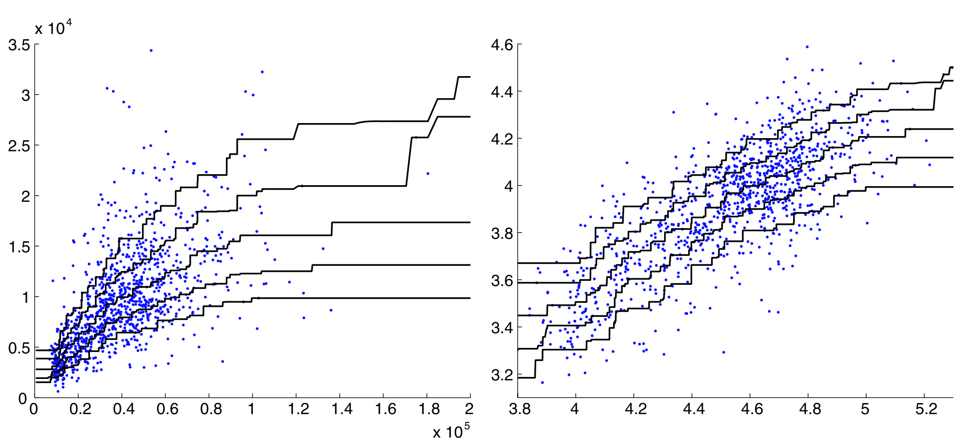

FIG. 2. *UK household data, raw (left) and log-transformed (right), with isotonic quantile curves.*

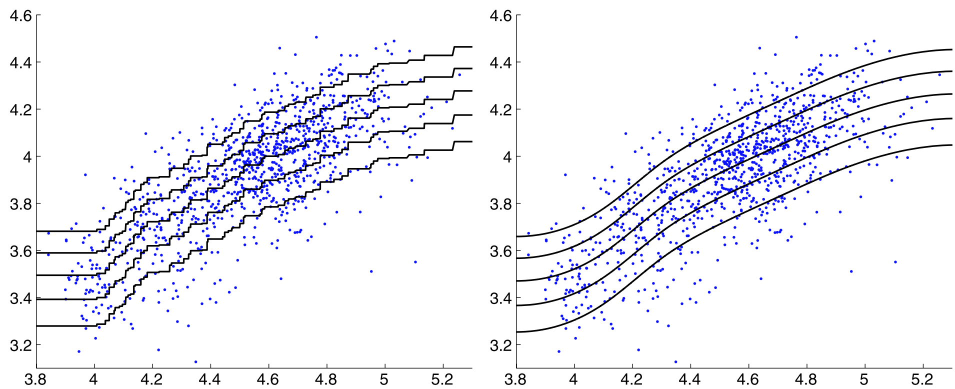

FIG. 3. *Log-transformed UK household data with isotonic fits (left) and spline fits (right) from our additive model.*

but with fewer irregularities such as big jumps which may be artifacts due to sampling error.

4. Proofs. 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 (DSS 2010). *Let $\phi \in \Phi$ be such that $\int e^{\phi(x)}\,dx = 1$. For real t define the level set $D_t := \{x \in \mathbb{R}^d : \phi(x) \geq t\}$. Then for $r < M \leq \max_{x\in\mathbb{R}^d} \phi(x)$,*

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

Another key ingredient for the proofs of Theorems 2.2 and 2.15 is a lemma on pointwise limits of sequences in Φ:

LEMMA 4.2 (DSS 2010). *Let $\bar{\phi}$ and $\phi_1, \phi_2, \phi_3, \ldots$ be functions in Φ such that $\phi_n \leq \bar{\phi}$ for all $n \in \mathbb{N}$. Further suppose that the set*

$$C := \Big\{x \in \mathbb{R}^d : \liminf_{n\to\infty} \phi_n(x) > -\infty\Big\}$$

is nonempty. Then there exist a subsequence $(\phi_{n(k)})_k$ of $(\phi_n)_n$ and a function $\phi \in \Phi$ such that $C \subseteq \mathrm{dom}(\phi) = \{\phi > -\infty\}$ and

$$\lim_{k\to\infty, x\to y} \phi_{n(k)}(x) = \phi(y) \qquad \text{for all } y \in \mathrm{interior}(\mathrm{dom}(\phi)),$$

$$\limsup_{k\to\infty, x\to y} \phi_{n(k)}(x) \leq \phi(y) \leq \bar{\phi}(y) \qquad \text{for all } y \in \mathbb{R}^d.$$

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

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

For the remainder of this proof suppose that $\int \|x\| Q(dx) < \infty$ and that csupp(Q) has nonempty interior. Since the concave function $h(x) = -\|x\|$ satisfies $\int h\,dQ > -\infty$, we have $L(Q) > -\infty$. When maximizing $L(\phi, Q)$ over all $\phi \in \Phi$ we may and do restrict our attention to functions $\phi \in \Phi$ such that $\int e^{\phi(x)}\,dx = 1$ (see end of Section 1) and $\mathrm{dom}(\phi) = \{\phi > -\infty\} \subseteq \mathrm{csupp}(Q)$. For if $\mathrm{dom}(\phi) \not\subset \mathrm{csupp}(Q)$, replacing $\phi(x)$ with $-\infty$ for all $x \notin \mathrm{csupp}(Q)$ would also increase $L(\phi, Q)$ strictly. Let $\Phi(Q)$ be the family of all $\phi \in \Phi$ with these properties.

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

$$
\begin{aligned}
L(\phi, Q) = \int \phi\,dQ &\leq -cM\,Q(\mathbb{R}^d \setminus D_{-cM}) + M\,Q(D_{-cM}) \\
&= -(c+1)M\left(\frac{c}{c+1} - Q(D_{-cM})\right).
\end{aligned}
$$

According to Lemma 4.1,

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

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

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

whence

$$
L(\phi, Q) \to -\infty \qquad \text{as } \max_{x \in \mathbb{R}^d} \phi(x) \to \infty.
$$

Note also that $L(\phi, Q) \leq \max_{x \in \mathbb{R}^d} \phi(x)$ for any $\phi \in \Phi(Q)$. These considerations show that $L(Q)$ is finite and, for suitable constants $M_o < M_*$, equals the supremum of $L(\phi, Q)$ over all $\phi \in \Phi(Q)$ such that $M_o \leq \max_x \phi(x) \leq M_*$.

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

$$
\inf_{n \geq 1} \phi_n(x_o) > -\infty \qquad \text{for any } x_o \in \mathrm{interior}(\mathrm{csupp}(Q)). \tag{5}
$$

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

$$\begin{aligned}\int \phi_n \, dQ &\leq \phi_n(x_o) + \big(M_n - \phi_n(x_o)\big) Q\{\phi_n \geq \phi_n(x_o)\} \\ &\leq \phi_n(x_o) + \big(M_n - \phi_n(x_o)\big) h(Q, x_o) \\ &\leq \phi_n(x_o)\big(1 - h(Q, x_o)\big) + \max(M_n, 0)\end{aligned}$$

with $h(Q, x_o) < 1$ defined in Lemma 2.13. In the case of $\phi_n(x_o) = M_n$ these inequalities are true as well. Thus

$$\phi_n(x_o) \geq -\frac{\max(M_n, 0) - L(\phi_n, Q)}{1 - h(Q, x_o)} \geq -\frac{\max(M_*, 0) - L(\phi_1, Q)}{1 - h(Q, x_o)},$$

which establishes (5). Combining (5) with $\phi_n \leq M_*$, we may deduce from Lemma 3.3 of Schuhmacher, Hüsler and Dümbgen (2009) that there exist constants a and $b > 0$ such that

(6) $$\phi_n(x) \leq a - b\|x\| \qquad \text{for all } n \in \mathbb{N}, x \in \mathbb{R}^d.$$

The inequalities (5) and (6) and Lemma 4.2 with $C \supset$ interior(csupp(Q)) and $\bar{\phi}(x) := a - b\|x\|$ imply existence of a function $\psi \in \Phi$ and a subsequence $(\phi_{n(k)})_k$ of $(\phi_n)_n$ such that $\psi = -\infty$ on $\mathbb{R}^d \setminus \text{csupp}(Q)$ and

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

Since the boundary of csupp(Q) 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 $x \mapsto a - b\|x\| - \phi_{n(k)}(x)$ yields

$$\limsup_{k\to\infty} \int \phi_{n(k)} \, dQ \leq \int \psi \, dQ.$$

Hence

$$L(Q) \geq L(\psi, Q) \geq \limsup_{k\to\infty} L\big(\phi_{n(k)}, Q\big) = L(Q)$$

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

Uniqueness of the maximizer ψ follows essentially from strict convexity of the exponential function: if $\tilde{\psi} \in \Phi(Q)$ with $L(\tilde{\psi}, Q) > -\infty$, then $L((1-t)\psi + t\tilde{\psi}, Q)$ is strictly concave in $t \in [0, 1]$, unless Leb$\{\psi \neq \tilde{\psi}\} = 0$. But for $\psi, \tilde{\psi} \in \Phi(Q)$, the latter requirement is equivalent to $\psi = \tilde{\psi}$ everywhere. □

In our proofs of Theorems 2.7 and 2.15 we utilize a special approximation scheme for functions in Φ:

LEMMA 4.3 (DSS 2010). *For any function $\phi \in \Phi$ with nonempty domain and any parameter $\varepsilon > 0$ set*

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

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

PROOF OF THEOREM 2.7. Let P be the distribution corresponding to F. Suppose first that $\phi = \psi(\cdot|Q)$. Then it follows from (4) and Fubini's theorem that

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

Moreover, for any $x \in \mathbb{R}$, the function $s \mapsto (s-x)^+$ is convex so that (3) and Fubini's theorem yield

$$0 \le \int_{\mathbb{R}} (s-x)^+ (Q-P)(ds) = -\int_{-\infty}^{x} (F-G)(t)\, dt.$$

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

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

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

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

with

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

One can easily verify that $\phi + tH_\delta$ is upper semicontinuous and concave whenever $0 < t \le \phi'(x-\delta) - \phi'(x+\delta)$. In case of $t < -\inf_{u\in\mathbb{R}} \phi'(u)$ it is also coercive. Thus it follows from (2) that

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

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

Now suppose that the distribution function F with log-density $\phi \in \Phi$ satisfies the integral (in)equalities stated in Theorem 2.7. Let $\Delta : \mathbb{R} \to \mathbb{R}$ be Lipschitz-continuous with constant L, so for arbitrary $x, y \in \mathbb{R}$ with $x < y$,

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

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

$$\int \Delta\, d(Q-P) = \int_{\mathbb{R}} \Delta'(t)(F-G)(t)\,dt.$$

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

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

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

$$\int \phi^{(\varepsilon)}\, d(Q-P) = 0.$$

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

$$\int \psi^{(\varepsilon)}\, d(Q-P) \le 0.$$

Since $\phi^{(\varepsilon)} \downarrow \phi$ and $\psi^{(\varepsilon)} \downarrow \psi$ as $\varepsilon \downarrow 0$, and since $\int \phi\, dP$ and $\int \psi\, dQ$ exist in $\mathbb{R}$, we can deduce from monotone convergence that $\int \phi\, d(Q-P) = 0 \geq \int \psi\, d(Q-P)$. Since $\int e^{\phi(x)}\, dx = \int e^{\psi(x)}\, dx = 1$, this entails that

$$L(\phi, Q) = L(\phi, P) \geq L(\psi, P) \geq L(\psi, Q),$$

where the first displayed inequality follows from log-concavity of P with log-density ϕ. Thus $\phi = \psi$. $\square$

Theorem 2.14 and the second part of Theorem 2.15 are a consequence of the following result:

THEOREM 4.4. *Let $(Q_n)_n$ be a sequence of distributions in $\mathcal{Q}_o$ such that $Q_n \to_w Q \in \mathcal{Q}_o$, $L(Q_n) \to \lambda \in [-\infty, \infty]$ and $\int \|x\| Q_n(dx) \to \gamma \in [0, \infty]$ as $n \to \infty$. Then $\gamma \geq \int \|x\| Q(dx)$, and $\lambda > -\infty$ if and only if $\gamma < \infty$. Moreover,*

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

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

$$\lim_{n\to\infty, x\to y} f_n(x) = f(y) \qquad \text{for all } y \in \mathbb{R}^d \setminus \partial\{f > 0\},$$

$$\limsup_{n\to\infty, x\to y} f_n(x) \leq f(y) \qquad \text{for } y \in \partial\{f > 0\},$$

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

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

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

$$\liminf_{n\to\infty} \int h\, dQ_n \geq \int h\, dQ.$$

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

$$\lim_{n\to\infty} \int f\, dQ_n = \int f\, dQ$$

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

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

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

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

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

From now on suppose that $\lambda > -\infty$, and without loss of generality let $L(Q_n) > -\infty$ for all $n \in \mathbb{N}$. We have to show that $\gamma < \infty$ and that $\lambda \leq L(Q)$ with equality if and only if $\gamma = \int \|x\| Q(dx)$. To this end we analyze the functions $\psi_n := \psi(\cdot|Q_n)$ and their maxima $M_n := \max_{x\in\mathbb{R}^d} \psi_n(x)$. First of all,

$$(M_n)_n \text{ is bounded.} \tag{7}$$

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

$$\limsup_{n\to\infty} Q_n(C_n) < 1$$

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 we refer to the proof of Lemma 2.1 in [DSS 2010]: there exist a simplex $\tilde{\Delta} = \mathrm{conv}(\tilde{x}_0, \ldots, \tilde{x}_d)$ with positive Lebesgue measure and open sets $U_0, U_1, \ldots, U_d$ with $Q(U_j) \geq \eta > 0$ for $0 \leq j \leq d$, such that $\tilde{\Delta} \subset C$ for any convex set C with $C \cap U_j \neq \varnothing$ for $0 \leq j \leq d$. But $\liminf_n Q_n(U_j) \geq Q(U_j) \geq \eta$ for all j. Hence $\mathrm{Leb}(C_n) < \mathrm{Leb}(\tilde{\Delta})$ entails that $Q_n(C_n) \leq 1 - \min_{0\leq j\leq d} Q_n(U_j) \leq 1 - \eta + o(1)$ as $n \to \infty$.

Another key property of the functions ψ_n is that

$$\liminf_{n\to\infty} \psi_n(x_o) > -\infty \qquad \text{for any } x_o \in \mathrm{interior}(\mathrm{csupp}(Q)). \tag{8}$$

For

$$L(Q_n) = \int \psi_n\, dQ_n \leq \psi_n(x_o) + \big(M_n - \psi_n(x_o)\big) h(Q_n, x_o),$$

whence as $n \to \infty$,

$$\psi_n(x_o) \geq -\frac{\max(M_n, 0) - L(Q_n)}{1 - h(Q_n, x_o)} \geq -\frac{\limsup_{\ell\to\infty} \max(M_\ell, 0) - \lambda}{1 - h(Q, x_o)} + o(1)$$

by virtue of Lemma 2.13. Combining (5) with (7) we may again deduce that there exist constants a and $b > 0$ such that

$$\psi_n(x) \leq a - b\|x\| \qquad \text{for all } n \in \mathbb{N}, x \in \mathbb{R}^d. \tag{9}$$

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

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

In particular,

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

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 $X_n \sim Q_n$ and $X \sim Q$ 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

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

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

It remains to analyze the case $\gamma = \int \|x\| Q(dx) < \infty$. Here $\lambda \le L(\tilde{\psi}, Q) \le L(Q)$, and it remains to show that $\lambda \ge L(Q)$ which would entail that $\tilde{\psi}$ equals the unique maximizer $\psi := \psi(\cdot|Q)$. With the approximations $\psi^{(1)} \ge \psi^{(\varepsilon)} \ge \psi$, $0 < \varepsilon \le 1$, introduced in Lemma 4.3, it follows from their Lipschitz-continuity and Lemma 4.5 that $\lambda = \lim_{n\to\infty} L(\psi_n, Q_n)$ is not smaller than

$$\lim_{n\to\infty} L\big(\psi^{(\varepsilon)}, Q_n\big) = L\big(\psi^{(\varepsilon)}, Q\big) = \int \psi^{(\varepsilon)}\, dQ - \int \exp\big(\psi^{(\varepsilon)}(x)\big)\, dx + 1.$$

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

$$\lambda \ge \lim_{\varepsilon\downarrow 0} L\big(\psi^{(\varepsilon)}, Q\big) = L(\psi, Q) = L(Q).$$

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

$$\lim_{n\to\infty, x\to y} f_n(x) = f(y) \qquad \text{for all } y \in \mathbb{R}^d \setminus \partial\{f > 0\},$$

$$\limsup_{n\to\infty, x\to y} f_n(x) \le f(y) \qquad \text{for all } y \in \partial\{f > 0\}.$$

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 $(Q_n)_n$. But elementary considerations outlined in [DSS 2010] show that this is sufficient. $\square$

PROOF OF THEOREM 2.15. The assertions of this theorem are essentially covered by Theorem 4.4 as long as $Q \in \mathcal{Q}_o \cap \mathcal{Q}^1$. It only remains to show that $L(Q_n) \to \infty$ if $D_1(Q_n, Q) \to 0$ for some $Q \in \mathcal{Q}^1 \setminus \mathcal{Q}_o$. Thus $\int \|x\| Q(dx) < \infty$ and $Q(H) = 1$ 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 $k \ge 1$ we define $\phi_k \in \Phi$ via

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

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 $\phi_k(x) = \log(k) - \|x\|$ for $x \in H$. Thus

$$L(\phi_k, Q_n) \to L(\phi_k, Q) = \log(k) - \int \|x\| Q(dx) + \int e^{-\|x\|}\,dx.$$

Since the right-hand side may be arbitrarily large, $\lim_{n\to\infty} L(Q_n) = \infty$. $\square$

PROOF OF THEOREM 3.2. Note that $\mathbf{v} \mapsto \hat{Q}_\mathbf{v}$ 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 $\mathbf{Y} \notin \mathcal{M}(\mathbf{x})$, none of the distributions $\hat{Q}_{m(\mathbf{x})}$, $m \in \mathcal{M}$, degenerates to a Dirac measure. According to Theorem 2.15, the mapping $\mathbf{v} \mapsto L(\hat{Q}_\mathbf{v})$ is thus continuous from $\mathcal{M}(\mathbf{x})$ into $\mathbb{R}$.

When proving existence of a maximizer, as explained in Section 3.1, we may restrict our attention to the closed subset $\mathcal{M}(\mathbf{x}, \bar{Y}) := \{\mathbf{v} \in \mathcal{M}(\mathbf{x}) : \bar{v} = \bar{Y}\}$ of $\mathcal{M}(\mathbf{x})$, where generally $\bar{w}$ denotes the arithmetic mean $n^{-1}\sum_{i=1}^n w_i$ for a vector $\mathbf{w} \in \mathbb{R}^n$. But for $\mathbf{v} \in \mathcal{M}(\mathbf{x}, \bar{Y})$,

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

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

$$L(\hat{Q}_\mathbf{v}) \to -\infty \qquad \text{as } \|\mathbf{v}\| \to \infty, \mathbf{v} \in \mathcal{M}(\mathbf{x}, \bar{Y}),$$

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

PROOF OF THEOREM 3.5. The proof that $Q \star R \in \mathcal{Q}_o \cap \mathcal{Q}^1$ is elementary and omitted here. By affine equivariance (Remark 2.4), we may and do assume that $\int y R(dy) = 0$. Now let $\psi := \psi(\cdot|Q)$ and $\tilde{\psi} := \psi(\cdot|Q \star R)$. Then

$$L(Q \star R) = \int\int \tilde{\psi}(x+y) Q(dx) R(dy) = \int \tilde{\psi}_R \, dQ,$$

where

$$\tilde{\psi}_R(x) := \int \tilde{\psi}(x+y) R(dy) \le \tilde{\psi}(x)$$

by Jensen's inequality. Hence

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

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

$$L(Q) = L(Q \star R) \le L(\tilde{\psi}_R, Q) \le L(Q),$$

that is, $\tilde{\psi}_R = \psi = \tilde{\psi}$ and

$$\psi(x) = \int \psi(x+y) R(dy) \qquad \text{for all } x \in \mathbb{R}^d. \tag{10}$$

It remains to be shown that (10) entails $R = \delta_0$. 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

$$M_o = \psi(x(u)) = \int \psi\big(x(u) + y\big) R(dy)$$

implies that $R\{y : u^\top y > 0\} = 0$. Since u is an arbitrary unit vector, this entails that $\operatorname{csupp}(R) = \{0\}$, that is, $R = \delta_0$. □

PROOF OF THEOREM 3.6. Assumptions (A.2) and (A.3) imply that the empirical distribution $\hat{Q}_n := \hat{Q}_{n,\mu_n}$ of the true errors ε_{ni} satisfies both $D_{\mathrm{BL}}(\hat{Q}_n, Q) \to_p 0$ and $\int |t| \hat{Q}_n(dt) \to_p \int |t| Q(dt)$. Thus $D_1(\hat{Q}_n, Q) \to_p 0$.

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

$$D_1(\hat{Q}_n, Q) + \sup_{m \in \mathcal{M} : \|(m-\mu_n)(\mathbf{x}_n)\|_n \le c} D_{\mathrm{BL}}(\hat{Q}_{n,m}, Q_{n,m}) \to 0. \tag{11}$$

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

$$\int |\hat{f}_n(x) - f(x)|\,dx \to 0 \quad \text{and} \quad \|(\hat{\mu}_n - \mu_n)(\mathbf{x}_n)\|_n \to 0. \tag{12}$$

Note that we replaced f_n with $f = \exp \circ \psi(\cdot|Q)$ because $\int |f_n(x) - f(x)|\,dx$ tends to 0.

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 $\int t\,\hat{Q}_{n,m}(dt) = 0$, that is, $\int t\,R_{(\mu_n - m)(\mathbf{x}_n)}(dt) = -\int t\,\hat{Q}_n(dt)$ converges to 0. Since $\{m(\mathbf{x}_n) : m \in \hat{\mathcal{M}}_n\}$ 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 $L(\hat{Q}_{n,m})$ over all $m \in \hat{\mathcal{M}}_n$ does exist. It is possible that $L(\hat{Q}_{n,\hat{\mu}_n}) = \infty$, but if we can show that $D_1(\hat{Q}_{n,\hat{\mu}_n}, Q) \to 0$, then $\hat{f}_n$ exists for sufficiently large n, too. Thus we may rephrase (12) as

$$D_1(\hat{M}_n, Q) \to 0 \quad \text{and} \quad \int |t|\hat{R}_n(dt) \to 0, \tag{13}$$

where $\hat{M}_n := \hat{Q}_{n,\hat{\mu}_n}$ and $\hat{R}_n := R_{(\mu_n - \hat{\mu}_n)(\mathbf{x}_n)}$.

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

$$L(\hat{M}_n) \geq L(\hat{Q}_{n,\check{\mu}_n}) = L(\hat{Q}_n) \to L(Q) \tag{14}$$

by Theorem 2.15. On the other hand

$$\int |t|\hat{M}_n(dt) = \frac{1}{n}\sum_{i=1}^{n} |\varepsilon_{ni} + (\mu_n - \hat{\mu}_n)(x_{ni})| \geq \int |t|\hat{R}_n(dt) - \int |t|\hat{Q}_n(dt).$$

Thus, by Lemma 3.1, $\hat{\mu}_n$ satisfies $\int |t|\hat{R}_n(dt) \leq c$ for sufficiently large $n \in \mathbb{N}$, provided that c is larger than $\int |t|Q(dt) + \exp(-L(Q))$. In particular,

$$D_{\mathrm{BL}}(\hat{M}_n, Q_n \star \hat{R}_n) = D_{\mathrm{BL}}(\hat{M}_n, Q_{n,\hat{\mu}_n}) \to 0.$$

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

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

Since $(\hat{R}_n)_n$ is tight, to verify (13) we may consider a subsequence $(\hat{R}_{n(k)})_k$ that converges weakly to some distribution R as $k \to \infty$. Then $\hat{M}_{n(k)} \to_w Q \star R$, so

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

by Theorems 2.14 and 3.5. Because of (14) we even know that $L(\hat{M}_{n(k)}) \to L(Q \star R) = L(Q)$ as $k \to \infty$. Consequently, we may deduce from Theorems 2.14 and 3.5 that

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

It remains to be shown that $a = 0$ and $\lim_{k\to\infty} \int |t| \hat{R}_{n(k)}(dt) = 0$. Elementary arguments reveal that for arbitrary $r > 0$ and $n \in \mathbb{N}$,

$$\int |t| \hat{M}_n(dt) \geq \int \min(|t|, r) \hat{M}_n(dt) + \int |t| \hat{R}_n(dt) - \int \min(|t|, 2r) \hat{R}_n(dt) - \int (|t| - r)^+ \hat{Q}_n(dt).$$

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

$$\int \min(|t|, 2r) \hat{R}_{n(k)}(dt) + \int (|t| - r)^+ \hat{M}_{n(k)}(dt) + \int (|t| - r)^+ \hat{Q}_{n(k)}(dt) \to \int \min(|t|, 2r) R(dt) + \int (|t| - r)^+ Q \star R(dt) + \int (|t| - r)^+ Q(dt)$$

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

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,

$$D_{\mathrm{KS}}(Q, Q') := \sup_{t\in\mathbb{R}} |(Q' - Q)((-\infty, t])|,$$

of two distributions $Q, Q' \in \mathcal{Q}(1)$:

LEMMA 4.6 (DSS 2010). *Let Q and Q' be distributions on the real line. Then for arbitrary $r > 0$,*

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

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

$$\begin{aligned} &\sup_{m\in\mathcal{M}_n, r\in\mathbb{R}} |(\hat{Q}_{n,m} - Q_{n,m})((-\infty, r])| \\ &\quad = \sup_{b\in\mathbb{R}^{q(n)}, s\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \bigl(1\{Y_{ni} - b^\top x_{ni} \leq s\} - \Pr(Y_{ni} - b^\top x_{ni} \leq s) \bigr) \right| \\ &\quad \leq \sup_{H\in\mathcal{H}_n} |(\hat{M}_n - M_n)(H)|, \end{aligned}$$

where $\mathcal{H}_n$ denotes the family of all closed half-spaces in $\mathbb{R}^{q(n)+1}$ while $\hat{M}_n$ is the empirical distribution of the random vectors $(Y_{ni}, x_{ni}^\top)^\top \in \mathbb{R}^{q(n)+1}$, $1 \leq i \leq n$, and $M_n := \mathbb{E}\hat{M}_n$. 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

$$\mathbb{E} \sup_{m \in \mathcal{M}_n} D_{\mathrm{KS}}(\hat{Q}_{n,m}, Q_{n,m}) \leq C\sqrt{\frac{q(n)+3}{n}}$$

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

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

$$\lim_{n\to\infty} \mathbb{E} \sup_{m \in \mathcal{M}_n : \|(m-\mu_n)(\mathbf{x}_n)\|_n \leq c} D_{\mathrm{BL}}(\hat{Q}_{n,m}, Q_{n,m}) = 0. \qquad \square$$

Acknowledgments. Constructive comments by an Associate Editor and two referees are gratefully acknowledged.

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. MR0069229

BAGNOLI, M. and BERGSTROM, T. (2005). Log-concave probability and its applications. *Econometric Theory* **26** 445–469. MR2213177

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. MR2509075

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression*. Wiley, London.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103

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.

CULE, M. L., SAMWORTH, R. J. and STEWART, M. I. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. Roy. Statist. Soc. Ser. B* **72** 545–607.

DOKSUM, K., OZEKI, A., KIM, J. and NETO, E. C. (2007). Thinking outside the box: Statistical inference based on Kullback–Leibler empirical projections. *Statist. Probab. Lett.* **77** 1201–1213. MR2392791

DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827. MR1193313

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 61, IMSV, Univ. 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. MR2546798

DÜMBGEN, L., SAMWORTH, R. and SCHUHMACHER, D. (2010). Approximation by log-concave distributions with applications to regression. Technical Report 75, IMSV, Univ. Bern. Available at http://arxiv.org/abs/1002.3448.

GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415

KANTOROVIČ, L. V. and RUBINŠTEĬN, G. Š. (1958). On a space of completely additive functions. *Vestnik Leningrad. Univ.* **13** 52–59. MR0102006

KOENKER, R. and MIZERA, I. (2010). Quasi-convex density estimation. *Ann. Statist.* **38** 2998–3027.

MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Ann. Math. Statist.* **43** 508–515. MR0298812

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. IMS, Beachwood, OH. MR2459192

PATILEA, V. (2001). Convex models, MLE and misspecification. *Ann. Statist.* **29** 94–123. MR1833960

PFANZAGL, J. (1990). Large deviation probabilities for certain nonparametric maximum likelihood estimators. *Ann. Statist.* **18** 1868–1877. MR1074441

POLLARD, D. (1990). *Empirical Processes: Theory and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics* **2**. IMS, Hayward, CA. MR1089429

PRICE, K., STORN, R. and LAMPINEN, J. (2005). *Differential Evolution: A Practical Approach to Global Optimization*. Springer, Berlin. MR2191377

RUFIBACH, K. (2006). Log-concave density estimation and bump hunting for i.i.d. observations. Ph.D. thesis, Dept. Mathematics and Statistics, Univ. Bern.

SCHUHMACHER, D. and DÜMBGEN, L. (2010). Consistency of multivariate log-concave density estimators. *Statist. Probab. Lett.* **80** 376–380. MR2593576

SCHUHMACHER, D., HÜSLER, A. and DÜMBGEN, L. (2009). Multivariate log-concave distributions as a nearly parametric model. Technical Report 74, IMSV, Univ. Bern. Available at http://arxiv.org/abs/0907.0250.

SEREGIN, A. and WELLNER, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist.* **38** 3751–3781.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes, with Applications to Statistics*. Springer, New York. MR1385671

VILLANI, C. (2003). *Topics in Optimal Transportation. Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483

WALTHER, G. (2009). Inference and modeling with log-concave distributions. *Statist. Sci.* **24** 319–327.

L. DÜMBGEN
D. SCHUHMACHER
INSTITUTE OF MATHEMATICAL STATISTICS
AND ACTUARIAL SCIENCE
ALPENEGGSTRASSE 22
CH-3012 BERN
SWITZERLAND
E-MAIL: duembgen@stat.unibe.ch
dominic.schuhmacher@stat.unibe.ch

R. SAMWORTH
STATISTICAL LABORATORY
CENTRE FOR MATHEMATICAL SCIENCES
WILBERFORCE ROAD
CAMBRIDGE, CB3 0WB
UNITED KINGDOM
E-MAIL: r.samworth@statslab.cam.ac.uk