Bernoulli **27**(4), 2021, 2437–2472
https://doi.org/10.3150/20-BEJ1316

Local continuity of log-concave projection, with applications to estimation under model misspecification

RINA FOYGEL BARBER[1] and RICHARD J. SAMWORTH[2]

[1]*The University of Chicago Department of Statistics, George Herbert Jones Laboratory, Suite 222, 5747 S. Ellis Avenue, Chicago, IL 60637, United States of America.*
E-mail: rina@uchicago.edu; url: https://www.stat.uchicago.edu/~rina/
[2]*Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, United Kingdom.*
E-mail: r.samworth@statslab.cam.ac.uk; url: http://www.statslab.cam.ac.uk/~rjs57

The log-concave projection is an operator that maps a d-dimensional distribution P to an approximating log-concave density. It is known that, with suitable metrics on the underlying spaces, this projection is continuous, but not uniformly continuous. In this work, we prove a local uniform continuity result for log-concave projection – in particular, establishing that this map is locally Hölder-(1/4) continuous. A matching lower bound verifies that this exponent cannot be improved. We also examine the implications of this continuity result for the empirical setting – given a sample drawn from a distribution P, we bound the squared Hellinger distance between the log-concave projection of the empirical distribution of the sample, and the log-concave projection of P. In particular, this yields interesting statistical results for the misspecified setting, where P is not itself log-concave.

Keywords: Hellinger distance; Hölder continuity; log-concavity; maximum likelihood estimation; Wasserstein distance

1. Introduction

In nonparametric statistics and inference, many problems are formulated in terms of shape constraints. Examples include isotonic regression and convex regression (for supervised learning problems, placing constraints on the shape of the regression function relating the response to the covariates), and monotone or log-concave density estimation (for unsupervised learning problems, placing constraints on a distribution that is the target we wish to estimate).

Among these examples, log-concave density estimation is especially challenging in that it cannot be formulated as an L_2-projection onto a convex constraint set. Remarkably, projection onto the space of log-concave densities can still be uniquely defined, but unlike a convex projection, this operation is not uniformly continuous (Dümbgen, Samworth and Schuhmacher [15]) and its mathematical and statistical properties are therefore difficult to analyze. In this work, we examine the continuity properties of log-concave projection more closely to establish locally uniform convergence, and study the statistical implications of these results.

1.1. Background

We begin by establishing some notation used throughout the paper, and then give background on log-concave projection and its known properties.

1.1.1. *Notation*

Throughout the paper, $\|\cdot\|$ denotes the usual Euclidean norm. For a distribution P, we write $\mathbb{E}_P[\cdot]$ and $\mathbb{P}_P\{\cdot\}$ to denote expectation or probability taken with respect to a random variable or vector X drawn from distribution P, and $\mu_P := \mathbb{E}_P[X]$ denotes its mean. We will analogously write $\mathbb{E}_f[\cdot]$, $\mathbb{P}_f\{\cdot\}$, and μ_f for a density f. We say a distribution, density, or random vector is isotropic if it has zero mean and identity covariance matrix. Given $x \in \mathbb{R}^d$ and $r > 0$, we write $\mathbb{B}_d(x,r) := \{y \in \mathbb{R}^d : \|y - x\| \le r\}$ for the closed Euclidean ball of radius r centered at x, $\mathbb{B}_d(r) = \mathbb{B}_d(0,r)$ for the closed Euclidean ball of radius r centered at zero, and $\mathbb{S}_{d-1}(r) := \{y \in \mathbb{R}^d : \|y\| = r\}$ for the sphere of radius r centered at zero. For the unit ball and unit sphere we write $\mathbb{B}_d = \mathbb{B}_d(1)$ and $\mathbb{S}_{d-1} = \mathbb{S}_{d-1}(1)$. For $x \in \mathbb{R}$, $(x)_+$ denotes $\max\{x, 0\}$, and $(x)_-$ denotes $\max\{-x, 0\}$. For independent observations $X_1, \dots, X_n \in \mathbb{R}^d$, we will write $\widehat{P}_n$ to denote the empirical distribution. We write Leb_d for Lebesgue measure on $\mathbb{R}^d$.

The L_1-Wasserstein distance $\mathrm{d_W}$ is defined for two distributions P, Q on $\mathbb{R}^d$ as

$$\mathrm{d_W}(P, Q) := \inf\left\{\mathbb{E}_{\tilde{P}}\big[\|X - Y\|\big] : \begin{array}{c}\text{Distributions } \tilde{P} \text{ on } (X,Y) \in \mathbb{R}^d \times \mathbb{R}^d \\ \text{such that marginally } X \sim P \text{ and } Y \sim Q\end{array}\right\} \in [0, +\infty].$$

For any distributions P, Q on $\mathbb{R}^d$, this infimum is attained for some coupling $\tilde{P}$ (Villani [50], Theorem 4.1). We will also use the Hellinger distance $\mathrm{d_H}$, defined for densities f, g on $\mathbb{R}^d$ as

$$\mathrm{d}^2_{\mathrm{H}}(f, g) := \int_{\mathbb{R}^d} \big(\sqrt{f(x)} - \sqrt{g(x)}\big)^2 \, \mathrm{d}x.$$

The Hellinger distance is known to satisfy $0 \le \mathrm{d}^2_{\mathrm{H}}(f,g) \le \min\{2, \mathrm{d_{KL}}(f \,\|\, g)\}$ for any densities f, g, where $\mathrm{d_{KL}}(f \,\|\, g) := \mathbb{E}_f[\log(f(X)/g(X))]$ is the Kullback–Leibler divergence. Both $\mathrm{d_W}$ and $\mathrm{d_H}$ satisfy the triangle inequality, while $\mathrm{d_{KL}}$ does not.

1.1.2. *The log-concave projection*

For any $d \in \mathbb{N}$, let $\mathcal{P}_d$ denote the set of probability distributions P on $\mathbb{R}^d$ satisfying $\mathbb{E}_P[\|X\|] < \infty$ and $\mathbb{P}_P\{X \in H\} < 1$ for every hyperplane $H \subseteq \mathbb{R}^d$, that is, P does not place all its mass in any hyperplane. Further, let $\mathcal{F}_d$ denote the set of all upper semi-continuous, log-concave densities on $\mathbb{R}^d$. Then, by Dümbgen, Samworth and Schuhmacher [15], Theorem 2.2, there exists a well-defined projection $\psi^* : \mathcal{P}_d \to \mathcal{F}_d$, given by

$$\psi^*(P) := \operatorname*{argmax}_{f \in \mathcal{F}_d} \mathbb{E}_P\big[\log f(X)\big].$$

When $P \in \mathcal{P}_d$ has a (Lebesgue) density f_P satisfying $\mathbb{E}_{f_P}[|\log f_P(X)|] < \infty$, we can see that $\psi^*(P)$ is the (unique) minimizer over $f \in \mathcal{F}_d$ of the Kullback–Leibler divergence from f_P to f – since the KL divergence acts as a sort of distance, we can think of $f = \psi^*(P)$ as the “closest” log-concave density to f_P, which explains the use of the terminology ‘projection’ to describe this map. In particular, if f_P itself is log-concave, then $\psi^*(P) = f_P$.

To see the gain of defining ψ^* more broadly (i.e., on all distributions $P \in \mathcal{P}_d$, rather than only on distributions with densities), consider the empirical setting, where $\widehat{P}_n$ is the empirical distribution of a sample. Then the result of Dümbgen, Samworth and Schuhmacher [15], Theorem 2.2, tells us that, provided the convex hull of the data is d-dimensional, there exists a unique log-concave maximum likelihood estimator. We can therefore carry out log-concave density estimation via maximum likelihood in much the same way as if the class $\mathcal{F}_d$ were a standard parametric model. To understand the estimation properties of this procedure, suppose we metrise $\mathcal{P}_d$ with the L_1-Wasserstein distance $\mathrm{d_W}$, and metrise $\mathcal{F}_d$ with the Hellinger distance $\mathrm{d_H}$. Then, by Dümbgen, Samworth and Schuhmacher [15], Theorem 2.15, the map ψ^* is continuous. For the empirical distribution $\widehat{P}_n$ obtained by drawing a sample $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$, we therefore have

$$\mathrm{d_H}\big(\psi^*(\widehat{P}_n), \psi^*(P)\big) \overset{\text{a.s.}}{\to} 0.$$

(This follows from the above continuity result because, by Varadarajan's theorem (Dudley [13], Theorem 11.4.1) and the strong law of large numbers, it holds that $d_{\mathrm{W}}(\widehat{P}_n, P) \overset{\text{a.s.}}{\to} 0$.) Thus, if $P \in \mathcal{P}_d$ has a log-concave density, then the log-concave maximum likelihood estimator is strongly consistent – and moreover, even if the log-concavity is misspecified, then the estimator $\psi^*(\widehat{P}_n)$ still converges to the log-concave projection $\psi^*(P)$ of P. In this sense, then, the log-concave maximum likelihood estimator converges to the closest element of $\mathcal{F}_d$ to P, so can be regarded as robust to misspecification.

Despite these positive results establishing continuity and consistency of ψ^*, however, the situation appears much less promising when it comes to obtaining rates of convergence (e.g., via a Lipschitz-type property of the map). Indeed, we cannot hope for Lipschitz continuity of this map, since the review article by Samworth [39] gives the following example to show that ψ^* is not even uniformly continuous: let $P^{(n)} = \mathrm{Unif}[-1/n, 1/n]$ and $Q^{(n)} = \mathrm{Unif}[-1/n^2, 1/n^2]$. Then $d_{\mathrm{W}}(P^{(n)}, Q^{(n)}) \to 0$, but since $P^{(n)}$ and $Q^{(n)}$ have log-concave densities $f^{(n)} := \frac{n}{2}\mathbf{1}_{[-1/n,1/n]}$ and $g^{(n)} := \frac{n^2}{2}\mathbf{1}_{[-1/n^2,1/n^2]}$ respectively, we deduce that

$$\mathrm{d_H}\big(\psi^*\big(P^{(n)}\big), \psi^*\big(Q^{(n)}\big)\big) = \mathrm{d_H}\big(f^{(n)}, g^{(n)}\big) \not\to 0. \tag{1}$$

Summary of contributions. While we have seen that log-concave projection does not satisfy uniform continuity, a natural question is whether it may be possible to place further restrictions on the class $\mathcal{P}_d$ to obtain a result of this type. Moreover, from the statistical point of view, we would like to find a uniform rate of convergence for $\mathrm{d_H}(\psi^*(\widehat{P}_n), \psi^*(P))$, where $\widehat{P}_n$ is the empirical distribution of a sample of size n drawn from $P \in \mathcal{P}_d$, which again might require stronger assumptions than simply $P \in \mathcal{P}_d$.

The first main result of this paper (Theorem 2) reveals that the metric space map $\psi^* : (\mathcal{P}_d, d_{\mathrm{W}}) \to (\mathcal{F}_d, \mathrm{d_H})$ is *locally* Hölder-(1/4) continuous, which establishes a precise understanding of the continuity properties of log-concave projection. Theorem 4 establishes a matching lower bound, revealing that the exponent $1/4$ cannot be improved. Next, we specialise to the empirical setting, proving a bound on $\mathbb{E}_P[\mathrm{d}^2_{\mathrm{H}}(\psi^*(\widehat{P}_n), \psi^*(P))]$ in Theorem 5. For $d \geq 2$, this result is a straightforward consequence of combining our main result in Theorem 2 with the recent work of Lei [34], which bounds $\mathrm{d_W}(\widehat{P}_n, P)$ in expectation, while the case $d = 1$ requires a completely different approach. To the best of our knowledge, this work provides the first understanding of the range of possible rates of convergence of the log-concave maximum likelihood estimator in the misspecified setting.

1.2. Outline of paper

The remainder of the paper is organized as follows. In Section 2, we present our main results, establishing the local Hölder continuity of log-concave projection, and examining the empirical setting, as described above. We review prior work on log-concave projection and related problems in Section 3. The proofs of our main results are presented in Section 4, with technical details deferred to the Appendix.

2. Main results

As mentioned in Section 1, Dümbgen, Samworth and Schuhmacher [15], Theorem 2.15, show that the log-concave projection operator ψ^* satisfies continuity with respect to appropriate metrics:

$$\text{The log-concave projection } \psi^* : (\mathcal{P}_d, \mathrm{d_W}) \to (\mathcal{F}_d, \mathrm{d_H}) \text{ is a continuous map.} \tag{2}$$

Our main results examine the continuity of the log-concave projection operator ψ^* more closely, and establish local uniform continuity results. To do this, we first introduce, for any distribution P on $\mathbb{R}^d$ with $\mathbb{E}_P[\|X\|] < \infty$, the quantity

$$\epsilon_P := \inf_{u \in \mathbb{S}_{d-1}} \mathbb{E}_P\bigl[\bigl|u^\top(X - \mu_P)\bigr|\bigr].$$

The quantity ϵ_P can be thought of as a robust analogue of the minimum eigenvalue of the covariance matrix of the distribution P (note that its definition does not require P to have a finite second moment). We can also interpret ϵ_P as measuring the extent to which P avoids placing all its mass on a single hyperplane.

First, we verify that ϵ_P is positive for all $P \in \mathcal{P}_d$, and is Lipschitz with respect to the Wasserstein distance.

Proposition 1. *We have $\epsilon_P > 0$ for any $P \in \mathcal{P}_d$. Furthermore, $|\epsilon_P - \epsilon_Q| \le 2\mathrm{d_W}(P, Q)$ for any distributions P, Q on $\mathbb{R}^d$ with $\mathbb{E}_P[\|X\|], \mathbb{E}_Q[\|X\|] < \infty$.*

We now present our first main result, which shows that ϵ_P allows for a more detailed analysis of the continuity of the map ψ^*.

Theorem 2. *For any $d \ge 1$ and $P, Q \in \mathcal{P}_d$,*

$$\mathrm{d_H}\bigl(\psi^*(P), \psi^*(Q)\bigr) \le C_d \cdot \left[\frac{\mathrm{d_W}(P, Q)}{\max\{\epsilon_P, \epsilon_Q\}}\right]^{1/4},$$

where $C_d > 0$ depends only on d.

This upper bound immediately implies the continuity result (2), but more importantly, to the best of our knowledge, this is the first general, quantitative statement about the local continuity of log-concave projection. Another consequence is that, when $d = 1$, the uniform continuity counterexample in (1) is in some sense canonical: if $(P^{(n)})$ and $(Q^{(n)})$ are sequences in $\mathcal{P}_1$ satisfying $\mathrm{d_W}(P^{(n)}, Q^{(n)}) \to 0$ and $\liminf_{n\to\infty} \max\{\epsilon_{P^{(n)}}, \epsilon_{Q^{(n)}}\} > 0$, then $\mathrm{d_H}(\psi^*(P^{(n)}), \psi^*(Q^{(n)})) \to 0$.

2.1. Extension to affine transformations

By Dümbgen, Samworth and Schuhmacher [15], Remark 2.4, log-concave projection commutes with affine transformations; that is, if $\psi^*(P) = f$ then $\psi^*(\mathbf{A} \circ P) = \mathbf{A} \circ f$ for any invertible matrix $\mathbf{A}$, where $\mathbf{A} \circ P$ denotes the distribution obtained by drawing $X \sim P$ and returning $\mathbf{A}X$, and similarly $\mathbf{A} \circ f$ denotes the density of the random variable obtained by drawing X according to density f and returning $\mathbf{A}X$.

Turning to the terms appearing in Theorem 2, the Hellinger distance is invariant to affine transformations, but the terms on the right-hand side – namely, $\mathrm{d_W}(P, Q)$ and $\max\{\epsilon_P, \epsilon_Q\}$ – are not. By considering affine transformations, we obtain the following corollary to Theorem 2, which we state without further proof.

Corollary 3. *For any* $d \geq 1$ *and* $P, Q \in \mathcal{P}_d$,

$$\mathrm{d_H}\big(\psi^*(P), \psi^*(Q)\big) \leq C_d \cdot \inf_{\mathbf{A}\in\mathbb{R}^{d\times d}, \mathrm{rank}(\mathbf{A})=d} \left[\frac{\mathrm{d_W}(\mathbf{A}\circ P, \mathbf{A}\circ Q)}{\max\{\epsilon_{\mathbf{A}\circ P}, \epsilon_{\mathbf{A}\circ Q}\}}\right]^{1/4},$$

where $C_d > 0$ *depends only on* d.

2.2. A matching lower bound

To see that our main result in Theorem 2 is optimal in terms of its dependence on the Wasserstein distance $\mathrm{d_W}(P, Q)$ and on the terms ϵ_P, ϵ_Q, we now construct an explicit example to provide a matching lower bound.

Theorem 4. *Fix any* $d \geq 1$, $\epsilon > 0$, *and* $\delta > 0$. *Then there exist distributions* $P, Q \in \mathcal{P}_d$ *with* $\epsilon_P, \epsilon_Q \geq \epsilon$ *and* $\mathrm{d_W}(P, Q) \leq \delta$, *such that*

$$\mathrm{d_H}\big(\psi^*(P), \psi^*(Q)\big) \geq c_d \cdot \min\big\{1, (\delta/\epsilon)^{1/4}\big\},$$

where $c_d > 0$ *depends only on dimension* d.

The theorem will be proved using the following construction: Let $P \in \mathcal{P}_d$ be the uniform distribution on the sphere $\mathbb{S}_{d-1}(\rho)$, where $\rho \propto \epsilon$, and let $Q \in \mathcal{P}_d$ be the mixture distribution that, with probability $\beta \propto \delta/\epsilon$, draws uniformly from $\mathbb{S}_{d-1}(2\rho)$, and with probability $1-\beta$ draws uniformly from $\mathbb{S}_{d-1}(\rho)$. Then $\mathrm{d_W}(P, Q) = \rho\beta \propto \delta$, and we will see that $\mathrm{d_H}(\psi^*(P), \psi^*(Q)) \propto (\delta/\epsilon)^{1/4}$, as desired.

2.3. Bounds for empirical processes

Now let $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P \in \mathcal{P}_d$, with corresponding empirical distribution function $\widehat{P}_n$. Under an additional moment assumption on P, we consider the problem of bounding $\mathrm{d}^2_{\mathrm{H}}(\psi^*(\widehat{P}_n), \psi^*(P))$. However, to be fully precise, we need to consider the possibility that $\psi^*(\widehat{P}_n)$ may not be defined – specifically, if P places positive probability on some hyperplane $H \subseteq \mathbb{R}^d$, then it is possible that the empirical distribution $\widehat{P}_n$ may place all its mass on this hyperplane, in which case we have $\widehat{P}_n \notin \mathcal{P}_d$ and $\psi^*(\widehat{P}_n)$ is not defined. In a slight abuse of notation, for such a case we will interpret $\mathrm{d}^2_{\mathrm{H}}(\psi^*(\widehat{P}_n), \psi^*(P))$ as the maximum possible squared Hellinger distance (i.e., 2).

Theorem 5. *Fix any* $P \in \mathcal{P}_d$, *and assume that*

$$\mathbb{E}_P\big[\|X\|^q\big]^{1/q} \leq M_q$$

for some $q > 1$. *Let* $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P$ *for some* $n \geq 2$, *and let* $\widehat{P}_n$ *denote the corresponding empirical distribution. Then*

$$\mathbb{E}\big[\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(\widehat{P}_n), \psi^*(P)\big)\big] \leq C_{d,q} \cdot \sqrt{\frac{M_q}{\epsilon_P}} \cdot \frac{\log^{3/2} n}{n^{\min\{\frac{1}{2d}, \frac{1}{2}-\frac{1}{2q}\}}},$$

where $C_{d,q} > 0$ *depends only on* d *and* q.

Proof of Theorem 5. First, we consider the case $d \geq 2$. The result will follow by combining the bound (4), obtained from Theorem 2, together with a bound on the expected Wasserstein distance between $\widehat{P}_n$ and P (Lei [34]). Specifically, Lei [34], Theorem 3.1, establishes that[1]

$$\mathbb{E}\big[\mathrm{d_W}(\widehat{P}_n, P)\big] \leq \tilde{C}_q M_q \cdot \frac{\log^2 n}{n^{\min\{\frac{1}{2}, \frac{1}{d}, 1-\frac{1}{q}\}}} \tag{3}$$

for some $\tilde{C}_q > 0$ depending only on q. Furthermore, on the event that $\widehat{P}_n \in \mathcal{P}_d$ (i.e., $\widehat{P}_n$ does not place all its mass in any hyperplane), then by applying Theorem 2 with $Q = \widehat{P}_n$ we have

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(\widehat{P}_n), \psi^*(P)\big) \leq C_d^2 \cdot \frac{\mathrm{d}_{\mathrm{W}}^{1/2}(\widehat{P}_n, P)}{\max\{\epsilon_P^{1/2}, \epsilon_{\widehat{P}_n}^{1/2}\}}.$$

If instead $\widehat{P}_n$ does place all its mass in a hyperplane and so $\psi^*(\widehat{P}_n)$ is undefined, then in this case we have $\epsilon_{\widehat{P}_n} = 0$, and so by Proposition 1, $2\mathrm{d_W}(\widehat{P}_n, P) \geq |\epsilon_{\widehat{P}_n} - \epsilon_P| = \epsilon_P$. Recalling from above that we interpret $\mathrm{d}^2_{\mathrm{H}}(\psi^*(\widehat{P}_n), \psi^*(P))$ as equal to 2 in the case where $\widehat{P}_n \notin \mathcal{P}_d$, we can see that in either case, it holds that

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(\widehat{P}_n), \psi^*(P)\big) \leq \max\big\{C_d^2, \sqrt{8}\big\} \cdot \frac{\mathrm{d}_{\mathrm{W}}^{1/2}(\widehat{P}_n, P)}{\epsilon_P^{1/2}}. \tag{4}$$

Now, taking the expected value and combining the bounds (3) and (4), we obtain

$$\begin{aligned}
\mathbb{E}\big[\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(\widehat{P}_n), \psi^*(P)\big)\big] &\leq \mathbb{E}\left[\max\big\{C_d^2, \sqrt{8}\big\} \cdot \frac{\mathrm{d}_{\mathrm{W}}^{1/2}(\widehat{P}_n, P)}{\epsilon_P^{1/2}}\right] \\
&\leq \max\big\{C_d^2, \sqrt{8}\big\} \cdot \left[\frac{\mathbb{E}[\mathrm{d_W}(\widehat{P}_n, P)]}{\epsilon_P}\right]^{1/2} \\
&\leq \max\big\{C_d^2, \sqrt{8}\big\}\sqrt{\tilde{C}_q} \cdot \sqrt{\frac{M_q}{\epsilon_P}} \cdot \frac{\log n}{n^{\min\{\frac{1}{4}, \frac{1}{2d}, \frac{1}{2} - \frac{1}{2q}\}}}.
\end{aligned}$$

Choosing $C_{d,q} = \max\{C_d^2, \sqrt{8}\} \cdot \sqrt{\tilde{C}_q}$, this proves the desired result for the case $d \geq 2$.

For the case $d = 1$, the result cannot be proved with the same argument, as the exponent on n in the bound above is at best $1/4$, which does not lead to the desired scaling if $q > 2$. We establish the desired bound for $d = 1$ in Section 4.4, using a more technical argument. □

We remark that, if X is additionally assumed to be subexponential, then Lei [34], Corollary 5.2, establishes exponential tail bounds for $\mathrm{d_W}(\widehat{P}_n, P)$; under this stronger assumption, the results of Theorem 5 could then be strengthened to give a tail bound for $\mathrm{d}^2_{\mathrm{H}}(\psi^*(\widehat{P}_n), \psi^*(P))$, in place of the bound on expected value.

[1]In fact, Lei [34], Theorem 3.1, shows that the $\log^2 n$ term may be reduced to $(\log n)\mathbf{1}_{\{d=1,q=2\}} + (\log n)\mathbf{1}_{\{d=2,q>2\}} + (\log^2 n)\mathbf{1}_{\{d=2,q=2\}} + (\log n)\mathbf{1}_{\{d\geq 3,q=d/(d-1)\}}$. Since poly-logarithmic factors are not our primary concern in this work, however, we will present simpler bounds based on (3).

2.3.1. *Lower bounds for the empirical setting*

Our final main result studies the optimality of the power of n appearing in Theorem 5.

Theorem 6. *For any $d \geq 1$ and $q > 1$, there exist $\epsilon^*_d, c_d > 0$, depending only on d, such that*

$$\sup_{P \in \mathcal{P}_d : \mathbb{E}_P[\|X\|^q] \leq 1, \epsilon_P \geq \epsilon^*_d} \mathbb{E}\big[\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(\widehat{P}_n), \psi^*(P)\big)\big] \geq c_d \cdot n^{-\min\{\frac{2}{d+1}, \frac{1}{2} - \frac{1}{2q}\}}.$$

Ignoring a logarithmic factor in n, the first term, namely $n^{-\frac{2}{d+1}}$, is the known minimax rate for *any* estimator under the well-specified case where P is itself log-concave, for any $d \geq 2$ (Kim and Samworth [31], Kur, Dagan and Rakhlin [33]). The second term is a new result and will be proved via a misspecified construction where P is not log-concave: the distribution is given by $X = R \cdot U$, where U is drawn uniformly from the unit sphere $\mathbb{S}_{d-1}$, while the radius R is drawn independently with

$$R = \begin{cases} 1/2, & \text{with probability } 1 - 1/2n, \\ n^{1/q}, & \text{with probability } 1/2n. \end{cases}$$

The intuition is that, with positive probability, the empirical distribution $\widehat{P}_n$ (and, therefore, its log-concave projection $\psi^*(\widehat{P}_n)$), is supported on the ball of radius $1/2$; on the other hand, we will see in the proof that $\psi^*(P)$ places $\sim n^{-\frac{1}{2}+\frac{1}{2q}}$ mass outside this ball, leading to a lower bound on the Hellinger distance between these two log-concave projections.

A consequence of this last result in dimension $d = 1$ is that rates of convergence in log-concave density estimation can be much slower in the misspecified setting, with a minimax rate of $n^{-1/2}$ at best, as compared to the well-specified setting when P is assumed to have a log-concave density, where the corresponding rate is $n^{-4/5}$ (Kim and Samworth [31]).

2.3.2. *A gap for dimension $d \geq 2$*

Comparing the lower bound established in Theorem 6 with the upper bound given in Theorem 5, we see that for the case $d = 1$ the two bounds match, as they both scale as $n^{-\frac{1}{2}+\frac{1}{2q}}$ (ignoring poly-logarithmic factors). For $d \geq 2$, however, there is a gap – for sufficiently large q (i.e., a sufficiently strong moment condition), the upper bound scales as $n^{-\frac{1}{2d}}$ (up to poly-logarithmic factors) while the lower bound has the faster rate $n^{-\frac{2}{d+1}}$. We also remark that the optimal dependence of the minimax rate on d remains unknown as well.

3. Relationship with prior work

Log-concave density estimation is a central problem within the field of nonparametric inference under shape constraints. Entry points to the field include the book by Groeneboom and Jongbloed [22], as well as the 2018 special issue of the journal *Statistical Science* (Samworth and Sen [40]). Other important shape-constrained problems that could benefit from the perspective taken in this work include decreasing density estimation (Grenander [20], Prakasa Rao [38], Groeneboom [21], Birgé [4], Jankowski [29]), isotonic regression (Barlow *et al.* [2], Zhang [56], Chatterjee, Guntuboyina and Sen [7], Durot and Lopuhaä [16], Bellec [3], Yang and Barber [55], Han *et al.* [25]) and convex regression (Hildreth [28], Seijo and Sen [44], Cai and Low [5], Guntuboyina and Sen [23], Han and Wellner [26], Fang and Guntuboyina [17]), among many others. In these cases, the analysis is likely to be more straightforward, since the canonical least squares/maximum likelihood estimator can be characterised as an L_2-projection onto a convex set. By contrast, the class $\mathcal{F}_d$ is not convex, and the Kullback–Leibler projection ψ^* is considerably more involved.

Early work on log-concave density estimation includes Walther [52], Pal, Woodroofe and Meyer [36], Dümbgen and Rufibach [14], Walther [53], Cule, Samworth and Stewart [11], Cule and Samworth [10], Schuhmacher, Hüsler and Dümbgen [43], Samworth and Yuan [41] and Chen and Samworth [8]. Sometimes, the class is considered as a special case of the class of s-concave densities (Koenker and Mizera [32], Seregin and Wellner [45], Han and Wellner [27], Doss and Wellner [12], Han [24]). For the case of correct model specification, where P has density $f_P \in \mathcal{F}_d$ and $\widehat{f}_n := \psi^*(\widehat{P}_n)$, it is now known (Kim and Samworth [31], Kur, Dagan and Rakhlin [33]) that

$$\sup_{f_P \in \mathcal{F}_d} \mathbb{E}\big[\mathrm{d}_{\mathrm{H}}^2(\widehat{f}_n, f_P)\big] \leq K_d \cdot \begin{cases} n^{-4/5} & \text{when } d = 1 \\ n^{-2/(d+1)} \log n & \text{when } d \geq 2, \end{cases}$$

where $K_d > 0$ depends only on d, and that this risk bound is minimax optimal (up to the logarithmic factor when $d \geq 2$). See also Carpenter *et al.* [6] for an earlier result in the case $d \geq 4$, and Xu and Samworth [54] for an alternative approach to high-dimensional log-concave density estimation that seeks to evade the curse of dimensionality in the additional presence of symmetry constraints. It is further known that when $d \leq 3$, the log-concave maximum likelihood estimator can adapt to certain subclasses of log-concave densities, including log-concave densities whose logarithms are piecewise affine (Kim, Guntuboyina and Samworth [30], Feng *et al.* [18]). Although these recent works provide a relatively complete picture of the behaviour of the log-concave maximum likelihood estimator when the true distribution has a log-concave density, there is almost no prior work on risk bounds under model misspecification. The only exception of which we are aware is Kim, Guntuboyina and Samworth [30], Theorem 1, which considers a univariate case where the true distribution has a density that is very close to log-affine on its support.

One feature that distinguishes our contributions from earlier work on rates of convergence in log-concave density estimation in the correctly specified setting is that our arguments avoid entirely notions of bracketing entropy, as well as empirical process arguments that control the behaviour of M-estimators in terms of the entropy of a relevant function class (e.g., van der Vaart and Wellner [49], van de Geer [48]). It turns out that, for non-convex classes of densities, these ideas are not well suited to the misspecified setting.[2] Instead, our main tool is a detailed and delicate analysis of the Lipschitz approximations to concave functions introduced in Dümbgen, Samworth and Schuhmacher [15]. In their original usage, these were employed in conjunction with asymptotic results such as Skorokhod's representation theorem to derive the consistency and robustness results described above. By contrast, our analysis facilitates the direct inequality established in Theorem 2.

Another role of this work is to advocate for the benefits of regarding an estimator as a function of the empirical distribution, as opposed to the more conventional view where it is seen as a function on the sample space. The empirical distribution $\widehat{P}_n$ of a sample $X_1, \dots, X_n$ encodes all of the information in the data when we regard it as a multi-set $\{X_1, \dots, X_n\}$, that is, when we discard information in the ordering of the indices. It follows that any statistic $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \dots, X_n)$ that is invariant to permutation of its arguments can be thought of as a functional $\theta(\widehat{P}_n)$ of the empirical distribution. Frequently, the definition of θ can be extended to a more general class of distributions $\mathcal{P}$, and we may regard θ

[2]See Patilea [37], Proposition 4.1, for applications of entropy methods to studying rates of convergence of maximum likelihood estimators for *convex* classes of densities. However, the class of densities f that are log-concave is not a convex class; if we instead consider the class of concave log-densities (i.e., $\log f$, where f is a log-concave density), then this class is also not convex, because of the need for the exponentials of these log-densities to integrate to 1.

as a *projection* from $\mathcal{P}$ onto a model, or parameter space, Θ. This perspective, which was pioneered by Richard von Mises in the 1940s (von Mises [51]) and described in Serfling [46], Chapter 6, offers many advantages to the statistician. In particular, once the analytical properties (e.g., continuity, differentiability) of θ are understood, key statistical properties of the estimator (consistency, robustness to misspecification, rates of convergence), can often be deduced as simple corollaries of basic facts about the convergence of empirical distributions.

4. Proofs of upper bounds

In this section, we prove Theorem 2 (for arbitrary dimension d), and complete the proof of Theorem 5 (for the remaining case of dimension $d = 1$). In Section 4.1, we review some known properties of log-concave projection, and in Section 4.2 we establish a key lemma that will be used in both proofs. In Section 4.3, we complete the proof of Theorem 2, and in Section 4.4 we complete the proof of Theorem 5 for the remaining case $d = 1$.

4.1. Background on log-concave projection

We begin by reviewing some known properties of log-concave projection, and computing some new bounds.

4.1.1. *Moment inequalities*

The log-concave projection ψ^* is known to satisfy a useful convex ordering property (Dümbgen, Samworth and Schuhmacher [15], Eqn. (3)): for any $P \in \mathcal{P}_d$ and for $f = \psi^*(P)$,

$$\mathbb{E}_f\big[h(X)\big] \le \mathbb{E}_P\big[h(X)\big] \quad \text{for any convex function } h : \mathbb{R}^d \to (-\infty, \infty]. \tag{5}$$

In particular, this implies that

$$\mathbb{E}_f\big[\big|v^\top(X - \mu_P)\big|\big] \le \mathbb{E}_P\big[\big|v^\top(X - \mu_P)\big|\big] \quad \text{for all } v \in \mathbb{R}^d.$$

The following lemma establishes that, up to a constant, this inequality is tight for all vectors $v \in \mathbb{R}^d$.

Lemma 7. *Fix any $P \in \mathcal{P}_d$, and let $f = \psi^*(P)$. Then*

$$\mathbb{E}_f\big[\big|v^\top(X - \mu_P)\big|\big] \ge c_d \cdot \mathbb{E}_P\big[\big|v^\top(X - \mu_P)\big|\big] \quad \textit{for all } v \in \mathbb{R}^d,$$

where $c_d \in (0, 1]$ depends only on d.

By Dümbgen, Samworth and Schuhmacher [15], Eqn. (4), log-concave projection preserves the mean, that is,

$$\mu_P = \mathbb{E}_P[X] = \mathbb{E}_f[X].$$

We can also define the covariance matrix $\Sigma = \mathrm{Cov}_f(X)$, which is finite (since all moments of a log-concave distribution are finite) and strictly positive definite. Lemma 7 immediately implies bounds on the eigenvalues of Σ.

Corollary 8. *Fix any $P \in \mathcal{P}_d$, let $f = \psi^*(P)$, and let $\Sigma = \mathrm{Cov}_f(X)$ be the covariance matrix of the distribution with density f. Then for all $v \in \mathbb{R}^d$,*

$$c_d^2\{\mathbb{E}_P[|v^\top(X-\mu_P)|]\}^2 \le v^\top\Sigma v \le 16\{\mathbb{E}_P[|v^\top(X-\mu_P)|]\}^2,$$

where $c_d \in (0,1]$ is taken from Lemma 7. In particular, this implies that

$$\lambda_{\min}(\Sigma) \ge (c_d\epsilon_P)^2,$$

where $\lambda_{\min}(\Sigma)$ denotes the smallest eigenvalue of Σ.

Proof of Corollary 8. First, for the lower bound, by Lemma 7 and Cauchy–Schwarz,

$$c_d^2\{\mathbb{E}_P[|v^\top(X-\mu_P)|]\}^2 \le \{\mathbb{E}_f[|v^\top(X-\mu_P)|]\}^2 \le \mathbb{E}_f[|v^\top(X-\mu_P)|^2] = v^\top\Sigma v.$$

Next, for the upper bound,

$$v^\top\Sigma v = \mathbb{E}_f[|v^\top(X-\mu_P)|^2] \le 16\{\mathbb{E}_f[|v^\top(X-\mu_P)|]\}^2 \le 16\{\mathbb{E}_P[|v^\top(X-\mu_P)|]\}^2,$$

where the first inequality is due to Lovász and Vempala [35], Theorem 5.22, while the second is by (5) (Dümbgen, Samworth and Schuhmacher [15], Eqn. (3)). □

4.1.2. *A lower bound on a ball*

Next we show that for any P, its log-concave projection $f = \psi^*(P)$ is lower bounded on a ball of radius of order ϵ_P.

Lemma 9. *Fix any $P \in \mathcal{P}_d$, and let $f = \psi^*(P)$. Then there exist $b_d, r_d \in (0,1]$, depending only on d, such that*

$$f(x) \ge b_d \cdot \sup_{x' \in \mathbb{R}^d} f(x') \quad \textit{for all } x \in \mathbb{B}_d(\mu_P, r_d\epsilon_P).$$

Proof of Lemma 9. Let $\Sigma = \mathrm{Cov}_f(X)$, and define the isotropic, log-concave density $g(x) = f(\Sigma^{1/2}x + \mu_P)\det^{1/2}(\Sigma)$. By Lovász and Vempala [35], Theorem 5.14(a) and (b),

$$\inf_{x:\|x\|\le 1/9} g(x) \ge b_d \sup_{x \in \mathbb{R}^d} g(x),$$

where $b_d \in (0,1]$ depends only on d. This immediately implies that

$$f(x) \ge b_d \sup_{x' \in \mathbb{R}^d} f(x') \quad \text{for all } x \in \mathbb{R}^d \text{ with } \|\Sigma^{-1/2}(x-\mu_P)\| \le 1/9.$$

But $\|\Sigma^{-1/2}(x-\mu_P)\| \le \lambda_{\min}^{-1/2}(\Sigma)\|x-\mu_P\| \le \|x-\mu_P\|/(c_d\epsilon_P)$ by Corollary 8, so the result holds with $r_d = c_d/9$. □

4.2. Key lemma: The Lipschitz majorization

Let

$$\Phi_d := \left\{\phi : \mathbb{R}^d \to [-\infty,\infty) : \begin{array}{c}\phi \text{ is a proper concave, upper semi-continuous function,}\\ \text{and } \phi(x) \to -\infty \text{ as } \|x\| \to \infty\end{array}\right\},$$

and define the function $\phi^* : \mathcal{P}_d \to \Phi_d$ that maps a distribution P to the log-density $\phi = \phi^*(P)$ given by $\phi(x) = \log[\psi^*(P)](x)$. Dümbgen, Samworth and Schuhmacher [15], Theorem 2.2, establishes that the log-density $\phi = \phi^*(P)$ maximizes $\ell(\phi, P) := \mathbb{E}_P[\phi(X)] - \int_{\mathbb{R}^d} e^{\phi(x)}\,\mathsf{d}x + 1$ over Φ_d. We now show that this maximum can be nearly attained by a Lipschitz function. In particular, for any $\phi \in \Phi_d$ and any $L > 0$, define its L-Lipschitz majorization $\phi^L : \mathbb{R}^d \to \mathbb{R}$ by

$$\phi^L(x) := \sup_{y \in \mathbb{R}^d} \bigl\{\phi(y) - L\|x - y\|\bigr\}. \tag{6}$$

It can easily be verified that this function is concave, L-Lipschitz, and satisfies $\phi^L(x) \geq \phi(x)$ for all $x \in \mathbb{R}^d$. Furthermore, it holds that $\int_{\mathbb{R}^d} e^{\phi^L(x)}\,\mathsf{d}x < \infty$ (this follows from the fact that there exist constants $a \in \mathbb{R}$, $b > 0$ such that $\phi(y) \leq a - b\|y\|$ for all $y \in \mathbb{R}^d$ (Dümbgen, Samworth and Schuhmacher [15])), and moreover $\int_{\mathbb{R}^d} e^{\phi^L(x)}\,\mathsf{d}x > 0$.

Next we normalize to produce a log-density. For any $\phi \in \Phi_d$, we define

$$\tilde{\phi}^L(x) := \phi^L(x) - \log\biggl(\int_{\mathbb{R}^d} e^{\phi^L(x)}\,\mathsf{d}x\biggr). \tag{7}$$

The following result proves that, if $\phi = \phi^*(P)$, then for L sufficiently large, $\tilde{\phi}^L \in \Phi_d$ is nearly optimal for P (in the sense of maximizing $\ell(\cdot, P)$).

Lemma 10. *Fix any $P \in \mathcal{P}_d$, let $\phi = \phi^*(P)$, and let ϕ^L and $\tilde{\phi}^L$ be defined as in (6) and (7). Then for any $L \geq \frac{2d}{r_d \epsilon_P}$,*

$$\ell\bigl(\tilde{\phi}^L, P\bigr) \geq \ell\bigl(\phi^L, P\bigr) \geq \ell(\phi, P) - \frac{4d}{Lb_d r_d \epsilon_P},$$

where $r_d, b_d \in (0, 1]$ are taken from Lemma 9. In particular, this implies that

$$\mathbb{E}_P\bigl[\tilde{\phi}^L(X)\bigr] \geq \mathbb{E}_P\bigl[\phi(X)\bigr] - \frac{4d}{Lb_d r_d \epsilon_P}.$$

4.2.1. *Bounding the Hellinger distance*

Now we apply Lemma 10 to the problem of bounding Hellinger distance.

Corollary 11. *Fix any $P, Q \in \mathcal{P}_d$, and define $\epsilon = \min\{\epsilon_P, \epsilon_Q\} > 0$. Let $\phi_P = \phi^*(P)$ and $\phi_Q = \phi^*(Q)$, and let $f_P = \psi^*(P)$ and $f_Q = \psi^*(Q)$ be the corresponding density functions. Let ϕ_P^L and ϕ_Q^L be the L-Lipschitz majorizations of ϕ_P and ϕ_Q, respectively, as defined in (6), for some $L \geq \frac{2d}{r_d \epsilon}$, where $r_d \in (0, 1]$ is taken from Lemma 9. Then*

$$\mathrm{d}_{\mathrm{H}}^2(f_P, f_Q) \leq \frac{16d}{Lb_d r_d \epsilon} + \bigl(\mathbb{E}_P\bigl[\phi_P^L(X)\bigr] - \mathbb{E}_Q\bigl[\phi_P^L(X)\bigr]\bigr) + \bigl(\mathbb{E}_Q\bigl[\phi_Q^L(X)\bigr] - \mathbb{E}_P\bigl[\phi_Q^L(X)\bigr]\bigr),$$

where $b_d \in (0, 1]$ is taken from Lemma 9.

Proof of Corollary 11. Let $\tilde{\phi}_P^L, \tilde{\phi}_Q^L$ be defined as in (7), and let $\tilde{f}_P^L, \tilde{f}_Q^L$ be the corresponding densities, that is, $\tilde{f}_P^L(x) = e^{\tilde{\phi}_P^L(x)}$ and similarly for $\tilde{f}_Q^L$. We first calculate

$$\mathrm{d}_{\mathrm{KL}}\bigl(f_P \parallel \tilde{f}_P^L\bigr) = \mathbb{E}_{f_P}\bigl[\phi_P(X) - \tilde{\phi}_P^L(X)\bigr] \leq \mathbb{E}_P\bigl[\phi_P(X) - \tilde{\phi}_P^L(X)\bigr]$$

and

$$\mathrm{d_{KL}}\big(f_P \,\|\, \tilde{f}_Q^L\big) = \mathbb{E}_{f_P}\big[\phi_P(X) - \tilde{\phi}_Q^L(X)\big] \le \mathbb{E}_P\big[\phi_P(X) - \tilde{\phi}_Q^L(X)\big],$$

where the inequalities hold by Dümbgen, Samworth and Schuhmacher [15], Remark 2.3. The same bounds hold with the roles of P and Q reversed. Furthermore, by the triangle inequality,

$$\begin{aligned}
\mathrm{d}_\mathrm{H}^2(f_P, f_Q) &= \frac{1}{2}\mathrm{d}_\mathrm{H}^2(f_P, f_Q) + \frac{1}{2}\mathrm{d}_\mathrm{H}^2(f_P, f_Q) \\
&\le \frac{1}{2}\big\{\mathrm{d_H}\big(f_P, \tilde{f}_P^L\big) + \mathrm{d_H}\big(f_Q, \tilde{f}_P^L\big)\big\}^2 + \frac{1}{2}\big\{\mathrm{d_H}\big(f_P, \tilde{f}_Q^L\big) + \mathrm{d_H}\big(f_Q, \tilde{f}_Q^L\big)\big\}^2 \\
&\le \mathrm{d}_\mathrm{H}^2\big(f_P, \tilde{f}_P^L\big) + \mathrm{d}_\mathrm{H}^2\big(f_Q, \tilde{f}_P^L\big) + \mathrm{d}_\mathrm{H}^2\big(f_P, \tilde{f}_Q^L\big) + \mathrm{d}_\mathrm{H}^2\big(f_Q, \tilde{f}_Q^L\big) \\
&\le \mathrm{d_{KL}}\big(f_P \,\|\, \tilde{f}_P^L\big) + \mathrm{d_{KL}}\big(f_Q \,\|\, \tilde{f}_P^L\big) + \mathrm{d_{KL}}\big(f_P \,\|\, \tilde{f}_Q^L\big) + \mathrm{d_{KL}}\big(f_Q \,\|\, \tilde{f}_Q^L\big),
\end{aligned}$$

where the last step holds by the standard inequality relating KL divergence with Hellinger distance (i.e., $\mathrm{d}_\mathrm{H}^2 \le \mathrm{d_{KL}}$). Combining all these calculations, and then rearranging terms, we see that[3]

$$\begin{aligned}
\mathrm{d}_\mathrm{H}^2(f_P, f_Q) &\le \mathbb{E}_P\big[\phi_P(X) - \tilde{\phi}_P^L(X)\big] + \mathbb{E}_Q\big[\phi_Q(X) - \tilde{\phi}_P^L(X)\big] \\
&\quad + \mathbb{E}_P\big[\phi_P(X) - \tilde{\phi}_Q^L(X)\big] + \mathbb{E}_Q\big[\phi_Q(X) - \tilde{\phi}_Q^L(X)\big] \\
&= 2\big(\mathbb{E}_P\big[\phi_P(X) - \tilde{\phi}_P^L(X)\big] + \mathbb{E}_Q\big[\phi_Q(X) - \tilde{\phi}_Q^L(X)\big]\big) \\
&\quad + \big(\mathbb{E}_P\big[\tilde{\phi}_P^L(X)\big] - \mathbb{E}_Q\big[\tilde{\phi}_P^L(X)\big]\big) + \big(\mathbb{E}_Q\big[\tilde{\phi}_Q^L(X)\big] - \mathbb{E}_P\big[\tilde{\phi}_Q^L(X)\big]\big) \\
&= 2\big(\mathbb{E}_P\big[\phi_P(X) - \tilde{\phi}_P^L(X)\big] + \mathbb{E}_Q\big[\phi_Q(X) - \tilde{\phi}_Q^L(X)\big]\big) \\
&\quad + \big(\mathbb{E}_P\big[\phi_P^L(X)\big] - \mathbb{E}_Q\big[\phi_P^L(X)\big]\big) + \big(\mathbb{E}_Q\big[\phi_Q^L(X)\big] - \mathbb{E}_P\big[\phi_Q^L(X)\big]\big),
\end{aligned}$$

where the last step holds since $\tilde{\phi}_P^L, \tilde{\phi}_Q^L$ are simply shifts of the functions ϕ_P^L, ϕ_Q^L, respectively. Finally, applying Lemma 10 concludes the proof. □

4.3. Completing the proof of Theorem 2

We will now apply Corollary 11 to prove Theorem 2, bounding $\mathrm{d}_\mathrm{H}^2(f_P, f_Q)$ in terms of the Wasserstein distance. Define

$$L = \sqrt{\frac{8d}{r_d b_d \min\{\epsilon_P, \epsilon_Q\} \mathrm{d_W}(P, Q)}},$$

where $r_d, b_d \in (0, 1]$ are taken from Lemma 9. Take a coupling (X, Y) of d-dimensional random vectors with marginal distributions $X \sim P$ and $Y \sim Q$, such that $\mathbb{E}[\|X - Y\|] = \mathrm{d_W}(P, Q)$, which is guaranteed to exist by Villani [50], Theorem 4.1. Then, since ϕ_P^L is L-Lipschitz, we have

$$\mathbb{E}\big[\phi_P^L(X)\big] - \mathbb{E}\big[\phi_P^L(Y)\big] \le \mathbb{E}\big[L\|X - Y\|\big] = L\mathrm{d_W}(P, Q),$$

[3] All expectations in this display are finite, because, e.g., $\sup_{x\in\mathbb{R}^d} \phi_P(x) = \sup_{x\in\mathbb{R}^d} \phi_P^L(x) < \infty$; moreover, $\mathbb{E}_P[\phi_P^L(X)] \ge \mathbb{E}_P[\phi_P(X)] > -\infty$ because $P \in \mathcal{P}_d$, and $\mathbb{E}_P[\phi_Q^L(X)] > -\infty$ because ϕ_Q^L is Lipschitz and P has a finite first moment.

and similarly

$$\mathbb{E}\big[\phi_Q^L(Y)\big] - \mathbb{E}\big[\phi_Q^L(X)\big] \leq L\mathrm{d_W}(P, Q).$$

If $L \geq \frac{2d}{r_d \min\{\epsilon_P, \epsilon_Q\}}$, then applying Corollary 11, we have

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq \frac{16d}{Lb_d r_d \min\{\epsilon_P, \epsilon_Q\}} + 2L\mathrm{d_W}(P, Q) = \sqrt{\frac{128d\mathrm{d_W}(P, Q)}{r_d b_d \min\{\epsilon_P, \epsilon_Q\}}}.$$

If instead $L < \frac{2d}{r_d \min\{\epsilon_P, \epsilon_Q\}}$, then $\frac{db_d \mathrm{d_W}(P,Q)}{2r_d \min\{\epsilon_P, \epsilon_Q\}} > 1$. Since Hellinger distance is always bounded by $\sqrt{2}$, we then have

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq 2 \leq \sqrt{\frac{2db_d\mathrm{d_W}(P, Q)}{r_d \min\{\epsilon_P, \epsilon_Q\}}} \leq \sqrt{\frac{2d\mathrm{d_W}(P, Q)}{r_d b_d \min\{\epsilon_P, \epsilon_Q\}}},$$

where the last step holds trivially since $b_d \leq 1$. Thus, in either case, we have

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq \sqrt{\frac{128d}{r_d b_d}} \cdot \sqrt{\frac{\mathrm{d_W}(P, Q)}{\min\{\epsilon_P, \epsilon_Q\}}}.$$

We now split into cases. If $\mathrm{d_W}(P, Q) \leq \max\{\epsilon_P, \epsilon_Q\}/4$, then

$$\frac{\mathrm{d_W}(P, Q)}{\min\{\epsilon_P, \epsilon_Q\}} = \frac{\mathrm{d_W}(P, Q)}{\max\{\epsilon_P, \epsilon_Q\} - |\epsilon_P - \epsilon_Q|} \leq \frac{\mathrm{d_W}(P, Q)}{\max\{\epsilon_P, \epsilon_Q\} - 2\mathrm{d_W}(P, Q)} \leq \frac{2\mathrm{d_W}(P, Q)}{\max\{\epsilon_P, \epsilon_Q\}},$$

where the second step applies Proposition 1. If instead $\mathrm{d_W}(P, Q) > \max\{\epsilon_P, \epsilon_Q\}/4$, then we will instead use the trivial bound

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq 2 \leq 4\sqrt{\frac{\mathrm{d_W}(P, Q)}{\max\{\epsilon_P, \epsilon_Q\}}} \leq 4\sqrt{\frac{d}{r_d b_d}} \cdot \sqrt{\frac{\mathrm{d_W}(P, Q)}{\max\{\epsilon_P, \epsilon_Q\}}}$$

where the last step is trivial since $d \geq 1$ and $r_d, b_d \in (0, 1]$. Thus, in both cases, we have

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq 16\sqrt{\frac{d}{r_d b_d}} \cdot \sqrt{\frac{\mathrm{d_W}(P, Q)}{\max\{\epsilon_P, \epsilon_Q\}}}.$$

This proves the theorem, when we choose $C_d = 4(\frac{d}{r_d b_d})^{1/4}$.

4.4. Completing the proof of Theorem 5: The case $d = 1$

Before proving the theorem, we first state several supporting lemmas. First we state a deterministic result.

Lemma 12. *Let* $P, Q \in \mathcal{P}_1$ *satisfy* $\max\{\mathbb{E}_P[|X|^q]^{1/q}, \mathbb{E}_Q[|X|^q]^{1/q}\} \leq M_q$ *for some* $q > 1$. *Define*

$$\Delta_{\mathrm{CDF}}(P, Q) := \max\Big\{\sup_{t\in\mathbb{R}}\Big|\sqrt{\mathbb{P}_P\{X > t\}} - \sqrt{\mathbb{P}_Q\{X > t\}}\Big|, \sup_{t\in\mathbb{R}}\Big|\sqrt{\mathbb{P}_P\{X < t\}} - \sqrt{\mathbb{P}_Q\{X < t\}}\Big|\Big\}.$$

Then

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big) \le C_*\sqrt{\frac{M_q}{\max\{\epsilon_P,\epsilon_Q\}}}\cdot\big\{\Delta_{\mathrm{CDF}}(P,Q)\cdot\log\big(e/\Delta_{\mathrm{CDF}}(P,Q)\big)\big\}^{1-1/q},$$

for a universal constant $C_* > 0$.

Next, in order to prove Theorem 5, we will want to apply this result with $Q = \widehat{P}_n$, i.e., we want to bound $\Delta_{\mathrm{CDF}}(\widehat{P}_n, P)$. Let F denote the distribution function of P, and, for $t \in (0,1)$, let $F^{-1}(t) := \inf\{x : F(x) \ge t\}$. Then, with $U \sim \mathrm{Unif}[0,1]$, we know that $F^{-1}(U) \sim P$. We may therefore assume that $X_1,\dots,X_n$ are generated as $X_i = F^{-1}(U_i)$, where $U_1,\dots,U_n \stackrel{\mathrm{iid}}{\sim} \mathrm{Unif}[0,1]$. Since F^{-1} is monotonic, we have

$$\Delta_{\mathrm{CDF}}(\widehat{P}_n, P) \le \Delta_{\mathrm{CDF}}\big(\widehat{U}_n, \mathrm{Unif}[0,1]\big), \tag{8}$$

where $\widehat{U}_n$ is the empirical distribution of $U_1,\dots,U_n$. Therefore, it suffices to consider the case that P is the uniform distribution. We now apply results from Shorack and Wellner [47] to prove a tail bound on $\Delta_{\mathrm{CDF}}(\widehat{U}_n, \mathrm{Unif}[0,1])$.

Lemma 13. *Fix any* $n \ge 2$, *and let* $\widehat{U}_n$ *be the empirical distribution of* $U_1,\dots,U_n \stackrel{\mathrm{iid}}{\sim} \mathrm{Unif}[0,1]$. *Then, for any* $c > 0$,

$$\mathbb{P}\left\{\Delta_{\mathrm{CDF}}\big(\widehat{U}_n, \mathrm{Unif}[0,1]\big) \le c'\sqrt{\frac{\log n}{n}}\right\} \ge 1 - n^{-c},$$

where $c' > 0$ *depends only on* c.

With these lemmas in place, we are now in a position to prove Theorem 5. Let $M_{q,n} = (\frac{1}{n}\sum_{i=1}^n |X_i|^q)^{1/q}$ and $\Delta = \Delta_{\mathrm{CDF}}(\widehat{P}_n, P)$. If $\widehat{P}_n \in \mathcal{P}_1$ (that is, $\widehat{P}_n$ does not place all its mass on a single point), then we have

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(\widehat{P}_n),\psi^*(P)\big) \le \min\left\{2, C_*\sqrt{\frac{\max\{M_q, M_{q,n}\}}{\max\{\epsilon_P,\epsilon_{\widehat{P}_n}\}}}\cdot\big(\Delta\log(e/\Delta)\big)^{1-1/q}\right\} \tag{9}$$

by applying Lemma 12 with $Q = \widehat{P}_n$. On the other hand, if $\widehat{P}_n$ does place all its mass on one point, then recall that $\psi^*(\widehat{P}_n)$ is not defined but we take $\mathrm{d}_{\mathrm{H}}^2(\psi^*(\widehat{P}_n),\psi^*(P)) = 2$ by convention. For this case, we can trivially calculate

$$\Delta \ge \min\big\{\sqrt{\mathbb{P}_P\{X > \mu_P\}}, \sqrt{\mathbb{P}_P\{X < \mu_P\}}\big\}.$$

We will now need an additional lemma.

Lemma 14. *Fix any* $P \in \mathcal{P}_1$ *and any* $q > 1$. *Suppose* $M_q = \mathbb{E}_P[|X|^q]^{1/q} < \infty$. *Then*

$$\min\big\{\mathbb{P}_P\{X > \mu_P\}, \mathbb{P}_P\{X < \mu_P\}\big\} \ge \left(\frac{\epsilon_P}{4M_q}\right)^{\frac{q}{q-1}}.$$

This implies

$$\Delta \geq \left(\frac{\epsilon_P}{4M_q}\right)^{\frac{q}{2(q-1)}}$$

for the case where $\widehat{P}_n \notin \mathcal{P}_1$ (i.e., $\widehat{P}_n$ is supported on a single point). Since also $\Delta \leq 1$ by definition, this means that

$$\sqrt{\frac{\max\{M_q, M_{q,n}\}}{\epsilon_P}} \cdot \big(\Delta \log(e/\Delta)\big)^{1-1/q} \geq \frac{1}{2} = \frac{\mathrm{d}_{\mathrm{H}}^2(\psi^*(\widehat{P}_n), \psi^*(P))}{4}.$$

Combining this with (9) for the case $\widehat{P}_n \in \mathcal{P}_1$, we see that

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(\widehat{P}_n), \psi^*(P)\big) \leq \min\left\{2, \max\{C_*, 4\}\sqrt{\frac{\max\{M_q, M_{q,n}\}}{\epsilon_P}} \cdot \big(\Delta \log(e/\Delta)\big)^{1-1/q}\right\}$$

holds for both cases.

Next, we will combine this calculation with Lemma 13, applied with $c = 1/2$. Let c' be the constant from Lemma 13. First, if $c'\sqrt{\frac{\log n}{n}} > 1$, then

$$\begin{aligned}\mathbb{E}\big[\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(\widehat{P}_n), \psi^*(P)\big)\big] \leq 2 \leq 2\left(c'\sqrt{\frac{\log n}{n}}\right)^{1-1/q} &\leq \frac{2c'^{1-1/q}}{(\log 2)^{1-1/q}} \frac{\log^{\frac{3}{2}(1-1/q)} n}{n^{\frac{1}{2}-\frac{1}{2q}}} \\ &\leq \frac{2c'^{1-1/q}}{(\log 2)^{1-1/q}} \cdot \sqrt{\frac{2M_q}{\epsilon}} \cdot \frac{\log^{\frac{3}{2}(1-1/q)} n}{n^{\frac{1}{2}-\frac{1}{2q}}},\end{aligned}$$

where the last step holds since

$$\epsilon_P = \mathbb{E}_P\big[|X - \mu_P|\big] \leq \mathbb{E}_P\big[|X|\big] + |\mu_P| \leq 2\mathbb{E}_P\big[|X|\big] \leq 2\big\{\mathbb{E}_P\big[|X|^q\big]\big\}^{1/q} \leq 2M_q. \tag{10}$$

If instead $c'\sqrt{\frac{\log n}{n}} \leq 1$, then we have

$$\begin{aligned}&\mathbb{E}\big[\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(\widehat{P}_n), \psi^*(P)\big)\big] \\ &\leq \mathbb{E}\left[\min\left\{2, \max\{C_*, 4\}\sqrt{\frac{\max\{M_q, M_{q,n}\}}{\max\{\epsilon_P, \epsilon_{\widehat{P}_n}\}}} \cdot \big(\Delta \log(e/\Delta)\big)^{1-1/q}\right\}\right] \\ &\leq 2\mathbb{P}\left\{\Delta > c'\sqrt{\frac{\log n}{n}}\right\} + \mathbb{E}\left[\max\{C_*, 4\}\sqrt{\frac{M_q + M_{q,n}}{\epsilon_P}} \cdot \left\{c'\sqrt{\frac{\log n}{n}} \log\left(\frac{e}{c'\sqrt{\frac{\log n}{n}}}\right)\right\}^{1-1/q}\right] \\ &\leq 2n^{-1/2} + \max\{C_*, 4\}\sqrt{\frac{M_q + \mathbb{E}[M_{q,n}]}{\epsilon_P}} \cdot \left\{c'\sqrt{\frac{\log n}{n}} \log\left(\frac{e}{c'\sqrt{\frac{\log n}{n}}}\right)\right\}^{1-1/q} \\ &\leq 2n^{-1/2} + \max\{C_*, 4\}\sqrt{\frac{2M_q}{\epsilon_P}} \cdot \left\{c'\sqrt{\frac{\log n}{n}} \log\left(\frac{e}{c'\sqrt{\frac{\log n}{n}}}\right)\right\}^{1-1/q}\end{aligned}$$

$$\leq \sqrt{\frac{2M_q}{\epsilon_P}} \cdot \left[2n^{-1/2} + \max\{C_*, 4\}\left\{c'\sqrt{\frac{\log n}{n}} \log\left(\frac{e}{c'\sqrt{\frac{\log n}{n}}}\right)\right\}^{1-1/q}\right],$$

where the third-to-last step applies Jensen's inequality, the second-to-last step holds because $\mathbb{E}[M_{q,n}] \leq M_q$, and the last step holds by (10). After simplifying, we obtain

$$\mathbb{E}\big[\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(\widehat{P}_n), \psi^*(P)\big)\big] \leq C_{1,q}\sqrt{\frac{M_q}{\epsilon_P}} \cdot \frac{\log^{\frac{3}{2}(1-1/q)} n}{n^{\frac{1}{2}-\frac{1}{2q}}}$$

for all $n \geq 2$ when $C_{1,q}$ is chosen appropriately. This completes the proof of Theorem 5 for the case $d = 1$.

Appendix: Additional proofs

A.1. Proof of Proposition 1

First fix any distribution P on $\mathbb{R}^d$ with $\mathbb{E}_P[\|X\|] < \infty$. Observe that $u \mapsto \mathbb{E}_P[|u^\top(X - \mu_P)|]$ is a continuous function on $\mathbb{S}_{d-1}$, since for any $u, v \in \mathbb{S}_{d-1}$, we have

$$\begin{aligned}\big|\mathbb{E}_P\big[\big|u^\top(X-\mu_P)\big|\big] - \mathbb{E}_P\big[\big|v^\top(X-\mu_P)\big|\big]\big| &\leq \mathbb{E}_P\big[\big|(u-v)^\top(X-\mu_P)\big|\big] \\ &\leq \|u-v\| \cdot \mathbb{E}_P\big[\|X-\mu_P\|\big] \leq \|u-v\| \cdot 2\mathbb{E}_P\big[\|X\|\big],\end{aligned}$$

and $\mathbb{E}_P[\|X\|] < \infty$ by assumption. Therefore, $u \mapsto \mathbb{E}_P[|u^\top(X - \mu_P)|]$ must attain its infimum, that is,

$$\epsilon_P = \inf_{u \in \mathbb{S}_{d-1}} \mathbb{E}_P\big[\big|u^\top(X-\mu_P)\big|\big] = \mathbb{E}_P\big[\big|u_0^\top(X-\mu_P)\big|\big]$$

for some $u_0 \in \mathbb{S}_{d-1}$.

Next, suppose $P \in \mathcal{P}_d$. We will show that $\epsilon_P > 0$. As above, we have $\epsilon_P = \mathbb{E}_P[|u_0^\top(X - \mu_P)|]$ for some $u_0 \in \mathbb{S}_{d-1}$. If $\epsilon_P = 0$, then this implies that $u_0^\top(X - \mu_P) = 0$ with probability 1, meaning that P places all its mass on a single hyperplane $H = \{x \in \mathbb{R}^d : u_0^\top x = u_0^\top \mu_P\}$. This contradicts the assumption $P \in \mathcal{P}_d$, thus proving the first claim.

Finally, consider distributions P, Q on $\mathbb{R}^d$ with $\mathbb{E}_P[\|X\|], \mathbb{E}_Q[\|X\|] < \infty$. By Villani [50], Theorem 4.1, we can find a pair of d-dimensional random vectors X and Y such that marginally $X \sim P$, $Y \sim Q$ and $\mathbb{E}[\|X - Y\|] = \mathrm{d}_{\mathrm{W}}(P, Q)$. Let u_0 be defined as above, so that $\epsilon_P = \mathbb{E}[|u_0^\top(X - \mu_P)|]$. Then

$$\begin{aligned}\epsilon_Q - \epsilon_P &= \inf_{u \in \mathbb{S}_{d-1}} \mathbb{E}\big[\big|u^\top(Y-\mu_Q)\big|\big] - \mathbb{E}\big[\big|u_0^\top(X-\mu_P)\big|\big] \\ &\leq \mathbb{E}\big[\big|u_0^\top(Y-\mu_Q)\big|\big] - \mathbb{E}\big[\big|u_0^\top(X-\mu_P)\big|\big] \\ &\leq \mathbb{E}\big[\big|u_0^\top(X-Y)\big|\big] + \big|u_0^\top(\mu_P-\mu_Q)\big| \\ &\leq \mathbb{E}\big[\|X-Y\|\big] + \|\mu_P-\mu_Q\| \\ &\leq 2\mathbb{E}\big[\|X-Y\|\big] \\ &= 2\mathrm{d}_{\mathrm{W}}(P, Q).\end{aligned}$$

An identical argument proves the reverse bound, and we deduce that $|\epsilon_P - \epsilon_Q| \leq 2\mathrm{d_W}(P, Q)$, as desired.

A.2. Proof of Lemma 7

Let $\Sigma = \mathrm{Cov}_f(X)$ and define an isotropic log-concave density g on $\mathbb{R}^d$ by $g(x) = f(\Sigma^{1/2}x + \mu_P)\det^{1/2}(\Sigma)$. Note that, if $X \sim f$, then $\Sigma^{-1/2}(X - \mu_P) \sim g$. Hence,

$$\begin{aligned}\mathbb{E}_f\big[\big|v^\top(X-\mu_P)\big|\big] &= \mathbb{E}_f\big[\big|\big(\Sigma^{1/2}v\big)^\top\big(\Sigma^{-1/2}(X-\mu_P)\big)\big|\big] = \mathbb{E}_g\big[\big|\big(\Sigma^{1/2}v\big)^\top X\big|\big] \\ &\geq \frac{1}{4}\big(\mathbb{E}_g\big[\big(\big(\Sigma^{1/2}v\big)^\top X\big)^2\big]\big)^{1/2} = \frac{1}{4}\big\|\Sigma^{1/2}v\big\|,\end{aligned}$$

where the inequality applies Lovász and Vempala [35], Theorem 5.22, and the last step holds because g is isotropic.

Next, define a distribution Q obtained by drawing $X \sim P$ and then taking the affine transformation $\Sigma^{-1/2}(X - \mu_P)$. By definition of Q, we have

$$\begin{aligned}\mathbb{E}_P\big[\big|v^\top(X-\mu_P)\big|\big] &= \mathbb{E}_P\big[\big|\big(\Sigma^{1/2}v\big)^\top\big(\Sigma^{-1/2}(X-\mu_P)\big)\big|\big] \\ &= \mathbb{E}_Q\big[\big|\big(\Sigma^{1/2}v\big)^\top X\big|\big] \leq \big\|\Sigma^{1/2}v\big\| \cdot \mathbb{E}_Q\big[\|X\|\big].\end{aligned}$$

Since log-concave projection commutes with affine transformations, we have

$$\psi^*(Q) = g,$$

which is an isotropic log-concave density. Lemma 15 below establishes that $\mathbb{E}_Q[\|X\|] \leq a_d$, where $a_d > 0$ depends only on d. Therefore, we have proved that, for any $v \in \mathbb{R}^d$,

$$\mathbb{E}_P\big[\big|v^\top(X-\mu_P)\big|\big] \leq \big\|\Sigma^{1/2}v\big\| \cdot a_d$$

while

$$\mathbb{E}_f\big[\big|v^\top(X-\mu_P)\big|\big] \geq \frac{1}{4}\big\|\Sigma^{1/2}v\big\|.$$

Setting $c_d = \frac{1}{4a_d}$ establishes the desired result.

A.2.1. *Supporting lemma for Lemma 7*

Lemma 15. *There exists $a_d > 0$, depending only on d, such that, for any isotropic log-concave density f on $\mathbb{R}^d$ and any $P \in \mathcal{P}_d$ with $\psi^*(P) = f$,*

$$\mathbb{E}_P\big[\|X\|\big] \leq a_d.$$

Proof of Lemma 15. By Fresen [19], Lemma 13, since f is an isotropic log-concave density, it holds that

$$f(x) \leq e^{\beta_d - \alpha_d\|x\|} \quad \text{for all } x \in \mathbb{R}^d,$$

where $\alpha_d > 0$ and $\beta_d \in \mathbb{R}$ depend only on d. We can therefore calculate

$$\mathbb{E}_P\big[\log f(X)\big] \leq \mathbb{E}_P\big[\beta_d - \alpha_d\|X\|\big] = \beta_d - \alpha_d\mathbb{E}_P\big[\|X\|\big].$$

On the other hand, consider the log-concave density

$$g(x) = \left(\frac{d^d}{\mathbb{E}_P[\|X\|]^d (d-1)! S_{d-1}}\right) \cdot \exp\left\{-\frac{d\|x\|}{\mathbb{E}_P[\|X\|]}\right\},$$

where S_{d-1} denotes the surface area of the unit sphere $\mathbb{S}_{d-1}$ in $\mathbb{R}^d$ (with $S_0 = 2$). We have

$$\mathbb{E}_P\bigl[\log g(X)\bigr] = \log\left(\frac{d^d}{\mathbb{E}_P[\|X\|]^d (d-1)! S_{d-1}}\right) - d.$$

But, since $f = \psi^*(P)$, it must hold that

$$\mathbb{E}_P\bigl[\log f(X)\bigr] \geq \mathbb{E}_P\bigl[\log g(X)\bigr],$$

and so

$$\beta_d - \alpha_d \mathbb{E}_P\bigl[\|X\|\bigr] \geq \log\left(\frac{(d/e)^d}{(d-1)! S_{d-1}}\right) - d \log \mathbb{E}_P\bigl[\|X\|\bigr].$$

The result follows. □

A.3. Proof of Lemma 10

We will prove below that, when $L \geq \frac{2d}{r_d \epsilon_P}$, the function $\phi^L(x) = \sup_{y \in \mathbb{R}^d}\{\phi(x) - L\|x - y\|\}$ satisfies

$$\int_{\mathbb{R}^d} e^{\phi^L(x)}\, \mathsf{d}x \leq 1 + \frac{4d}{L b_d r_d \epsilon_P}. \tag{11}$$

Assuming this holds, we then have

$$\begin{aligned}\ell\bigl(\phi^L, P\bigr) &= \mathbb{E}_P\bigl[\phi^L(X)\bigr] - \int_{\mathbb{R}^d} e^{\phi^L(x)}\, \mathsf{d}x + 1 \geq \mathbb{E}_P\bigl[\phi^L(X)\bigr] - \frac{4d}{L b_d r_d \epsilon_P} \\ &\geq \mathbb{E}_P\bigl[\phi(X)\bigr] - \frac{4d}{L b_d r_d \epsilon_P} = \ell(\phi, P) - \frac{4d}{L b_d r_d \epsilon_P},\end{aligned}$$

where the last inequality holds since $\phi^L \geq \phi$ pointwise. Finally, normalizing to $\tilde{\phi}^L$ can only improve the objective function, since

$$\ell\bigl(\tilde{\phi}^L, P\bigr) = \mathbb{E}_P\bigl[\tilde{\phi}^L(X)\bigr] = \mathbb{E}_P\bigl[\phi^L(X)\bigr] - \log\left(\int_{\mathbb{R}^d} e^{\phi^L(x)}\, \mathsf{d}x\right) \geq \ell\bigl(\phi^L, P\bigr),$$

because $\log t \leq t - 1$ for all $t > 0$.

From this point on, we only need to prove (11) in order to complete the proof of the lemma. For any $x \in \mathbb{R}^d$, we will write y_x to denote a point attaining the supremum, that is, $\phi^L(x) = \phi(y_x) - L\|x - y_x\|$ (Lemma 16 below verifies the existence and measurability of such a map $x \mapsto y_x$).

We now derive the desired bound (11). We have

$$\int_{\mathbb{R}^d} e^{\phi^L(x)}\, \mathsf{d}x = \int_{\mathbb{R}^d} e^{\phi(y_x)} \cdot e^{-L\|x - y_x\|}\, \mathsf{d}x$$

$$= \int_{\mathbb{R}^d} \left(\int_{-\infty}^{\phi(y_x)} e^t \,\mathrm{d}t \right) \cdot \left(\int_{L\|x-y_x\|}^{\infty} e^{-s} \,\mathrm{d}s \right) \mathrm{d}x$$

$$= \int_{-\infty}^{M_\phi} \int_0^\infty e^{t-s} \left(\int_{\mathbb{R}^d} \mathbf{1}\big\{\phi(y_x) \geq t, \|x - y_x\| \leq s/L\big\} \,\mathrm{d}x \right) \mathrm{d}s \,\mathrm{d}t,$$

where the last step follows by Fubini's theorem, and where $M_\phi = \sup_{x \in \mathbb{R}^d} \phi(x)$ (note that we must have $M_\phi < \infty$ by definition of Φ_d). We now examine this indicator function. For $t \in \mathbb{R}$ define the super-level set $D_t = \{x : \phi(x) \geq t\}$. Note that D_t is convex for any t by concavity of ϕ, and furthermore is bounded since ϕ is a log-density. Moreover, we can observe that D_t has non-empty interior for any $t < M_\phi$, since ϕ is concave and is a log-density.

Now, for any compact, convex set $C \subseteq \mathbb{R}^d$ and any $\delta > 0$, define the δ-neighborhood of C by

$$\mathrm{Nbd}(C, \delta) := \big\{x \in \mathbb{R}^d : \mathrm{dist}(x, C) \leq \delta\big\},$$

where $\mathrm{dist}(x, C) := \min_{y \in C} \|x - y\|$. (If C is the empty set, then this neighborhood is also defined to be the empty set.) If $x \in \mathbb{R}^d$ is such that $\phi(y_x) \geq t$, then $y_x \in D_t$, and if, furthermore, $\|x - y_x\| \leq s/L$, then

$$x \in \mathrm{Nbd}(D_t, s/L).$$

Hence,

$$\int_{\mathbb{R}^d} e^{\phi^L(x)} \,\mathrm{d}x \leq \int_{-\infty}^{M_\phi} \int_0^\infty e^{t-s} \cdot \mathrm{Leb}_d\big(\mathrm{Nbd}(D_t, s/L)\big) \,\mathrm{d}s \,\mathrm{d}t.$$

On the other hand, we have

$$\int_{-\infty}^{M_\phi} \int_0^\infty e^{t-s} \cdot \mathrm{Leb}_d(D_t) \,\mathrm{d}s \,\mathrm{d}t = \int_{-\infty}^{M_\phi} e^t \cdot \mathrm{Leb}_d(D_t) \,\mathrm{d}t = \int_{-\infty}^{M_\phi} e^t \left(\int_{\mathbb{R}^d} \mathbf{1}\big\{\phi(x) \geq t\big\} \,\mathrm{d}x \right) \mathrm{d}t$$

$$= \int_{\mathbb{R}^d} \int_{-\infty}^{\phi(x)} e^t \,\mathrm{d}t \,\mathrm{d}x = \int_{\mathbb{R}^d} e^{\phi(x)} \,\mathrm{d}x = 1, \tag{12}$$

by again applying Fubini's theorem. Therefore, to prove (11), we only need to show that

$$\int_{-\infty}^{M_\phi} \int_0^\infty e^{t-s} \cdot \mathrm{Leb}_d\big(\mathrm{Nbd}(D_t, s/L) \backslash D_t\big) \,\mathrm{d}s \,\mathrm{d}t \leq \frac{4d}{L b_d r_d \epsilon_P}. \tag{13}$$

Next we will use a basic result about neighborhoods of convex sets – Lemma 17 verifies that

$$\delta \mapsto \frac{\mathrm{Leb}_d(\mathrm{Nbd}(C, \delta) \backslash C)}{\delta}$$

is a non-decreasing function for any compact, convex set $C \subseteq \mathbb{R}^d$ with non-empty interior. Therefore, for any $t < M_\phi$, it holds that

$$\mathrm{Leb}_d\big(\mathrm{Nbd}(D_t, s/L) \backslash D_t\big) \leq \frac{2d}{L r_d \epsilon_P} \cdot \mathrm{Leb}_d\left(\mathrm{Nbd}\left(D_t, \frac{s r_d \epsilon_P}{2d} \right) \backslash D_t \right)$$

since we have assumed $L \geq \frac{2d}{r_d \epsilon_P}$. We also have $D_t \subseteq D_{t + \log b_d}$, where $b_d \in (0, 1]$ is the constant appearing in Lemma 9, and so

$$\mathrm{Leb}_d\left(\mathrm{Nbd}\left(D_t, \frac{s r_d \epsilon_P}{2d} \right) \backslash D_t \right) \leq \mathrm{Leb}_d\left(\mathrm{Nbd}\left(D_t, \frac{s r_d \epsilon_P}{2d} \right) \right) \leq \mathrm{Leb}_d\left(\mathrm{Nbd}\left(D_{t + \log b_d}, \frac{s r_d \epsilon_P}{2d} \right) \right).$$

Recall from Lemma 9 that $D_{M_\phi+\log b_d}$ contains $\mathbb{B}_d(\mu_P, r_d\epsilon_P)$. Therefore, for any $t < M_\phi$, $D_{t+\log b_d} \supseteq D_{M_\phi+\log b_d}$ also contains this ball, and so

$$\begin{aligned}\mathrm{Nbd}\left(D_{t+\log b_d}, \frac{sr_d\epsilon_P}{2d}\right) &= D_{t+\log b_d} + \frac{s}{2d}\cdot\mathbb{B}_d(\mu_P, r_d\epsilon_P)\\ &\subseteq D_{t+\log b_d} + \frac{s}{2d}\cdot D_{t+\log b_d}\\ &= \left(1+\frac{s}{2d}\right)\cdot D_{t+\log b_d},\end{aligned}$$

where for two sets $A, B \subseteq \mathbb{R}^d$, we write $A + B := \{x + y : x \in A, y \in B\}$ to denote their Minkowski sum. Therefore,

$$\mathrm{Leb}_d\left(\mathrm{Nbd}\left(D_{t+\log b_d}, \frac{sr_d\epsilon_P}{2d}\right)\right) \le \mathrm{Leb}_d(D_{t+\log b_d})\cdot\left(1+\frac{s}{2d}\right)^d \le \mathrm{Leb}_d(D_{t+\log b_d})\cdot e^{s/2}$$

for any $t < M_\phi$. Combining this with our work above, we obtain

$$\mathrm{Leb}_d\big(\mathrm{Nbd}(D_t, s/L)\backslash D_t\big) \le \frac{2d}{Lr_d\epsilon_P}\cdot\mathrm{Leb}_d(D_{t+\log b_d})\cdot e^{s/2} \tag{14}$$

for any $t < M_\phi$. Therefore,

$$\begin{aligned}&\int_{-\infty}^{M_\phi}\int_0^\infty e^{t-s}\cdot\mathrm{Leb}_d\big(\mathrm{Nbd}(D_t, s/L)\backslash D_t\big)\,\mathsf{d}s\,\mathsf{d}t\\ &\quad\le \int_{-\infty}^{M_\phi}\int_0^\infty e^{t-s}\cdot\frac{2d}{Lr_d\epsilon_P}\cdot\mathrm{Leb}_d(D_{t+\log b_d})\cdot e^{s/2}\,\mathsf{d}s\,\mathsf{d}t\\ &\quad= \frac{2d}{Lr_d\epsilon_P}\cdot\left(\int_{-\infty}^{M_\phi} e^t\cdot\mathrm{Leb}_d(D_{t+\log b_d})\,\mathsf{d}t\right)\cdot\left(\int_0^\infty e^{-s}\cdot e^{s/2}\,\mathsf{d}s\right)\\ &\quad= \frac{4d}{Lr_d\epsilon_P}\cdot\int_{-\infty}^{M_\phi} e^t\cdot\mathrm{Leb}_d(D_{t+\log b_d})\,\mathsf{d}t\\ &\quad= \frac{4d}{Lb_dr_d\epsilon_P}\cdot\int_{-\infty}^{M_\phi} e^{t+\log b_d}\cdot\mathrm{Leb}_d(D_{t+\log b_d})\,\mathsf{d}t\\ &\quad= \frac{4d}{Lb_dr_d\epsilon_P}\cdot\int_{-\infty}^{M_\phi+\log b_d} e^t\cdot\mathrm{Leb}_d(D_t)\,\mathsf{d}t\\ &\quad\le \frac{4d}{Lb_dr_d\epsilon_P}\cdot\int_{-\infty}^{M_\phi} e^t\cdot\mathrm{Leb}_d(D_t)\,\mathsf{d}t\\ &\quad= \frac{4d}{Lb_dr_d\epsilon_P},\end{aligned}$$

where for the last step we again apply (12). This completes the proof of Lemma 10.

A.3.1. Supporting lemmas for Lemma 10

Lemma 16. *For any $x \in \mathbb{R}^d$ and any $\phi \in \Phi_d$, there exists a Borel measurable map $x \mapsto y_x$ such that y_x attains* $\sup_{y \in \mathbb{R}^d}\{\phi(y) - L\|x - y\|\}$.

Proof of Lemma 16. Let $M_\phi := \sup_{x \in \mathbb{R}^d} \phi(x)$, and let $x_\phi \in \operatorname{argmax}_{x \in \mathbb{R}^d} \phi(x)$ (note that, by definition of $\Phi_d \ni \phi$, M_ϕ must be finite, and x_ϕ must exist). Define

$$\mathcal{Y} = \{y \in \mathbb{R}^d : \phi(y) \geq \phi(y') - L\|y - y'\| \text{ for all } y' \in \mathbb{R}^d\}.$$

Note that $\mathcal{Y}$ is non-empty, since trivially $x_\phi \in \mathcal{Y}$.

Next define $h : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ as $h(x, y) = \phi(y) - L\|x - y\|$. For each $x \in \mathbb{R}^d$, define

$$S(x) = \mathcal{Y} \cap \mathbb{B}_d(x, \|x - x_\phi\|).$$

Note that, for any x, we have $x_\phi \in S(x)$ by definition.

Now we will apply Aliprantis and Border [1], Theorem 18.19, which guarantees the existence of a Borel measurable function $x \mapsto y_x \in S(x)$ such that, for each x,

$$y_x \in \operatorname*{argmax}_{y \in S(x)} h(x, y),$$

as long as we verify the following conditions:

- $\mathbb{R}^d$ is a measurable space, and $\mathcal{Y}$ is a separable metrizable space. This holds trivially.
- h is a Carathéodory function (i.e., $x \mapsto h(x, y)$ is measurable for any $y \in \mathcal{Y}$, and $y \mapsto h(x, y)$ is continuous for almost every $x \in \mathbb{R}^d$). It holds trivially that $x \mapsto h(x, y)$ is measurable. To check that $y \mapsto h(x, y)$ is continuous for any fixed x, it is sufficient to verify that ϕ is continuous on $\mathcal{Y}$. In fact, examining the definition of $\mathcal{Y}$, we can see that ϕ is L-Lipschitz on $\mathcal{Y}$ by definition, thus ensuring continuity.
- $S(x)$ is non-empty and compact for any $x \in \mathbb{R}^d$. We have already seen that $x_\phi \in S(x)$ for all x. To check compactness, it is sufficient to verify that $\mathcal{Y}$ is closed, which follows immediately from the definition of $\mathcal{Y}$ along with the fact that ϕ is upper semi-continuous (by definition of $\phi \in \Phi_d$).
- In the terminology of Aliprantis and Border [1], the correspondence $\mathcal{X} \twoheadrightarrow \mathcal{Y}$, mapping $x \mapsto S(x) \subseteq \mathcal{Y}$, is weakly measurable, meaning that the set $X_A := \{x \in \mathbb{R}^d : S(x) \cap A \neq \varnothing\}$ is measurable for any open subset $A \subseteq \mathcal{Y}$. Aliprantis and Border [1], Lemma 18.2, establishes that, since $\mathcal{Y}$ is metrizable, this is implied by the stronger condition that X_A is measurable for every *closed* subset $A \subseteq \mathcal{Y}$, so we will check this stronger condition.

  Let $A \subseteq \mathcal{Y}$ be a closed subset. Consider any $x, x_1, x_2, \ldots \in \mathbb{R}^d$ such that $x_i \in X_A$ for all $i \geq 1$ and such that $\lim_{i\to\infty} x_i = x$. Let $R = \sup_i \|x_i - x_\phi\|$, which is finite since the sequence converges. This means that $S(x_i) \subseteq \mathbb{B}_d(x_\phi, 2R)$ for all i. For each i, $x_i \in X_A$ implies that $S(x_i) \cap A \neq \varnothing$, and so we can find some $y_i \in S(x_i) \cap A \subseteq \mathbb{B}_d(x_\phi, 2R)$. Therefore, we can find some convergent subsequence, that is, $i_1, i_2, \ldots$ such that $\lim_{j\to\infty} y_{i_j} = y$ for some $y \in \mathbb{R}^d$. By assumption, A is a closed subset of $\mathcal{Y}$, and we have already shown that $\mathcal{Y}$ is a closed subset of $\mathbb{R}^d$. Therefore, $A \subseteq \mathbb{R}^d$ is closed, and so we must have $y \in A$. Now we check that $y \in S(x)$. We know that $y \in A \subseteq \mathcal{Y}$, and so we only need to check that $y \in \mathbb{B}_d(x, \|x - x_\phi\|)$. This holds because, for each $j \geq 1$, $y_{i_j} \in S(x_{i_j}) \subseteq \mathbb{B}_d(x_{i_j}, \|x_{i_j} - x_\phi\|)$, and so

  $$\|y - x\| = \lim_{j\to\infty} \|y_{i_j} - x_{i_j}\| \leq \lim_{j\to\infty} \|x_{i_j} - x_\phi\| = \|x - x_\phi\|.$$

  We have now seen that $y \in S(x) \cap A$, proving that $S(x) \cap A \neq \varnothing$ and so $x \in X_A$. Therefore, we have established that X_A is closed, and is therefore measurable.

Finally, we check that, for any x,

$$\sup_{y\in\mathbb{R}^d}\{\phi(y)-L\|x-y\|\}=\sup_{y\in S(x)}\{\phi(y)-L\|x-y\|\}.$$

First, for any $y\notin\mathbb{B}_d(x,\|x-x_\phi\|)$, we have $\|x-y\|>\|x-x_\phi\|$, and so since $\phi(y)\le\phi(x_\phi)$ by definition of x_ϕ, it holds that

$$\phi(y)-L\|x-y\|<\phi(x_\phi)-L\|x-x_\phi\|.$$

Therefore,

$$\sup_{y\in\mathbb{R}^d}\{\phi(y)-L\|x-y\|\}=\sup_{y\in\mathbb{B}_d(x,\|x-x_\phi\|)}\{\phi(y)-L\|x-y\|\}.$$

Next, since ϕ is upper semi-continuous, the supremum on the right-hand side is attained, that is, there exists some $y_1\in\mathbb{B}_d(x,\|x-x_\phi\|)$ such that

$$\phi(y_1)-L\|x-y_1\|=\sup_{y\in\mathbb{B}_d(x,\|x-x_\phi\|)}\{\phi(y)-L\|x-y\|\}=\sup_{y\in\mathbb{R}^d}\{\phi(y)-L\|x-y\|\}.$$

Now we verify that $y_1\in\mathcal{Y}$. To see this, fix any $y'\in\mathbb{R}^d$. Then

$$\phi(y')-L\|x-y'\|\le\sup_{y\in\mathbb{R}^d}\{\phi(y)-L\|x-y\|\}=\phi(y_1)-L\|x-y_1\|$$

and so

$$\phi(y_1)\ge\phi(y')-L\|x-y'\|+L\|x-y_1\|\ge\phi(y')-L\|y_1-y'\|.$$

Since this holds for all $y'\in\mathbb{R}^d$, we have established that $y_1\in\mathcal{Y}$. Therefore, $y_1\in S(x)$, which verifies $\sup_{y\in\mathbb{R}^d}\{\phi(y)-L\|x-y\|\}=\sup_{y\in S(x)}\{\phi(y)-L\|x-y\|\}$. □

Lemma 17. *Let $C\subseteq\mathbb{R}^d$ be any compact, convex set with non-empty interior. Then*

$$\delta\mapsto\frac{\mathrm{Leb}_d(\mathrm{Nbd}(C,\delta)\backslash C)}{\delta}$$

is a non-decreasing function of $\delta>0$.

Proof of Lemma 17. This result follows immediately from Steiner's formula (Schneider [42], Chapter 4), which states that for all $\epsilon\ge 0$,

$$\mathrm{Leb}_d(\mathrm{Nbd}(C,\epsilon))=\mathrm{Leb}_d(C)+\sum_{k=1}^d V_{d-k}(C)\cdot\mathrm{Leb}_k(\mathbb{B}_k)\cdot\epsilon^k,$$

where $V_{d-k}(C)\ge 0$ is the $(d-k)$-th intrinsic volume of C. Rearranging, we have

$$\frac{\mathrm{Leb}_d(\mathrm{Nbd}(C,\epsilon)\backslash C)}{\epsilon}=\sum_{k=1}^d V_{d-k}(C)\cdot\mathrm{Leb}_k(\mathbb{B}_k)\cdot\epsilon^{k-1},$$

which is a non-decreasing function of ϵ. □

A.4. Proof of Lemma 12

First, we consider the bounded case. Suppose that P and Q are both supported on $[-R, R]$ for some $R > 0$. Write $\Delta = \Delta_{\mathrm{CDF}}(P, Q)$ and $\epsilon = \min\{\epsilon_P, \epsilon_Q\}$. Let $r_1, b_1 \in (0, 1]$ be the universal constants defined in Lemma 9 (for dimension $d = 1$), and fix any $L \geq \frac{4}{r_1\epsilon}$. By Corollary 11, we have

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P), \psi^*(Q)\big) \leq \frac{16}{Lb_1r_1\epsilon} + \big(\mathbb{E}_P\big[\phi_P^L(X)\big] - \mathbb{E}_Q\big[\phi_P^L(X)\big]\big) + \big(\mathbb{E}_Q\big[\phi_Q^L(X)\big] - \mathbb{E}_P\big[\phi_Q^L(X)\big]\big).$$

Now we bound the two differences. For any $\phi \in \Phi_d$ define $M_\phi = \sup_{x\in\mathbb{R}^d} \phi(x)$ (note that M_ϕ is finite by definition of Φ_d). We note that $M_{\phi_P} = M_{\phi_P^L}$ by definition of ϕ_P^L, and that $\phi_P^L(X) \geq M_{\phi_P} - 2LR$ with probability 1 under either P or Q, since the distributions are supported on $[-R, R]$ and so ϕ_P must attain its maximum somewhere in this range. We then have

$$\begin{aligned}\mathbb{E}_P\big[\phi_P^L(X)\big] - \mathbb{E}_Q\big[\phi_P^L(X)\big] &= \mathbb{E}_Q\big[M_{\phi_P} - \phi_P^L(X)\big] - \mathbb{E}_P\big[M_{\phi_P} - \phi_P^L(X)\big] \\ &= \int_0^{2LR} \big(\mathbb{P}_Q\big\{M_{\phi_P} - \phi_P^L(X) \geq t\big\} - \mathbb{P}_P\big\{M_{\phi_P} - \phi_P^L(X) \geq t\big\}\big)\,\mathrm{d}t.\end{aligned}$$

It is trivial to verify that

$$\big|\sqrt{\mathbb{P}_P\{X \notin C\}} - \sqrt{\mathbb{P}_Q\{X \notin C\}}\big| \leq \Delta\sqrt{2}$$

for any convex set (i.e., an interval) $C \subseteq \mathbb{R}$, by definition of Δ (this follows from the fact that $|\sqrt{a+c} - \sqrt{b+d}|^2 \leq |\sqrt{a} - \sqrt{b}|^2 + |\sqrt{c} - \sqrt{d}|^2$ for any $a, b, c, d \geq 0$). Since ϕ_P^L is concave, the set $\{x : M_{\phi_P} - \phi_P^L(x) < t\}$ is convex, and so

$$\mathbb{P}_Q\big\{M_{\phi_P} - \phi_P^L(X) \geq t\big\} \leq \Big(\sqrt{\mathbb{P}_P\big\{M_{\phi_P} - \phi_P^L(X) \geq t\big\}} + \Delta\sqrt{2}\Big)^2$$

and so, since it also holds that $\phi_P^L \geq \phi_P$ pointwise, we have

$$\mathbb{P}_Q\big\{M_{\phi_P} - \phi_P^L(X) \geq t\big\} - \mathbb{P}_P\big\{M_{\phi_P} - \phi_P^L(X) \geq t\big\} \leq \Delta\sqrt{8} \cdot \sqrt{\mathbb{P}_P\big\{M_{\phi_P} - \phi_P(X) \geq t\big\}} + 2\Delta^2.$$

Lemma 18 below will establish that, for $t \geq \frac{8R}{r_1\epsilon}$, we have $\mathbb{P}_P\{M_{\phi_P} - \phi_P(X) \geq t\} \leq \frac{32}{b_1r_1\epsilon} \cdot \frac{R}{t^2}$. Applying this bound, we have

$$\begin{aligned}&\mathbb{E}_P\big[\phi_P^L(X)\big] - \mathbb{E}_Q\big[\phi_P^L(X)\big] \\ &\quad\leq \int_0^{2LR} \big(\Delta\sqrt{8} \cdot \sqrt{\mathbb{P}_P\big\{M_{\phi_P} - \phi_P(X) \geq t\big\}} + 2\Delta^2\big)\,\mathrm{d}t \\ &\quad= \Delta\sqrt{8} \int_0^{2LR} \sqrt{\mathbb{P}_P\big\{M_{\phi_P} - \phi_P(X) \geq t\big\}}\,\mathrm{d}t + 4LR\Delta^2 \\ &\quad= \Delta\sqrt{8}\bigg(\int_0^{\frac{8R}{r_1\epsilon}} \sqrt{\mathbb{P}_P\big\{M_{\phi_P} - \phi_P(X) \geq t\big\}}\,\mathrm{d}t + \int_{\frac{8R}{r_1\epsilon}}^{2LR} \sqrt{\mathbb{P}_P\big\{M_{\phi_P} - \phi_P(X) \geq t\big\}}\,\mathrm{d}t\bigg) + 4LR\Delta^2 \\ &\quad\leq \Delta\sqrt{8}\sqrt{\frac{8R}{r_1\epsilon}} \cdot \bigg(\int_0^{\frac{8R}{r_1\epsilon}} \mathbb{P}_P\big\{M_{\phi_P} - \phi_P(X) \geq t\big\}\,\mathrm{d}t\bigg)^{1/2} + \Delta\sqrt{8}\int_{\frac{8R}{r_1\epsilon}}^{2LR} \sqrt{\frac{32}{b_1r_1\epsilon} \cdot \frac{R}{t^2}}\,\mathrm{d}t + 4LR\Delta^2\end{aligned}$$

$$\leq \Delta\sqrt{8}\sqrt{\frac{8R}{r_1\epsilon}} \cdot \sqrt{\mathbb{E}_P\big[M_{\phi_P} - \phi_P(X)\big]} + \Delta\sqrt{8}\sqrt{\frac{32R}{b_1 r_1 \epsilon}}\log(Lr_1\epsilon/4) + 4LR\Delta^2$$

$$\leq \Delta\sqrt{8}\sqrt{\frac{8Rh_1}{r_1\epsilon}} + \Delta\sqrt{8}\sqrt{\frac{32R}{b_1 r_1 \epsilon}}\log(Lr_1\epsilon/4) + 4LR\Delta^2,$$

where the last step applies Lemma 19 below, which will establish that $\mathbb{E}_P[\phi(X)] \geq M_\phi - h_1$ for a universal constant h_1. By symmetry the same bound holds for $\mathbb{E}_Q[\phi^L_Q(X)] - \mathbb{E}_P[\phi^L_Q(X)]$. Combining all our work so far, then,

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq \frac{16}{Lb_1 r_1 \epsilon} + 2\left\{\Delta\sqrt{8}\left(\sqrt{\frac{8Rh_1}{r_1\epsilon}} + \sqrt{\frac{32R}{b_1 r_1\epsilon}}\log(Lr_1\epsilon/4)\right) + 4LR\Delta^2\right\}.$$

Next, we split into cases. If $\frac{1}{\Delta\sqrt{R\epsilon}} \geq \frac{4}{r_1\epsilon}$, then setting $L = \frac{1}{\Delta\sqrt{R\epsilon}}$ we apply this bound to obtain

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq C'\Delta\sqrt{R/\epsilon}\max\left\{1, \log\left(\frac{1}{\Delta\sqrt{R/\epsilon}}\right)\right\},$$

for a universal constant C'. Since $\epsilon \leq 2R$ by definition, and $\Delta \leq 1$, we can relax this to

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq C'\Delta\sqrt{R/\epsilon}\log(e/\Delta).$$

If instead $\frac{1}{\Delta\sqrt{R\epsilon}} < \frac{4}{r_1\epsilon}$, then

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq 2 \leq \frac{8}{r_1}\Delta\sqrt{R/\epsilon}.$$

Therefore, combining both cases, we have

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \leq C''\Delta\sqrt{\frac{R}{\min\{\epsilon_P, \epsilon_Q\}}}\log(e/\Delta) \tag{15}$$

for a universal constant $C'' = \max\{C', 8/r_1\}$. Next, we will need to relate $\min\{\epsilon_P, \epsilon_Q\}$ with $\max\{\epsilon_P, \epsilon_Q\}$. Without loss of generality, suppose that $\mu_P \geq \mu_Q$. We then have

$$\begin{aligned}
\frac{\epsilon_Q}{2} &= \frac{1}{2}\mathbb{E}_Q\big[|X - \mu_Q|\big] = \mathbb{E}_Q\big[(X - \mu_Q)_+\big] \\
&\geq \mathbb{E}_Q\big[(X - \mu_P)_+\big] = \int_{\mu_P}^R \mathbb{P}_Q\{X > t\}\,\mathrm{d}t \\
&\geq \int_{\mu_P}^R \mathbb{P}_P\{X > t\} - 2\Delta\sqrt{\mathbb{P}_P\{X > t\}}\,\mathrm{d}t \\
&\geq \int_{\mu_P}^R \mathbb{P}_P\{X > t\}\,\mathrm{d}t - 2\Delta\sqrt{R - \mu_P}\sqrt{\int_{\mu_P}^R \mathbb{P}_P\{X > t\}\,\mathrm{d}t} \\
&\geq \mathbb{E}_P\big[(X - \mu_P)_+\big] - 2\Delta\sqrt{2R}\sqrt{\mathbb{E}_P\big[(X - \mu_P)_+\big]}
\end{aligned}$$

$$= \frac{\epsilon_P}{2} - 2\Delta\sqrt{R \cdot \epsilon_P},$$

where the final inequality follows because $|\mu_P| \le R$. We can similarly calculate

$$\frac{\epsilon_P}{2} = \frac{1}{2}\mathbb{E}_P\big[|X - \mu_P|\big] = \mathbb{E}_P\big[(X - \mu_P)_-\big] \ge \frac{\epsilon_Q}{2} - 2\Delta\sqrt{R \cdot \epsilon_Q}.$$

Combining these two bounds, then,

$$\max\{\epsilon_P, \epsilon_Q\} = \min\{\epsilon_P, \epsilon_Q\} + |\epsilon_P - \epsilon_Q| \le \min\{\epsilon_P, \epsilon_Q\} + 4\Delta_{\mathrm{CDF}}(P, Q) \cdot \sqrt{R \cdot \max\{\epsilon_P, \epsilon_Q\}}. \quad (16)$$

Now we work with the general case, where P, Q may not have bounded support. Fix any $R > 0$. For any $x \in \mathbb{R}$ define

$$[x]_R := \begin{cases} -R, & x < -R, \\ x, & |x| \le R, \\ R, & x > R, \end{cases} \quad (17)$$

the truncation of x to the range $[-R, R]$. Let $[P]_R$ denote the distribution of $[X]_R$ when $X \sim P$, and same for $[Q]_R$. Lemma 20 below calculates that $\mathrm{d_W}(P, [P]_R) \le \frac{M_q^q}{R^{q-1}}$. Applying Theorem 2 to compare the distributions P and $[P]_R$, then, we have

$$\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*\big([P]_R\big)\big) \le C_1^2\sqrt{\frac{\mathrm{d_W}(P, [P]_R)}{\max\{\epsilon_P, \epsilon_{[P]_R}\}}} \le C_1^2\sqrt{\frac{M_q^q}{\epsilon_{[P]_R}R^{q-1}}},$$

and the same bound holds with Q in place of P. Therefore, by the triangle inequality,

$$\begin{aligned}
&\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) \\
&\quad \le \big\{\mathrm{d_H}\big(\psi^*(P), \psi^*\big([P]_R\big)\big) + \mathrm{d_H}\big(\psi^*(Q), \psi^*\big([Q]_R\big)\big) + \mathrm{d_H}\big(\psi^*\big([P]_R\big), \psi^*\big([Q]_R\big)\big)\big\}^2 \\
&\quad \le 3\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*\big([P]_R\big)\big) + 3\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(Q), \psi^*\big([Q]_R\big)\big) + 3\mathrm{d}^2_{\mathrm{H}}\big(\psi^*\big([P]_R\big), \psi^*\big([Q]_R\big)\big) \\
&\quad \le 6C_1^2\sqrt{\frac{M_q^q}{\min\{\epsilon_{[P]_R}, \epsilon_{[Q]_R}\}R^{q-1}}} + 3\mathrm{d}^2_{\mathrm{H}}\big(\psi^*\big([P]_R\big), \psi^*\big([Q]_R\big)\big). \quad (18)
\end{aligned}$$

We now need to apply the bound (15) to the bounded distributions $[P]_R$ and $[Q]_R$, in order to bound this last term. Combining (15) with (18), we obtain

$$\begin{aligned}
\mathrm{d}^2_{\mathrm{H}}\big(\psi^*(P), \psi^*(Q)\big) &\le 6C_1^2\sqrt{\frac{M_q^q}{\min\{\epsilon_{[P]_R}, \epsilon_{[Q]_R}\}R^{q-1}}} \\
&\quad + 3C''\Delta_{\mathrm{CDF}}\big([P]_R, [Q]_R\big)\sqrt{\frac{R}{\min\{\epsilon_{[P]_R}, \epsilon_{[Q]_R}\}}}\log\big(e/\Delta_{\mathrm{CDF}}\big([P]_R, [Q]_R\big)\big).
\end{aligned}$$

Now fix

$$R = M_q\big\{\Delta_{\mathrm{CDF}}\big([P]_R, [Q]_R\big)\log\big(e/\Delta_{\mathrm{CDF}}\big([P]_R, [Q]_R\big)\big)\big\}^{-2/q}.$$

This yields

$$\begin{aligned}&\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big)\\&\quad\le C'_*\sqrt{\frac{M_q}{\min\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}}}\cdot\left\{\Delta_{\mathrm{CDF}}\big([P]_R,[Q]_R\big)\log\left(\frac{e}{\Delta_{\mathrm{CDF}}([P]_R,[Q]_R)}\right)\right\}^{1-1/q},\end{aligned}$$

when the universal constant $C'_* > 0$ is chosen appropriately. Next, it holds trivially that $\Delta_{\mathrm{CDF}}([P]_R, [Q]_R) \le \Delta_{\mathrm{CDF}}(P,Q)$, and since $t \mapsto t\log(e/t)$ is increasing on $t \in (0,1]$, we therefore have

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big)\le C'_*\sqrt{\frac{M_q}{\min\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}}}\cdot\big\{\Delta_{\mathrm{CDF}}(P,Q)\log\big(e/\Delta_{\mathrm{CDF}}(P,Q)\big)\big\}^{1-1/q}.$$

Finally, we need to lower bound $\epsilon_{[P]_R}$ and $\epsilon_{[Q]_R}$. First, we relate $\min\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}$ to $\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}$. Applying (16) from above, along with the fact that $\Delta_{\mathrm{CDF}}([P]_R,[Q]_R)\le\Delta_{\mathrm{CDF}}(P,Q)$, we have

$$\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}\le\min\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}+4\Delta_{\mathrm{CDF}}(P,Q)\sqrt{R\cdot\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}}.$$

If $8\Delta_{\mathrm{CDF}}(P,Q)\sqrt{R}\le\sqrt{\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}}$, then this proves that

$$\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}\le 2\min\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}$$

and so

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big)\le C'_*\sqrt{\frac{2M_q}{\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}}}\cdot\big\{\Delta_{\mathrm{CDF}}(P,Q)\log\big(e/\Delta_{\mathrm{CDF}}(P,Q)\big)\big\}^{1-1/q}.$$

If instead $8\Delta_{\mathrm{CDF}}(P,Q)\sqrt{R}>\sqrt{\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}}$, then we have

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big)\le 2\le\frac{16\Delta_{\mathrm{CDF}}(P,Q)\sqrt{R}}{\sqrt{\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}}}.$$

Plugging in the definition of R and combining both cases, we obtain

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big)\le C''_*\sqrt{\frac{M_q}{\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}}}\cdot\big(\Delta_{\mathrm{CDF}}(P,Q)\log\big(e/\Delta_{\mathrm{CDF}}(P,Q)\big)\big)^{1-1/q}$$

for an appropriately chosen universal constant C''_*. The last step is to relate $\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}$ to $\max\{\epsilon_P,\epsilon_Q\}$. Applying Proposition 1 together with the bound on $\mathrm{d}_{\mathrm{W}}(P,[P]_R)$ from Lemma 20, we have

$$\epsilon_{[P]_R}\ge\epsilon_P-2\mathrm{d}_{\mathrm{W}}\big(P,[P]_R\big)\ge\epsilon_P-2\cdot\frac{M_q^q}{R^{q-1}},$$

and the same bound holds for Q in place of P. If $\frac{2M_q^q}{R^{q-1}}\le\frac{\max\{\epsilon_P,\epsilon_Q\}}{2}$, then

$$\max\{\epsilon_{[P]_R},\epsilon_{[Q]_R}\}\ge\frac{\max\{\epsilon_P,\epsilon_Q\}}{2},$$

and so we obtain

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big) \le C_*''\sqrt{\frac{2M_q}{\max\{\epsilon_P,\epsilon_Q\}}}\cdot\big\{\Delta_{\mathrm{CDF}}(P,Q)\log\big(e/\Delta_{\mathrm{CDF}}(P,Q)\big)\big\}^{1-1/q}.$$

If instead $\frac{2M_q^q}{R^{q-1}} > \frac{\max\{\epsilon_P,\epsilon_Q\}}{2}$, then it trivially holds that

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big) \le 2 \le 2\sqrt{\frac{4M_q^q}{\max\{\epsilon_P,\epsilon_Q\}R^{q-1}}}.$$

Plugging in the definition of R, and combining the two cases, we obtain

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big) \le C_*\sqrt{\frac{M_q}{\max\{\epsilon_P,\epsilon_Q\}}}\cdot\big\{\Delta_{\mathrm{CDF}}(P,Q)\log\big(e/\Delta_{\mathrm{CDF}}(P,Q)\big)\big\}^{1-1/q}$$

for appropriately chosen universal constant C_*, which completes the proof of Lemma 12.

A.4.1. Supporting lemmas for Lemma 12

Lemma 18. *Let $P\in\mathcal{P}_d$ and let $\phi=\phi^*(P)$. Let $M_\phi := \sup_{x\in\mathbb{R}^d}\phi(x)$ and let $x_\phi\in\operatorname{argmax}_{x\in\mathbb{R}^d}\phi(x)$ (which is guaranteed to exist by definition of $\Phi_d\ni\phi$). Fix any $R>0$ and $t\ge\frac{8dR}{r_d\epsilon_P}$, where $r_d\in(0,1]$ is taken from Lemma 9. Then*

$$\mathbb{P}_P\big\{\phi(X)\le M_\phi - t \text{ and } \|X-x_\phi\|\le 2R\big\} \le \frac{32d}{b_d r_d\epsilon_P}\cdot\frac{R}{t^2},$$

where $b_d\in(0,1]$ is taken from Lemma 9.

Proof of Lemma 18. First, for any x with $\|x-x_\phi\|\le 2R$,

$$\phi^{t/4R}(x) = \sup_{y\in\mathbb{R}^d}\left\{\phi(y)-\frac{t}{4R}\|y-x\|\right\} \ge \phi(x_\phi)-\frac{t}{4R}\|x-x_\phi\| \ge M_\phi - \frac{t}{2}.$$

Hence, if $\phi(x)\le M_\phi - t$ and $\|x-x_\phi\|\le 2R$, then

$$\phi^{t/4R}(x)-\phi(x)\ge\frac{t}{2}.$$

Moreover, by definition of $\phi=\phi^*(P)$, since $\phi^{t/4R}\in\Phi_d$, it holds that

$$\begin{aligned}\mathbb{E}_P\big[\phi(X)\big] = \ell(\phi,P) \ge \ell\big(\phi^{t/4R},P\big) &= \mathbb{E}_P\big[\phi^{t/4R}(X)\big] - \int_{\mathbb{R}^d} e^{\phi^{t/4R}(x)}\,\mathrm{d}x + 1\\ &\ge \mathbb{E}_P\big[\phi^{t/4R}(X)\big] - \frac{4d}{\frac{t}{4R}b_d r_d\epsilon_P},\end{aligned}$$

where the last step holds by (11) as calculated in the proof of Lemma 10, noting that $\frac{t}{4R}\ge\frac{2d}{r_d\epsilon_P}$. We deduce that

$$\mathbb{P}_P\big\{\phi(X)\le M_\phi - t \text{ and } \|X-x_\phi\|\le 2R\big\} \le \mathbb{P}_P\left\{\phi^{t/4R}(X)-\phi(X)\ge\frac{t}{2}\right\}$$

$$\leq \frac{\mathbb{E}_P[\phi^{t/4R}(X) - \phi(X)]}{t/2} \leq \frac{\frac{4d}{\frac{t}{4R} b_d r_d \epsilon_P}}{t/2} = \frac{32d}{b_d r_d \epsilon_P} \cdot \frac{R}{t^2},$$

as required. □

Lemma 19. *Fix any $P \in \mathcal{P}_d$ and let $\phi = \phi^*(P)$. Then*

$$\mathbb{E}_P\big[\phi(X)\big] \geq M_\phi - h_d,$$

where $M_\phi = \sup_{x \in \mathbb{R}^d} \phi(x)$ and where $h_d \geq 0$ depends only on d.

Proof of Lemma 19. Write $\mathbb{E}_\phi[\cdot]$ to denote the expectation with respect to the distribution with log-density ϕ. Let $\mu_\phi := \mathbb{E}_\phi[X]$ be the mean and $\Sigma := \mathbb{E}_\phi[(X - \mu_\phi)(X - \mu_\phi)^\top]$ the covariance of this distribution. Let $\bar{\phi}$ denote the log-density of the isotropic, log-concave random vector $\Sigma^{-1/2}(X - \mu_\phi)$, where X has log-density ϕ. Let $M_{\bar{\phi}} := \sup_{x \in \mathbb{R}^d} \bar{\phi}(x)$.

Since $x \mapsto \phi(x) + \frac{1}{2}\{M_\phi - \phi(x)\}$ is concave and coercive, it holds by Dümbgen, Samworth and Schuhmacher [15], Remark 2.3, that

$$\mathbb{E}_P\big[M_\phi - \phi(X)\big] \leq \mathbb{E}_\phi\big[M_\phi - \phi(X)\big].$$

Next, we can trivially verify that

$$\mathbb{E}_\phi\big[M_\phi - \phi(X)\big] = \mathbb{E}_{\bar{\phi}}\big[M_{\bar{\phi}} - \bar{\phi}(X)\big]$$

since the log-densities ϕ and $\bar{\phi}$ are related via the linear transformation on random variables above. Furthermore,

$$\mathbb{E}_{\bar{\phi}}\big[M_{\bar{\phi}} - \bar{\phi}(X)\big] = M_{\bar{\phi}} - \int_{\mathbb{R}^d} e^{\bar{\phi}(y)} \cdot \bar{\phi}(y)\,\mathrm{d}y \leq M_{\bar{\phi}} + \frac{d}{2}\log(2\pi e),$$

where the last step holds since $\bar{\phi}$ is the log-density of an isotropic distribution on $\mathbb{R}^d$, and so its entropy is bounded by that of the standard d-dimensional Gaussian (e.g., Cover and Thomas [9], Theorem 9.6.5). Finally, by Lovász and Vempala [35], Theorem 5.14(e), $M_{\bar{\phi}} \leq m_d$ where $m_d \in \mathbb{R}$ depends only on the dimension d. Therefore, combining everything,

$$\mathbb{E}_P\big[M_\phi - \phi(X)\big] \leq m_d + \frac{d}{2}\log(2\pi e),$$

which proves the desired bound. □

Lemma 20. *Let $P \in \mathcal{P}_1$ satisfy $\mathbb{E}_P[|X|^q]^{1/q} \leq M_q$, for some $q > 1$. Let $[P]_R$ be the distribution of $[X]_R$ when $X \sim P$ (where the truncation $[X]_R$ is defined as in (17)). Then*

$$\mathrm{d_W}\big(P, [P]_R\big) \leq \frac{M_q^q}{R^{q-1}}.$$

Proof of Lemma 20. Drawing $X \sim P$, note that $(X, [X]_R)$ is a coupling of the distributions P and $[P]_R$. Hence,

$$\mathrm{d_W}\big(P, [P]_R\big) \leq \mathbb{E}_P\big[\big|X - [X]_R\big|\big] = \mathbb{E}_P\big[\big(|X| - R\big)_+\big] \leq \mathbb{E}_P\left[\frac{|X|^q}{R^{q-1}}\right] \leq \frac{M_q^q}{R^{q-1}},$$

as required. □

A.5. Proof of Lemma 13

Write $\widehat{U}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{U_i \le t\}$. First, we calculate

$$\Delta_{\mathrm{CDF}}\big(\widehat{U}_n, \mathrm{Unif}[0,1]\big) = \max\Big\{\underbrace{\sup_{t\in[0,1]}\Big|\sqrt{1-\widehat{U}_n(t)} - \sqrt{1-t}\Big|}_{=\Delta_0}, \underbrace{\sup_{t\in[0,1]}\Big|\sqrt{\widehat{U}_n(t)} - \sqrt{t}\Big|}_{=:\Delta_1}\Big\},$$

by observing that

$$\sup_{t\in[0,1]}\left|\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{U_i < t\}} - \sqrt{t}\right| = \sup_{t\in[0,1]}\left|\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{U_i \le t\}} - \sqrt{t}\right|$$

(i.e., the supremum is unchanged by replacing $<$ with $\le$). We can further write

$$\Delta_1 = \max\Big\{\underbrace{\sup_{t\in[0,\frac{\log n}{n}]}\Big|\sqrt{\widehat{U}_n(t)} - \sqrt{t}\Big|}_{=:\Delta_{1,0}}, \underbrace{\sup_{t\in[\frac{\log n}{n},1-\frac{\log n}{n}]}\Big|\sqrt{\widehat{U}_n(t)} - \sqrt{t}\Big|}_{=:\Delta_{1,1}}, \underbrace{\sup_{t\in[1-\frac{\log n}{n},1]}\Big|\sqrt{\widehat{U}_n(t)} - \sqrt{t}\Big|}_{=:\Delta_{1,2}}\Big\}.$$

We have

$$\Delta_{1,0} = \sup_{t\in[0,\frac{\log n}{n}]}\Big|\sqrt{\widehat{U}_n(t)} - \sqrt{t}\Big| \le \sqrt{\frac{\log n}{n}} + \sqrt{\widehat{U}_n\left(\frac{\log n}{n}\right)} \le 2\sqrt{\frac{\log n}{n}} + \Delta_{1,1},$$

and

$$\Delta_{1,2} = \sup_{t\in[1-\frac{\log n}{n},1]}\Big|\sqrt{\widehat{U}_n(t)} - \sqrt{t}\Big| \le \sqrt{\frac{\log n}{n}} + \left(1 - \sqrt{\widehat{U}_n\left(1 - \frac{\log n}{n}\right)}\right) \le 2\sqrt{\frac{\log n}{n}} + \Delta_{1,1}.$$

Furthermore,

$$\Delta_{1,1} = \sup_{t\in[\frac{\log n}{n},1-\frac{\log n}{n}]}\Big|\sqrt{\widehat{U}_n(t)} - \sqrt{t}\Big| = \sup_{t\in[\frac{\log n}{n},1-\frac{\log n}{n}]}\frac{|\widehat{U}_n(t) - t|}{\sqrt{\widehat{U}_n(t)} + \sqrt{t}} \le \sup_{t\in[\frac{\log n}{n},1-\frac{\log n}{n}]}\frac{|\widehat{U}_n(t) - t|}{\sqrt{t}}.$$

Combining these calculations, we have

$$\Delta_1 \le 2\sqrt{\frac{\log n}{n}} + \sup_{t\in[\frac{\log n}{n},1-\frac{\log n}{n}]}\frac{|\widehat{U}_n(t) - t|}{\sqrt{t}}.$$

Similarly we can calculate

$$\Delta_0 \le 2\sqrt{\frac{\log n}{n}} + \sup_{t\in[\frac{\log n}{n},1-\frac{\log n}{n}]}\frac{|\widehat{U}_n(t) - t|}{\sqrt{1-t}},$$

and so we have

$$\Delta_{\mathrm{CDF}}\big(\widehat{U}_n, \mathrm{Unif}[0,1]\big) \le 2\sqrt{\frac{\log n}{n}} + \sup_{t\in[\frac{\log n}{n}, 1-\frac{\log n}{n}]} \frac{|\widehat{U}_n(t)-t|}{\sqrt{\min\{t, 1-t\}}}$$

$$= 2\sqrt{\frac{\log n}{n}} + \max\Bigg\{\sup_{t\in[\frac{\log n}{n},\frac{1}{2}]} \frac{|\widehat{U}_n(t)-t|}{\sqrt{t}}, \sup_{t\in[\frac{1}{2},1-\frac{\log n}{n}]} \frac{|\widehat{U}_n(t)-t|}{\sqrt{1-t}}\Bigg\}.$$

Next, Shorack and Wellner [47], Proposition 11.1.1 (part (10)) + Inequality 11.2.1, (applied with $q(t)=\sqrt{t}$, with $a=\frac{\log n}{n}$, and with $b=\delta=\frac{1}{2}$) establishes that, for any $\lambda>0$,

$$\mathbb{P}\Bigg\{\sup_{t\in[\frac{\log n}{n},\frac{1}{2}]} \frac{|\widehat{U}_n(t)-t|}{\sqrt{t}} \ge \frac{\lambda}{\sqrt{n}}\Bigg\} \le 12\int_{\frac{\log n}{n}}^{1/2} \frac{1}{t}\cdot \exp\Bigg\{-\frac{\lambda^2}{8(1+\frac{\lambda}{3\sqrt{\log n}})}\Bigg\}\,\mathrm{d}t,$$

as long as n satisfies $\frac{\log n}{n}\le\frac{1}{4}$ (which holds for $n>8$; for $n\le 8$, by taking $c'\ge 2$ we can ensure that the lemma's claim is trivial, since $\Delta_{\mathrm{CDF}}(\widehat{U}_n,\mathrm{Unif}[0,1])\le 1$ deterministically). Furthermore, clearly we see that $\sup_{t\in[\frac{\log n}{n},\frac{1}{2}]}\frac{|\widehat{U}_n(t)-t|}{\sqrt{t}}$ and $\sup_{t\in[\frac{1}{2},1-\frac{\log n}{n}]}\frac{|\widehat{U}_n(t)-t|}{\sqrt{1-t}}$ are equal in distribution. Therefore, we have

$$\mathbb{P}\Bigg\{\Delta_{\mathrm{CDF}}\big(\widehat{U}_n,\mathrm{Unif}[0,1]\big) \ge 2\sqrt{\frac{\log n}{n}} + \frac{\lambda}{\sqrt{n}}\Bigg\} \le 24\log\bigg(\frac{n}{2\log n}\bigg)\cdot\exp\Bigg\{-\frac{\lambda^2}{8(1+\frac{\lambda}{3\sqrt{\log n}})}\Bigg\}$$

for any $\lambda>0$. Taking $\lambda = 5(c+2)\sqrt{\log n}$, we can calculate $\exp\{-\frac{\lambda^2}{8(1+\frac{\lambda}{3\sqrt{\log n}})}\}\le\exp\{-(c+2)\log n\} = n^{-(c+2)}$, and so we have

$$\mathbb{P}\Bigg\{\Delta_{\mathrm{CDF}}\big(\widehat{U}_n,\mathrm{Unif}[0,1]\big) \ge 2\sqrt{\frac{\log n}{n}} + 5(c+2)\sqrt{\frac{\log n}{n}}\Bigg\} \le 24\log\bigg(\frac{n}{2\log n}\bigg)\cdot n^{-(c+2)} \le n^{-c}$$

where the last step holds since we have assumed that $n>8$. This proves the lemma with $c'=5c+12$.

A.6. Proof of Lemma 14

We have

$$\begin{aligned}
\epsilon_P &= \mathbb{E}_P\big[|X-\mu_P|\big] \\
&= 2\mathbb{E}_P\big[(X-\mu_P)_+\big] \\
&\le 2\mathbb{E}_P\big[|X-\mu_P|\cdot \mathbf{1}\{X>\mu_P\}\big] \\
&\le 2\mathbb{E}_P\big[|X-\mu_P|^q\big]^{1/q}\mathbb{E}_P\big[\mathbf{1}\{X>\mu_P\}^{\frac{q}{q-1}}\big]^{\frac{q-1}{q}} \\
&\le 2\big(\mathbb{E}_P\big[|X|^q\big]^{1/q} + \big(|\mu_P|^q\big)^{1/q}\big)\cdot\mathbb{P}_P\{X>\mu_P\}^{\frac{q-1}{q}} \\
&\le 4M_q\cdot\mathbb{P}_P\{X>\mu_P\}^{\frac{q-1}{q}}.
\end{aligned}$$

Therefore,

$$\mathbb{P}_P\{X > \mu_P\} \geq \left(\frac{\epsilon_P}{4M_q}\right)^{\frac{q}{q-1}}.$$

Similarly, the same bound holds for $\mathbb{P}_P\{X < \mu_P\}$.

A.7. Proofs of lower bounds (Theorems 4 and 6)

We begin with some preliminary calculations that we will use for the constructions for both theorems. Fix any $0 < \rho_0 < \rho_1$ and any $\beta \in (0, \rho_0/\rho_1]$. Let P be the mixture distribution drawing

$$X \sim \begin{cases} \text{Unif}\big(\mathbb{S}_{d-1}(\rho_0)\big), & \text{with probability } 1-\beta, \\ \text{Unif}\big(\mathbb{S}_{d-1}(\rho_1)\big), & \text{with probability } \beta. \end{cases} \tag{19}$$

Defining

$$s_d = \mathbb{E}\big[|V_1|\big] \quad \text{for } V = (V_1, \dots, V_d) \sim \text{Unif}(\mathbb{S}_{d-1}), \tag{20}$$

we can calculate

$$\epsilon_P = (1-\beta)\rho_0 \cdot s_d + \beta\rho_1 \cdot s_d \geq s_d \rho_0.$$

We will apply Lemma 18 to this distribution P and the log-density $\phi = \phi^*(P)$ of its log-concave projection. Observe that ϕ is spherically symmetric around 0, and is constant over $\|x\| \leq \rho_0$ – in particular, this means that $\phi(x) = M_\phi$ for all $\|x\| \leq \rho_0$, where $M_\phi = \sup_{x \in \mathbb{R}^d} \phi(x)$ as before. Next, let $t_* \geq 0$ be the value of $M_\phi - \phi(x)$ for points x with $\|x\| = \rho_1$ (since ϕ is spherically symmetric, this is well defined). We now split into cases. If $t_* \geq \frac{8d\rho_1}{r_d s_d \rho_0}$, then applying Lemma 18 with $R = \rho_1/2$, $x_\phi = 0$, and $t = t_*$, we obtain

$$\beta \leq \mathbb{P}_P\big\{\phi(X) \leq M_\phi - t_* \text{ and } \|X\| \leq \rho_1\big\} \leq \frac{16d}{b_d r_d s_d \rho_0} \cdot \frac{\rho_1}{t_*^2},$$

which proves that

$$t_* \leq \sqrt{\frac{16d}{b_d r_d s_d} \cdot \frac{\rho_1}{\rho_0 \beta}}.$$

If this case does not hold, then we instead have $t_* < \frac{8d\rho_1}{r_d s_d \rho_0}$, so combining the two cases,

$$t_* \leq \max\left\{\sqrt{\frac{16d}{b_d r_d s_d} \cdot \frac{\rho_1}{\rho_0\beta}}, \frac{8d}{r_d s_d} \cdot \frac{\rho_1}{\rho_0}\right\} \leq \max\left\{\sqrt{\frac{16d}{b_d r_d s_d}}, \frac{8d}{r_d s_d}\right\} \cdot \sqrt{\frac{\rho_1}{\rho_0\beta}},$$

where the last step comes from our assumption on β. Therefore,

$$\phi(x) \geq \phi(0) - \max\left\{\sqrt{\frac{16d}{b_d r_d s_d}}, \frac{8d}{r_d s_d}\right\} \cdot \sqrt{\frac{\rho_1}{\rho_0\beta}}$$

for $\|x\| = \rho_1$ while

$$\phi(x) = \phi(0)$$

for $\|x\| \le \rho_0$. By concavity of ϕ, then

$$\phi(x) \ge \phi(0) - \max\left\{\sqrt{\frac{16d}{b_d r_d s_d}}, \frac{8d}{r_d s_d}\right\}$$

for all x with $\|x\| \le \rho_0 + (\rho_1 - \rho_0) \cdot \sqrt{\frac{\rho_0 \beta}{\rho_1}}$. Therefore, for any density f supported on $\mathbb{B}_d(\rho_0)$, it holds that

$$\begin{aligned} \mathrm{d}_{\mathrm{H}}^2\big(f, \psi^*(P)\big) &\ge \int_{\mathbb{R}^d} e^{\phi(0) - \max\{\sqrt{\frac{16d}{b_d r_d s_d}}, \frac{8d}{r_d s_d}\}} \cdot \mathbf{1}\left\{\rho_0 < \|x\| < \rho_0 + (\rho_1 - \rho_0) \cdot \sqrt{\frac{\rho_0\beta}{\rho_1}}\right\} \mathrm{d}x \\ &= e^{\phi(0) - \max\{\sqrt{\frac{16d}{b_d r_d s_d}}, \frac{8d}{r_d s_d}\}} \cdot \mathrm{Leb}_d\big(\mathbb{B}_d\big(\rho_0 + (\rho_1 - \rho_0) \cdot \sqrt{\rho_0\beta/\rho_1}\big) \backslash \mathbb{B}_d(\rho_0)\big) \\ &\ge e^{\phi(0) - \max\{\sqrt{\frac{16d}{b_d r_d s_d}}, \frac{8d}{r_d s_d}\}} \cdot \rho_0^{d-1} \cdot (\rho_1 - \rho_0) \cdot \sqrt{\frac{\rho_0\beta}{\rho_1}} \cdot S_{d-1}, \end{aligned}$$

where as before S_{d-1} denotes the surface area of $\mathbb{S}_{d-1}$. Finally, we need to place a lower bound on $\phi(0)$. By Corollary 8, we know that the covariance matrix Σ of the distribution with log-density ϕ has operator norm bounded as

$$\|\Sigma\|_{\mathrm{op}} \le 16\big((1-\beta)\rho_0 + \beta\rho_1\big)^2.$$

Furthermore, $\tilde{\phi}(x) = \frac{1}{2}\log\det(\Sigma) + \phi(\Sigma^{1/2}x)$ is an isotropic concave log-density, and so $\tilde{\phi}(0) \ge c'_d$ where $c'_d > 0$ depends only on d, by Lovász and Vempala [35], Theorem 5.14(d). Therefore,

$$\phi(0) \ge c'_d - \frac{d}{2}\log(16) - d\log\big((1-\beta)\rho_0 + \beta\rho_1\big).$$

We conclude that

$$\mathrm{d}_{\mathrm{H}}^2\big(f, \psi^*(P)\big) \ge c''_d \cdot \rho_0^{d-1} \cdot (\rho_1 - \rho_0) \cdot \sqrt{\frac{\rho_0\beta}{\rho_1}} \cdot \big((1-\beta)\rho_0 + \beta\rho_1\big)^{-d}, \tag{21}$$

where c''_d depends only on d.

A.7.1. *Completing the proof of Theorem 4*

To prove Theorem 4, let P be the distribution constructed in (19) with

$$\rho_0 = \epsilon/s_d,\ \rho_1 = 2\epsilon/s_d,\ \beta = \min\left\{\frac{s_d\delta}{\epsilon}, \frac{1}{2}\right\},$$

where s_d is defined as in (20). Let

$$Q = \mathrm{Unif}\big(\mathbb{S}_{d-1}(\rho_0)\big).$$

Clearly $\epsilon_P \ge \epsilon_Q = s_d\rho_0 = \epsilon$, and $\mathrm{d}_{\mathrm{W}}(P, Q) = \beta(\rho_1 - \rho_0) \le \delta$, thus satisfying the conditions of the theorem. Since Q is supported on $\mathbb{B}_d(\rho_0)$, $\psi^*(Q)$ is also supported on this ball. Then applying our calculation (21), and plugging in our choices of ρ_0, ρ_1, β, after simplifying we have

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(P),\psi^*(Q)\big) \ge c_d'' \cdot \frac{2^d}{3^d} \cdot \sqrt{\min\left\{\frac{s_d\delta}{2\epsilon},\frac{1}{4}\right\}}.$$

This completes the proof of the theorem, when c_d is chosen appropriately.

A.7.2. Completing the proof of Theorem 6

The first term in the lower bound, that is, $\sup_{P\in\mathcal{P}_d:\mathbb{E}_P[\|X\|^q]\le 1,\epsilon_P\ge\epsilon_d^*}\mathbb{E}[\mathrm{d}_{\mathrm{H}}^2(\psi^*(\widehat{P}_n),\psi^*(P))] \ge c_d n^{-\frac{2}{d+1}}$, holds by Kim and Samworth [31], Theorem 1, which establishes this as the minimax rate (for $d\ge 2$) over distributions P that are log-concave (we can verify that the distribution P constructed in their proof satisfies the conditions $\mathbb{E}_P[\|X\|^q]\le 1$, $\epsilon_P\ge\epsilon_d^*$, for appropriately chosen ϵ_d^*). If instead $d=1$, then the first term cannot be the minimum.

Next, to prove the second term in the lower bound, we consider a mixture model. Let P be the distribution constructed in (19) with

$$\rho_0=\frac{1}{2},\qquad \rho_1=n^{1/q},\qquad \beta=\frac{1}{2n}.$$

Then clearly $\mathbb{E}_P[\|X\|^q]\le 1$, and $\epsilon_P\ge\frac{1}{2}s_d$, so $\epsilon_P\ge\epsilon_d^*$ for an appropriately chosen ϵ_d^*. Now, with probability at least $1/2$, the observations $X_1,\dots,X_n$ are all drawn from the first component of the mixture model, that is, $\psi^*(\widehat{P}_n)$ is supported on $\mathbb{B}_d(1/2)$. On this event, applying (21) and plugging in our choices of ρ_0, ρ_1, β, after simplifying we have

$$\mathrm{d}_{\mathrm{H}}^2\big(\psi^*(\widehat{P}_n),\psi^*(P)\big) \ge c_d''' \cdot n^{-\frac{1}{2}+\frac{1}{2q}},$$

where c_d''' depends only on d. This establishes the second term in the lower bound claimed in Theorem 6, and thus completes the proof of the theorem.

Acknowledgements

The authors thank the anonymous reviewers and Oliver Feng for helpful comments.

Funding

R.F.B. was supported by the National Science Foundation via grant DMS-1654076 and by an Alfred P. Sloan fellowship. R.J.S. was supported by EPSRC grants EP/P031447/1 and EP/N031938/1.

References

[1] Aliprantis, C.D. and Border, K.C. (2006). *Infinite Dimensional Analysis*: *A Hitchhiker's Guide*, 3rd ed. Berlin: Springer. MR2378491

[2] 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 Series in Probability and Mathematical Statistics*. London–Sydney: Wiley. MR0326887

[3] Bellec, P.C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383 https://doi.org/10.1214/17-AOS1566

[4] Birgé, L. (1989). The Grenander estimator: A nonasymptotic approach. *Ann. Statist.* **17** 1532–1549. MR1026298 https://doi.org/10.1214/aos/1176347380

[5] Cai, T.T. and Low, M.G. (2015). A framework for estimation of convex functions. *Statist. Sinica* **25** 423–456. MR3379081

[6] Carpenter, T., Diakonikolas, I., Sidiropoulos, A. and Stewart, A. (2018). Near-optimal sample complexity bounds for maximum likelihood estimation of multivariate log-concave densities. In *COLT* 2018.

[7] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878 https://doi.org/10.1214/15-AOS1324

[8] Chen, Y. and Samworth, R.J. (2013). Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica* **23** 1373–1398. MR3114718

[9] Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*. *Wiley Series in Telecommunications*. New York: Wiley. A Wiley-Interscience Publication. MR1122806 https://doi.org/10.1002/0471200611

[10] Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.* **4** 254–270. MR2645484 https://doi.org/10.1214/09-EJS505

[11] Cule, M., Samworth, R. and Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. MR2758237 https://doi.org/10.1111/j.1467-9868.2010.00753.x

[12] Doss, C.R. and Wellner, J.A. (2016). Global rates of convergence of the MLEs of log-concave and s-concave densities. *Ann. Statist.* **44** 954–981. MR3485950 https://doi.org/10.1214/15-AOS1394

[13] Dudley, R.M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge: Cambridge Univ. Press. Revised reprint of the 1989 original. MR1932358 https://doi.org/10.1017/CBO9780511755347

[14] 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 https://doi.org/10.3150/08-BEJ141

[15] Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730. MR2816336 https://doi.org/10.1214/10-AOS853

[16] Durot, C. and Lopuhaä, H.P. (2018). Limit theory in monotone function estimation. *Statist. Sci.* **33** 547–567. MR3881208 https://doi.org/10.1214/18-STS664

[17] Fang, B. and Guntuboyina, A. (2019). On the risk of convex-constrained least squares estimators under misspecification. *Bernoulli* **25** 2206–2244. MR3961246 https://doi.org/10.3150/18-BEJ1051

[18] Feng, O.Y., Guntuboyina, A., Kim, A.K.H. and Samworth, R.J. (2021). Adaptation in multivariate log-concave density estimation. *Ann. Statist.* **49** 129–153. MR4206672 https://doi.org/10.1214/20-AOS1950

[19] Fresen, D. (2013). A multivariate Gnedenko law of large numbers. *Ann. Probab.* **41** 3051–3080. MR3127874 https://doi.org/10.1214/12-AOP804

[20] Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153 (1957). MR0093415 https://doi.org/10.1080/03461238.1956.10414944

[21] Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 539–555. Belmont, CA: Wadsworth. MR0822052

[22] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation Under Shape Constraints*: *Estimators, Algorithms and Asymptotics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. New York: Cambridge Univ. Press. MR3445293 https://doi.org/10.1017/CBO9781139020893

[23] Guntuboyina, A. and Sen, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379–411. MR3405621 https://doi.org/10.1007/s00440-014-0595-3

[24] Han, Q. (2019). Global empirical risk minimizers with "shape constraints" are rate optimal in general dimensions. arXiv preprint arXiv:1905.12823.

[25] Han, Q., Wang, T., Chatterjee, S. and Samworth, R.J. (2019). Isotonic regression in general dimensions. *Ann. Statist.* **47** 2440–2471. MR3988762 https://doi.org/10.1214/18-AOS1753

[26] Han, Q. and Wellner, J.A. (2016). Multivariate convex regression: Global risk bounds and adaptation. arXiv preprint arXiv:1601.06844.

[27] Han, Q. and Wellner, J.A. (2016). Approximation and estimation of s-concave densities via Rényi divergences. *Ann. Statist.* **44** 1332–1359. MR3485962 https://doi.org/10.1214/15-AOS1408
[28] Hildreth, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49** 598–619. MR0065093
[29] Jankowski, H. (2014). Convergence of linear functionals of the Grenander estimator under misspecification. *Ann. Statist.* **42** 625–653. MR3210981 https://doi.org/10.1214/13-AOS1196
[30] Kim, A.K.H., Guntuboyina, A. and Samworth, R.J. (2018). Adaptation in log-concave density estimation. *Ann. Statist.* **46** 2279–2306. MR3845018 https://doi.org/10.1214/17-AOS1619
[31] Kim, A.K.H. and Samworth, R.J. (2016). Global rates of convergence in log-concave density estimation. *Ann. Statist.* **44** 2756–2779. MR3576560 https://doi.org/10.1214/16-AOS1480
[32] Koenker, R. and Mizera, I. (2010). Quasi-concave density estimation. *Ann. Statist.* **38** 2998–3027. MR2722462 https://doi.org/10.1214/10-AOS814
[33] Kur, G., Dagan, Y. and Rakhlin, A. (2019). The log-concave maximum likelihood estimator is optimal in high dimensions. arXiv preprint arXiv:1903.05315v3.
[34] Lei, J. (2020). Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli* **26** 767–798. MR4036051 https://doi.org/10.3150/19-BEJ1151
[35] Lovász, L. and Vempala, S. (2007). The geometry of logconcave functions and sampling algorithms. *Random Structures Algorithms* **30** 307–358. MR2309621 https://doi.org/10.1002/rsa.20135
[36] Pal, J.K., Woodroofe, M. and Meyer, M. (2007). Estimating a Polya frequency function$_2$. In *Complex Datasets and Inverse Problems*. *Institute of Mathematical Statistics Lecture Notes – Monograph Series* **54** 239–249. Beachwood, OH: IMS. MR2459192 https://doi.org/10.1214/074921707000000184
[37] Patilea, V. (2001). Convex models, MLE and misspecification. *Ann. Statist.* **29** 94–123. MR1833960 https://doi.org/10.1214/aos/996986503
[38] Prakasa Rao, B.L.S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36. MR0267677
[39] Samworth, R.J. (2018). Recent progress in log-concave density estimation. *Statist. Sci.* **33** 493–509. MR3881205 https://doi.org/10.1214/18-STS666
[40] Samworth, R.J. and Sen, B. (2018). Editorial: Special issue on "Nonparametric inference under shape constraints". *Statist. Sci.* **33** 469–472. MR3881203 https://doi.org/10.1214/18-STS673
[41] Samworth, R.J. and Yuan, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.* **40** 2973–3002. MR3097966 https://doi.org/10.1214/12-AOS1060
[42] Schneider, R. (2014). *Convex Bodies: The Brunn–Minkowski Theory*, expanded ed. *Encyclopedia of Mathematics and Its Applications* **151**. Cambridge: Cambridge Univ. Press. MR3155183
[43] Schuhmacher, D., Hüsler, A. and Dümbgen, L. (2011). Multivariate log-concave distributions as a nearly parametric model. *Stat. Risk Model.* **28** 277–295. MR2838319 https://doi.org/10.1524/stnd.2011.1073
[44] Seijo, E. and Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.* **39** 1633–1657. MR2850215 https://doi.org/10.1214/10-AOS852
[45] Seregin, A. and Wellner, J.A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist.* **38** 3751–3781. MR2766867 https://doi.org/10.1214/10-AOS840
[46] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0595165
[47] Shorack, G.R. and Wellner, J.A. (2009). *Empirical Processes with Applications to Statistics*. *Classics in Applied Mathematics* **59**. Philadelphia, PA: SIAM. Reprint of the 1986 original [MR0838963]. MR3396731 https://doi.org/10.1137/1.9780898719017.ch1
[48] van de Geer, S.A. (2000). *Empirical Processes in M-Estimation*. Cambridge: Cambridge Univ. Press.
[49] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2
[50] Villani, C. (2009). *Optimal Transport: Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Berlin: Springer. MR2459454 https://doi.org/10.1007/978-3-540-71050-9
[51] von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Stat.* **18** 309–348. MR0022330 https://doi.org/10.1214/aoms/1177730385

[52] Walther, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc.* **97** 508–513. MR1941467 https://doi.org/10.1198/016214502760047032
[53] Walther, G. (2009). Inference and modeling with log-concave distributions. *Statist. Sci.* **24** 319–327. MR2757433 https://doi.org/10.1214/09-STS303
[54] Xu, M. and Samworth, R.J. (2021). High-dimensional nonparametric density estimation via symmetry and shape constraints. *Ann. Statist.* **49** 650–672. https://doi.org/10.1214/20-AOS1972
[55] Yang, F. and Barber, R.F. (2019). Contraction and uniform convergence of isotonic regression. *Electron. J. Stat.* **13** 646–677. MR3914177 https://doi.org/10.1214/18-ejs1520
[56] Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898 https://doi.org/10.1214/aos/1021379864

Received July 2020 and revised December 2020