Risk Bounds for Mixture Density Estimation on Compact Domains via the h-Lifted Kullback–Leibler Divergence

Anonymous authors
Paper under double-blind review

Abstract

We consider the problem of estimating probability density functions based on sample data, using a finite mixture of densities from some component class. To this end, we introduce the h-lifted Kullback–Leibler (KL) divergence as a generalization of the standard KL divergence and a criterion for conducting risk minimization. Under a compact support assumption, we prove an $\mathcal{O}(1/\sqrt{n})$ bound on the expected estimation error when using the h-lifted KL divergence, which extends the results of Rakhlin et al. (2005, ESAIM: Probability and Statistics, Vol. 9) and Li & Barron (1999, Advances in Neural Information Processing Systems, Vol. 12) to permit the risk bounding of density functions that are not strictly positive. We develop a procedure for the computation of the corresponding maximum h-lifted likelihood estimators (h-MLLEs) using the Majorization-Maximization framework and provide experimental results in support of our theoretical bounds.

1 Introduction

Let $(\Omega, \mathfrak{A}, \mathbf{P})$ be an abstract probability space and let $X : \Omega \to \mathcal{X}$ be a random variable taking values in the measurable space $(\mathcal{X}, \mathfrak{F})$, where $\mathcal{X}$ is a compact metric space equipped with its Borel σ-algebra $\mathfrak{F}$. Suppose that we observe an independent and identically distributed (i.i.d.) sample of random variables $\mathbf{X}_n = (X_i)_{i\in[n]}$, where $[n] = \{1, \dots, n\}$, and that each X_i arises from the same data generating process as X, characterized by the probability measure $F \ll \mu$ on $(\mathcal{X}, \mathfrak{F})$, with density function $f = \mathrm{d}F/\mathrm{d}\mu$, for some σ-finite μ.

In this work, we are concerned with the estimating f via a data dependent double-index sequence of estimators $(f_{k,n})_{k,n\in\mathbb{N}}$, where

$$\begin{aligned} f_{k,n} \in \mathcal{C}_k &= \mathrm{co}_k\left(\mathcal{P}\right) \\ &= \left\{ f_k\left(\cdot;\psi_k\right) = \sum_{j=1}^{k} \pi_j \varphi\left(\cdot;\theta_j\right) \mid \varphi\left(\cdot;\theta_j\right) \in \mathcal{P},\, \pi_j \geq 0,\, j \in [k],\, \sum_{j=1}^{k} \pi_j = 1 \right\}, \end{aligned}$$

for each $k, n \in \mathbb{N}$, and where

$$\mathcal{P} = \left\{ \varphi\left(\cdot;\theta\right) : \mathcal{X} \to \mathbb{R}_{\geq 0} \mid \theta \in \Theta \subset \mathbb{R}^d \right\},$$

$\psi_k = (\pi_1, \dots, \pi_k, \theta_1, \dots, \theta_k)$, and $d \in \mathbb{N}$. To ensure the measurability and existence of various optima, we shall assume that φ is *Caratheodory* in the sense that $\varphi\left(\cdot;\theta\right)$ is $(\mathcal{X}, \mathfrak{F})$-measurable, for each $\theta \in \Theta$, and $\varphi\left(X;\cdot\right)$ is continuous for each $X \in \mathcal{X}$.

In the definition above, we can identify the set $\mathcal{C}_k = \mathrm{co}_k\left(\mathcal{P}\right)$ as the set of density functions that can be written as a convex combination of k elements of $\mathcal{P}$, where $\mathcal{P}$ is often called the space of component density functions. We then interpret $\mathcal{C}_k$ as the class of k-component finite mixtures of densities of class $\mathcal{P}$, as studied, for example, by McLachlan & Peel (2004); Nguyen et al. (2020; 2022b).

1.1 Risk bounds for mixture density estimation

We are particularly interested in oracle bounds of the form

$$\mathbf{E}\left\{\ell\left(f, f_{k,n}\right)\right\} - \ell\left(f, \mathcal{C}\right) \leq \rho\left(k, n\right), \tag{1}$$

where $(p, q) \mapsto \ell(p, q) \in \mathbb{R}_{\geq 0}$ is a loss function on pairs of density functions. We also define the density-to-class loss

$$\ell\left(f, \mathcal{C}\right) = \inf_{q \in \mathcal{C}} \ell\left(f, q\right), \ \mathcal{C} = \mathrm{cl}\left(\bigcup_{k \in \mathbb{N}} \mathrm{co}_k\left(\mathcal{P}\right)\right),$$

where cl$(\cdot)$ is the closure. Here, we identify $(k, n) \mapsto \rho(k, n)$ as a characterization of the rate at which the left-hand side of (1) converges to zero as k and n increase. Our present work follows the research of Li & Barron (1999), Rakhlin et al. (2005) and Klemelä (2007) (see also Klemelä 2009, Ch. 19). In Li & Barron (1999) and Rakhlin et al. (2005), the authors consider the case where $\ell(p, q)$ is taken to be the Kullback–Leibler (KL) divergence

$$\mathrm{KL}\left(p \,||\, q\right) = \int p \log \frac{p}{q} \mathrm{d}\mu,$$

and $f_{k,n} = f_k\left(\cdot; \psi_{k,n}\right)$ is a maximum likelihood estimator (MLE), where

$$\psi_{k,n} \in \underset{\psi_k \in \mathcal{S}_k \times \Theta^k}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \log f_k\left(X_i; \psi_k\right),$$

is a function of $\mathbf{X}_n$, with $\mathcal{S}_k$ denoting the probability simplex in $\mathbb{R}^k$.

Under the assumption that $f, f_k \geq a$, for some $a > 0$ and each $k \in [n]$ (i.e., strict positivity), Li & Barron (1999) obtained the bound

$$\mathbf{E}\left\{\mathrm{KL}\left(f \,||\, f_{k,n}\right)\right\} - \mathrm{KL}\left(f \,||\, \mathcal{C}\right) \leq c_1 \frac{1}{k} + c_2 \frac{k \log\left(c_3 n\right)}{n},$$

for constants $c_1, c_2, c_3 > 0$, which was then improved by Rakhlin et al. (2005) who obtain the bound

$$\mathbf{E}\left\{\mathrm{KL}\left(f \,||\, f_{k,n}\right)\right\} - \mathrm{KL}\left(f \,||\, \mathcal{C}\right) \leq c_1 \frac{1}{k} + c_2 \frac{1}{\sqrt{n}},$$

for constants $c_1, c_2 > 0$ (constants $(c_j)_{j \in \mathbb{N}}$ are typically different between expressions).

Alternatively, Klemelä (2007) takes $\ell(p, q)$ to be the squared $L_2(\mu)$ norm distance (i.e., the least-squares loss):

$$\ell\left(p, q\right) = \left\|p - q\right\|_{2,\mu}^2,$$

where $\|p\|_{2,\mu}^2 = \int_{\mathcal{X}} |p|^2 \mathrm{d}\mu$, for each $p \in L_2(\mu)$, and choose $f_{k,n}$ as minimizers of the $L_2(\mu)$ empirical risk: i.e., $f_{k,n} = f_k\left(\cdot; \psi_{k,n}\right)$, where

$$\psi_{k,n} \in \underset{\psi_k \in \mathcal{S}_k \times \Theta^k}{\arg\min} -\frac{2}{n} \sum_{i=1}^{n} f_k\left(\cdot; \psi_k\right) + \left\|f_k\left(\cdot; \psi_k\right)\right\|_{2,\mu}^2. \tag{2}$$

Here, Klemelä (2007) establish the bound

$$\mathbf{E}\left\|f - f_{k,n}\right\|_{2,\mu}^2 - \inf_{q \in \mathcal{C}} \left\|f - q\right\|_{2,\mu}^2 \leq c_1 \frac{1}{k} + c_2 \frac{1}{\sqrt{n}},$$

$c_1, c_2 > 0$, without the lower bound assumption on f, f_k above, even permitting $\mathcal{X}$ to be unbounded. Via the main results of Naito & Eguchi (2013), the bound above can be generalized to the U-divergences, which includes the special $L_2(\mu)$ norm distance as a special case.

On the one hand, the sequence of MLEs required for the results of Li & Barron (1999) and Rakhlin et al. (2005) are typically computable, for example, via the usual expectation–maximization approach (cf. McLachlan & Peel 2004, Ch. 2). This contrasts with the computation of least-squares density estimators of form

(2), which requires evaluations of the typically intractable integral expressions: $\|f_k(\cdot;\psi_k)\|_2^2$. However, the least-squares approach of Klemelä (2007) permits the analysis using families $\mathcal{P}$ of usual interest, such as normal distributions and beta distributions, the latter of which being compactly supported but having densities that cannot be bounded away from zero without restrictions, and thus do not satisfy the regularity conditions of Li & Barron (1999) and Rakhlin et al. (2005).

1.2 Main contributions

We propose the following h-lifted KL divergence, as a generalization of the standard KL divergence to address the computationally tractable estimation of density functions which do not satisfy the regularity conditions of Li & Barron (1999) and Rakhlin et al. (2005). The use of the h-lifted KL divergence has the possibility to advance theories based on the standard KL divergence in statistical machine learning. To this end, let $h : \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a function in $L_1(\mu)$, and define the h-lifted KL divergence by:

$$\mathrm{KL}_h(p\,||\,q) = \int_{\mathcal{X}} \{p+h\} \log \frac{p+h}{q+h} \mathrm{d}\mu. \tag{3}$$

In the sequel, we shall show that KL_h is a Bregman divergence on the space of probability density functions, as per Csiszár (1995).

Assume that h is a probability density function, and let $\mathbf{Y}_n = (Y_i)_{i\in[n]}$ be a an i.i.d. sample, independent of $\mathbf{X}_n$, where each $Y_i : \Omega \to \mathcal{X}$ is a random variable with probability measure on $(\mathcal{X}, \mathfrak{F})$, characterized by the density h with respect to μ. Then, for each k and n, let $f_{k,n}$ be defined via the maximum h-lifted likelihood estimator (h-MLLE): $f_{k,n} = f_k(\cdot;\psi_{k,n})$, where

$$\psi_{k,n} \in \underset{\psi_k \in \mathcal{S}_k \times \Theta^k}{\arg\max}\ \frac{1}{n}\sum_{i=1}^{n} \left(\log\{f_k(X_i;\psi_k) + h(X_i)\} + \log\{f_k(Y_i;\psi_k) + h(Y_i)\}\right). \tag{4}$$

The primary aim of this work is to show that

$$\mathbf{E}\{\mathrm{KL}_h(f\,||\,f_{k,n})\} - \mathrm{KL}_h(f\,||\,\mathcal{C}) \leq c_1\frac{1}{k} + c_2\frac{1}{\sqrt{n}} \tag{5}$$

for some constants $c_1, c_2 > 0$, without requiring the strict positivity assumption that $f, f_k \geq a > 0$.

This result is a compromise between the works of Li & Barron (1999) and Rakhlin et al. (2005), and Klemelä (2007), as it applies to a broader space of component densities $\mathcal{P}$, and because the required h-MLLEs, (4), can be efficiently computed via minorization–maximization (MM) algorithms (see, e.g., Lange 2016). We shall discuss this assertion in Section 6.

1.3 Relevant literature

Our work largely follows the approach of Li & Barron (1999), which have was extended upon by Rakhlin et al. (2005) and Klemelä (2007). All three texts use approaches based on the availability of greedy algorithms for maximizing convex functions with convex functional domains. In this work, we shall make use of the proof techniques of Zhang (2003). Related results in this direction can be found in DeVore & Temlyakov (2016) and Temlyakov (2016). Making the same boundedness assumption as Rakhlin et al. (2005), Dalalyan & Sebbar (2018) obtain refined oracle inequalities under the additional assumption that the class $\mathcal{P}$ is finite. Numerical implementations of greedy algorithms for estimating finite mixtures of Gaussian densities were studied by Vlassis & Likas (2002) and Verbeek et al. (2003).

The h-MLLE as an optimization objective can be compared to other similar modified likelihood estimators, such as the L_q likelihood of Ferrari & Yang (2010) and Qin & Priebe (2013), the β-likelihood of Basu et al. (1998) and Fujisawa & Eguchi (2006), penalized likelihood estimators, such as maximum a posteriori estimators of Bayesian models, or f-separable Bregman distortion measures of Kobayashi & Watanabe (2024; 2021).

The practical computation of the h-MLLEs, (4), is made possible via the MM algorithm framework of Lange (2016), see also Hunter & Lange (2004), Wu & Lange (2010), and Nguyen (2017) for further details. Such algorithms have well-studied global convergence properties and can be modified for mini-batch and stochastic settings (see, e.g., Razaviyayn et al., 2013 and Nguyen et al., 2022a).

A related and popular setting of investigations is that of model selection, where the objects of interest are single-index sequences $(f_{k_n,n})_{n\in\mathbb{N}}$, and where the aim is to obtain finite-sample bounds for losses of the form $\ell\left(f_{k_n,n}, f\right)$, where each $k_n \in \mathbb{N}$ is a data dependent function, often obtained by optimizing some penalized loss criterion, as described in Massart (2007), Koltchinskii (2011, Ch. 6), and Giraud (2021, Ch. 2). In the context of finite mixtures, examples of such analyses can be found in the works of Maugis & Michel (2011) and Maugis-Rabusseau & Michel (2013). A comprehensive bibliography of model selection results for finite mixtures and related statistical models can be found in Nguyen et al. (2022c).

1.4 Organization of paper

The manuscript proceeds as follows. In the following section we formally characterize the h-lifted KL divergence as a Bregman divergence, and establish some of its properties. In Section 4, we prove new risk bounds of the form (1) in terms of the h-lifted KL divergence. We discuss the computation of the h-lifted likelihood estimator of form (4) in Section 6, followed by the presentation of some empirical results regarding the convergence of (1) in terms of both k and n. Additional technical results are also included in the Appendices.

2 The h-lifted KL divergence and its properties

In this section we formally define the h-lifted KL divergence on the space of density functions and establish some of its properties.

Definition 1 (h-lifted KL divergence)**.** *Let f, g, and h be probability density functions on the space $\mathcal{X}$, where $h > 0$. The h-lifted* KL *divergence from g to f is defined as follows:*

$$\mathrm{KL}_h\left(f\,||\,g\right) = \int_{\mathcal{X}} \{f+h\} \log\frac{f+h}{g+h} \mathrm{d}\mu = \mathbf{E}_f\left\{\log\frac{f+h}{g+h}\right\} + \mathbf{E}_h\left\{\log\frac{f+h}{g+h}\right\}.$$

2.1 $\mathbf{KL}_h$ as a Bregman divergence

Let $\phi : \mathcal{I} \to \mathbb{R}$, $\mathcal{I} = (0, \infty)$ be a strictly convex function that is continuously differentiable. The Bregman divergence between scalars $d_\phi : \mathcal{I} \times \mathcal{I} \to \mathbb{R}_{\geq 0}$ generated by the function ϕ is given by:

$$d_\phi(p, q) = \phi(p) - \phi(q) - \phi'(q)(p - q),$$

where $\phi'(q)$ denotes the derivative of ϕ at q.

Bregman divergences possess several useful properties, including the following list:

1. Non-negativity: $d_\phi(p, q) \geq 0$ for all $p, q \in \mathcal{I}$ with equality if and only if $p = q$;
2. Asymmetry: $d_\phi(p, q) \neq d_\phi(q, p)$ in general;
3. Convexity: $d_\phi(p, q)$ is convex in p for every fixed $q \in \mathcal{I}$.
4. Linearity: $d_{c_1\phi_1 + c_2\phi_2}(p, q) = c_1\, d_{\phi_1}(p, q) + c_2\, d_{\phi_2}(p, q)$ for $c_1, c_2 \geq 0$.

The properties for Bregman divergences between scalars can be extended to density functions and other functional spaces, as established in Frigyik et al. (2008) and Stummer & Vajda (2012), for example. We also direct the interested reader to the works of Pardo (2006), Basu et al. (2011), and Amari (2016).

The class of h-lifted KL divergences constitute a generalization of the usual KL divergence and are a subset of the Bregman divergences over the space of density functions that are considered by Csiszár (1995). Namely, let $\mathcal{P}$ be a convex set of probability densities with respect to the measure μ on $\mathcal{X}$. The Bregman divergence $D_\phi : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ between densities $p, q \in \mathcal{P}$ can be constructed as follows:

$$D_\phi(p \,||\, q) = \int_{\mathcal{X}} d_\phi \left(p(x), q(x)\right) \mathrm{d}\mu(x).$$

The h-lifted KL divergence KL_h as a Bregman divergence is generated by the function $\phi(u) = (u+h)\log(u+h) - (u+h) + 1$. This assertion is demonstrated in Appendix A.

2.2 Advantages of the h-lifted KL divergence

When the standard KL divergence is employed in the density estimation problem, it is common to restrict consideration of density functions to those bounded away from zero by some positive constant. That is, one typically considers the smaller class of so-called *admissible* target densities $\mathcal{P}_\alpha \subset \mathcal{P}$ (cf. Meir & Zeevi, 1997), where

$$\mathcal{P}_\alpha = \left\{\varphi(\cdot; \theta) \in \mathcal{P} \mid \varphi(\cdot;\theta) \geq \alpha > 0\right\}.$$

Without this restriction, the standard KL divergence can be unbounded, even for functions with bounded L_1 norms. For example, let p and q be densities of beta distributions on the support $\mathcal{X} = [0, 1]$. That is, suppose that $p, q \in \mathcal{P}_{\mathrm{beta}}$, respectively characterized by parameters $\theta_p = (a_p, b_p)$ and $\theta_q = (a_q, b_q)$, where

$$\mathcal{P}_{\mathrm{beta}} = \left\{ x \mapsto \beta\left(x;\theta\right) = \frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} x^{a-1}\left(1-x\right)^{b-1}, \theta = (a,b) \in \mathbb{R}^2_{>0} \right\}. \tag{6}$$

Then, from Gil et al. (2013), the KL divergence between p and q is given by:

$$\begin{aligned}\mathrm{KL}\left(p \,||\, q\right) = \log\left\{\frac{\Gamma\left(a_q\right)\Gamma\left(b_q\right)}{\Gamma\left(a_q+b_q\right)}\right\} - \log\left\{\frac{\Gamma\left(a_p\right)\Gamma\left(b_p\right)}{\Gamma\left(a_p+b_p\right)}\right\} \\ + \left(a_p - a_q\right)\left\{\psi\left(a_p\right) - \psi\left(a_p+b_p\right)\right\} + \left(b_p - b_q\right)\left\{\psi\left(b_p\right) - \psi\left(a_p+b_p\right)\right\},\end{aligned}$$

where $\psi : \mathbb{R}_{>0} \to \mathbb{R}$ is the digamma function. Next, suppose that $a_p = b_q$ and $a_q = b_p = 1$, which leads to the simplification

$$\mathrm{KL}\left(p \,||\, q\right) = \left(a_p - 1\right)\left\{\psi\left(a_p\right) - \psi(1)\right\}.$$

Since ψ is strictly increasing, we observe that the right-hand side diverges as $a_p \to \infty$. Thus, the KL divergence between beta distributions is unbounded. The h-lifted KL divergence in contrast does not suffer from this problem, and does not require the restriction to $\mathcal{P}_\alpha$. This allows us to consider cases where $p, q \in \mathcal{P}$ are not bounded away from 0.

Proposition 2. $\mathrm{KL}_h\left(p \,||\, q\right)$ *is bounded for all continuous densities* $p, q \in \mathcal{P}$.

Proof. Let $\tilde{p} = p + h$ and $\tilde{q} = q + h$. Since h is positive, there exists some $\tilde{q}_*$ such that $\tilde{q}_* = \inf_{x\in\mathcal{X}}\left\{q(x) + h(x)\right\} > 0$. Similarly, since $\mathcal{X}$ is compact, there exists some positive $\tilde{p}^*$ such that $0 < \tilde{p}^* = \sup_{x\in\mathcal{X}}\left\{p(x) + h(x)\right\} < \infty$. Define $M = \sup_{x\in\mathcal{X}} \log\{\tilde{p}(x)/\tilde{q}(x)\}$. Then $M < \infty$, and

$$\mathrm{KL}_h\left(p \,||\, q\right) = \int_{\mathcal{X}} \tilde{p} \log\frac{\tilde{p}}{\tilde{q}} \mathrm{d}\mu \leq \sup_{x\in\mathcal{X}} \log\frac{\tilde{p}}{\tilde{q}} \int_{\mathcal{X}} \tilde{p} \mathrm{d}\mu = 2M < \infty.$$

■

Let $L_p(f, g)$ denote the standard L_p norm,

$$L_p(f, g) = \left\{\int_{\mathcal{X}} \left|f(x) - g(x)\right|^p \mathrm{d}\mu(x)\right\}^{1/p}.$$

As remarked previously, Klemelä (2007) established empirical risk bounds in terms of the L_2 norm distance. Following results from Meir & Zeevi (1997), characterizing the relationship between the KL divergence in terms of the L_2 norm distance, we establish the corresponding relationship between the h-lifted KL divergence and the L_2 norm distance.

Proposition 3. *For probability density functions f, g, and h, where h is such that $h(x) \geq \gamma > 0$ for all $x \in \mathcal{X}$, the following inequality holds:*

$$\mathrm{KL}_h\left(f \,||\, g\right) \leq \gamma^{-1} L_2^2(f, g).$$

Proof. Defining $\tilde{f}$ and $\tilde{g}$ as above, we have

$$\mathrm{KL}_h\left(f \,||\, g\right) = \int_{\mathcal{X}} \tilde{f} \log \frac{\tilde{f}}{\tilde{g}} \mathrm{d}\mu \leq \int_{\mathcal{X}} \tilde{f} \left(\frac{\tilde{f}}{\tilde{g}} - 1\right) \mathrm{d}\mu = \int_{\mathcal{X}} \frac{(f-g)^2}{\tilde{g}} \mathrm{d}\mu \leq \gamma^{-1} L_2^2(f, g),$$

■

The following section is devoted to establishing some technical definitions and instrumental results which will be utilized in later sections.

3 Preliminaries

Recall that we are interested in bounds of the form (1). Note that $\mathcal{P}$ is a subset of the linear space

$$\mathcal{V} = \mathrm{cl}\left(\bigcup_{k \in \mathbb{N}} \left\{\sum_{j=1}^{k} \varpi_j \varphi\left(\cdot; \theta_j\right) \mid \varphi\left(\cdot; \theta_j\right) \in \mathcal{P}, \varpi_j \in \mathbb{R}, j \in [k]\right\}\right),$$

and hence we can apply the following result, paraphrased from Zhang (2003, Thm. II.1).

Lemma 4. *Let κ be a differentiable and convex function on $\mathcal{V}$, and let $\left(\bar{f}_k\right)_{k \in \mathbb{N}}$ be a sequence of functions obtained by Algorithm 1. If*

$$\sup_{p, q \in \mathcal{C}, \pi \in (0,1)} \frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \kappa\left(\left(1-\pi\right) p + \pi q\right) \leq \mathfrak{M} < \infty,$$

then, for each $k \in \mathbb{N}$,

$$\kappa\left(\bar{f}_k\right) - \inf_{p \in \mathcal{C}} \kappa\left(p\right) \leq \frac{2\mathfrak{M}}{k+2}.$$

Algorithm 1 Algorithm for computing a greedy approximation sequence.

```
Input: $\bar{f}_0 \in \mathcal{P}$
for $k \in \mathbb{N}$ do
    Compute $\left(\bar{\pi}_k, \bar{\theta}_k\right) = \operatorname*{arg\,min}_{(\pi,\theta) \in [0,1] \times \Theta} \kappa\left(\left(1-\pi\right) \bar{f}_{k-1} + \pi \varphi\left(\cdot; \theta\right)\right)$
    Define $\bar{f}_k = \left(1 - \bar{\pi}_k\right) \bar{f}_{k-1} + \bar{\pi}_k \varphi\left(\cdot; \bar{\theta}_k\right)$
end for
```

We are interested in two choices for κ:

$$\kappa\left(p\right) = \mathrm{KL}_h\left(f \,||\, p\right) \tag{7}$$

and its sample counterpart,

$$\kappa_n\left(p\right) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f\left(X_i\right) + h\left(X_i\right)}{p\left(X_i\right) + h\left(X_i\right)} + \frac{1}{n} \sum_{i=1}^{n} \log \frac{f\left(Y_i\right) + h\left(Y_i\right)}{p\left(Y_i\right) + h\left(Y_i\right)}, \tag{8}$$

where $\left(X_i\right)_{i \in [n]}$ and $\left(Y_i\right)_{i \in [n]}$ are realisations of X and Y, respectively. For choice (7), by the dominated convergence theorem, we observe that

$$\frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \kappa\left(\left(1-\pi\right) p + \pi q\right)$$

$$= \mathbf{E}_f \left\{ \frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \log \frac{f+h}{(1-\pi)\,p+\pi q+h} \right\} + \mathbf{E}_h \left\{ \frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \log \frac{f+h}{(1-\pi)\,p+\pi q+h} \right\}$$

$$= \mathbf{E}_f \left\{ \frac{(p-q)^2}{[(1-\pi)\,p+\pi q+h]^2} \right\} + \mathbf{E}_h \left\{ \frac{(p-q)^2}{[(1-\pi)\,p+\pi q+h]^2} \right\}.$$

Suppose that each $\varphi(\cdot;\theta) \in \mathcal{P}$ is bounded from above by $c < \infty$. Then, since $p, q \in \mathcal{C}$ are non-negative functions, we have the fact that $(p-q)^2 \le c^2$. If we further have $a \le h$ for some $a > 0$, then $[(1-\pi)\,p+\pi q+h]^2 \ge a^2$, which implies that

$$\frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \kappa\left((1-\pi)\,p+\pi q\right) \le 2 \times \frac{c^2}{a^2}$$

for every $p, q \in \mathcal{C}$ and $\pi \in (0,1)$, and thus

$$\sup_{p,q\in\mathcal{C},\pi\in(0,1)} \frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \kappa\left((1-\pi)\,p+\pi q\right) \le \frac{2c^2}{a^2} < \infty.$$

Similarly, for case (8), we have

$$\frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \kappa_n\left((1-\pi)\,p+\pi q\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \log \frac{f(x_i)+h(x_i)}{(1-\pi)\,p(x_i)+\pi q(x_i)+h(x_i)}$$

$$+ \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \log \frac{f(y_i)+h(y_i)}{(1-\pi)\,p(y_i)+\pi q(y_i)+h(y_i)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{(p(x_i)-q(x_i))^2}{[(1-\pi)\,p(x_i)+\pi q(x_i)+h(x_i)]^2} + \frac{1}{n}\sum_{i=1}^{n} \frac{(p(y_i)-q(y_i))^2}{[(1-\pi)\,p(y_i)+\pi q(y_i)+h(y_i)]^2}.$$

By the same argument, as for κ, we have the fact that $(p(x)-q(x))^2 \le c^2$, for every $p, q \in \mathcal{C}$ and every $x \in \mathcal{X}$, and furthermore $[(1-\pi)\,p(x)+\pi q(x)+h(x)]^2 \ge a^2$, for any $\pi \in (0,1)$. Thus,

$$\sup_{p,q\in\mathcal{C},\pi\in(0,1)} \frac{\mathrm{d}^2}{\mathrm{d}\pi^2} \kappa\left((1-\pi)\,p+\pi q\right) \le \frac{2c^2}{a^2} < \infty,$$

as required. We therefore obtain the following result.

Proposition 5. *Let $\Box$ denote either κ, the KL$_h$ divergence (7), or κ_n, the sample KL$_h$ divergence (8), and assume that $h \ge a$ and $\varphi(\cdot;\theta) \le c$, for each $\theta \in \Theta$. Then,*

$$\Box\left(\bar{f}_k\right) - \inf_{p\in\mathcal{C}} \Box(p) \le \frac{4a^{-2}c^2}{k+2},$$

for each $k \in \mathbb{N}$, where $\left(\bar{f}_k\right)_{k\in\mathbb{N}}$ is obtained as per Algorithm 1.

Notice that sequences $\left(\bar{f}_k\right)_{k\in\mathbb{N}}$ obtained via Lemma 4 are greedy approximation sequences, and that $\bar{f}_k \in \mathcal{C}_k$, for each $k \in \mathbb{N}$. Let $(f_k)_{k\in\mathbb{N}}$ be the sequence of minimizers defined by

$$f_k = \underset{\psi_k\in\mathcal{S}_k\times\Theta^k}{\arg\min}\, \mathrm{KL}_h\left(f \,||\, f_k(\cdot;\psi_k)\right), \tag{9}$$

and let $(f_{k,n})_{k\in\mathbb{N}}$ be the sequence of h-MLLEs, as per (4). Then, by definition, we have the fact that $\kappa(f_k) \le \kappa\left(\bar{f}_k\right)$ and $\kappa(f_{k,n}) \le \kappa\left(\bar{f}_k\right)$, for κ set as (7) or (8), respectively. Thus, we have the following result.

Proposition 6. *For the KL$_h$ divergence (7), under the assumption that $h \ge a$ and $\varphi(\cdot;\theta) \le c$, for each $\theta \in \Theta$, we have*

$$\kappa(f_k) - \inf_{p\in\mathcal{C}} \kappa(p) \le \frac{4a^{-2}c^2}{k+2} \tag{10}$$

for each $k \in \mathbb{N}$, where $(f_k)_{k\in\mathbb{N}}$ is the sequence of minimizers defined via (9)*. Furthermore, for the sample KL$_h$ divergence* (8)*, under the same assumptions as above, we have*

$$\kappa_n\left(f_{k,n}\right) - \inf_{p\in\mathcal{C}} \kappa_n\left(p\right) \leq \frac{4a^{-2}c^2}{k+2}, \tag{11}$$

for each $k \in \mathbb{N}$, where $(f_{k,n})_{k\in\mathbb{N}}$ are h-MLLEs defined via (4)*.*

As is common in many statistical learning/uniform convergence results (e.g., Bartlett & Mendelson, 2002, Koltchinskii & Panchenko, 2004), we employ the use of Rademacher processes and associated bounds. Let $(\varepsilon_i)_{i\in[n]}$ be i.i.d. Rademacher random variables, that is $\mathbf{P}(\varepsilon_i = -1) = \mathbf{P}(\varepsilon_i = 1) = 1/2$, that are independent of $(X_i)_{i\in[n]}$. The Rademacher process, indexed by a class of real measurable functions $\mathcal{S}$, is defined as the quantity

$$R_n(s) = \frac{1}{n}\sum_{i=1}^{n} s(X_i)\varepsilon_i,$$

for $s \in \mathcal{S}$. The Rademacher complexity of the class $\mathcal{S}$ is given by $\mathcal{R}_n(\mathcal{S}) = \mathbf{E}\sup_{s\in\mathcal{S}}|R_n(s)|$.

In the subsequent section, we make use of the following result regarding the supremum of convex functions:

Lemma 7 (Rockafellar, 1997, Thm. 32.2)**.** *Let η be a convex function on a linear space $\mathcal{T}$, and let $\mathcal{S} \subset \mathcal{T}$ be an arbitrary subset. Then,*

$$\sup_{p\in\mathcal{S}} \eta\left(p\right) = \sup_{p\in\mathrm{co}(\mathcal{S})} \eta\left(p\right).$$

In particular, we use the fact that since a linear functional of convex combinations achieves its maximum value at vertices, the Rademacher complexity of $\mathcal{S}$ is equal to the Rademacher complexity of co($\mathcal{S}$) (see Lemma 19). We consequently obtain the following result.

Lemma 8. *Let $(\varepsilon_i)_{i\in[n]}$ be i.i.d. Rademacher random variables, independent of $(X_i)_{i\in[n]}$ and $\mathcal{P}$ be defined as above. The sets $\mathcal{C}$ and $\mathcal{P}$ will have equal complexity, $\mathcal{R}_n(\mathcal{C}) = \mathcal{R}_n(\mathcal{P})$, and the supremum of the Rademacher process indexed by $\mathcal{C}$ is equal to the supremum on the basis functions of $\mathcal{P}$:*

$$\mathbf{E}_\varepsilon \sup_{g\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^{n} g(X_i)\varepsilon_i\right| = \mathbf{E}_\varepsilon \sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i;\theta)\varepsilon_i\right|.$$

Proof. Follows immediately from Lemma 7. ■

4 Main results

Here we provide explicit statements regarding the convergence rates claimed in (5). We assume that f is bounded above by some constant c and that the lifting function h is bounded above and below by constants a and b, respectively.

Theorem 9. *Let h be a positive density satisfying $0 < a \leq h(x) \leq b$, for all $x \in \mathcal{X}$. For any target density f satisfying $0 \leq f(x) \leq c$, for all $x \in \mathcal{X}$ and where $f_{k,n}$ is the minimizer of* KL_h *over k-component mixtures, the following inequality holds:*

$$\mathbf{E}\left\{\mathrm{KL}_h\left(f \,||\, f_{k,n}\right)\right\} - \mathrm{KL}_h\left(f \,||\, \mathcal{C}\right) \leq \frac{u_1}{k+2} + \frac{u_2}{\sqrt{n}}\int_0^c \log^{1/2} N(\mathcal{P}, \varepsilon/2, \|\cdot\|_\infty)\mathrm{d}\varepsilon + \frac{u_3}{\sqrt{n}},$$

where u_1, u_2, and u_3 are positive constants that depend on some or all of a, b, and c.

Corollary 10. *Let $\mathcal{X}$ and Θ be compact and assume the following Lipschitz condition holds: for each $x \in \mathcal{X}$, and for each $\theta, \tau \in \Theta$,*

$$\left|\varphi\left(x;\theta\right) - \varphi\left(x;\tau\right)\right| \leq \Phi\left(x\right)\left\|\theta - \tau\right\|_1, \tag{12}$$

for some function $\varPhi : \mathcal{X} \to \mathbb{R}_{\geq 0}$, where $\|\varPhi\|_\infty = \sup_{x\in\mathcal{X}}|\varPhi(x)| < \infty$. Then the bound in Theorem 9 becomes

$$\mathbf{E}\left\{\mathrm{KL}_h\left(f\,||\,f_{k,n}\right)\right\} - \mathrm{KL}_h\left(f\,||\,\mathcal{C}\right) \leq \frac{c_1}{k+2} + \frac{c_2}{\sqrt{n}},$$

where c_1 and c_2 are positive constants.

5 Proofs

We first present a result establishing a uniform concentration bound for the h-lifted log-likelihood ratios, which is instrumental in the proof of Theorem 9. Our proofs broadly follow the structure of Rakhlin et al. (2005), modified as needed for the use of KL_h.

Assume that $0 \leq \varphi(\cdot;\theta) < c$ for some $c \in \mathbb{R}_{>0}$. For brevity, we adopt the notation: $\|T(g)\|_{\mathcal{C}} = \sup_{g\in\mathcal{C}}|T(g)|$.

Theorem 11. *Let $X_1, \ldots, X_n$ be an i.i.d. sample of size n drawn from a fixed density f such that $0 \leq f(x) \leq c$ for all $x \in \mathcal{X}$, and let h be a positive density with $0 < a \leq h(x) \leq b$ for all $x \in \mathcal{X}$. Then with probability at least $1 - \mathrm{e}^{-t}$,*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\log\frac{g(X_i)+h(X_i)}{f(X_i)+h(X_i)} - \mathbf{E}_f\log\frac{g+h}{f+h}\right\|_{\mathcal{C}} \leq \frac{w_1}{\sqrt{n}}\,\mathbf{E}\int_0^c \log^{1/2} N(\mathcal{P},\varepsilon,d_{n,x})\mathrm{d}\varepsilon + \frac{w_2}{\sqrt{n}} + w_3\sqrt{\frac{t}{n}},$$

where w_1, w_2, and w_3 are constants that each depend on some or all of a, b, and c, and $N(\mathcal{P},\varepsilon,d_{n,x})$ is the ε-covering number of $\mathcal{P}$ with respect to the following empirical L_2 metric

$$d^2_{n,x}(\varphi_1,\varphi_2) = \frac{1}{n}\sum_{i=1}^{n}(\varphi_1(X_i) - \varphi_2(X_i))^2.$$

Remark 12. *The bound on the term*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\log\frac{g(Y_i)+h(Y_i)}{f(Y_i)+h(Y_i)} - \mathbf{E}_h\log\frac{g+h}{f+h}\right\|_{\mathcal{C}}$$

is the same as the above, except where the empirical distance $d_{n,x}$ is replaced by $d_{n,y}$, defined in the same way as $d_{n,x}$ but with Y_i replacing X_i.

Proof. [of Theorem 11]. Fix h and define the following quantities: $\tilde{g} = g + h$, $\tilde{f} = f + h$, $\tilde{\mathcal{C}} = \mathcal{C} + h$,

$$m_i = \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}, \quad m_i' = \log\frac{\tilde{g}(X_i')}{\tilde{f}(X_i')}, \quad Z(x_1,\ldots,x_n) = \left\|\frac{1}{n}\sum_{i=1}^{n}\log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)} - \mathbf{E}\log\frac{\tilde{g}}{\tilde{f}}\right\|_{\tilde{\mathcal{C}}}.$$

We first apply McDiarmid's inequality (Lemma 21) to the random variable Z. The bound on the martingale difference is given by

$$\begin{aligned}|Z(X_1,\ldots,X_i,\ldots,X_n) - Z(X_1,\ldots,X_i',\ldots,X_n)| &= \Bigg|\left\|\mathbf{E}\log\frac{\tilde{g}}{\tilde{f}} - \frac{1}{n}(m_1 + ... + m_i + ... + m_n)\right\|_{\tilde{\mathcal{C}}} \\ &\quad - \left\|\mathbf{E}\log\frac{\tilde{g}}{\tilde{f}} - \frac{1}{n}(m_1 + ... + m_i' + ... + m_n)\right\|_{\tilde{\mathcal{C}}}\Bigg| \\ &\leq \frac{1}{n}\left\|\log\frac{\tilde{g}(X_i')}{\tilde{f}(X_i')} - \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}\right\|_{\tilde{\mathcal{C}}} \\ &\leq \frac{1}{n}\left(\log\frac{c+b}{a} - \log\frac{a}{c+b}\right) \\ &= \frac{1}{n}2\log\frac{c+b}{a} = c_i.\end{aligned}$$

The chain of inequalities holds because of the triangle inequality and the properties of the supremum. By Lemma 21, we have

$$\mathbf{P}(Z - \mathbf{E}\, Z > \varepsilon) \leq \exp\left\{ -\frac{n\varepsilon^2}{(\sqrt{2}\log\frac{c+b}{a})^2} \right\},$$

so

$$\mathbf{P}(Z \leq \varepsilon + \mathbf{E}\, Z) \geq 1 - \exp\left\{ -\frac{n\varepsilon^2}{(\sqrt{2}\log\frac{c+b}{a})^2} \right\},$$

where it follows from $t = n\varepsilon^2/(\sqrt{2}\log\frac{c+b}{a})^2$ that $\varepsilon = \sqrt{2}\log\left(\frac{c+b}{a}\right)\sqrt{\frac{t}{n}}$. Therefore with probability at least $1 - \mathrm{e}^{-t}$,

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)} - \mathbf{E}_f \log\frac{\tilde{g}}{\tilde{f}} \right\|_{\tilde{\mathcal{C}}} \leq \mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)} - \mathbf{E}_f \log\frac{\tilde{g}}{\tilde{f}} \right\|_{\tilde{\mathcal{C}}} + \sqrt{2}\log\left(\frac{c+b}{a}\right)\sqrt{\frac{t}{n}}.$$

Let $(\varepsilon_i)_{i\in[n]}$ be i.i.d. Rademacher random variables, independent of $(X_i)_{i\in[n]}$. By Lemma 22,

$$\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)} - \mathbf{E}_f \log\frac{\tilde{g}}{\tilde{f}} \right\|_{\tilde{\mathcal{C}}} \leq 2\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}\varepsilon_i \right\|_{\tilde{\mathcal{C}}}.$$

By combining the results above, the following inequality holds with probability at least $1 - \mathrm{e}^{-t}$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)} - \mathbf{E}_f \log\frac{\tilde{g}}{\tilde{f}} \right\|_{\tilde{\mathcal{C}}} \leq 2\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}\varepsilon_i \right\|_{\tilde{\mathcal{C}}} + \sqrt{2}\log\left(\frac{c+b}{a}\right)\sqrt{\frac{t}{n}}.$$

Now let $p_i = \frac{\tilde{g}(X_i)}{\tilde{f}(X_i)} - 1$, such that $\frac{a}{c+b} \leq p_i + 1 \leq \frac{c+b}{a}$ holds for all $i \in [n]$. Additionally, let $\eta(p) = \log(p+1)$ so that $\eta(0) = 0$ and note that for $p \in \left[\frac{a}{c+b} - 1, \frac{c+b}{a} - 1\right]$, the derivative of $\eta(p)$ is maximal at $p^* = \frac{a}{c+b} - 1$, and equal to $\eta^{'}(p^*) = (c+b)/a$. Therefore, $\frac{a}{b+c}\log(p+1)$ is 1-Lipschitz. By Lemma 20 applied to $\eta(p)$,

$$\begin{aligned}
2\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}\varepsilon_i \right\|_{\tilde{\mathcal{C}}} &= 2\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \eta(p_i)\varepsilon_i \right\|_{\tilde{\mathcal{C}}} \\
&\leq \frac{4(c+b)}{a}\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}\varepsilon_i - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \right\|_{\tilde{\mathcal{C}}} \\
&\leq \frac{4(c+b)}{a}\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}\varepsilon_i \right\|_{\tilde{\mathcal{C}}} + \frac{4(c+b)}{a}\mathbf{E}_\varepsilon\left| \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \right| \\
&\leq \frac{4(c+b)}{a}\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}\varepsilon_i \right\|_{\tilde{\mathcal{C}}} + \frac{4(c+b)}{a}\frac{1}{\sqrt{n}},
\end{aligned}$$

where the final inequality follows from the following result, proved in Haagerup (1981):

$$\mathbf{E}_\varepsilon\left| \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \right| \leq \left(\mathbf{E}_\varepsilon\left\{ \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \right\}^2 \right)^{1/2} = \frac{1}{\sqrt{n}}.$$

Now, let $\xi_i(\tilde{g}_i) = a \cdot \tilde{g}(X_i)/\tilde{f}(X_i)$, and note that

$$|\xi_i(u_i) - \xi_i(v_i)| = \frac{a}{|\tilde{f}(X_i)|}|u(X_i) - v(X_i)| \leq |u(X_i) - v(X_i)|.$$

By again applying Lemma 20, we have

$$\frac{4(c+b)}{a}\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{g}(X_i)}{\tilde{f}(X_i)}\varepsilon_i \right\|_{\tilde{\mathcal{C}}} \leq \frac{8(c+b)}{a^2}\,\mathbf{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \tilde{g}(X_i)\varepsilon_i \right\|_{\tilde{\mathcal{C}}}$$

$$\leq \frac{8(c+b)}{a^2}\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n} g(X_i)\varepsilon_i\right\|_{\mathcal{C}} + \frac{8(c+b)}{a^2}\mathbf{E}\left|\frac{1}{n}\sum_{i=1}^{n} h(X_i)\varepsilon_i\right|$$
$$\leq \frac{8(c+b)}{a^2}\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n} g(X_i)\varepsilon_i\right\|_{\mathcal{C}} + \frac{8(c+b)}{a^2}\frac{b}{\sqrt{n}}.$$

Applying Lemmas 8 and 23, the following inequality holds for some constant $K > 0$:

$$\mathbf{E}_\varepsilon \sup_{g\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i)\varepsilon_i\right| = \mathbf{E}_\varepsilon \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i;\theta)\varepsilon_i\right| \leq \frac{K}{\sqrt{n}}\mathbf{E}\int_0^c \log^{1/2} N(\mathcal{P},\varepsilon,d_{n,x})\mathrm{d}\varepsilon, \tag{13}$$

and combining the results together, the following inequality holds with probability at least $1-\mathrm{e}^{-t}$:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\log\frac{\tilde{g}(X_i)}{\tilde{f}(X_i)} - \mathbf{E}_f\log\frac{\tilde{g}}{\tilde{f}}\right\|$$
$$\leq \frac{8(c+b)K}{a^2\sqrt{n}}\mathbf{E}\int_0^c \log^{1/2} N(\mathcal{P},\varepsilon,d_{n,x})\mathrm{d}\varepsilon + \frac{(8b+4a)(c+b)}{a^2\sqrt{n}} + \sqrt{2}\log\left(\frac{c+b}{a}\right)\sqrt{\frac{t}{n}},$$
$$= \frac{w_1}{\sqrt{n}}\mathbf{E}\int_0^c \log^{1/2} N(\mathcal{P},\varepsilon,d_{n,x})\mathrm{d}\varepsilon + \frac{w_2}{\sqrt{n}} + w_3\sqrt{\frac{t}{n}},$$

where w_1, w_2, and w_3 are constants that each depend on some or all of a, b, and c. ■

Remark 13. *From Lemma 23 we have that $\sigma_n^2 := \sup_{f\in\mathscr{F}} P_n f^2$. To make explicit why $2\sigma_n = \left(\sup_{g\in\mathcal{C}} P_n g^2\right)^{1/2} = 2c$, let $\mathscr{F} = \mathcal{C}$ and observe*

$$\sigma_n^2 = \sup_{g\in\mathcal{C}} P_n g^2 = \sup_{g\in\mathcal{C}}\frac{1}{n}\sum_{i=1}^{n} g(X_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n} c^2 = c^2.$$

Since our basis functions $\varphi(\cdot,\theta)$ are bounded by c, everything greater than c will have value 0 and hence the change from 2c to c is inconsequential. However, it can also be motivated by the fact that $\varphi(\cdot,\theta)$ are positive functions.

As highlighted in Remark 12, the full result of Theorem 11 relies on the empirical L_2 distances $d_{n,x}$ and $d_{n,y}$. In the result of Theorem 9, we make use of the following result to bound $d_{n,x}$ and $d_{n,y}$.

Proposition 14. *By combining Lemmas 16 and 17, the following inequality holds:*

$$\log N(\mathcal{P},\varepsilon,\|\cdot\|) \leq \log N_{[]}(\mathcal{P},\varepsilon,\|\cdot\|) \leq \log N(\mathcal{P},\varepsilon/2,\|\cdot\|_\infty),$$

where $N_{[]}(\mathcal{P},\varepsilon,\|\cdot\|)$ is the ε-bracketing number of $\mathcal{P}$. Therefore, we have that

$$\log N(\mathcal{P},\varepsilon,d_{n,x}) \leq \log N(\mathcal{P},\varepsilon/2,\|\cdot\|_\infty),$$

and

$$\log N(\mathcal{P},\varepsilon,d_{n,y}) \leq \log N(\mathcal{P},\varepsilon/2,\|\cdot\|_\infty).$$

With this result, we can now prove Theorem 9.

Proof. [of Theorem 9] The notation is the same as in the proof of Theorem 11. The values of the constants may change from line to line.

$$\mathrm{KL}_h\left(f\,||\,f_{k,n}\right) - \mathrm{KL}_h\left(f\,||\,f_k\right)$$
$$= \mathbf{E}_f\log\frac{\tilde{f}}{\tilde{f}_{k,n}} + \mathbf{E}_h\log\frac{\tilde{f}}{\tilde{f}_{k,n}} - \mathbf{E}_f\log\frac{\tilde{f}}{\tilde{f}_k} - \mathbf{E}_h\log\frac{\tilde{f}}{\tilde{f}_k}$$

$$
\begin{aligned}
&= \mathbf{E}_f \log \frac{\tilde f}{\tilde f_{k,n}} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_{k,n}(X_i)} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_{k,n}(X_i)} \\
&\quad + \mathbf{E}_h \log \frac{\tilde f}{\tilde f_{k,n}} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_{k,n}(Y_i)} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_{k,n}(Y_i)} \\
&\quad - \mathbf{E}_f \log \frac{\tilde f}{\tilde f_k} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_k(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_k(X_i)} \\
&\quad - \mathbf{E}_h \log \frac{\tilde f}{\tilde f_k} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_k(Y_i)} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_k(Y_i)} \\
&= \left(\mathbf{E}_f \log \frac{\tilde f}{\tilde f_{k,n}} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_{k,n}(X_i)}\right) + \left(\mathbf{E}_h \log \frac{\tilde f}{\tilde f_{k,n}} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_{k,n}(Y_i)}\right) \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_k(X_i)} - \mathbf{E}_f \log \frac{\tilde f}{\tilde f_k}\right) + \left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_k(Y_i)} - \mathbf{E}_h \log \frac{\tilde f}{\tilde f_k}\right) \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_{k,n}(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_k(X_i)}\right) + \left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_{k,n}(Y_i)} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_k(Y_i)}\right) \\
&\le 2 \sup_{\tilde g \in \tilde{\mathcal{C}}} \left| \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde g(X_i)}{\tilde f(X_i)} - \mathbf{E}_f \log \frac{\tilde g}{\tilde f} \right| + 2 \sup_{\tilde g \in \tilde{\mathcal{C}}} \left| \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde g(Y_i)}{\tilde f(Y_i)} - \mathbf{E}_h \log \frac{\tilde g}{\tilde f} \right| \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_{k,n}(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(X_i)}{\tilde f_k(X_i)}\right) + \left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_{k,n}(Y_i)} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f(Y_i)}{\tilde f_k(Y_i)}\right) \\
&\le 2\,\mathbf{E}\left\{ \frac{w_1^x}{\sqrt{n}} \int_0^c \log^{1/2} N(\mathcal{P}, \varepsilon, d_{n,x}) \mathrm{d}\varepsilon \right\} + \frac{w_2^x}{\sqrt{n}} + w_3^x \sqrt{\frac{t}{n}} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f_k(X_i)}{\tilde f_{k,n}(X_i)} \\
&\quad + 2\,\mathbf{E}\left\{ \frac{w_1^y}{\sqrt{n}} \int_0^c \log^{1/2} N(\mathcal{P}, \varepsilon, d_{n,y}) \mathrm{d}\varepsilon \right\} + \frac{w_2^y}{\sqrt{n}} + w_3^y \sqrt{\frac{t}{n}} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f_k(Y_i)}{\tilde f_{k,n}(Y_i)} \\
&\le \frac{w_1}{\sqrt{n}} \int_0^c \log^{1/2} N(\mathcal{P}, \varepsilon/2, \|\cdot\|_\infty) \mathrm{d}\varepsilon + \frac{w_2}{\sqrt{n}} + w_3 \sqrt{\frac{t}{n}} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f_k(X_i)}{\tilde f_{k,n}(X_i)} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f_k(Y_i)}{\tilde f_{k,n}(Y_i)},
\end{aligned}
$$

with probability at least $1 - \mathrm{e}^{-t}$, by Theorem 11. Now, we can use (11) from Proposition 6 applied to the target density f_k, obtaining the following:

$$
\begin{aligned}
\mathrm{KL}_h\left(f_k \,||\, f_{k,n}\right) &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f_k(X_i)}{\tilde f_{k,n}(X_i)} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\tilde f_k(Y_i)}{\tilde f_{k,n}(Y_i)} \\
&\le \frac{4a^{-2}c^2}{k+2} + \inf_{p \in \mathcal{C}} \mathrm{KL}_h\left(f_k \,||\, p\right).
\end{aligned}
$$

Since by definition we have that $f_k \in \mathcal{C}$, $\inf_{p\in\mathcal{C}} \mathrm{KL}_h\left(f_k \,||\, p\right) = 0$, and so with probability at least $1 - \mathrm{e}^{-t}$ we have:

$$
\begin{aligned}
&\mathrm{KL}_h\left(f \,||\, f_{k,n}\right) - \mathrm{KL}_h\left(f \,||\, f_k\right) \\
&\qquad\qquad \le \frac{w_1}{\sqrt{n}} \int_0^c \log^{1/2} N(\mathcal{P}, \varepsilon/2, \|\cdot\|_\infty) \mathrm{d}\varepsilon + \frac{w_2}{\sqrt{n}} + w_3 \sqrt{\frac{t}{n}} + \frac{w_4}{k+2}.
\end{aligned} \tag{14}
$$

We can write the overall error as the sum of the approximation and estimation errors as follows. The former is bounded by (10), and the latter is bounded as above in (14). Therefore, with probability at least $1 - \mathrm{e}^{-t}$,

$$
\mathrm{KL}_h\left(f \,||\, f_{k,n}\right) - \mathrm{KL}_h\left(f \,||\, \mathcal{C}\right) = \left[\mathrm{KL}_h\left(f \,||\, f_k\right) - \mathrm{KL}_h\left(f \,||\, \mathcal{C}\right)\right] + \left[\mathrm{KL}_h\left(f \,||\, f_{k,n}\right) - \mathrm{KL}_h\left(f \,||\, f_k\right)\right]
$$

$$\leq \frac{w_4}{k+2} + \frac{w_1}{\sqrt{n}} \int_0^c \log^{1/2} N(\mathcal{P}, \varepsilon/2, \|\cdot\|_\infty) \mathrm{d}\varepsilon + \frac{w_2}{\sqrt{n}} + w_3 \sqrt{\frac{t}{n}}.$$

As in Rakhlin et al. (2005), we rewrite the above probabilistic statement as a statement in terms of expectations. To this end, let

$$\mathcal{A} \coloneqq \frac{w_4}{k+2} + \frac{w_1}{\sqrt{n}} \int_0^c \log^{1/2} N(\mathcal{P}, \varepsilon/2, \|\cdot\|_\infty) \mathrm{d}\varepsilon + \frac{w_2}{\sqrt{n}},$$

and $\mathcal{Z} \coloneqq \mathrm{KL}_h\left(f \,||\, f_{k,n}\right) - \mathrm{KL}_h\left(f \,||\, \mathcal{C}\right)$. We have shown $\mathbf{P}\left(\mathcal{Z} \geq \mathcal{A} + w_3\sqrt{\frac{t}{n}}\right) \leq \mathrm{e}^{-t}$. Since $\mathcal{Z} \geq 0$,

$$\mathbf{E}\{\mathcal{Z}\} = \int_0^{\mathcal{A}} \mathbf{P}(\mathcal{Z} > s)\mathrm{d}s + \int_{\mathcal{A}}^\infty \mathbf{P}(\mathcal{Z} > s)\mathrm{d}s \leq \mathcal{A} + \int_0^\infty \mathbf{P}(\mathcal{Z} > \mathcal{A} + s)\mathrm{d}s.$$

Setting $s = w_3\sqrt{\frac{t}{n}}$, we have $t = w_5 n s^2$ and $\mathbf{E}\{\mathcal{Z}\} \leq \mathcal{A} + \int_0^\infty e^{-w_5 n s^2}\mathrm{d}s \leq \mathcal{A} + \frac{w}{\sqrt{n}}$. Hence,

$$\mathbf{E}\left\{\mathrm{KL}_h\left(f \,||\, f_{k,n}\right)\right\} - \mathrm{KL}_h\left(f \,||\, \mathcal{C}\right) \leq \frac{c_1}{k+2} + \frac{c_2}{\sqrt{n}} \int_0^c \log^{1/2} N\left(\mathcal{P}, \varepsilon/2, \|\cdot\|_\infty\right) \mathrm{d}\varepsilon + \frac{c_3}{\sqrt{n}},$$

where c_1, c_2, and c_3 are constants that depend on some or all of a, b, and c. ■

Remark 15. *The approximation error characterises the suitability of the class $\mathcal{C}$, i.e., how well functions in $\mathcal{C}$ are able to estimate a target f which does not necessarily lie in $\mathcal{C}$. The estimation error characterises the error arising from the estimation of the target f on the basis of the finite sample of size n.*

Proof. [of Corollary 10] Let $\mathcal{X}$ and Θ be compact and assume the Lipshitz condition given in (12). If $\varphi(x;\cdot)$ is continuously differentiable, then

$$\begin{aligned}\left|\varphi\left(x;\theta\right) - \varphi\left(x;\tau\right)\right| &\leq \sum_{k=1}^d \left|\frac{\partial \varphi\left(x;\cdot\right)}{\partial \theta_k}\left(\theta_k^*\right)\right| \left|\theta_k - \tau_k\right| \\ &\leq \sup_{\theta^* \in \Theta} \left\|\frac{\partial \varphi\left(x;\cdot\right)}{\partial \theta}\left(\theta^*\right)\right\|_1 \left\|\theta - \tau\right\|_1.\end{aligned}$$

Setting

$$\Phi\left(x\right) = \sup_{\theta^* \in \Theta} \left\|\frac{\partial \varphi\left(x;\cdot\right)}{\partial \theta}\left(\theta^*\right)\right\|_1,$$

we have $\|\Phi\|_\infty < \infty$. From Lemma 18, we obtain the fact that

$$\log N_{[]}\left(\mathcal{P}, 2\varepsilon\|\Phi\|_\infty, \|\cdot\|_\infty\right) \leq \log N\left(\Theta, \varepsilon, \|\cdot\|_\infty\right),$$

which by the change of variable $\delta = 2\varepsilon\|\Phi\|_\infty \implies \varepsilon = \delta/2\|\Phi\|_\infty$ implies

$$\log N_{[]}\left(\mathcal{P}, \varepsilon/2, \|\cdot\|_\infty\right) \leq \log N\left(\Theta, \frac{\varepsilon}{4\|\Phi\|_\infty}, \|\cdot\|_1\right).$$

Since $\Theta \subset \mathbb{R}^d$, using the fact that a Euclidean set of radius r has covering number

$$N\left(r, \varepsilon\right) \leq \left(\frac{3r}{\varepsilon}\right)^d,$$

we have

$$\log N\left(\Theta, \frac{\varepsilon}{4\|\Phi\|_\infty}, \|\cdot\|_1\right) \leq d \log\left[\frac{12\|\Phi\|_\infty \mathrm{diam}\left(\Theta\right)}{\varepsilon}\right].$$

So

$$\int_0^c \sqrt{\log N\left(\Theta, \frac{\varepsilon}{4\|\Phi\|_\infty}, \|\cdot\|_1\right)} \mathrm{d}\varepsilon \leq \int_0^c \sqrt{d \log\left[\frac{12\|\Phi\|_\infty \mathrm{diam}\left(\Theta\right)}{\varepsilon}\right]} \mathrm{d}\varepsilon,$$

and since $c < \infty$, the integral is finite, as required. ■

6 Numerical experiments

In this section, we discuss the computability and computation of KL_h estimation problems and provide empirical evidence towards the rates obtained in Theorem 9. Namely, we seek to develop a methodology for computing h-MLLEs, and to use numerical experiments to demonstrate that the sequence of expected h-lifted KL divergences between some density f and a sequence of k-component mixture densities from a suitable class $\mathcal{P}$, estimated using n observations does indeed decrease at rates proportional to $1/k$ and $1/\sqrt{n}$, as k and n increase.

6.1 Minorization–Maximization algorithm

One solution for computing problems of kind (4) is to employ an MM algorithm. To do so, we first write the objective of (4) as

$$L_{h,n}\left(\psi_k\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\log\left\{\sum_{j=1}^{k}\pi_j\varphi\left(X_i;\theta_j\right)+h\left(X_i\right)\right\}+\log\left\{\sum_{j=1}^{k}\pi_j\varphi\left(Y_i;\theta_j\right)+h\left(Y_i\right)\right\}\right),$$

where $\psi_k\in\Psi_k=\mathcal{S}_k\times\Theta^k$. We then require the definition of a minorizer Q_n for $L_{h,n}$ on the space Ψ_k, where $Q_n:\Psi_k\times\Psi_k\to\mathbb{R}$ is a function with the properties:

(i) $Q_n\left(\psi_k,\psi_k\right)=L_{h,n}\left(\psi_k\right)$, and

(ii) $Q_n\left(\psi_k,\chi_k\right)\leq L_{h,n}\left(\psi_k\right)$,

for each $\psi_k,\chi_k\in\Psi_k$. In this context, given a fixed χ_k, the minorizer $Q_n\left(\cdot,\chi_k\right)$ should possess properties that simplify it compared to the original objective $L_{h,n}$. These properties should make the minorizer more tractable and might include features such as parametric separability, differentiability, convexity, among others.

In order to build an appropriate minorizer for $L_{h,n}$, we make use of the so-called Jensen's inequality minorizer, as detailed in Lange (2016, Sec. 4.3), applied to the logarithm function. This construction results in a minorizer of the form

$$\begin{aligned}Q_n\left(\psi_k,\chi_k\right) \;=\;& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\left\{\tau_j\left(X_i;\chi_k\right)\log\pi_j+\tau_j\left(X_i;\chi_k\right)\log\varphi\left(X_i;\theta_j\right)\right\}\\ &+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\left\{\tau_j\left(Y_i;\chi_k\right)\log\pi_j+\tau_j\left(Y_i;\chi_k\right)\log\varphi\left(Y_i;\theta_j\right)\right\}\\ &+\frac{1}{n}\sum_{i=1}^{n}\left\{\gamma\left(X_i;\chi_k\right)\log h\left(X_i\right)+\gamma\left(Y_i;\chi_k\right)\log h\left(Y_i\right)\right\}\\ &-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\left\{\tau_j\left(X_i;\chi_k\right)\log\tau_j\left(X_i;\chi_k\right)+\tau_j\left(Y_i;\chi_k\right)\log\tau_j\left(Y_i;\chi_k\right)\right\}\\ &-\frac{1}{n}\sum_{i=1}^{n}\left\{\gamma\left(X_i;\chi_k\right)\log\gamma\left(X_i;\chi_k\right)+\gamma\left(Y_i;\chi_k\right)\log\gamma\left(Y_i;\chi_k\right)\right\}\end{aligned}$$

where

$$\gamma\left(X_i;\psi_k\right)=h\left(X_i\right)/\left\{\sum_{j=1}^{k}\pi_j\varphi\left(X_i;\theta_j\right)+h\left(X_i\right)\right\}$$

and

$$\tau_j\left(X_i;\psi_k\right)=\pi_j\varphi\left(X_i;\theta_j\right)/\left\{\sum_{j=1}^{k}\pi_j\varphi\left(X_i;\theta_j\right)+h\left(X_i\right)\right\}.$$

Observe that $Q_n(\cdot, \chi_k)$ now takes the form of a sum-of-logarithms, as opposed to the more challenging log-of-sum form of $L_{h,n}$. This change produces a functional separation of the elements of ψ_k.

Using Q_n, we then define the MM algorithm via the parameter sequence $\left(\psi_k^{(s)}\right)_{s\in\mathbb{N}}$, where

$$\psi_k^{(s)} = \underset{\psi_k \in \Psi_k}{\arg\max}\ Q_n\left(\psi_k, \psi_k^{(s-1)}\right), \tag{15}$$

for each $s > 0$, and where $\psi_k^{(0)}$ is user chosen and is typically referred to as the initialization of the algorithm. Notice that for each s, (15) is a simpler optimization problem than (4). Writing $\psi_k^{(s)} = \left(\pi_1^{(s)}, \dots, \pi_k^{(s)}, \theta_1^{(s)}, \dots, \theta_k^{(s)}\right)$, we observe that (15) simplifies to the separated expressions:

$$\pi_j^{(s)} = \frac{\sum_{i=1}^n \left\{\tau_j\left(X_i; \psi_k^{(s-1)}\right) + \tau_j\left(Y_i; \psi_k^{(s-1)}\right)\right\}}{\sum_{i=1}^n \sum_{l=1}^k \left\{\tau_l\left(X_i; \psi_k^{(s-1)}\right) + \tau_l\left(Y_i; \psi_k^{(s-1)}\right)\right\}}$$

and

$$\theta_j^{(s)} = \underset{\theta_j \in \Theta}{\arg\max}\ \frac{1}{n}\sum_{i=1}^n \left\{\tau_j\left(X_i; \psi_k^{(s-1)}\right)\log\varphi\left(X_i; \theta_j\right) + \tau_j\left(Y_i; \psi_k^{(s-1)}\right)\log\varphi\left(Y_i; \theta_j\right)\right\},$$

for each $j \in [k]$.

A noteworthy property of the MM sequence $\left(\psi_k^{(s)}\right)_{s\in\mathbb{N}}$ is that it generates an increasing sequence of objective values, due to the chain of inequalities

$$L_{h,n}\left(\psi_k^{(s-1)}\right) = Q_n\left(\psi_k^{(s-1)}, \psi_k^{(s-1)}\right) \le Q_n\left(\psi_k^{(s)}, \psi_k^{(s-1)}\right) \le L_{h,n}\left(\psi_k^{(s)}\right),$$

where the equality is due to property (i) of Q_n, the first in equality is due to the definition of $\psi_k^{(s)}$, and the second inequality is due to property (ii) of Q_n. This provides a kind of stability and regularity to the sequence $\left(L_{h,n}\left(\psi_k^{(s)}\right)\right)_{s\in\mathbb{N}}$.

Of course, we can provide stronger guarantees under additional assumptions. Namely, assume that (iii) $\Psi_k \subset \mathit{\Psi}_k$, where $\mathit{\Psi}_k$ is an open set in a finite dimensional Euclidean space on which $L_{h,n}$ and $Q_n(\cdot, \chi_k)$ is differentiable, for each $\chi_k \in \Psi_k$. Then, under assumptions (i)–(iii) regarding $L_{h,n}$ and Q_n, and due to the compactness of Ψ_k and the continuity of Q_n on $\Psi_k \times \Psi_k$, Razaviyayn et al. (2013, Cor. 1) implies that $\left(\psi_k^{(s)}\right)_{s\in\mathbb{N}}$ converges to the set of stationary points of $L_{h,n}$ in the sense that

$$\lim_{s\to\infty} \inf_{\psi_k^* \in \Psi_k^*} \left\|\psi_k^{(s)} - \psi_k^*\right\|_2 = 0$$

where,

$$\Psi_k^* = \left\{\psi_k^* \in \Psi_k : \left.\frac{\partial L_{h,n}}{\partial \psi_k}\right|_{\psi_k = \psi_k^*} = 0\right\}.$$

More concisely, we say that the sequence $\left(\psi_k^{(s)}\right)_{s\in\mathbb{N}}$ globally converges to the set of stationary points Ψ_k^*.

6.2 Experimental setup

Towards the task of demonstrating empirical evidence of the rates in Theorem 9, we consider the family of beta distributions on the unit interval $\mathcal{X} = [0,1]$ as our base class (i.e., (6)) to estimate a pair of target densities

$$f_1(x) = \frac{1}{2}\chi_{[0,2/5]}(x) + \frac{1}{2}\chi_{[3/5,1]}(x),$$

and

$$f_2(x) = \chi_{[0,1]}(x) \begin{cases} 2 - 4x & \text{if } x \le 1/2, \\ -2 + 4x & \text{if } x > 1/2, \end{cases}$$

where $\chi_{\mathcal{A}}$ is the characteristic function that takes value 1 if $x \in \mathcal{A}$ and 0, otherwise. Note that neither f_1 nor f_2 are in $\mathcal{C}$. In particular, $f_1(x) = 0$ when $x \in (2/5, 3/5)$, and $f_2(x) = 0$ when $x = 1/2$, and hence neither densities are bounded away from 0, on $\mathcal{X}$. Thus, the theory of Rakhlin et al. (2005) cannot be applied to provide bounds for the expected KL divergence between MLEs of beta mixtures and the pair of targets. We visualize f_1 and f_2 in Figure 1.

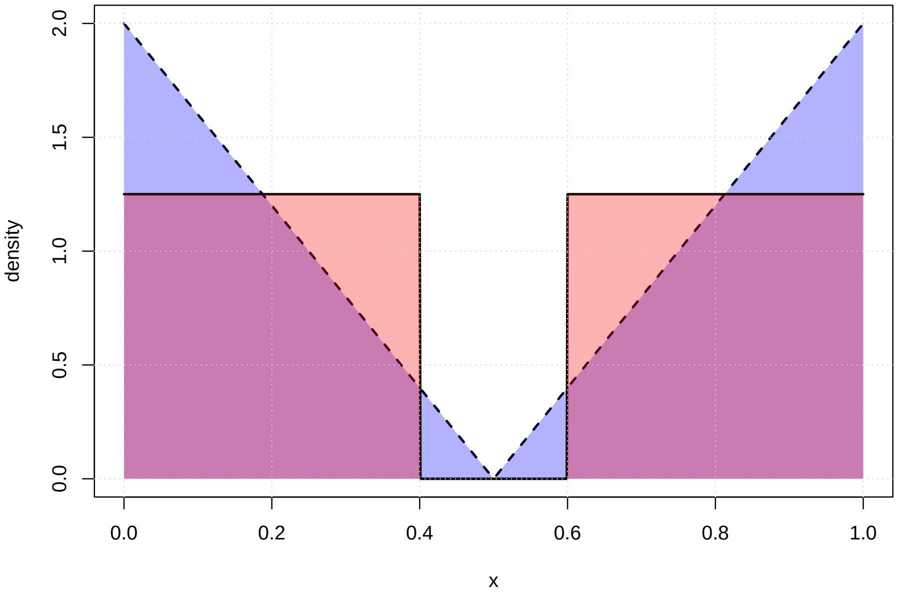

Figure 1: Simulation target densities f_1 (solid line) and f_2 (dashed line).

To observe the rate of decrease of the h-lifted KL divergence between the targets and respective sequences of h-MLLEs, we conduct two experiments **E1** and **E2**. In **E1**, our target density is set to f_1 and $h_1 = \beta(\cdot; 1/2, 1/2)$. For each $n \in \{2^{10}, \ldots, 2^{15}\}$ and $k \in \{2, \ldots, 8\}$, we independently simulate $\mathbf{X}_n$ and $\mathbf{Y}_n$ with each X_i and Y_i ($i \in [n]$), i.i.d., from the distributions characterized by f_1 and h_1, respectively. In **E2**, we target f_2 with h-MLLEs over the same ranges of k and n, but with $h_2 = \beta(\cdot; 1, 1)$–the density of the uniform distribution. For each k and n, we simulate $\mathbf{X}_n$ and $\mathbf{Y}_n$ respectively from distributions characterized by f_2 and h_2.

In both experiments, we simulate $r = 50$ replicates of each (k, n)-scenario and compute the corresponding h-MLLEs, $(f_{k,n,l})_{l \in [r]}$, using the previously described MM algorithm. For each $l \in [r]$, we compute the corresponding negative log h-lifted likelihood between the target f and $f_{k,n,l}$:

$$K_{k,n,l} = -\int_{\mathcal{X}} (f + h) \log (f_{k,n,l} + h) \, \mathrm{d}\mu$$

to assess the rates, and note that

$$\mathrm{KL}_h (f \,||\, f_{k,n,l}) = \int_{\mathcal{X}} (f + h) \log (f + h) \, \mathrm{d}\mu + K_{k,n,l},$$

where the prior term is a constant with respect to k and n.

To analyze the sample of $7 \times 6 \times 50 = 2100$ observations of relationship between the values $(K_{k,n,l})_{l\in[r]}$ and the corresponding values of k and n, we use non-linear least squares (Amemiya, 1985, Sec. 4.3) to fit the regression relationship:

$$\mathbf{E}\left[K_{k,n,l}\right] = a_0 + \frac{a_1}{(k+2)^{b_1}} + \frac{a_2}{n^{b_2}}. \tag{16}$$

We obtain 95% asymptotic confidence intervals for the estimates of the regression parameters $a_0, a_1, a_2, b_1, b_2 \in \mathbb{R}$, under the assumption of potential mis-specification of (16), by using the sandwich estimator for the asymptotic covariance matrix (cf. White 1982). We include the code for these experiments in Appendix B.

6.3 Results

We report the estimates along with 95% asymptotic confidence intervals for the parameters of (16) for **E1** and **E2** in Table 1. Plots of the average negative log h-lifted likelihood values by sample sizes n and numbers of components k are provided in Figure 2.

Table 1: Estimates of parameters for fitted relationships (with 95% confidence intervals) between negative log h-lifted likelihood values, sample size and number of mixture components for experiments **E1** and **E2**.

E1	a_0	a_1	a_2	b_1	b_2
Est.	-1.68	0.73	6.80	1.87	0.99
95% CI	$(-1.68, -1.67)$	$(0.68, 0.78)$	$(1.24, 12.36)$	$(1.81, 1.93)$	$(0.87, 1.11)$
E2	a_0	a_1	a_2	b_1	b_2
Est.	-1.47	1.49	6.75	4.36	1.07
95% CI	$(-1.48, -1.47)$	$(0.58, 2.41)$	$(2.17, 11.32)$	$(3.91, 4.81)$	$(0.97, 1.16)$

From Table 1, we observe that $\mathbf{E}\left[K_{k,n,l}\right]$ decreases with both n and k in both simulations, and that the rates at which the averages decrease are faster than anticipated by Theorem 9, with respect to both n and k. We can visually confirm the decreases in the estimate of $\mathbf{E}\left[K_{k,n,l}\right]$ via Figure 2. In both **E1** and **E2**, the rate of decrease over the assessed range of n is approximately proportional to $1/n$, as opposed to the anticipated rate of $1/\sqrt{n}$, whereas the rate of decrease in k is far larger, at approximately $1/k^{1.87}$ for **E1** and $1/k^{4.36}$ for **E2**.

These observations provide empirical evidence towards the fact that the rate of decrease of $\mathbf{E}\left[K_{k,n,l}\right]$ is at least $1/k$ and $1/\sqrt{n}$, respectively, for k and n, at least over the simulation scenarios. These fast rates of fit over small values of n and k may be indicative of a diminishing returns of fit phenomenon, as discussed in Cadez & Smyth (2000) or the so-called elbow phenomenon (see, e.g., Ritter 2014, Sec. 4.2), whereupon the rate of decrease in average loss for small values of k is fast and becomes slower as k increases, converging to some asymptotic rate. This is also the reason why we do not include the outcomes when $k = 1$, as the drop in $\mathbf{E}\left[K_{k,n,l}\right]$ between $k = 1$ and $k = 2$ is so dramatic that it makes our simulated data ill-fitted by any model of form (16). As such, we do not view Theorem 9 as being pessimistic in light of these phenomena, as it applies uniformly over all values of k and n.

7 Conclusion

The estimation of probability densities using finite mixtures from some base class $\mathcal{P}$ appears often in machine learning and statistical inference as a natural method for modeling underlying data generating processes. In this work, we sought to provide novel generalization bounds for such mixture estimators. To this end, we introduce the family of h-lifted KL divergences for densities on compact supports, within the family of Bregman divergences, which correspond to risk functions that can be bounded, even when densities in the class $\mathcal{P}$ are not bounded away from zero, unlike the standard KL divergence. Unlike the least-squares loss, the corresponding maximum h-likelihood estimation problem can be computed via an MM algorithm, mirroring the availability of EM algorithms for the maximum likelihood problem corresponding to the KL divergence. Along with our derivations of generalization bounds that achieve the same rates as the best-known bounds

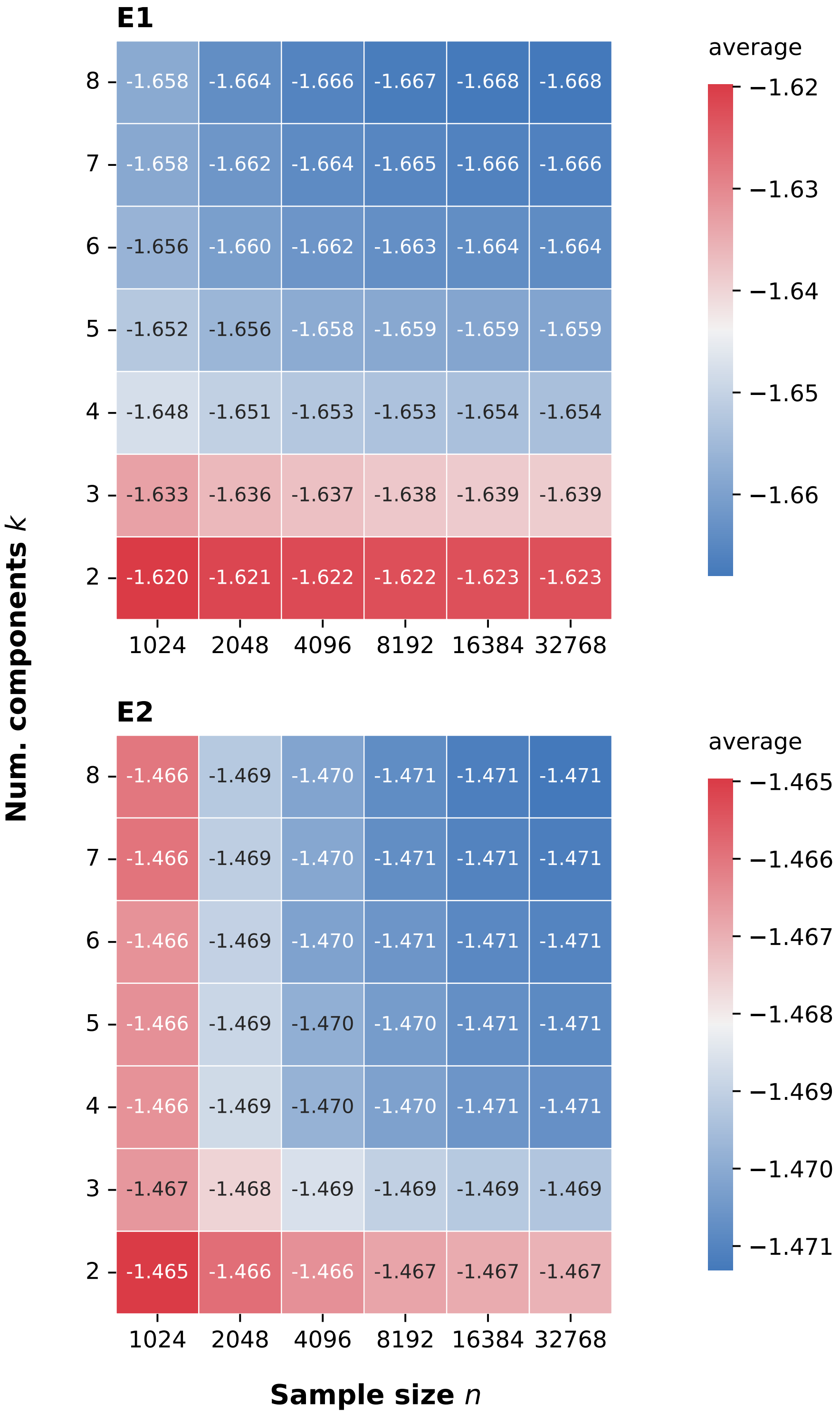

Figure 2: Average negative log h-lifted likelihood values by sample sizes n and numbers of components k for experiments **E1** and **E2**.

for the KL divergence and least square loss, we also provide numerical evidence towards the correctness of these bounds in the case when $\mathcal{P}$ corresponds to beta distribution densities.

Aside from beta distributions, mixture densities on compact supports that can be analysed under our framework appear frequently in the literature. For supports on compact Euclidean subset, examples include mixtures of Dirichlet distributions (Fan et al., 2012) and bivariate binomial distributions (Papageorgiou & David, 1994). Alternatively, one can consider distributions on compact Euclidean manifolds, such as mixtures of Kent (Peel et al., 2001) distributions and von Mises–Fisher distributions (Banerjee et al., 2005, Ng & Kwong, 2022). We defer investigating the practical performance of the maximum h-lifted likelihood estimators and accompanying theory for such models to future work.

References

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194. Springer, New York, 2016.

Takeshi Amemiya. *Advanced econometrics*. Harvard university press, 1985.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical Inference: The Minimum Distance Approach*. CRC press, Boca Raton, 2011.

Igor Cadez and Padhraic Smyth. Model complexity, goodness of fit and diminishing returns. *Advances in Neural Information Processing Systems*, 13, 2000.

Imre Csiszár. Generalized projections for non-negative functions. In *Proceedings of 1995 IEEE International Symposium on Information Theory*, pp. 6. IEEE, 1995.

Arnak S Dalalyan and Mehdi Sebbar. Optimal kullback–leibler aggregation in mixture density estimation by maximum likelihood. *Mathematical Statistics and Learning*, 1(1):1–35, 2018.

Ronald A DeVore and Vladimir N Temlyakov. Convex optimization on banach spaces. *Foundations of Computational Mathematics*, 16(2):369–394, 2016.

Wentao Fan, Nizar Bouguila, and Djemel Ziou. Variational learning for finite dirichlet mixture models and applications. *IEEE transactions on neural networks and learning systems*, 23:762–774, 2012.

Davide Ferrari and Yuhong Yang. Maximum lq-likelihood method. *Annals of Statistics*, 38:573–583, 2010.

Béla A Frigyik, Santosh Srivastava, and Maya R Gupta. Functional bregman divergence and bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54(11):5130–5139, 2008.

Hironori Fujisawa and Shinto Eguchi. Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference*, 136(11):3989–4011, 2006.

Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

Christophe Giraud. *Introduction to high-dimensional statistics*. CRC Press, 2021.

Uffe Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

Jussi S Klemelä. Density estimation with stagewise optimization of the empirical risk. *Machine Learning*, 67:169–195, 2007.

Jussi S Klemelä. *Smoothing of multivariate data: density estimation and visualization.* John Wiley & Sons, 2009.

Masahiro Kobayashi and Kazuho Watanabe. Generalized Dirichlet-process-means for f-separable distortion measures. *Neurocomputing*, 458:667–689, 2021. ISSN 0925-2312.

Masahiro Kobayashi and Kazuho Watanabe. Unbiased Estimating Equation and Latent Bias under f-Separable Bregman Distortion Measures. *IEEE Transactions on Information Theory*, 2024.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008.* Lecture Notes in Mathematics. Springer, 2011. ISBN 9783642221460.

Vladimir Koltchinskii and Dmitry Panchenko. Rademacher processes and bounding the risk of function learning. *arXiv: Probability*, pp. 443–457, 2004.

Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference.* Springer Series in Statistics. Springer New York, 2007. ISBN 9780387749785.

Kenneth Lange. *MM optimization algorithms.* SIAM, 2016.

Jonathan Li and Andrew Barron. Mixture Density Estimation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003.* Springer, 2007.

Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68, 2011.

Cathy Maugis-Rabusseau and Bertrand Michel. Adaptive density estimation for clustering with gaussian mixtures. *ESAIM: Probability and Statistics*, 17:698–724, 2013.

Colin McDiarmid. On the method of bounded differences. In J.Editor Siemons (ed.), *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, London Mathematical Society Lecture Note Series, pp. 148–188. Cambridge University Press, 1989.

Colin McDiarmid. Concentration. In Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed (eds.), *Probabilistic Methods for Algorithmic Discrete Mathematics*, pp. 195–248. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-662-12788-9.

Geoffrey J McLachlan and David Peel. *Finite Mixture Models.* John Wiley & Sons, 2004.

Ronny Meir and Assaf Zeevi. Density estimation through convex combinations of densities: Approximation and estimation bounds. *Neural Networks*, 10:99–109, 02 1997.

Kanta Naito and Shinto Eguchi. Density estimation with minimization of U-divergence. *Machine Learning*, 90(1):29–57, January 2013.

Tin Lok James Ng and Kwok-Kun Kwong. Universal approximation on the hypersphere. *Communications in Statistics-Theory and Methods*, 51:8694–8704, 2022.

Hien D Nguyen. An introduction to majorization-minimization algorithms for machine learning and statistical estimation. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2):e1198, 2017.

Hien D Nguyen, Florence Forbes, Gersende Fort, and Olivier Cappé. An online minorization-maximization algorithm. In *17th Conference of the International Federation of Classification Societies*, 2022a.

TrungTin Nguyen, Hien D Nguyen, Faicel Chamroukhi, and Geoffrey J McLachlan. Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7: 1750861, 2020.

TrungTin Nguyen, Faicel Chamroukhi, Hien D Nguyen, and Geoffrey J McLachlan. Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*, pp. 1–12, 2022b.

TrungTin Nguyen, Hien D Nguyen, Faicel Chamroukhi, and Florence Forbes. A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models. *Electronic Journal of Statistics*, 16(2):4742–4822, 2022c.

H Papageorgiou and Katerina M David. On countable mixtures of bivariate binomial distributions. *Biometrical journal*, 36(5):581–601, 1994.

Leandro Pardo. *Statistical Inference Based on Divergence Measures*. CRC Press, Boca Raton, 2006.

David Peel, William J Whiten, and Geoffrey J McLachlan. Fitting mixtures of kent distributions to aid in joint set identification. *Journal of the American Statistical Association*, 96:56–63, 2001.

Yichen Qin and Carey E Priebe. Maximum L_q-likelihood estimation via the expectation-maximization algorithm: a robust estimation of mixture models. *Journal of the American Statistical Association*, 108 (503):914–928, 2013.

Alexander Rakhlin, Dmitry Panchenko, and Sayan Mukherjee. Risk bounds for mixture density estimation. *ESAIM: PS*, 9:220–229, 2005.

Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Gunter Ritter. *Robust cluster analysis and variable selection*. CRC Press, 2014.

Ralph Tyrrell Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Wolfgang Stummer and Igor Vajda. On bregman distances and divergences of probability measures. *IEEE Transactions on Information Theory*, 58(3):1277–1288, 2012.

Vladimir N Temlyakov. Convergence and rate of convergence of some greedy algorithms in convex optimization. *Proceedings of the Steklov Institute of Mathematics*, 293:325–337, 2016.

Sara van de Geer. *Estimation and Testing Under Sparsity: École d'Été de Probabilités de Saint-Flour XLV – 2015*. Lecture Notes in Mathematics. Springer International Publishing, 2016. ISBN 9783319327747.

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN 9780387946405.

Jakob J Verbeek, Nikos Vlassis, and Ben Kröse. Efficient greedy learning of gaussian mixture models. *Neural computation*, 15(2):469–485, 2003.

Nikos Vlassis and Aristidis Likas. A greedy em algorithm for gaussian mixture learning. *Neural processing letters*, 15:77–87, 2002.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, pp. 1–25, 1982.

Tong Tong Wu and Kenneth Lange. The mm alternative to em. *Statistical Science*, 25(4):492–505, 2010.

Tong Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, 2003.

A The h-lifted KL divergence as a Bregman divergence

Let $\tilde{u} = u + h$, so that $\phi(u) = \tilde{u}\log(\tilde{u}) - \tilde{u} + 1$. Then $\phi'(u) = \log(\tilde{u})$, and

$$\begin{aligned} D_\phi(p \,||\, q) &= \int_X \{\tilde{p}\log(\tilde{p}) - \tilde{p} - 1\} - \{\tilde{q}\log(\tilde{q}) - \tilde{q} - 1\} - \log(\tilde{q})(p - q)\mathrm{d}\mu \\ &= \int_{\mathcal{X}} \tilde{p}\log(\tilde{p}) - \tilde{q}\log(\tilde{q}) - p\log(\tilde{q}) + q\log(\tilde{q})\mathrm{d}\mu \\ &= \int_{\mathcal{X}} \{p + h\}\log(\tilde{p}) - \{q + h\}\log(\tilde{q}) - p\log(\tilde{q}) + q\log(\tilde{q})\mathrm{d}\mu \\ &= \int_{\mathcal{X}} \{p + h\}\log\frac{p + h}{q + h}\mathrm{d}\mu = \mathrm{KL}_h\left(p \,||\, q\right). \end{aligned}$$

B Experimental results

The code for all simulations and analyses of Experiment 1 and 2 is available in both the `R` and `Python` programming languages. The experiments and analyses presented were conducted in `R`. The figures were created in `Python`. The code repository is available here: `https://github.com/XXXX`.

C Technical results

Here we collect some technical results that are required in our proofs but appear elsewhere in the literature. In some places, notation may be modified from the original text to keep with the established conventions herein.

Lemma 16 (Kosorok, 2007. Lem 9.18)**.** *Let $N(\mathscr{F}, \varepsilon, \|\cdot\|)$ denote the ε-covering number of $\mathscr{F}$, and $N_{[]}(\mathscr{F}, \varepsilon, \|\cdot\|)$ the ε-bracketing number of $\mathscr{F}$. Let $\|\cdot\|$ be any norm on $\mathscr{F}$. Then*

$$N(\mathscr{F}, \varepsilon, \|\cdot\|) \le N_{[]}(\mathscr{F}, \varepsilon, \|\cdot\|)$$

for all $\varepsilon > 0$.

Lemma 17 (Kosorok, 2007. Lem 9.22)**.** *For any norm $\|\cdot\|$ dominated by $\|\cdot\|_\infty$ and any class of functions $\mathscr{F}$,*

$$\log N_{[]}(\mathscr{F}, 2\varepsilon, \|\cdot\|) \le \log N(\mathscr{F}, \varepsilon, \|\cdot\|_\infty),$$

for all $\varepsilon > 0$.

Lemma 18 (Kosorok, 2007. Thm 9.23)**.** *Let $\mathscr{F} = \{f_t : t \in T\}$ be a function class satisfying*

$$|f_s(x) - f_t(x)| \le d(s, t)F(x),$$

for some metric d on T, some fixed function F on $\mathcal{X}$, and for all $x \in \mathcal{X}$ and $s, t \in T$. Then, for any norm $\|\cdot\|$,

$$N_{[]}(\mathscr{F}, 2\varepsilon\|F\|, \|\cdot\|) \le N(T, \varepsilon, d).$$

Lemma 19 (Shalev-Shwartz & Ben-David (2014), Lem 26.7)**.** *Let A be a subset of $\mathbb{R}^m$ and let*

$$A' = \left\{ \sum_{j=1}^{n} \alpha_j \mathbf{a}_j \mid n \in \mathbb{N}, \mathbf{a}_j \in A, \alpha_j \ge 0, \|\alpha\|_1 = 1 \right\}.$$

Then, $\mathcal{R}_n(A') = \mathcal{R}_n(A)$, i.e., both A and A' have the same Rademacher complexity.

Lemma 20 (van de Geer, 2016, Thm. 16.2)**.** *Let $(X_i)_{i\in[n]}$ be non-random elements of $\mathcal{X}$ and let $\mathscr{F}$ be a class of real-valued functions on $\mathcal{X}$. If $\varphi_i : \mathbb{R} \to \mathbb{R}$, $i \in [n]$, are functions vanishing at zero that satisfy for all $u, v \in \mathbb{R}$,*

$$|\varphi_i(u) - \varphi_i(v)| \le |u - v|,$$

then we have

$$\mathbf{E}\left\{ \left\| \sum_{i=1}^{n} \varphi_i(f(X_i))\varepsilon_i \right\|_{\mathscr{F}} \right\} \le 2\mathbf{E}\left\{ \left\| \sum_{i=1}^{n} f(X_i)\varepsilon_i \right\|_{\mathscr{F}} \right\}.$$

Lemma 21 (McDiarmid, 1998, Thm. 3.1 or McDiarmid, 1989)**.** *Suppose $(X_i)_{i\in[n]}$ are independent random variables and let $Z = g(X_1, \ldots, X_n)$, for some function g. If g satisfies the bounded difference condition, that is there exists constant c_j such that for all $j \in [n]$ and all $x_1, \ldots, x_j, x'_j, \ldots, x_n$,*

$$|g(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n) - g(x_1, \ldots, x_{j-1}, x'_j, x_{j+1}, \ldots, x_n)| \leq c_j,$$

then

$$\mathbf{P}(Z - \mathbf{E}Z \geq t) \leq \exp\left\{\frac{-2t^2}{\sum_{j=1}^n c_j^2}\right\}.$$

Lemma 22 (van der Vaart & Wellner, 1996, Lem. 2.3.1)**.** *Let $\mathfrak{R}(f) = \mathbf{E}f$ and $\mathfrak{R}_n(f) = n^{-1}\sum_{i=1}^n f(X_i)$. If $\Phi : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is a convex function, then the following inequality holds for any class of measurable functions $\mathscr{F}$:*

$$\mathbf{E}\Phi\left(\|\mathfrak{R}(f) - \mathfrak{R}_n(f)\|_{\mathscr{F}}\right) \leq \mathbf{E}\Phi\left(2\,\|R_n(f)\|_{\mathscr{F}}\right),$$

where $R_n(f)$ is the Rademacher process indexed by $\mathscr{F}$. In particular, since the identity map is convex,

$$\mathbf{E}\left\{\|\mathfrak{R}(f) - \mathfrak{R}_n(f)\|_{\mathscr{F}}\right\} \leq 2\mathbf{E}\left\{\|R_n(f)\|_{\mathscr{F}}\right\}.$$

Lemma 23 (Koltchinskii, 2011, Thm. 3.11)**.** *Let d_n be the empirical distance*

$$d_n^2(f_1, f_2) = \frac{1}{n}\sum_{i=1}^n (f_1(X_i) - f_2(X_i))^2$$

and denote by $N(\mathscr{F}, \varepsilon, d_n)$ the ε-covering number of $\mathscr{F}$. Let $\sigma_n^2 := \sup_{f\in\mathscr{F}} P_n f^2$. Then the following inequality holds

$$\mathbf{E}\left\{\|R_n(f)\|_{\mathscr{F}}\right\} \leq \frac{K}{\sqrt{n}}\mathbf{E}\int_0^{2\sigma_n} \log^{1/2} N(\mathscr{F}, \varepsilon, d_n)\mathrm{d}\varepsilon$$

for some constant $K > 0$.