Expected redundancy of mixtures from unconstrained families

W. D. Brinda and Andrew R. Barron

In a groundbreaking paper, Jones [1992] proved that the integrated squared error between a function in a Hilbert space and the best k-term linear combination greedily selected from a spanning set decays with order $1/k$ as long as a certain L^1-type norm is finite. Implications for neural network approximation of sigmoidal functions were worked out in detail by Barron [1993]; bounds for greedily estimating neural nets from data were given in Barron [1994]. These developments were significant for two main reasons: they showed that good approximation is possible without the number of nodes growing exponentially with the dimension of the function's domain, and they provided a more feasible optimization algorithm (greedily, one node at a time) for defining the nodes.

Under the advisement of Andrew Barron, Jonathan Li established analogous $1/k$ rates of approximation error and risk bounds for greedy k-component mixture *density estimation*. Their work is detailed in Li's doctoral thesis (Li [1999]) and summarized by Li and Barron [2000]. However, their inequality requires the family to have a uniformly bounded density ratio. As a result, it does not apply to familiar families, including Gaussian mixtures. In such cases, Li and Barron [2000] advocate truncating the distributions and restricting the parameter space to a compact subset of $\mathbb{R}^d$. We will demonstrate that $1/k$ rates can hold without a uniformly bounded density ratio; in particular, we prove such a result for expected redundancy rate of a greedy maximum likelihood estimator (MLE) for Gaussian mixtures.

The minimum description length (MDL) community introduced the notion of a *two-part code* for specifying data X. First, $\mathcal{L}(\theta)$ nats are used to specify distribution P_θ, then $\log \frac{1}{p_\theta(X)}$ are used to efficiently specify the data with respect to that distribution. The *redundancy* of P_θ for $X \sim P$ is the length of a two-part code minus $\log \frac{1}{p(X)}$, the length that would be used by the data-generating distribution P. In the case of $X^n \overset{iid}{\sim} P$, we often divide the redundancy by the sample size to define the *redundancy rate*

$$\frac{1}{n}\left[\sum_{i=1}^{n} \log \frac{p(X_i)}{p_\theta(X_i)} + \mathcal{L}(\theta)\right].$$

For an estimator $\hat{\theta}$, the expected redundancy rate

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\log\frac{p(X_i)}{p_{\hat{\theta}}(X_i)}+\mathcal{L}(\hat{\theta})\right]$$

can be related to statistical risk. Barron and Cover [1991] proved that if $\mathcal{L}$ is large enough, an estimator's Bhattacharyya risk is bounded by its expected redundancy rate. [Brinda, 2018, Chap 2] showed that even if $\mathcal{L}$ is small, the risk can be bounded by expected redundancy rate plus a corrective term that is often manageable.

For a penalized MLE with penalty $\mathcal{L}$, the expected redundancy rate is bounded by a quantity called an *index of resolvability* of the model for the data-generating distribution.[1]

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\log\frac{p(X_i)}{p_{\hat{\theta}}(X_i)}+\mathcal{L}(\hat{\theta})\right] &= \frac{1}{n}\mathbb{E}\min_{\theta}\left[\sum_{i=1}^{n}\log\frac{p(X_i)}{p_{\theta}(X_i)}+\mathcal{L}(\theta)\right]\\ &\leq \frac{1}{n}\min_{\theta}\mathbb{E}\left[\sum_{i=1}^{n}\log\frac{p(X_i)}{p_{\theta}(X_i)}+\mathcal{L}(\theta)\right]\\ &= \min_{\theta}\left[D(P\|P_{\theta})+\frac{\mathcal{L}(\theta)}{n}\right]\end{aligned}$$

For a more extensive overview, see Barron et al. [2008].

In Section 1, we will prove bounds on the expected redundancy and the approximation error of greedy maximizers of likelihood. The expected redundancy of the true maximizer is no greater than the expected redundancy of a greedy maximizer, so the bounds apply to ordinary penalized MLEs as well. Section 2 uses one of the expected redundancy results to bound the risk of a penalized MLE for Gaussian mixtures. For simplicity, we will use mixtures of spherically symmetric components all having the same scale, which we will call Gaussian radial basis mixtures (GRBMs).

Proofs of lemmas and theorems are at the end.

1 Expected redundancy of mixtures

Suppose $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ is a family of probability densities on a measurable space $\mathcal{X}$ with respect to a σ-finite dominating measure. Let Q be a probability measure on Γ whose domain σ-algebra is fine enough that $(\mu, x) \mapsto \phi_\mu(x)$ is

[1] Brinda and Klusowski [Submitted to Bernoulli in 2018] presents essentially the same results as [Brinda, 2018, Chap 2] but states them for resolvability rather than redundancy.

product-measurable.[2] Let $\bar{\phi}_Q$ denote the integral transform of Q defined by

$$\begin{aligned}\bar{\phi}_Q(x) &:= \int_\Gamma \phi_\mu(x) dQ(\mu) \\ &= \mathbb{E}_{\mu\sim Q}\phi_\mu(x).\end{aligned}$$

Tonelli's Theorem allows us to conclude that $\bar{\phi}_Q$ is measurable, and, by interchanging integrals, that $\bar{\phi}_Q$ must be a probability density as well. The corresponding probability measure on $\mathcal{X}$ is denoted $\bar{\Phi}_Q$ and is called the Q *mixture (over* Φ*)*.

We let $\mathcal{C}(\Phi)$ denote set of all such integral transforms of probability measures (each defined on a sufficiently fine σ-algebra of Γ); this set is convex. Notice that $\mathcal{C}(\Phi)$ includes all discrete mixtures from Φ. Importantly, $\mathcal{C}(\Phi)$ also includes all of the other well-defined "mixtures" such as *continuous mixtures*, as allowed by the nature of Γ.

Given any "target" probability measure P on $\mathcal{X}$, the greedy algorithm of Barron and Li constructs a sequence of approximating mixtures

$$p_{\theta^{(P)}_{k+1}} = (1-\alpha_{k+1})p_{\theta^{(P)}_k} + \alpha_{k+1}\phi_{\mu^{(P)}_{k+1}}.$$

The mixture components $\theta^{(P)}_k = \{\mu^{(P)}_1, \ldots, \mu^{(P)}_k\}$ are greedily chosen according to

$$\begin{aligned}\mu^{(P)}_1 &:= \operatorname*{argmax}_{\mu\in\Gamma} \mathbb{E}_{X\sim P} \log \phi_\mu(X), \qquad \text{followed by} \\ \mu^{(P)}_{j+1} &:= \operatorname*{argmax}_{\mu\in\Gamma} \mathbb{E}_{X\sim P} \log[(1-\alpha_{j+1})p_{\theta^{(P)}_j}(X) + \alpha_{j+1}\phi_\mu(X)].\end{aligned}$$

We will assume throughout this paper that a maximizer exists at each step; it need not be unique.

We will use the term "Barron's weights" to refer to the sequence $\alpha_j = 2/(j+1)$. Barron and Li suggest using either these weights or finding the optimal weights at each step.[3] After k steps, the weight of component $j \in \{1, \ldots, k\}$ is $\alpha_j \prod_{i=j+1}^{k}(1-\alpha_i)$; with Barron's weights, this simplifies to $\frac{2j}{k(k+1)}$. We will provide results for this choice of weights and also for the choice $\alpha_j = 1/j$ which results in an equal-weighted mixture.

Theorem 1.1 is a variant on Li's Lemma 5.9 that will make it possible for us to avoid requiring a lower bound on the densities being mixed. For any $A \subseteq \Gamma$ and probability measure Q on Γ, we define

$$b^{(A)}_Q(P) := \mathbb{E}_{X\sim P}\left[\left(1 + \sup_{\mu^*\in A} \log \frac{\sup_{\mu\in\Gamma}\phi_\mu(X)}{\phi_{\mu^*}(X)}\right)\frac{\mathbb{E}_{\mu\sim Q}\phi^2_\mu(X)}{[\bar{\phi}_Q(X)]^2}\right].$$

[2] By the theory of Carathéodory functions, if $\mathcal{X}$ is a separable metrizable space and each density $\phi_\mu : \mathcal{X} \to \mathbb{R}^+$ is continuous, then product-measurability is guaranteed as long as the domain σ-algebra is fine enough that $\mu \mapsto \phi_\mu(x)$ is measurable for every $x \in \mathcal{X}$ — see [Aliprantis and Border, 2006, Lem 4.51].

[3] Technically, Li presented the slightly different sequence $\alpha_2 = 2/3$ and $\alpha_j = 2/j$ thereafter. The sequence $2/(j+1)$ also works and is a bit simpler.

In particular, the quantities of current interest to us will have the greedy selections $\theta_k^{(P)}$ as the set A. We use $b_Q^{(k)}(P)$ as shorthand for $b_Q^{(\theta_k^P)}(P)$.

Theorem 1.1. *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities with respect to a σ-finite dominating measure, and let Q be a probability measure on Γ for which $(\mu, x) \mapsto \phi_\mu(x)$ is product-measurable. Let $P_{\theta_1^{(P)}}, P_{\theta_2^{(P)}}, \ldots$ be the sequence of mixtures from Φ that greedily maximize $\mathbb{E}_{X\sim P} \log p_{\theta_1}(X)$, $\mathbb{E}_{X\sim P} \log p_{\theta_2}(X)$, ... If either Barron's weights or optimal weights were used, then*

$$\mathbb{E}_{X\sim P} \log \frac{\bar{\phi}_Q(X)}{p_{\theta_k^{(P)}}(X)} \leq \frac{b_Q^{(k)}(P)}{k}.$$

Alternatively, if equal weights were used, then

$$\mathbb{E}_{X\sim P} \log \frac{\bar{\phi}_Q(X)}{p_{\theta_k^{(P)}}(X)} \leq \frac{(1+\log k)\, b_Q^{(k)}(P)}{k}.$$

After stating some of the interesting consequences this theorem, we will explore ways of bounding $b_Q^{(k)}(P)$ in specific contexts.

Corollary 1.2 uses Theorem 1.1 to bound the approximation error of greedy k-component mixtures in terms of any specific mixture over the family.

Corollary 1.2. *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities with respect to a σ-finite dominating measure, and let Q be a probability measure on Γ for which $(\mu, x) \mapsto \phi_\mu(x)$ is product-measurable. Let $P_{\theta_1^{(P)}}, P_{\theta_2^{(P)}}, \ldots$ be the sequence of mixtures from Φ that greedily maximize $\mathbb{E}_{X\sim P} \log p_{\theta_1}(X)$, $\mathbb{E}_{X\sim P} \log p_{\theta_2}(X)$, ... If either Barron's weights or optimal weights were used, then*

$$D(P\|P_{\theta_k^{(P)}}) \leq D(P\|\bar{\Phi}_Q) + \frac{b_Q^{(k)}(P)}{k}.$$

Alternatively, if equal weights were used, then

$$D(P\|P_{\theta_k^{(P)}}) \leq D(P\|\bar{\Phi}_Q) + \frac{(1+\log k)\, b_Q^{(k)}(P)}{k}.$$

The above result holds for any legitimate mixing distribution Q, so it holds for the infimum:

$$D(P\|P_{\theta_k^{(P)}}) \leq \inf_Q \{D(P\|\bar{\Phi}_Q) + \frac{b_Q^{(k)}(P)}{k}\}.$$

We will focus on conclusions for which the first term achieves its infimum so that our approximation error bound explicitly exhibits the divergence from the

target to the set of all mixtures. To that end, we define[4]

$$b_{\Phi}^{(k)}(P) := \liminf_{\epsilon \to 0} \left\{ b_{Q}^{(k)}(P) : Q \text{ s.t. } D(P\|\bar{\Phi}_Q) \leq D(P\|\mathcal{C}(\Phi)) + \epsilon \right\}.$$

This quantity can also be thought of as the smallest possible limit of $b_{Q_n}^{(k)}(P)$ among the sequences (Q_n) for which $D(P\|\bar{\Phi}_{Q_n})$ approaches the infimum relative entropy $D(P\|\mathcal{C}(\Phi))$.

Corollary 1.3. *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities with respect to a σ-finite dominating measure. Let $P_{\theta_1^{(P)}}$, $P_{\theta_2^{(P)}}$, ... be the sequence of mixtures from Φ that greedily maximize $\mathbb{E}_{X\sim P} \log p_{\theta_1}(X)$, $\mathbb{E}_{X\sim P} \log p_{\theta_2}(X)$, ... If either Barron's weights or optimal weights were used, then*

$$D(P\|P_{\theta_k^{(P)}}) \leq D(P\|\mathcal{C}(\Phi)) + \frac{b_{\Phi}^{(k)}(P)}{k}.$$

Alternatively, if equal weights were used, then

$$D(P\|P_{\theta_k^{(P)}}) \leq D(P\|\mathcal{C}(\Phi)) + \frac{(1+\log k)\, b_{\Phi}^{(k)}(P)}{k}.$$

The MDL method for bounding risk penalized likelihood estimation (which was the topic of Chapter **??**) is neatly stated in terms of the model's relative entropy approximation error. In truth, the method works for more general estimators and only needs a bound on the expected coding redundancy, which Corollary 1.4 bounds using Theorem 1.1. Throughout the remainder of this section, let P_n denote the random empirical distribution of $X^n \overset{iid}{\sim} P$; the notation $\hat{\theta}_j := \theta_j^{(P_n)}$ comes naturally.

Corollary 1.4. *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities with respect to a σ-finite dominating measure, and let Q be a probability measure on Γ for which $(\mu, x) \mapsto \phi_\mu(x)$ is product-measurable. Let $P_{\hat{\theta}_1}, P_{\hat{\theta}_2}, \ldots$ be the sequence of mixtures from Φ that greedily maximize the iid likelihood. If either Barron's weights or optimal weights were used, then*

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} \frac{1}{n} \sum_i \log \frac{p(X_i)}{p_{\hat{\theta}_k}(X_i)} \leq D(P\|\bar{\Phi}_Q) + \frac{\mathbb{E}\, b_{Q}^{(k)}(P_n)}{k}$$

and

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} \frac{1}{n} \sum_i \log \frac{p(X_i)}{p_{\hat{\theta}_k}(X_i)} \leq D(P\|\mathcal{C}(\Phi)) + \frac{\mathbb{E}\, b_{\Phi}^{(k)}(P_n)}{k}.$$

[4]This definition and other similar ones to come are analogous to that of [Li, 1999, Cor 3.3.1].

Alternatively, if equal weights were used, then

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} \frac{1}{n} \sum_i \log \frac{p(X_i)}{p_{\hat{\theta}_k}(X_i)} \leq D(P \| \bar{\Phi}_Q) + \frac{(1 + \log k) \mathbb{E}\, b_Q^{(k)}(P_n)}{k}$$

and

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} \frac{1}{n} \sum_i \log \frac{p(X_i)}{p_{\hat{\theta}_k}(X_i)} \leq D(P \| \mathcal{C}(\Phi)) + \frac{(1 + \log k) \mathbb{E}\, b_\Phi^{(k)}(P_n)}{k}.$$

Note that the expected redundancy bounds of Corollary 1.4 hold for the true maximum likelihood estimator as well, since it produces larger log likelihood values than the greedy algorithm does.

The above corollaries become useful once a bound for $b_Q^{(k)}(P)$ has been established. Theorem 1.5 does so by generalizing Li's approach. First, we define the point-wise density ratio supremum $s_\Phi(x) := \sup_{\mu_1, \mu_2 \in \Gamma} \frac{\phi_{\mu_1}(x)}{\phi_{\mu_2}(x)}$.

Theorem 1.5. *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities, and let Q be a probability measure on Γ. Then both $b_Q^{(k)}(P)$ and $\mathbb{E} b_Q^{(k)}(P_n)$ are bounded by*

$$\mathbb{E}_{X \sim P} \left[(1 + \log s_\Phi(X)) \frac{\mathbb{E}_{\mu \sim Q} \phi_\mu^2(X)}{\bar{\phi}_Q^2(X)} \right].$$

A uniform bound on the density ratio provides a constant bound on s_Φ. In that case, $(1 + \log \sup s_\Phi)\, c_Q^2(P)$ works as a bound, where

$$c_Q^2(P) := \mathbb{E}_{X \sim P} \frac{\mathbb{E}_{\mu \sim Q} \phi_\mu^2(X)}{\bar{\phi}_Q^2(X)};$$

likewise $(1 + \log \sup s_\Phi)\, c_\Phi^2(P)$ works in the infimum version of the bound, where

$$c_\Phi^2(P) := \lim_{\epsilon \to 0} \inf \left\{ c_Q^2(P) : Q \text{ s.t. } D(P \| Q) \leq D(P \| \mathcal{C}(\Phi)) + \epsilon \right\}.$$

These are essentially the bounds given in Li [1999]. Section 3.2 of that dissertation discusses $c_Q^2(P)$, pointing out that it is 1 plus an expected coefficient of variation; his Lemma 3.1 shows that $c_Q^2(\bar{\Phi}_Q)$ is bounded by the number of components of $\bar{\Phi}_Q$ if it is a discrete mixture from the model.

Li's results rely on a uniform bound for the density ratio, whereas Theorem 1.5 allows the density ratio to be bounded as a function of x and incorporates this function into a complexity constant for P.

For GRBMs with component means in an unbounded $\Gamma \subseteq \mathbb{R}^d$ there is no *uniform* bound, but in that case

$$\begin{aligned} \log s_\Phi(x) &= \frac{1}{2\sigma^2} \sup_{\mu \in \Gamma} \|x - \mu\|^2 \\ &\leq \frac{\|x - \mathbb{E}X\|^2 + \sup_{\mu \in \Gamma} \|\mu - \mathbb{E}X\|^2}{\sigma^2}. \end{aligned}$$

This leads us to define a weighted version of $c_Q^2(P)$ that arises in the GRBM bounds.

$$C_Q^2(P) := \mathbb{E}_{X\sim P} \frac{\|X - \mathbb{E}X\|^2}{\sigma^2} \frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X)}{\bar{\phi}_Q^2(X)}$$

By comparison to the proof of [Li, 1999, Lem 3.1], it is easily seen that if $\bar{\Phi}_Q$ is a discrete mixture of components $\phi_1, \ldots, \phi_k$, then

$$C_Q^2(\bar{\Phi}_Q) \leq \frac{1}{\sigma^2} \sum_{j=1}^{k} \mathbb{E}_{X_j\sim\phi_j} \|X_j - \mathbb{E}_{X\sim\bar{\Phi}_Q} X\|^2.$$

When the parameter space is bounded, Corollary 1.6 states a bound that follows from Theorem 1.5.

Corollary 1.6. *Let $\Phi := \{N(\mu, \sigma^2 I_d) : \mu \in \Gamma \subseteq B(a, r)\}$, and let Q be a probability measure on Γ with domain at least as fine as the Borel σ-field. Then both $b_Q^{(k)}(P)$ and $\mathbb{E}\, b_Q^{(k)}(P_n)$ are bounded by*

$$(1 + \tfrac{2r^2 + 2\|a - \mathbb{E}X\|^2}{\sigma^2})\, c_Q^2(P) + 2\, C_Q^2(P)$$

where $X \sim P$. Additionally, both $b_\Phi^{(k)}(P)$ and $\mathbb{E}\, b_\Phi^{(k)}(P_n)$ are bounded by

$$\liminf_{\epsilon\to 0} \left\{ [(1 + \tfrac{2r^2 + 2\|a - \mathbb{E}X\|^2}{\sigma^2}) c_Q^2(P) + 2C_Q^2(P)] : Q \text{ s.t. } D(P\|\bar{\Phi}_Q) \leq D(P\|\mathcal{C}(\Phi)) + \epsilon \right\}.$$

In conjunction with the previous corollaries, Corollary 1.6 enables us to bound the approximation error and expected redundancy of GRBMs with constrained component means.

Without constraining the parameter space, we can still bound the expected redundancy of GRBM maximum likelihood estimation by using Corollary 1.4 with Theorem 1.7 which uses the fact for the GRBM model all selected component means must be in the convex hull of the data points. The bound involves the $\mathbb{L}^p$-norm $\|Y\|_p := (\mathbb{E}\|Y\|^p)^{1/p}$.

Theorem 1.7. *Let $\Phi := \{N(\mu, \sigma^2 I_d) : \mu \in \mathbb{R}^d\}$, and let Q be a probability measure on $\mathbb{R}^d$ with domain at least as fine as the Borel σ-field. Then for any $z \geq 1$, $\mathbb{E}\, b_Q^{(k)}(P_n)$ is bounded by*

$$n^{1/z} \left[(1 + \tfrac{\|X - \mathbb{E}X\|_{2z}^2}{\sigma^2})\, c_Q^2(P) + 2\, C_Q^2(P) \right],$$

and $\mathbb{E}\, b_\Phi^{(k)}(P_n)$ is bounded by

$$n^{1/z} \liminf_{\epsilon\to 0} \left\{ [(1 + \tfrac{\|X - \mathbb{E}X\|_{2z}^2}{\sigma^2}) c_Q^2(P) + 2C_Q^2(P)] : Q \text{ s.t. } D(P\|\bar{\Phi}_Q) \leq D(P\|\mathcal{C}(\Phi)) + \epsilon \right\}.$$

Furthermore, if P has the subgaussianity property that $\mathbb{E}_{X\sim P} e^{t\|X - \mathbb{E}X\|} \leq e^{\sigma_P^2 t^2/2}$ for all $t \geq 0$, then $\mathbb{E}\, b_Q^{(k)}(P_n)$ is bounded by

$$(1 + \log n) \left[(1 + \tfrac{5\sigma_P^2}{\sigma^2})\, c_Q^2(P) + 2\, C_Q^2(P) \right],$$

and $\mathbb{E}\, b_\Phi^{(k)}(P_n)$ *is bounded by*

$$(1+\log n)\liminf_{\epsilon\to 0}\left\{[(1+\tfrac{5\sigma_P^2}{\sigma^2})c_Q^2(P)+2C_Q^2(P)] : Q \text{ s.t. } D(P\|\bar{\Phi}_Q)\le D(P\|\mathcal{C}(\Phi))+\epsilon\right\}.$$

2 Risk of Gaussian radial basis mixtures

Using the generalized MDL risk bound approach from [Brinda, 2018, Chap 2] with the expected redundancy bound derived in 1.7, we derive the following risk bound for GRBM estimation.[5]

Theorem 2.1. *Let* $\Phi := \{N(\mu,\sigma^2 I_d) : \mu\in\mathbb{R}^d\}$ *represent the Gaussian location family with covariance* $\sigma^2 I_d$*. Let* $\hat{\theta}=(\hat{k},\{\hat{\mu}_1,\dots\hat{\mu}_k\})$ *index the equal-weighted GRBM that maximizes (or greedily maximizes) log-likelihood minus penalty* $\mathbb{L}(\theta)=3dk\log 4nk$*. If there exists* $\sigma_P>0$ *for which* $\mathbb{E}_{X\sim P}e^{t\|X-\mathbb{E}X\|}\le e^{\sigma_P^2 t^2/2}$ *for all* $t\ge 0$*, then*

$$\mathbb{E}D_B(P,P_{\hat{\theta}})\le D(P\|\mathcal{C}(\Phi))+\frac{12d(1+\log n)^2}{\sqrt{n}}\left[\eta_\Phi^2(P)+\sigma_P^2+\tfrac{1}{\sigma^2}+\mathbb{E}\|\widetilde{X}-\mathbb{E}X\|+1\right]$$

where the distribution of $\widetilde{X}$ *has density proportional to* $\sqrt{p}$ *and*

$$\eta_\Phi^2(P):=\liminf_{\epsilon\to 0}\{(1+\tfrac{\sigma_P^2}{\sigma^2})c_Q^2(P)+C_Q^2(P) : Q \text{ s.t. } D(P\|\bar{\Phi}_Q)\le D(P\|\mathcal{C}(\Phi))+\epsilon\}.$$

Furthermore, $D_B(P,P_{\hat{\theta}})$ *minus*

$$\frac{1}{n}\sum_i\log\frac{p(X_i)}{p_{\hat{\theta}}(X_i)}+\frac{3d\hat{k}\log 4n\hat{k}}{n}+\frac{3d}{\sqrt{n}}\left[\max_i\|X_i-\mathbb{E}X\|^2+1/\sigma^2+\mathbb{E}\|\widetilde{X}-\mathbb{E}X\|+1\right]$$

is stochastically less than an exponential random variable with rate $n/2$*.*

Proofs

First, we will establish an iteration lemma similar to [Li, 1999, Lem 5.6] that enables us to deal with *equal-weighted* greedy mixtures.

Lemma 2.2. *Let* (B_k) *be a non-negative and non-decreasing sequence of real numbers. If* (D_k) *is a sequence such that*

$$D_{k+1}\le\tfrac{k}{k+1}D_k+\tfrac{1}{(k+1)^2}B_{k+1}.$$

then

$$D_k\le\frac{D_1+B_k\log k}{k}.$$

[5] The proof of Theorem 2.1 shows that the constant factors and the dependence on dimension are better than stated here. The inequality presented by the theorem was chosen for simplicity.

Proof. The inequality is trivial for $k = 1$. For $k \geq 2$, the stated consequence follows from the fact that

$$D_k \leq \frac{D_1 + B\sum_{j=2}^{k} 1/j}{k} \tag{1}$$

because the harmonic sum is bounded by the logarithm. We prove (1) by induction, assuming $B_k = B$ is fixed for all k. For $k = 2$,

$$D_2 \leq \frac{D_1 + B/2}{2}$$

as required. Next, assuming (1) holds for D_k,

$$\begin{aligned} D_{k+1} &\leq \tfrac{k}{k+1} D_k + \tfrac{1}{(k+1)^2} B \\ &\leq \frac{D_1 + B\sum_{j=2}^{k} 1/j}{k+1} + \frac{B/(k+1)}{k+1} \\ &= \frac{D_1 + B\sum_{j=2}^{k+1} 1/j}{k+1}. \end{aligned}$$

Now suppose rather than a fixed B, we have non-decreasing (B_k). To get the desired result for any particular k, simply invoke the fixed version with $B = B_k$ which is at least as large as the sequence's previous terms. □

An crucial function in Li [1999] is

$$\zeta(z) := \frac{z - 1 - \log z}{(z-1)^2}.$$

Li's Lemma 5.4 provides a convenient bound; the following lemma is a slight variant on that bound.

Lemma 2.3. *For any $t \geq 0$,*

$$\zeta(\tfrac{t}{3}) \leq 1 + \log(\tfrac{1}{t} \vee 1).$$

Proof. It is easy to verify that if $t \geq 1$, then $\zeta(\frac{t}{3})$ is less than 1, which is the value on the right side.

Next, we derive a rough bound that will provide the desired result for small values of t. Assuming $z \leq 1$,

$$\begin{aligned} \zeta(z) &:= \frac{z - 1 - \log z}{(z-1)^2} \\ &= \log \tfrac{1}{z} + \frac{z - 1 - (2z - z^2)\log z}{(z-1)^2} \\ &\leq \log \tfrac{1}{z} + \frac{z - 1 - 2z \log z}{(z-1)^2} \end{aligned}$$

Assuming further that $z = .1$, the numerator of the second term is no greater than $.1 - 1 + .2 \log 10 \approx -.44$; the denominator inflates the term, making it more negative. For any $z \le .1$, the second term's numerator will be less than that of the $z = .1$ case (because $z \log z$ is monotonic on $[0, .1]$). Thus for $z \le .1$, the second term is bounded by $1 - \log 3 \approx -.10$. This verifies that the proposed inequality works for $t \le .3$.

For the intermediate region $t \in (.3, 1)$, draw a plot to see that $\zeta(\frac{t}{3})$ is less than $1 - \log t$. □

Proof of ***Theorem 1.1****.* First, follow the proof of [Li, 1999, Lem 5.8] except use our Lemma 2.3 to bound $\zeta((1-\alpha)\frac{p_{\theta_{k-1}^{(P)}}}{\bar{\phi}_Q})$, which differs only slightly from Li's Lemma 5.4. Since ζ is decreasing ([Li, 1999, Lem 5.3]), the bound for $\alpha = 2/3$ also works for any smaller value of α.

$$\begin{aligned}
\zeta\left((1-\alpha)\frac{p_{\theta_{k-1}^{(P)}}}{\bar{\phi}_Q}\right) &\le \zeta\left(\frac{p_{\theta_{k-1}^{(P)}}}{3\bar{\phi}_Q}\right) \\
&\le 1 + \log \frac{\bar{\phi}_Q \vee p_{\theta_{k-1}^{(P)}}}{p_{\theta_{k-1}^{(P)}}} \\
&= 1 + \log \frac{\bar{\phi}_Q \vee \sum_j \lambda_j \phi_{\mu_j^{(P)}}}{\sum_j \lambda_j \phi_{\mu_j^{(P)}}} \\
&\le 1 + \log \frac{\sum_j \lambda_j (\bar{\phi}_Q \vee \phi_{\mu_j^{(P)}})}{\sum_j \lambda_j \phi_{\mu_j^{(P)}}} \\
&\le 1 + \max_{j \in \{1,\ldots,k-1\}} \log \frac{\bar{\phi}_Q \vee \phi_{\mu_j^{(P)}}}{\phi_{\mu_j^{(P)}}}
\end{aligned}$$

by the log-sum inequality. The numerator is bounded by $\sup_{(\mu,x)} \phi_\mu(x)$.

Combine this with the proof of [Li, 1999, Lem 5.9] to see the iterative inequality

$$\mathbb{E}_{X\sim P} \log \frac{\phi_Q(X)}{p_{\theta_{k+1}^{(P)}}(X)} \le (1-\alpha)\mathbb{E}_{X\sim P} \log \frac{\phi_Q(X)}{p_{\theta_k^{(P)}}(X)} + \alpha^2 b_Q^{(k)}(P).$$

The initial term is

$$\begin{aligned}
\mathbb{E}_{X\sim P} \log \frac{\phi_Q(X)}{p_{\theta_1^{(P)}}(X)} &= \mathbb{E}_{X\sim P} \log \frac{\phi_Q(X)}{\phi_{\mu_1^{(P)}}(X)} \\
&\le \mathbb{E}_{X\sim P}\left(1 + \log \frac{\phi_Q(X)}{\phi_{\mu_1^{(P)}}(X)}\right) \\
&\le \mathbb{E}_{X\sim P}\left[\left(1 + \log \frac{\phi_Q(X)}{\phi_{\mu_1^{(P)}}(X)}\right) \frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X)}{\bar{\phi}_Q^2(X)}\right] \\
&= b_Q^{(1)}(P)
\end{aligned}$$

because $\frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2}{\bar{\phi}_Q^2} \geq 1$ point-wise.

$b_Q^{(k)}(P)$ is a non-negative and non-decreasing sequence as k increases. If Barron's weights are used then [Li, 1999, Lem 5.6] applies. If optimal weights are used at any step, then it results in a smaller expected log likelihood ratio than the Barron weight does, so the inequality still holds.

The result for equal weights follows from Lemma 2.2 using the fact that the initial term is bounded by $b_Q^{(1)}(P)$ which is in turn bounded by $b_Q^{(k)}(P)$. □

*Proof of **Theorem 1.5**.* For $b_Q^{(k)}(P)$, the result is immediate from the definitions. For the expected empirical version of the inequality,

$$\begin{aligned}
\mathbb{E}b_Q^{(k)}(P_n) &:= \mathbb{E}_{X^n \overset{iid}{\sim} P} \tfrac{1}{n} \sum_i \left[\left(1 + \max_{\hat{\mu}\in\hat{\theta}_k} \log \frac{\sup_\mu \phi_\mu(X_i)}{\phi_{\hat{\mu}}(X_i)} \right) \frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X_i)}{\bar{\phi}_Q^2(X_i)} \right] \\
&\leq \mathbb{E}_{X^n \overset{iid}{\sim} P} \tfrac{1}{n} \sum_i \left[(1 + \log s_\Phi(X_i)) \frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X_i)}{\bar{\phi}_Q^2(X_i)} \right] \\
&= \tfrac{1}{n} \sum_i \mathbb{E}_{X_i\sim P} \left[(1 + \log s_\Phi(X_i)) \frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X_i)}{\bar{\phi}_Q^2(X_i)} \right].
\end{aligned}$$

□

Lemma 2.4. *Let $X, X_1, \ldots, X_n \overset{iid}{\sim} P$. For any non-negative functions g and h,*

$$\mathbb{E}\,\frac{1}{n}\sum_i \left[g(X_i) \max_j h(X_j) \right] \leq \mathbb{E}g(X)h(X) + [\mathbb{E}g(X)]\mathbb{E}\max_i h(X_i).$$

Proof.

$$\begin{aligned}
\mathbb{E}\,\frac{1}{n}\sum_i g(X_i)\max_j h(X_j) &\leq \mathbb{E}\,\frac{1}{n}\sum_i g(X_i)[h(X_i) + \max_{j\neq i} h(X_j)] \\
&= \mathbb{E}\,\frac{1}{n}\left[\sum_i g(X_i)h(X_i) + \sum_i g(X_i)\max_{j\neq i} h(X_j) \right] \\
&= \mathbb{E}\,g(X)h(X) + [\mathbb{E}g(X_1)][\mathbb{E}\max_{i\leq n-1} h(X_i)]
\end{aligned}$$

□

Proof of ***Theorem 1.7****.* For the GRBM model,

$$\begin{aligned} b_Q^{(k)}(P) &:= \mathbb{E}_{X\sim P}\left[\left(1+\max_{\hat{\mu}\in\hat{\theta}_k}\log\frac{\sup_\mu \phi_\mu(X)}{\phi_{\hat{\mu}}(X)}\right)\frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X)}{[\bar{\phi}_Q(X)]^2}\right] \\ &= \mathbb{E}_{X\sim P}\left[\left(1+\max_{\hat{\mu}\in\hat{\theta}_k}\frac{\|X-\hat{\mu}\|^2}{2\sigma^2}\right)\frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X)}{[\bar{\phi}_Q(X)]^2}\right] \\ &\le \mathbb{E}_{X\sim P}\left[\left(1+\frac{\|X-\mathbb{E}X\|^2}{\sigma^2}+\max_{\hat{\mu}\in\hat{\theta}_k}\frac{\|\hat{\mu}-\mathbb{E}X\|^2}{\sigma^2}\right)\frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X)}{[\bar{\phi}_Q(X)]^2}\right]. \end{aligned}$$

Therefore, with $X, X_1, \ldots, X_n \overset{iid}{\sim} P$,

$$\mathbb{E}b_Q^{(k)}(P_n) \le \mathbb{E}_{X^n\overset{iid}{\sim}P}\frac{1}{n}\sum_i\left[\left(1+\frac{\|X_i-\mathbb{E}X\|^2}{\sigma^2}+\max_{\hat{\mu}\in\hat{\theta}_k}\frac{\|\hat{\mu}-\mathbb{E}X\|^2}{\sigma^2}\right)\frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X_i)}{[\bar{\phi}_Q(X_i)]^2}\right].$$

The likelihood maximizing (or greedily maximizing) component means must be in the convex hull of the data points; otherwise, moving a proposed component mean toward its projection onto the convex hull would increase the likelihood of every data point. Furthermore, the farthest point to any convex polytope always occurs at a corner point; every corner point of the data's convex hull is itself a data point. Thus,

$$\max_{\hat{\mu}\in\hat{\theta}_k}\|\hat{\mu}-\mathbb{E}X\| \le \max_j\|X_j-\mathbb{E}X\|.$$

By Lemma 2.4,

$$\mathbb{E}\, b_Q^{(k)}(P_n) \le \left(1+\frac{\mathbb{E}\max_i\|X_i-\mathbb{E}X\|^2}{\sigma^2}\right)c_Q^2(P)+2C_Q^2(P)$$

Lemmas 2.6 and 2.7 below complete the proof by bounding the expected maximum squared deviation. □

The following lemma provides a general pattern for bounding an expected sample maximum. We present it here along with a standard proof for the reader's convenience.

Lemma 2.5. *If $X, X_1, \ldots, X_n \overset{iid}{\sim} P$, then for any convex, increasing, non-negative function f,*

$$\mathbb{E}\max_i X_i \le f^{-1}\left(n\mathbb{E}f(X)\right).$$

Proof.

$$\begin{aligned} f(\mathbb{E}\max_i X_i) &\le \mathbb{E}f(\max_i X_i) \\ &= \mathbb{E}\max_i f(X_i) \\ &\le \mathbb{E}\sum_i f(X_i) \\ &= n\mathbb{E}f(X) \end{aligned}$$

□

Lemma 2.6. *Let $X, X_1, \ldots, X_n \overset{iid}{\sim} P$. For any $z \geq 1$,*

$$\mathbb{E} \max_i \|X_i - \mathbb{E}X\|^2 \leq n^{1/z} (\mathbb{E}\|X - \mathbb{E}X\|^{2z})^{1/z}.$$

Proof. Use Lemma 2.5 with $f(x) = x^z$. □

Lemma 2.7. *Let $X, X_1, \ldots, X_n \overset{iid}{\sim} P$. If there exists $\sigma_P > 0$ such that $\mathbb{E}e^{t\|X - \mathbb{E}X\|} \leq e^{\sigma_P^2 t^2/2}$ for all $t \geq 0$, then*

$$\mathbb{E} \max_i \|X_i - \mathbb{E}X_1\|^2 \leq \tfrac{2(e+1)}{e-1} \sigma_P^2 (1 + \log n) < 5\, \sigma_P^2 (1 + \log n).$$

Proof. Using Lemma 2.5 with $f(x) = e^{x/2z}$,

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} \max_i \|X_i - \mathbb{E}X\|^2 \leq 2z \log \left(n \mathbb{E} e^{\|X_i - \mathbb{E}X\|^2/2z} \right).$$

Using the Taylor series representation and a common subgaussian moment bound (e.g. [Rivasplata, 2012, Prop 3.2]),

$$\begin{aligned}
\mathbb{E}e^{\|X_i - \mathbb{E}X\|^2/2z} &= 1 + \sum_{k \geq 1} \frac{\mathbb{E}\|X_i - \mathbb{E}X\|^{2k}/(2z)^k}{k!} \\
&\leq 1 + 2 \sum_{k \geq 1} (\sigma_P^2/z)^k \\
&= 1 + \frac{2}{1 - \sigma^2/z} \\
&= e
\end{aligned}$$

when we use $\frac{e+1}{e-1} \sigma_P^2$ for z. □

Lemma 2.8 formalizes a self-evident observation about reweighting a density toward a point. The stochastic inequality implies an inequality for the expectations, which is used for Theorem 2.1. It also implies a stochastic inequality (and therefore expectation inequality) for the squared norms, which is used for an example in Section **??**.

Lemma 2.8. *Suppose $g : \mathbb{R}^d \to \mathbb{R}$ is a unimodal measurable function that is spherically symmetric about its peak at μ. Let U be a random vector with Lebesgue density q, and let W be a random vector with density proportional to the product qg. Then*

$$\|W - \mu\| \overset{st}{\leq} \|U - \mu\|.$$

Proof. Define B_ϵ to be the closed ball of radius ϵ centered at μ, and define g_ϵ to be the value of g on the boundary of B_ϵ. Consider the ratio $\mathbb{P}(W \in B_\epsilon)/\mathbb{P}(W \notin B_\epsilon)$; the normalizing constant $\int qg\, d\gamma$ cancels out. Then because of the assumed shape of g, the numerator integrand is lower bounded by qg_ϵ, while the denominator integrand is upper bounded by qg_ϵ. Canceling the common g_ϵ gives

$$\frac{\mathbb{P}(W \in B_\epsilon)}{\mathbb{P}(W \notin B_\epsilon)} \geq \frac{\mathbb{P}(U \in B_\epsilon)}{\mathbb{P}(U \notin B_\epsilon)}$$

Because $\frac{x}{1-x}$ is a monotonic transformation, we have $\mathbb{P}(W \in B_\epsilon) \geq \mathbb{P}(U \in B_\epsilon)$, true for any ϵ, which implies the desired stochastic inequality. □

Lemma 2.9. *Let $\theta = (\mu_1, \ldots, \mu_k)$ with each $\mu_j \in \mathbb{R}^d$ indexing a component mean of an equal-weighted k-component GRBM P_θ. Let $\delta = (\delta_1, \ldots, \delta_k)$ where each $\delta_j \in \mathbb{R}^d$ has norm bounded by a. Then*

$$|D_B(P, P_{\theta+\delta}) - D_B(P, P_\theta)| \leq 2ka \left[a + \mathbb{E}\|\widetilde{X} - \mathbb{E}X\| + \max_j \|\mu_j - \mathbb{E}X\| \right]$$

where $X \sim P$ and $\widetilde{X}$ has density proportional to $\sqrt{p}$. Additionally, if each δ_j is random with expectation zero, then

$$\mathbb{E} \log \frac{1}{p_{\theta+\delta}(x)} - \log \frac{1}{p_\theta(x)} \leq a^2/2\sigma^2.$$

Proof. The deviation is bounded by the supremum absolute value of the derivative along the path from θ to $\theta + \delta$. (Let p denote the part of the density of P that is continuous with respect to Lebesgue measure.)

$$\begin{aligned}
\frac{d}{dt} D_B(P, P_{\theta+t\delta}) &= \frac{d}{dt} - 2 \log \int \sqrt{p(x)} (1/2\pi\sigma^2)^{d/4} \sqrt{\tfrac{1}{k} \sum_j e^{-\|x-(\mu_j+t\delta_j)\|^2/2\sigma^2}}\, dx \\
&= -2 \int \frac{\sqrt{p(x)} \sum_j e^{-\|x-(\mu_j+t\delta_j)\|^2/2\sigma^2} \delta_j'(x - (\mu_j + t\delta_j))}{\sqrt{\sum_i e^{-\|x-(\mu_i+t\delta_i)\|^2/2\sigma^2}} \int \sqrt{p(y)} \sqrt{\sum_i e^{-\|y-(\mu_i+t\delta_i)\|^2/2\sigma^2}}\, dy} dx
\end{aligned}$$

Use Cauchy-Schwarz to bound its absolute value.

$$\begin{aligned}
\left|\frac{d}{dt}D_B(P, P_{\theta+t\delta})\right| &\le 2\sum_j \int \frac{\sqrt{p(x)}e^{-\|x-(\mu_j+t\delta_j)\|^2/2\sigma^2}\|\delta_j\|\|x-(\mu_j+t\delta_j)\|}{\sqrt{\sum_i e^{-\|x-(\mu_i+t\delta_i)\|^2/2\sigma^2}}\int\sqrt{p(y)}\sqrt{\sum_i e^{-\|y-(\mu_i+t\delta_i)\|^2/2\sigma^2}}dy}dx \\
&\le 2\int\sum_j \frac{\sqrt{p(x)}e^{-\|x-(\mu_j+t\delta_j)\|^2/2\sigma^2}\|\delta_j\|\|x-(\mu_j+t\delta_j)\|}{\sqrt{e^{-\|x-(\mu_j+t\delta_j)\|^2/2\sigma^2}}\int\sqrt{p(y)}\sqrt{e^{-\|y-(\mu_j+t\delta_j)\|^2/2\sigma^2}}dy}dx \\
&= 2\sum_j \int \frac{\int\sqrt{p(x)}e^{-\|x-(\mu_j+t\delta_j)\|^2/4\sigma^2}\|\delta_j\|\|x-(\mu_j+t\delta_j)\|}{\int\sqrt{p(y)}e^{-\|y-(\mu_j+t\delta_j)\|^2/4\sigma^2}dy}dx \\
&\le 2\sum_j \|\delta_j\|\,\mathbb{E}_{\widetilde{X}\sim\sqrt{p}}\|\widetilde{X}-(\mu_j+t\delta_j)\| \\
&\le 2\sum_j \|\delta_j\|\left[\mathbb{E}_{\widetilde{X}\sim\sqrt{p}}\|\widetilde{X}-\mathbb{E}X\|+\|\mu_j-\mathbb{E}X\|+\|\delta_j\|\right]
\end{aligned}$$

by Lemma 2.8.

For the second part, we use [Brinda, 2018, Corollary B.0.7], which is a form of Hölder's inequality.

$$\begin{aligned}
\mathbb{E}-\log p_{\theta+\delta}(x) &= \mathbb{E}-\log\tfrac{1}{k}\sum_j e^{-\|x-(\mu_j+\delta_j)\|^2/2\sigma^2} \\
&\le -\log\tfrac{1}{k}\sum_j e^{-\mathbb{E}\|x-(\mu_j+\delta_j)\|^2/2\sigma^2} \\
&= -\log\tfrac{1}{k}\sum_j e^{-(\|x-\mu_j\|^2+\mathbb{E}\|\delta_j\|^2)/2\sigma^2} \\
&\le -\log\tfrac{1}{k}\sum_j e^{-\|x-\mu_j\|^2/2\sigma^2}+a^2/2\sigma^2
\end{aligned}$$

□

Proof of ***Theorem 2.1****.* Invoke [Brinda, 2018, Thm 2.2.1] with pseudo-penalty

$$\begin{aligned}
L(\theta) &= \tfrac{\sqrt{n}}{k}\sum_j \|\mu_j-\mathbb{E}X\|^2 \\
&\le \sqrt{n}\max_j\|\mu_j-\mathbb{E}X\|^2.
\end{aligned}$$

Because the [both greedy and true] likelihood-maximizing component means are in the convex hull of the data, each $\|\mu_j-\mathbb{E}X\|$ is bounded by $\max_i\|X_i-\mathbb{E}X\|$. Lemma 2.7 implies

$$\frac{\mathbb{E}L(\hat{\theta})}{n} \le \frac{(1+\log n)5\sigma_P^2}{\sqrt{n}}.$$

The summation part of [Brinda, 2018, Thm 2.2.1] can be handled by using integration grids ${\Theta_\epsilon}^{(k)} \subseteq \Theta^{(k)} = \mathbb{R}^{dk}$, as described in [Brinda, 2018, Sec 2.2].[6]

$$\sum_{k\geq 1} e^{-\frac{1}{2}\mathbb{L}(k)} \sum_{\theta_\epsilon \in {\Theta_\epsilon}^{(k)}} e^{-\frac{\sqrt{n}}{2k}\|\mu_j - \mathbb{E}X\|^2} = \sum_{k\geq 1} e^{-\frac{1}{2}\mathbb{L}(k)} \left(\frac{\sqrt{2\pi k}}{\epsilon n^{1/4}}\right)^{dk}. \tag{2}$$

Any penalty of at least $2dk\log(2\sqrt{2\pi k}/\epsilon n^{1/4})$ results in a summation no greater than 1.

The continuous optimization result is achieved by bounding the discrepancy from the grid within each model of the model class. Define $\hat{\theta}_k \in \mathbb{R}^{dk}$ to index the MLE (or greedy MLE) within $\Theta^{(k)}$. As demonstrated in [Brinda, 2018, Sec 2.2], we lower bound the infimum over the grid by an expectation for random $\hat{\theta}_k + \delta^{(k)}$ using a distribution for $\delta^{(k)} = (\delta_1, \ldots, \delta_k)$ on neighboring grid-points that has mean $\hat{\theta}_k$. The pseudo-penalty's contribution to expected discrepancy is

$$\begin{aligned}\tfrac{1}{n}[\mathbb{E}L(\hat{\theta}_k + \delta^{(k)}) - L(\hat{\theta}_k)] &= \tfrac{1}{n}[\tfrac{\sqrt{n}}{k}\mathbb{E}\|\delta^{(k)}\|^2] \\ &\leq 4\epsilon^2 d/\sqrt{n}\end{aligned}$$

using the bias-variance decomposition of the random $\delta^{(k)} \in \mathbb{R}^{dk}$ and the fact that each $\|\delta_j\| \leq 2\epsilon\sqrt{d}$.

The two remaining expected discrepancy terms are bounded by Lemma 2.9. First, the expected discrepancy of D_B is bounded by

$$4k\epsilon\sqrt{d}\left[2\epsilon\sqrt{d} + \mathbb{E}\|\widetilde{X} - \mathbb{E}X\| + \max_j \|X_i - \mathbb{E}X\|\right].$$

To further bound the maximum deviation term, use $z \leq (1+z^2)/2$ along with Lemma 2.7. Finally, the log-likelihood discrepancy is bounded by

$$2\epsilon^2 d/\sigma^2.$$

Let $\epsilon = \frac{1}{2.23k\sqrt{n}}$. (Note that if we knew a Bhattacharyya divergence discrepancy bound proportional to $1/\epsilon^2$, then we could use $\epsilon = n^{-1/4}$; in that case, the penalty would not need to involve n.)

One can confirm that the penalty is large enough to eliminate the summation term:

$$\begin{aligned}2dk\log(2\sqrt{2\pi k}/\epsilon n^{1/4}) &= 2dk\log(4.46\sqrt{2\pi}k^{3/2}n^{1/4}) \\ &< 3dk\log 5nk.\end{aligned}$$

Thus, after rounding up, we have established that

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta,\mathbb{L}}(P) + \frac{d(1+\log n)}{\sqrt{n}}\left[10\sigma_P^2 + \tfrac{1}{\sigma^2} + 2\mathbb{E}\|\widetilde{X} - \mathbb{E}X\| + 3.1\right].$$

[6] We will find that we want ϵ to depend on k; we will use increasingly refined discretizations for the more complex models.

Finally, we bound the expected redundancy using Theorem 1.7 then bound the infimum over k by comparison to the particular choice $k = \lceil \sqrt{n} \rceil \leq \sqrt{2n}$.

$$\begin{aligned}
\mathcal{R}^{(n)}_{\Theta,\mathbb{L}}(P) &= \mathbb{E}_{X^n \overset{iid}{\sim} P}\left[\frac{1}{n}\sum_i \log \frac{p(X_i)}{p_{\hat{\theta}}(X_i)} + \frac{\mathbb{L}(\hat{\theta})}{n}\right] \\
&= \mathbb{E}\min_k \left[\frac{1}{n}\sum_i \log \frac{p(X_i)}{p_{\hat{\theta}_k}(X_i)} + \frac{\mathbb{L}(k)}{n}\right] \\
&\leq \inf_k \left[\mathbb{E}\frac{1}{n}\sum_i \log \frac{p(X_i)}{p_{\hat{\theta}_k}(X_i)} + \frac{\mathbb{L}(k)}{n}\right] \\
&\leq \inf_k \left[D(P\|\mathcal{C}(\Phi)) + \frac{(1+\log k)(1+\log n)\eta^2_\Phi(P)}{k} + \frac{\mathbb{L}(k)}{n}\right] \\
&\leq D(P\|\mathcal{C}(\Phi)) + \frac{(1+\log\lceil\sqrt{n}\rceil)(1+\log n)\eta^2_\Phi(P)}{\lceil\sqrt{n}\rceil} + \frac{\mathbb{L}(\lceil\sqrt{n}\rceil)}{n} \\
&\leq D(P\|\mathcal{C}(\Phi)) + \frac{(1+\log n)^2\eta^2_\Phi(P)}{\sqrt{n}} + \frac{\mathbb{L}(\sqrt{2n})}{n} \\
&\leq D(P\|\mathcal{C}(\Phi)) + \frac{\eta^2_\Phi(P)}{\sqrt{n}} + \frac{8.3d(1+\log n)^2}{\sqrt{n}}
\end{aligned}$$

For the probabilistic result, compare to the proof of [Brinda, 2018, Thm 2.1.3]. To get the constant factor 3, we used $z \leq .45 + .56z^2$ for the norm of $\|X_i - \mathbb{E}X\|^2$. □

References

Charalambos D Aliprantis and Kim Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer Science & Business Media, 2006.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

Andrew R Barron, Cong Huang, Jonathan Q Li, and Xi Luo. The MDL Principle, Penalized Likelihoods, and Statistical Risk. *Feschrift for Jorma Rissanen*, 2008.

W. D. Brinda. *Adaptive Estimation with Gaussian Radial Basis Mixtures*. PhD thesis, Yale University, 2018.

W. D. Brinda and Jason M. Klusowski. Hölder's identity. Submitted to Bernoulli in 2018.

Lee K Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The annals of Statistics*, pages 608–613, 1992.

Jonathan Q Li. *Estimation of Mixture Models*. PhD thesis, Yale University, 1999.

Jonathan Q Li and Andrew R Barron. Mixture density estimation. In *Advances in Neural Information Processing Systems 12*, pages 279–285. The MIT Press, 2000.

Omar Rivasplata. Subgaussian random variables: An expository note. unpublished, available online, 2012.