# Expected redundancy of mixtures from unconstrained families

W. D. Brinda*[1], Jason M. Klusowski†[2], and Andrew R. Barron‡[1]

[1]Department of Statistics and Data Science, Yale University
[2]Department of Statistics and Biostatistics, Rutgers – New Brunswick

**Abstract**

By considering a rich class of models with appropriately devised penalties, density estimators can be designed to naturally adapt to the complexity revealed by the data. This paper explores approximation and estimation properties of Gaussian mixtures to perform this type of adaptive estimation. For simplicity and clarity of exposition, we use equal weights and fixed radial covariance, a model that we will call Gaussian radial basis mixtures (GRBMs).

The usual formulation of the minimum description length (MDL) risk bound does not apply to unpenalized maximum likelihood estimation or procedures with exceedingly small penalties. However, using techniques from Brinda and Klusowski [2018] that generalize the MDL redundancy risk bound method of Barron and Cover [1991] to arbitrary penalties, we extend the mixture redundancy bounds and approximation error of Li [1999] to the case of unconstrained parameter spaces. These results together allow us to establish an exact risk bound bound of order $(\log n)^2/\sqrt{n}$ on the statistical risk of penalized maximum likelihood GRBM estimation (or a greedily obtained variant) with a prescribed penalty on the number of parameters and no penalty on the sizes of those parameters.

Our works also extends the order $1/k$ relative entropy approximation error of $k$-component mixtures established by Li, who required that the component mixtures come from families that have a positive infimum density. Most densities of interest, including Gaussians, have an infimum of zero, and therefore do not satisfy Li's condition, though a truncated version does. We show that the desired bound on expected redundancy rate does hold for Gaussians and other elliptical distributions if one uses a different definition for the data-generating distribution's "complexity" constant.

## 1 Introduction

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.

*william.brinda@yale.edu
†jason.klusowski@rutgers.edu
‡andrew.barron@yale.edu

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 R. 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; see also Rakhlin et al. [2005] for some improvements. However, all these works require 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 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 $P_\theta$, then $\log \frac{1}{p_\theta(X)}$ are used to efficiently specify the data with respect to that distribution. The *redundancy* of $P_\theta$ 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]

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

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

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

In Section 2, 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 3 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). Although, in Section 4, we point to generalizations to other mixture distributions, including elliptical ones.

Proofs of lemmas and theorems are at the end.

## 2 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 $\sigma$-finite dominating measure. Let $Q$ be a probability measure on $\Gamma$ whose domain $\sigma$-algebra is fine enough that $(\mu, x) \mapsto \phi_\mu(x)$ is 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* $\Phi$*).*

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

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

$$p_{\theta_{k+1}^{(P)}} = (1-\alpha_{k+1})p_{\theta_k^{(P)}} + \alpha_{k+1}\phi_{\mu_{k+1}^{(P)}}.$$

The mixture components $\theta_k^{(P)} = \{\mu_1^{(P)}, \ldots, \mu_k^{(P)}\}$ are greedily chosen according to

$$\begin{aligned}\mu_1^{(P)} &:= \operatorname*{argmax}_{\mu\in\Gamma} \mathbb{E}_{X\sim P} \log \phi_\mu(X), \qquad \text{followed by} \\ \mu_{j+1}^{(P)} &:= \operatorname*{argmax}_{\mu\in\Gamma} \mathbb{E}_{X\sim P} \log[(1-\alpha_{j+1})p_{\theta_j^{(P)}}(X) + \alpha_{j+1}\phi_\mu(X)].\end{aligned}$$

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

[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 $\sigma$-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.

$k$ 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 2.1 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 $\Gamma$, we define

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

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 2.1.** *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities with respect to a $\sigma$-finite dominating measure, and let $Q$ be a probability measure on $\Gamma$ 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 $\Phi$ 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 2.2 uses Theorem 2.1 to bound the approximation error of greedy $k$-component mixtures in terms of any specific mixture over the family.

**Corollary 2.2.** *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities with respect to a $\sigma$-finite dominating measure, and let $Q$ be a probability measure on $\Gamma$ 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 $\Phi$ 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]

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

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 2.3.** *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities with respect to a $\sigma$-finite dominating measure. Let $P_{\theta_1^{(P)}}$, $P_{\theta_2^{(P)}}$, ... be the sequence of mixtures from $\Phi$ 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 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 2.4 bounds using Theorem 2.1. 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 2.4.** *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities with respect to a $\sigma$-finite dominating measure, and let $Q$ be a probability measure on $\Gamma$ 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 $\Phi$ 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 2.4 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 2.5 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 2.5.** *Let $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ be a family of probability densities, and let $Q$ be a probability measure on $\Gamma$. 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 $s_\Phi$. 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

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

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 2.5 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 2.6 states a bound that follows from Theorem 2.5.

**Corollary 2.6.** *Let $\Phi := \{N(\mu, \sigma^2 I_d) : \mu \in \Gamma \subseteq B(a, r)\}$, and let $Q$ be a probability measure on $\Gamma$ with domain at least as fine as the Borel $\sigma$-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 2.6 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 2.4 with Theorem 2.7 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 2.7.** *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 $\sigma$-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*

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

# 3 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 2.7, we derive the following risk bound for GRBM estimation.[5]

**Theorem 3.1.** *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 $\mathbb{E}_{X\sim P} e^{t\|X - \mathbb{E}X\|} \leq e^{\sigma_P^2 t^2/2}$ for all $t \geq 0$, then*

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

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

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

*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 $2/n$. If additionally $D(X\|X + x) \leq C_P\|x\|^\alpha$ for all $x$ and some positive constants $\alpha > 0$ and $C_P > 0$, then*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \sigma^\alpha C_P \mathbb{E}_{Z\sim N(0,I_d)}\|Z\|^\alpha + \frac{12d(1+\log n)^2}{\sqrt{n}}\left[\eta_\Phi^2(P) + \sigma_P^2 + \tfrac{1}{\sigma^2} + \mathbb{E}\|\widetilde{X} - \mathbb{E}X\| + 1\right],$$

*and if $\sigma \asymp (\log n)^{-1/8}$, then*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \to 0, \quad n \to +\infty.$$

*In other words, the penalized MLE that minimizes the Kullback-Leibler divergence to the truth $P$ is consistent even under misspecification.*

# 4 Discussion

The proof techniques used to obtain the statements in Theorem 3.1 can readily be adapted to handle mixtures of non-isotropic Gaussians, i.e., when $\Phi = \{N(\mu, \Sigma) : \mu \in \mathbb{R}^d\}$, where $\Sigma$ is a positive definite variance-covariance matrix,[6] and reach similar conclusions. There is also the opportunity to conduct mixture modeling beyond Gaussian. For example, the overall philosophy of our results remain valid when $\Phi$ is a location family of densities $x \mapsto h(\|x - \mu\|_H)$, where $h$ is a nonnegative, strictly decreasing function and $H$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$ and norm $\|x\|_H = \sqrt{\langle x, x\rangle_H}$. Examples of such distributions include the

[5] The proof of Theorem 3.1 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.

[6] In fact, the variance-covariance matrix need not be the same in each mixture component.

multivariate $t$-distribution, multivariate Laplace distribution, multivariate logistic distribution, multivariate Cauchy distribution, or any other elliptical distribution satisfying our assumption on $h$. An analogous condition to sub-Gaussian $\|X - \mathbb{E}X\|$ in our risk bounds is that $[h(2\|X - \mathbb{E}X\|_H)]^{-\lambda}$ be integrable for some $\lambda > 0$. We will leave a thorough treatment of these extensions for future consideration.

# 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. See also [Sancetta, 2013, Lemma 2] for a similar conclusion.

**Lemma 4.1.** *Let $(B_k)$ be a non-negative and non-decreasing sequence of real numbers. If $(D_k)$ is a sequence such that*

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

*then*

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

*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. □

A 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 4.2.** *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 $z \leq .1$, 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 $z \leq .1$, the second term is bounded by $1 - \log 3 \approx -.10$. This verifies that the proposed inequality works for $t \leq .3$.

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 2.1**.* First, follow the proof of [Li, 1999, Lem 5.8] except use our Lemma 4.2 to bound $\zeta((1-\alpha)\frac{p_{\theta^{(P)}_{k-1}}}{\bar{\phi}_Q})$, which differs only slightly from Li's Lemma 5.4. Since $\zeta$ is decreasing ([Li, 1999, Lem 5.3]), the bound for $\alpha = 2/3$ also works for any smaller value of $\alpha$.

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

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

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

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 $\frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2}{\bar{\phi}_Q^2} \ge 1$ 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 4.1 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 2.5**.* For $b_Q^{(k)}(P)$, the result is immediate from the definitions. For the expected empirical version of the inequality,

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

□

**Lemma 4.3.** *Let $X, X_1, \ldots, X_n \overset{iid}{\sim} P$. For any non-negative functions $g$ and $h$,*

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

*Proof.*

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

□

**Lemma 4.4.** *Let $h : \mathbb{R}^+ \mapsto \mathbb{R}^+$ be a strictly decreasing, nonnegative function. Let $H$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$ and norm $\|x\|_H = \sqrt{\langle x, x \rangle_H}$ and suppose that $\phi(x) = h(\|x\|_H)$ is a density with respect to a dominating measure on $H$. Consider the shift (location) family of densities $\Phi = \{\phi_\mu(x) = h(\|x - \mu\|_H) : \mu \in \Gamma\}$. Let $\hat{\mu}_1, \ldots, \hat{\mu}_k$ be the component means from the MLE (or greedily obtained MLE) of a $k$-component mixture of densities from $\Phi$ from an iid sample $X_1, \ldots, X_n$ and suppose conv$(\{X_1, \ldots, X_n\}) \subset \Gamma$. Then for each $j = 1, \ldots, k$, $\hat{\mu}_j$ belongs to the convex hull of the data $X_1, \ldots, X_n$. Furthermore, if $H$ is $d$-dimensional, there exists $A_j \subset [n]$ with $|A_j| \leq d + 1$ such that $\hat{\mu}_j = \sum_{i \in A_j} \lambda_i X_i$, where the $\lambda_i$ are nonnegative and sum to one.*

*Proof.* The log-likelihood of the model takes the form

$$\log \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \left( \sum_{j=1}^{k} \pi_j \phi_{\mu_j}(X_i) \right).$$

If $\mu$ does not belong to $\mathcal{P}_n = \text{conv}(\{X_1, \ldots, X_n\})$, the convex polytope of the data $X_1, \ldots, X_n$, then by the Hilbert projection theorem, there is a unique point $\tilde{\mu} = \text{proj}_{\mathcal{P}_n}(\mu)$ in $\mathcal{P}_n$ such that $\|\tilde{\mu} - x\|_H = \|\text{proj}_{\mathcal{P}_n}(\mu - x)\|_H < \|\mu - x\|_H$ for all $x \in \mathcal{P}_n$, i.e., the orthogonal projection onto a closed, convex set is a subcontractive linear operator. In particular, this means that $\phi_{\tilde{\mu}}(X_i) > \phi_\mu(X_i)$ for all $i = 1, \ldots, n$ and consequently the log-likelihood is increased. Thus, $\mu$ must belong to $\mathcal{P}_n$. The representation of each $\hat{\mu}_j$ as the convex combination of no more than $d + 1$ data points follows from Carathéodory's representation theorem for finite dimensional vector spaces. □

Specializing Lemma 4.4 to the isotropic Gaussian, i.e., $h$ is proportional to the univariate standard normal density and $\|\cdot\|_H$ is the Euclidean norm in $\mathbb{R}^d$ scaled by $1/\sigma$, we obtain the result that the MLE (or greedily obtained version) component means of a Gaussian mixture can each be written as a convex combination of at most $d + 1$ data points. Perhaps even more surprising is that this result continues to hold when the component distributions are Gaussian with general variance-covariance matrix $\Sigma$, since in this case, $x^\top \Sigma^{-1} y$ defines an inner product in $\mathbb{R}^d$ whenever $\Sigma^{-1}$ is positive definite.

Finally, we remark that the result need not hold for general normed vector spaces (e.g., some $L^p$ spaces on $\mathbb{R}^d$), since the nearest-point projection, even when it exists and is unique, may fail to be subcontractive. Hence for such norms, the likelihood maximizing component means may not lie in the convex hull of the data.

**Lemma 4.5.** *Suppose $X \sim P$ and $D(X\|X + x) \leq C_P \|x\|^\alpha$ for all $x$ and some positive constants $\alpha > 0$ and $C_P > 0$. Let $\Phi = \{\phi_\mu(x) = \phi(x - \mu) : \mu \in \Gamma\}$ and suppose supp$(X) \subset \Gamma$. If $Y$ is independent of $X$ and has density $\phi$, then*

$$D(P\|\mathcal{C}(\Phi)) \leq D(X\|X + Y) \leq C_P \mathbb{E}\|Y\|^\alpha. \tag{2}$$

*In particular, if $\phi$ is the Gaussian density in $\mathbb{R}^d$ with zero mean and variance-covariance matrix $\sigma^2 I_d$, then*

$$D(P\|\mathcal{C}(\Phi)) \leq \sigma^\alpha C_P \mathbb{E}_{Z \sim N(0, I_d)} \|Z\|^\alpha.$$

*Proof.* The first inequality in (2) is because the density of $Z = X + Y$, i.e., $z \mapsto \mathbb{E}_{X \sim P} \phi(z - X)$ belongs to $\mathcal{C}(\Phi)$, with mixing measure $P$. Suppose $X$ has density $p$. By the convexity of relative entropy, we have that

$$D(X\|Z) = \mathbb{E}_{X \sim P} \log \frac{p(X)}{\mathbb{E}_Y p(X - Y)} \leq \mathbb{E}_{X \sim P} \mathbb{E}_Y \log \frac{p(X)}{p(X - Y)}.$$

By Fubini's theorem and the assumption $D(X\|X+x) \le C_P\|x\|^\alpha$ for all $x$, we have

$$\mathbb{E}_{X\sim P}\mathbb{E}_Y \log\frac{p(X)}{p(X-Y)} = \mathbb{E}_Y\mathbb{E}_{X\sim P}\log\frac{p(X)}{p(X-Y)} \le C_P\mathbb{E}\|Y\|^\alpha.$$

□

As a side note, when $X \sim N(\mu_P, \sigma_P^2 I_d)$, $D(X\|X+x) = \|x\|^2/(2\sigma_P^2)$, and so for this case, the assumption of Lemma 4.5 holds with $C_P = 1/(2\sigma_P^2)$ and $\alpha = 2$. On the other hand, if $X$ is the product distribution of $d$ univariate Laplace distributions with diversity parameter $b_P > 0$, then $D(X\|X+x) \le \|x\|_1/b_P \le \sqrt{d}\|x\|/b_P$ and hence the assumption of Lemma 4.5 holds with $C_P = \sqrt{d}/b_P$ and $\alpha = 1$.

*Proof of* ***Theorem 2.7****.* 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 $X, X_1, \ldots, X_n \overset{iid}{\sim} P$,

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

By Lemma 4.4, the likelihood maximizing (or greedily maximizing) component means must be in the convex hull of the data points. 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 4.3,

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

Lemmas 4.7 and 4.8 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 4.6.** *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 4.7.** *Let $X, X_1, \ldots, X_n \overset{iid}{\sim} P$. For any $z \ge 1$,*

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

*Proof.* Use Lemma 4.6 with $f(x) = x^z$. □

**Lemma 4.8.** *Let $X, X_1, \ldots, X_n \overset{iid}{\sim} P$. If there exists $\sigma_P > 0$ such that $\mathbb{E}e^{t\|X-\mathbb{E}X\|} \le e^{\sigma_P^2 t^2/2}$ for all $t \ge 0$, then*

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

*Proof.* First, note that for $t < 1/(2\sigma_P^2)$, the assumption on the moment generating function implies that

$$\begin{aligned} \mathbb{E}e^{t\|X-\mathbb{E}X\|^2} &= \mathbb{E}\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma_P} e^{z\sqrt{2t/\sigma_P^2}\|X-\mathbb{E}X\| - z^2/(2\sigma_P^2)} dz \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma_P} \mathbb{E}e^{z\sqrt{2t/\sigma_P^2}\|X-\mathbb{E}X\| - z^2/(2\sigma_P^2)} dz \\ &\le \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma_P} e^{z^2 t - z^2/(2\sigma_P^2)} dz \\ &= \frac{1}{\sqrt{1-2t\sigma_P^2}}. \end{aligned}$$

Using Lemma 4.6 with $f(x) = e^{xt}$,

$$\begin{aligned} \mathbb{E}_{X^n \overset{iid}{\sim} P} \max_i \|X_i - \mathbb{E}X\|^2 &\le t^{-1}\log\left(n\mathbb{E}e^{t\|X_i-\mathbb{E}X\|^2}\right) \\ &\le t^{-1}\log\left(n(1-2t\sigma_P^2)^{-1/2}\right). \end{aligned}$$

The result follows from choosing $t = (1-e^{-2})/(2\sigma_P^2)$. □

Lemma 4.9 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 3.1. It also implies a stochastic inequality (and therefore expectation inequality) for the squared norms, which is used for an example in [Brinda, 2018, Section 2.2].

**Lemma 4.9.** *Let $h : \mathbb{R}^+ \mapsto \mathbb{R}^+$ be a decreasing, nonnegative function. Let $V$ be a normed vector space with norm $\|\cdot\|_V$ and set $g(x) = h(\|x - \mu\|_V)$. 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*

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

*Proof.* Define $B_\epsilon$ to be the closed ball of radius $\epsilon$ centered at $\mu$, and define $g_\epsilon$ to be the value of $g$ on the boundary of $B_\epsilon$. 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 $qg_\epsilon$, while the denominator integrand is upper bounded by $qg_\epsilon$. Canceling the common $g_\epsilon$ 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 $\epsilon$, which implies the desired stochastic inequality. □

**Lemma 4.10.** *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 $P_\theta$. 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 $\delta_j$ 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 $\theta$ 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.

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

by Lemma 4.9. ($\widetilde{X}\sim\sqrt{p}$ should be understood to mean the normalized version of $\sqrt{p}$.)

For the second part, we use [Brinda, 2018, Cor B.0.3], 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 3.1****.* Invoke [Brinda, 2018, Thm 2.2.1] with pseudo-penalty

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

By Lemma 4.4, both the 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 4.8 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].[7]

$$\sum_{k\geq 1} e^{-\frac{1}{2}\mathbb{L}(k)} \sum_{\theta_\epsilon \in {\Theta_\epsilon}^{(k)}} e^{-\frac{\sqrt{n}}{2k}\|\mu_j - \mathbb{E}X\|^2} = \sum_{k\geq 1} e^{-\frac{1}{2}\mathbb{L}(k)} \left(\frac{\sqrt{2\pi k}}{\epsilon n^{1/4}}\right)^{dk}. \tag{3}$$

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 4.10. 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 4.8. 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

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

where $\mathcal{R}$ denotes expected redundancy as used in Brinda [2018].

[7] We will find that we want $\epsilon$ to depend on $k$; we will use increasingly refined discretizations for the more complex models.

Finally, we bound the expected redundancy using Theorem 2.7 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 $\|X_i - \mathbb{E}X\|^2$.

The final conclusion of the theorem, which results from bounding $D(P\|\mathcal{C}(\Phi))$, follows from Lemma 4.5.

We now show that, if $\sigma \asymp (\log n)^{-1/8}$, then $\mathbb{E}D_B(P, P_{\hat{\theta}})$ tends to zero (albeit, at a slow polylogarithmic rate) as the sample size $n$ grows. This is based on showing that $\log \eta^2_\Phi(P) = O(\sigma^{-8})$. Recall that $\eta^2_\Phi(P)$ is defined as the limit infimum of $(1 + \frac{\sigma^2_P}{\sigma^2})c^2_Q(P) + C^2_Q(P)$ over all mixing distributions $Q$ that are arbitrarily close to $D(P\|\mathcal{C}(\Phi))$. We only show that the logarithm of $c^2_Q(P)$ scales as $\sigma^{-8}$, since $C^2_Q(P)$ is handled similarly. To this end, we work with $\mu \sim Q$ for which $\|\mu - \bar{\mu}\|$ and $\|\mu - \bar{\mu}\|^2$ are sub-Gaussian, where we set $\bar{\mu} := \mathbb{E}_{\mu\sim Q}\mu$. This is also satisfied if, for example, $\mu \sim Q$ has compact support. Note that

$$\begin{aligned}
c^2_Q(P) &= \mathbb{E}_{X\sim P}\frac{\mathbb{E}_{\mu\sim Q}\phi^2_\mu(X)}{\bar{\phi}^2_Q(X)} \\
&= \mathbb{E}_{X\sim P}\frac{\mathbb{E}_{\mu\sim Q}\phi^2_\mu(X)}{[\mathbb{E}_{\mu\sim Q}\phi_\mu(X)]^2} \\
&= \mathbb{E}_{X\sim P}\frac{\mathbb{E}_{\mu\sim Q}e^{-\|X-\mu\|^2/\sigma^2}}{[\mathbb{E}_{\mu\sim Q}e^{-\|X-\mu\|^2/(2\sigma^2)}]^2} \\
&\leq \mathbb{E}_{X\sim P}\frac{\mathbb{E}_{\mu\sim Q}e^{-\|X-\mu\|^2/\sigma^2}}{e^{-\mathbb{E}_{\mu\sim Q}\|X-\mu\|^2/\sigma^2}} \\
&= \mathbb{E}_{X\sim P}\mathbb{E}_{\mu\sim Q}e^{-\|X-\mu\|^2/\sigma^2 + \mathbb{E}_{\mu\sim Q}\|X-\mu\|^2/\sigma^2} \\
&= \mathbb{E}_{\mu\sim Q}\mathbb{E}_{X\sim P}e^{-\|X-\mu\|^2/\sigma^2 + \mathbb{E}_{\mu\sim Q}\|X-\mu\|^2/\sigma^2},
\end{aligned}$$

where we used Jensen's inequality and Tonelli's theorem for the iterated expectation exchange. Next, expanding squares in the exponent, we have

$$\mathbb{E}_{X\sim P} e^{-\|X-\mu\|^2/\sigma^2 + \mathbb{E}_{\mu\sim Q}\|X-\mu\|^2/\sigma^2} = \mathbb{E}_{X\sim P} e^{(2\langle X, \mu-\bar{\mu}\rangle + \mathbb{E}_{\mu\sim Q}\|\mu\|^2 - \|\mu\|^2)/\sigma^2}.$$

Since $\|X - \mathbb{E}X\|$ is sub-Gaussian, we have

$$\mathbb{E}_{X\sim P} e^{2\langle X, \mu-\bar{\mu}\rangle/\sigma^2} \leq e^{2\|\mu-\bar{\mu}\|^2\sigma_P^2/\sigma^4 + 2\langle \mathbb{E}X, \mu-\bar{\mu}\rangle/\sigma^2}.$$

Thus, it follows that

$$c_Q^2(P) \leq \mathbb{E}_{\mu\sim Q} e^{2\|\mu-\bar{\mu}\|^2\sigma_P^2/\sigma^4 + 2\langle \mathbb{E}X, \mu-\bar{\mu}\rangle/\sigma^2 + \mathbb{E}_{\mu\sim Q}\|\mu\|^2/\sigma^2 - \|\mu\|^2/\sigma^2}. \tag{4}$$

Finally, if $\|\mu - \bar{\mu}\|$ and $\|\mu - \bar{\mu}\|^2$ are both sub-Gaussian, then the logarithm of the upper bound in (4) is $O(\sigma^{-8})$.

□

# References


Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 3rd edition, 2006.

Andrew R. Barron. Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory*, 39(3):930–944, 1993.

Andrew R. Barron. Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning*, 14(1):113–143, 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 Li, and Xi Luo. *The MDL Principle, Penalized Likelihoods, and Statistical Risk*. Tampere International Center for Signal Processing. Tampere University Press, Tampere, Finland, 2008.

W. D. Brinda. *Adaptive Estimation with Gaussian Radial Basis Mixtures*. Thesis, 2018.

W. D. Brinda and Jason M. Klusowski. Finite-sample risk bounds for maximum likelihood estimation with arbitrary penalties. *IEEE Transactions on Information Theory*, 64(4): 2727–2741, 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*, 20(1):608–613, 1992.

Jonathan Q. Li. *Estimation of Mixture Models*. Thesis, 1999.

Jonathan Q. Li and Andrew R. Barron. Mixture Density Estimation. In S. A. Solla, T. K. Leen, and K. Muller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 279–285. MIT Press.

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

Alessio Sancetta. A recursive algorithm for mixture of densities estimation. *IEEE Transactions on Information Theory*, 59(10):6893–6906, 2013.