Abstract

Adaptive Estimation with Gaussian Radial Basis Mixtures

William David Brinda

2020

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 dissertation explores approximation, estimation, and computation properties of Gaussian mixtures to perform this type of adaptive estimation. For simplicity and clarity of exposition, we use equal weights and a fixed radial covariance, a model that we will call Gaussian radial basis mixtures (GRBMs). First, we generalize the MDL redundancy risk bound method of Barron and Cover [1991] to arbitrary penalties. Then we extend mixture redundancy bounds of Li [1999] to the case of unconstrained parameter spaces. These results together allow us to establish an exact risk bound for penalized likelihood GRBM estimation. Finally, simulations are performed to compare algorithms for optimizing the likelihood.

Adaptive Estimation with Gaussian Radial Basis Mixtures

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
William David Brinda

Dissertation Director: Andrew R. Barron

May, 2018

Contents

List of Figures

List of Tables

Acknowledgments

I'm eternally grateful to the faculty, staff, and students of the Yale Statistics Department. So many of you have helped me along the way or made my time here more pleasant; three have had a particularly profound impact. Professor Barron, you guided me toward questions that matter. Professor Chang, without your encouragement I wouldn't have lasted in this program. Jason, working with you has been stimulating and delightful.

Mom, Dad, Daniel, Nathan, Ray, Diann, my extended family, and my friends in the R&R Literary Society thank you for your support and patience throughout my many years far from home. I love all of you more than you know.

My beloved wife Sonya, we did it.

Chapter 1

Introduction

Gaussian mixtures are a popular family used by data analysts to estimate an unknown density when the data-generating mechanism is *a priori* thought to comprise k distinct sources, where k may be known or unknown. Each "component" Gaussian in the estimate is typically interpreted as the distribution of data coming from a corresponding source. The weight on a component is interpreted as the probability that a new observation comes from the corresponding source. The (penalized or unpenalized) maximum likelihood estimate (MLE) is often approximated by the Expectation Maximization (EM) algorithm. Alternatively, one may select the predictive mixture arising from a Bayesian posterior (which does not have a closed form) or a variational approximation thereof (which does have a closed form) as approximated by the mean field algorithm. Recently, algorithms have been developed to compute an approximate method of moments estimator.

Technically, a Gaussian kernel density estimate (KDE) is also a Gaussian mixture estimate that has $k = n$ equally weighted components with their means at the data points and all sharing the same fixed covariance. The KDE is not interpreted in terms of data-generating sources; its objective is only to approximate the overall data-generating density. The estimate is immediate, unlike traditional Gaussian mixture estimates which can be computationally challenging. However, KDEs suffer from the curse of dimensionality: volume increases exponentially with dimension, so the data points become increasingly sparse in high dimensions. In other words, for a KDE to perform well, it needs sample size increasing exponentially with dimension.

As a "compromise" between the two approaches, consider *Gaussian radial basis mixture* (GRBM) estimation. We restrict ourselves to equally-weighted mixtures from $\{N(\mu, \sigma^2 I_d) : \mu \in \mathbb{R}^d\}$ for a fixed known σ^2 with the sole aim of approximating the data-generating density well rather than trying to represent the data-generating mechanism in an interpretable way. Ordinary Gaussian mixture estimation algorithms are used, but they become slightly simpler to compute and to analyze with the GRBM restrictions of equal weights and fixed radial covariance. Notice that any rationally-weighted mixture of the radial basis functions can be achieved by an equally-weighted mixture if enough components are included. Compared to ordinary Gaussian mixture estimation, GRBM estimation substitutes quantity for quality, but it manages to avoid the curse of dimensionality that plagues its KDE cousin.

Over the course of this dissertation, we will establish new risk bounds for penalized maximum likelihood GRBM estimation and experiment with algorithms for performing the optimization.

Chapter 2 extends the minimum description length (MDL) method for bounding the statistical risk of penalized likelihood estimators on a countable model. The usual formulation of the MDL risk bound does not apply to unpenalized maximum likelihood estimation or procedures with exceedingly small penalties. We point out a more general inequality that holds for arbitrary penalties by adding a corrective term. In addition, this approach makes it possible to derive exact risk bounds of order $1/n$ for iid parametric models, which improves on the order $(\log n)/n$ resolvability bounds. We also describe how our bounds can be extended to penalized likelihood estimation over continuous models by comparison to a discrete grid, a pattern well-established for the resolvability bound; we demonstrate with the Gaussian location family.

An important term in the MDL risk bounds is the estimator's *expected redundancy*. Chapter 3 builds on the work of Li [1999] to bound expected redundancy and approximation error of mixtures. Li established order $1/k$ relative entropy approximation error of k-component mixtures 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 if one uses a different definition for the data-

generating distribution's "complexity" constant.

Chapter 4 uses the risk bound method of Chapter 2 together with an expected redundancy result from Chapter 3 to derive a bound of order $(\log n)/n$ on the statistical risk of penalized maximum likelihood GRBM estimation with a prescribed penalty on the number of parameters and no penalty on the sizes of those parameters.

In Chapter 5, we consider a variety of algorithms for initializing EM to find the likelihood maximizer of GRBMs. We first introduce a promising new internal annealing algorithm for approximately sampling from the normalized likelihood before describing the mean field procedure and a method of third moments. Finally, their performances for likelihood maximization are compared via simulation.

The three Appendix chapters make short commentaries that are relevant to earlier chapters, but may also be of more general interest in their own right. Chapter A describes the compensation and reverse compensation identities, which can be thought of as bias-variance decompositions for the relative entropy of a random distribution. Chapter B points out that Hölder's inequality can be generalized to an identity with an information-theoretic interpretation. Lastly, Chapter C provides a justification and formalism for treating measurability as a secondary concern when dealing with probabilities and expectations.

Within each chapter, the proofs are collected in a section at the end. All results labeled *lemma* or *theorem* have formal proofs, while *corollaries* are explained informally within the text if needed.

Chapter 2

Risk of penalized likelihood estimators[1]

A remarkably general method for bounding the statistical risk of penalized likelihood estimators comes from work on two-part coding, one of the minimum description length (MDL) approaches to statistical inference. Two-part coding MDL prescribes assigning code-lengths to a model (or model class) then selecting the distribution that provides the most efficient description of one's data [Rissanen, 1978]. The total description length has two parts: the part that specifies a distribution within the model (as well as a model within the model class if necessary) and the part that specifies the data with reference to the specified distribution. If the code-lengths are exactly Kraft-valid, this approach is equivalent to Bayesian maximum a posteriori (MAP) estimation, in that the two parts correspond to log reciprocal of prior and log reciprocal of likelihood respectively. More generally, one can call the part of the codelength specifying the distribution a *penalty* term; it is called the *complexity* in MDL literature.

Let $(\Theta, \mathcal{L})$ denote a discrete set indexing distributions along with a complexity function. With $X \sim P$, the (point-wise) *redundancy* of any $\theta \in \Theta$ is its two-part codelength minus $\log(1/p(X))$, the codelength one gets by using P as the coding distribution. [2] We define

1. Much of this chapter is adapted from [Brinda and Klusowski, 2018, Sec 1, 2, and Appendix]

2. For now, we mean that P governs the entirety of the data. The notion of sample size and iid assumptions are not essential to the bounds, as will be seen in the statement of Theorem 2.1.1. Specialization to iid data

an estimator's *expected redundancy* for P to be[3]

$$\mathcal{R}_{\hat{\theta},\mathcal{L}}(P) := \mathbb{E}_{X\sim P}\left[\log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta})\right]$$

or in the context of iid data $X^n \sim P^n$ and iid modeling $\{P_\theta^n : \theta \in \Theta\}$, its *expected redundancy rate* is denoted

$$\mathcal{R}^{(n)}_{\hat{\theta},\mathcal{L}}(P) := \frac{1}{n}\mathbb{E}_{X^n \overset{iid}{\sim} P}\left[\sum_i \log \frac{p(X_i)}{p_{\hat{\theta}}(X_i)} + \mathcal{L}(\hat{\theta})\right].$$

In penalized maximum likelihood estimation (for instance, in two-part MDL), the estimator is defined to be the minimizer of the quantity inside the expectation. In that case, we can bound the expected redundancy by moving the expectation through the minimum, defining

$$\mathcal{R}_{\Theta,\mathcal{L}}(P) := \inf_{\theta\in\Theta}\{D(P\|P_\theta) + \mathcal{L}(\theta)\}$$

and

$$\mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P) := \inf_{\theta\in\Theta}\left\{D(P\|P_\theta) + \frac{\mathcal{L}(\theta)}{n}\right\}.$$

There may be a $\theta^* \in \Theta$ that minimizes expected $D(P\|P_\theta) + \mathcal{L}(\theta)$; it is the average-case optimal representative from $(\Theta, \mathcal{L})$ when the true distribution is P. Its relative entropy plus penalty is an upper bound for the penalized maximum likelihood estimator's expected redundancy.

Barron and Cover [1991] showed that if the complexity function is large enough, then an estimator's statistical risk is bounded by its expected redundancy. In particular, the penalized likelihood estimator outperforms the best-case average representative; that result for iid modeling is stated in (2.2) below.[4]

will be discussed thereafter.

3. Brinda and Klusowski [2018] only introduces $\mathcal{R}_{\Theta,\mathcal{L}}(P)$ and deals with penalized maximum likelihood estimators. However, more generality is needed for our GRBM risk bound in Theorem 4.0.1.

4. Throughout the paper, we will refer to this inequality as "the resolvability bound," but realize that there are a variety of related resolvability bounds in other contexts. They involve comparing risk to a codelength and lead to bounds that are suboptimal by a $\log n$ factor.

There are a number of attractive features of the resolvability bound; we will highlight four. One of the most powerful aspects of the resolvability bound is the ease with which it can be used to devise adaptive estimation procedures for which the bound applies. For instance, to use a class of nested models rather than a single model, one only needs to tack on an additional penalty term corresponding to a codelength used to specify the selected model within the class.

Another nice feature is its generality: the inequality statement only requires that the data-generating distribution has finite relative entropy to some probability measure in the model.[5] In practice, the common assumptions of other risk bound methods, for instance, that the generating distribution belongs to the model, are unlikely to be exactly true.

A third valuable property of the bound is its exactness for finite samples. Many risk bound methods only provide asymptotic bounds. But such results do not imply anything exact for a data analyst with a specific sample.

Lastly, the resolvability bound uses a meaningful loss function: α-Renyi divergence [Rényi, 1961] with $\alpha \in (0, 1)$. For convenience, we specialize our discussion and our present work to Bhattacharyya divergence [Bhattacharyya, 1943] which is the $\frac{1}{2}$-Renyi divergence.

$$D_B(P, Q) := 2 \log \frac{1}{A(P, Q)}$$

where A denotes the Hellinger affinity

$$\begin{aligned} A(P, Q) &:= \int \sqrt{p(x)q(x)}dx \\ &= \mathbb{E}_{X \sim P} \sqrt{\frac{q(X)}{p(X)}}. \end{aligned}$$

Like relative entropy, D_B decomposes product measures into sums; that is,

$$A(P^n, Q^n) = A(P, Q)^n \qquad \text{thus} \qquad D_B(P^n, Q^n) = nD_B(P, Q).$$

Bhattacharyya divergence is bounded below by squared Hellinger distance (using $\log 1/z \geq$

5. Admittedly, the bound does not have the desired asymptotic behavior when the model is misspecified.

$1-z$) and above by relative entropy (using Jensen's inequality). Importantly, it has a strictly increasing relationship with squared Hellinger distance D_H, which is an f-divergence:

$$D_B = 2 \log \frac{1}{1 - D_H/2}$$

As such, it inherits desirable f-divergence properties such as the data processing inequality. Also, it is clear from the definition that D_B is parametrization-invariant. For many more properties of D_B, including its bound on total variation distance, see van Erven and Harremoës [2014].

Next, we make note of some of the limitations of the resolvability bound. One complaint is that it is for discrete parameter sets, while people generally want to optimize penalized likelihood over a continuous parameter space. In practice, one typically selects a parameter value that is rounded to a fixed precision, so in effect the selection is from a discretized space. However, for mathematical convenience, it is nice to have risk bounds for the theoretical optimizer. A method to extend the resolvability bound to continuous models was introduced by Barron et al. [2008]; in that paper, the method was specialized to estimation of a log density by linear combinations from a finite dictionary with an l_1 penalty on the coefficients. More recently, Chatterjee and Barron worked out the continuous extension for Gaussian graphical models (building on Luo [2009]) with l_1 penalty assuming the model is well-specified and for linear regression with l_0 penalty assuming the true error distribution is Gaussian. These results are explained in more detail by Chatterjee [2014], where the extension for the l_1 penalty for linear regression is also shown, again assuming the true error distribution is Gaussian.

Another limitation is that the resolvability bound needs a large enough penalty; it must have a finite Kraft sum. This paper provides a more general inequality that escapes such a requirement and therefore applies even to unpenalized maximum likelihood estimation. The resulting bound retains the four desirable properties we highlighted above, but loses the coding and resolvability interpretations.

Finally, the resolvability bounds for smooth parametric iid modeling are of order $(\log n)/n$ and cannot be improved, according to Rissanen [1986], whereas under regularity conditions

(for which Bhattacharyya divergence is locally equivalent to one-half relative entropy, according to Barron et al. [2008]) the optimal Bhattacharyya risk is of order $1/n$ [Barron and Hengartner, 1998]. Our variant on the resolvability method leads to the possibility of deriving exact bounds of order $1/n$.

Progress toward weakening the penalty requirements and establishing order $1/n$ risk bounds has previously come from a line of work starting with Zhang [2006]. He established a more general resolvability risk bound for "posterior" distributions on the parameter space. Implications for penalized MLEs come from forcing the "posteriors" to be point-masses. He derives risk bounds that have the form of $\mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P)$ plus a "corrective" term, which is comparable to the form of our results. Indeed, as we will point out, one of our corollaries nearly coincides with [Zhang, 2006, Thm 4.2] but works with arbitrary penalties.

The trick we employ is to introduce an arbitrary function L, which we call a *pseudo-penalty*, that adds to the penalty $\mathcal{L}$; strategic choices of pseudo-penalty can help to control the "penalty summation" over the model. The resulting risk bound has an additional $\mathbb{E}L(\hat{\theta})$ term that must be dealt with.

In Section 2.1, we prove our more general version of the resolvability bound inequality using a derivation closely analogous to the one by Li [1999]. We then explore corollaries that arise from various choices of pseudo-penalty. Section 2.2 extends this thinking to penalized likelihood over continuous models, following the technique from Barron et al. [2008]; a specific result is given for estimating Gaussian location. Chapter 4 explains how our approach applies in the context of adaptive modeling and demonstrates it for GRBMs.

Every result labeled a Theorem or Lemma has a formal proof at the end of this section. Any result labeled a Corollary is an immediate consequence of previously stated results and thus no formal proof is provided. For any random vector X, the notation $\mathbb{C}X$ means the covariance matrix, while $\mathbb{V}X$ represents its trace $\mathbb{E}\|X - \mathbb{E}X\|^2$. The notation $\lambda_j(\cdot)$ means the jth eigenvalue of the matrix argument. Whenever a capital letter has been introduced to represent a probability distribution, the corresponding lower-case letter will represent a density for the measure with respect to either Lebesgue or counting measure. The *penalized MLE* is the (random) parameter that maximizes log-likelihood minus penalty. The notation $D(P\|\Theta)$ represents the infimum relative entropy from P to distributions indexed by the

model Θ.

2.1 Models with countable cardinality

Let us begin with countable (e.g. discretized) models, which were the original context for the MDL penalized likelihood risk bounds. We will show that a generalization of that technique works for arbitrary penalties. The only assumption we need is that for any possible data, there exists a (not necessarily unique) minimizer of penalized likelihood.[6] This existence requirement will be implicit throughout our paper. Theorem 2.1.1 gives a general result that is agnostic about any structure within the data; the consequence for iid data with sample size n is pointed out after the proof.

Theorem 2.1.1. *Let $X \sim P$, and let $\hat{\theta}$ be an estimator over Θ indexing a countable model with penalty $\mathcal{L}$. Then for any $L : \Theta \to \mathbb{R}$,*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}_{\hat{\theta},\mathcal{L}}(P) + 2\log \sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]} + \mathbb{E}L(\hat{\theta}).$$

Suppose now that the data comprise n iid observations and are modeled as such; in other words, the data has the form $X^n \sim P^n$, and the model has the form $\{P_\theta^n : \theta \in \Theta\}$. Because $D_B(P^n, P_{\hat{\theta}}^n) = nD_B(P, P_{\hat{\theta}})$ and $D(P^n\|P_\theta^n) = nD(P\|P_\theta)$, we can divide both sides of Theorem 2.1.1 by n to reveal the role of sample size in this context:

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\hat{\theta},\mathcal{L}}(P) + \frac{2\log \sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]} + \mathbb{E}L(\hat{\theta})}{n}.$$

We will see three major advantages to Theorem 2.1.1. The most obvious is that it can handle cases in which the sum of exponential negative half penalties is infinite; unpenalized estimation, for example, has $\mathcal{L}$ identically zero. One consequence of this is that the resolvability method for minimax risk upper bounds can be extended to models that are not finitely covered by relative entropy balls. We will also find that Theorem 2.1.1 enables

6. We will say "the" penalized MLE, even though we do not require uniqueness; any scheme can be used for breaking ties.

us to derive exact risk bounds of order $1/n$ rather than the usual $(\log n)/n$ resolvability bounds.

In many cases, it is convenient to have only the L function in the summation. Substituting $L - \mathcal{L}$ as the pseudo-penalty in Theorem 2.1.1 gives us a corollary that moves $\mathcal{L}$ out of the summation.

Corollary 2.1.2. *Let $X \sim P$, and let $\hat{\theta}$ be an estimator over Θ indexing a countable model with penalty $\mathcal{L}$. Then for any $L : \Theta \to \mathbb{R}$,*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}_{\hat{\theta},\mathcal{L}}(P) + 2\log \sum_{\theta\in\Theta} e^{-\frac{1}{2}L(\theta)} + \mathbb{E}L(\hat{\theta}) - \mathbb{E}\mathcal{L}(\hat{\theta}).$$

The iid data and model version is

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P) + \frac{2\log \sum_{\theta\in\Theta} e^{-\frac{1}{2}L(\theta)} + \mathbb{E}L(\hat{\theta}) - \mathbb{E}\mathcal{L}(\hat{\theta})}{n}.$$

We will use the term *pseudo-penalty* for the function labeled L in *either* Theorem 2.1.1 or Corollary 2.1.2. Note that L is allowed to depend on P but not on the data.

A probabilistic loss bound can also be derived for the difference between the loss and the redundancy plus pseudo-penalty.

Theorem 2.1.3. *Let $X \sim P$, and let $\hat{\theta}$ be an estimator over Θ indexing a countable model with penalty $\mathcal{L}$. Then for any $L : \Theta \to \mathbb{R}$,*

$$P\left\{D_B(P, P_{\hat{\theta}}) - \left[\log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta}) + L(\hat{\theta})\right] \geq t\right\} \leq e^{-t/2} \sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]}.$$

For iid data $X^n \overset{iid}{\sim} P$ and an iid model, Theorem 2.1.3 implies

$$P\left\{D_B(P, P_{\hat{\theta}}) - \frac{1}{n}\left[\sum_i \log \frac{p(X_i)}{p_{\hat{\theta}}(X_i)} + \mathcal{L}(\hat{\theta}) + L(\hat{\theta})\right] \geq t\right\} \leq e^{-nt/2} \sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]}.$$

Several of our corollaries have $\mathcal{L}$ and L designed to make $\sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]} \leq 1$. In such cases, the difference between loss and the point-wise redundancy plus pseudo-penalty is stochastically less than an exponential random variable.

Often the countable model of interest is a discretization of a continuous model. Given any $\epsilon > 0$, an *ϵ-discretization* of $\mathbb{R}^d$ is $v + \epsilon\mathbb{Z}^d$, by which we mean $\{v + m\epsilon : m \in \mathbb{Z}^d\}$ for some $v \in \mathbb{R}^d$. An *ϵ-discretization* of $\Theta \subseteq \mathbb{R}^d$ is a set of the form $\Theta \cap (v + \epsilon\mathbb{Z}^d)$. A discussion of the behavior of $\mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P)$ in that context is provided later in this section.

2.1.1 Choices of pseudo-penalty

To derive useful consequences of the above results, we will explore some convenient choices of pseudo-penalty: zero, Bhattacharyya divergence, log reciprocal pmf of $\hat{\theta}$, quadratic forms, and the penalty. We specialize to the iid data and model setting for the remainder of this chapter to highlight the fact that many of the exact risk bounds we derive are of order $1/n$ in that case; we also specialize to penalized likelihood estimators, since practical bounds on expected redundancy are immediate. However, our main theorem for GRBMs (Theorem 4.0.1) relies on the expected redundancy version.

Zero as pseudo-penalty

Setting L to zero gives us the traditional resolvability bound, which we review in this section.

Corollary 2.1.4. *Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over Θ indexing a countable iid model with penalty $\mathcal{L}$. Then*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P) + \frac{2\log\sum_{\theta\in\Theta} e^{-\frac{1}{2}\mathcal{L}(\theta)}}{n}.$$

The usual statement of the resolvability bound [Barron et al., 2008] assumes $\mathcal{L}$ is at least twice a codelength function, so that it is large enough for the sum of exponential terms to be no greater than 1. That is,

$$\sum_{\theta\in\Theta} e^{-\frac{1}{2}\mathcal{L}(\theta)} \leq 1 \tag{2.1}$$

implies

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P). \tag{2.2}$$

The quantity on the right-hand side of (2.2) is called the *index of resolvability* of $(\Theta, \mathcal{L})$ for P at sample size n. Any corresponding minimizer $\theta^* \in \Theta$ is considered to index an average-case optimal representative for P at sample size n.

In fact, for any finite sum $z := \sum_{\theta\in\Theta} e^{-\frac{1}{2}\mathcal{L}(\theta)}$, the maximizer of the penalized likelihood is also the maximizer with penalty $\tilde{\mathcal{L}} := \mathcal{L} + 2\log z$. Thus one has a resolvability bound of the form (2.2) with the equivalent penalty $\tilde{\mathcal{L}}$, which satisfies (2.1) with equality.

Additionally, the resolvability bounds give an exact upper bound on the minimax risk for any model Θ that can be covered by finitely many relative entropy balls of radius ϵ^2; the log of the minimal covering number is called the *KL-metric entropy* $\mathcal{M}(\epsilon)$. These balls' center points are called a *KL-net*; we will denote the net by Θ_ϵ. With data $X^n \overset{iid}{\sim} P_{\theta^*}$ for any $\theta^* \in \Theta$, the MLE restricted to Θ_ϵ has the resolvability risk bound

$$\begin{aligned}
\mathbb{E}D_B(P_{\theta^*}, P_{\hat{\theta}}) &\leq \inf_{\theta\in\Theta_\epsilon} \left\{ D(P_{\theta^*}\|P_\theta) + \frac{2\mathcal{M}(\epsilon)}{n} \right\} \\
&= \inf_{\theta\in\Theta_\epsilon} D(P_{\theta^*}\|P_\theta) + \frac{2\mathcal{M}(\epsilon)}{n} \\
&\leq \epsilon^2 + \frac{2\mathcal{M}(\epsilon)}{n}.
\end{aligned}$$

If an explicit bound for $\mathcal{M}(\epsilon)$ is known, then the overall risk bound can be optimized over the radius ϵ — see for instance [Barron et al., 2008, Section 1.5].

Because this approach to upper bounding minimax risk requires twice-Kraft-valid code-lengths, it only applies to models that can be covered by finitely many relative entropy balls. However, Corollary 2.1.2 reveals new possibilities for establishing minimax upper bounds even if the cover is infinite. Given any L, one can use any constant penalty that is at least as large as $2\log\sum e^{-\frac{1}{2}L(\theta)} + \mathbb{E}L(\hat{\theta})$ where $\hat{\theta}$ is the unpenalized MLE on the net and the summation is taken over those points.[7] For a minimax result, one still needs this quantity to be uniformly bounded over all data-generating distribution $\theta^* \in \Theta$.

7. Putting $\mathcal{L} = 0$ into either Theorem 2.1.1 or Corollary 2.1.2 would give us the same idea.

Bhattacharyya divergence as pseudo-penalty

Important corollaries[8] to Theorems 2.1.1 and 2.1.2 come from setting the pseudo-penalty equal to $\alpha D_B(P, P_\theta)$; the expected pseudo-penalty is proportional to the risk, so that term can be subtracted from both sides. For the iid scenario, we also use the product property of Hellinger affinity: $A(P^n, P_\theta^n) = A(P, P_\theta)^n$.

Corollary 2.1.5. *Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over Θ indexing a countable iid model with penalty $\mathcal{L}$. Then for any $\alpha \in [0, 1]$,*

$$\mathbb{E} D_B(P, P_{\hat{\theta}}) \leq \frac{1}{1-\alpha} \left[\mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P) + \frac{2 \log \sum_{\theta \in \Theta} e^{-\frac{1}{2}\mathcal{L}(\theta)} A(P, P_\theta)^{\alpha n}}{n} \right].$$

Corollary 2.1.6. *Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over Θ indexing a countable iid model with penalty $\mathcal{L}$. Then for any $\alpha \in [0, 1]$,*

$$\mathbb{E} D_B(P, P_{\hat{\theta}}) \leq \frac{1}{1-\alpha} \left[\mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P) + \frac{2 \log \sum_{\theta \in \Theta} A(P, P_\theta)^{\alpha n} - \mathbb{E}\mathcal{L}(\hat{\theta})}{n} \right].$$

For simplicity, the corollaries throughout this subsection will use $\alpha = 1/2$.

Consider a penalized MLE selected from an ϵ-discretization of a continuous parameter space; as the sample size increases, one typically wants to shrink ϵ to make the grid more refined. Examining Corollaries 2.1.5 and 2.1.6, we see two opposing forces at work as n increases: the grid-points themselves proliferate, while the nth power depresses the terms in the summation. For more details, including application to location families, see [Brinda and Klusowski, 2018, Sec 2.2].

Log reciprocal pmf of $\hat{\theta}$ as pseudo-penalty

The Bhattacharyya pseudo-penalty had an expectation that was easy to handle; we only had to worry about the resulting log summation. Now we will select a pseudo-penalty with the opposite effect. We can eliminate Corollary 2.1.2's log summation term by letting L be twice a codelength function. The smallest resulting $\mathbb{E}L(\hat{\theta})$ comes from setting L to be

8. Our Corollary 2.1.5 was inspired by the very closely related result of [Zhang, 2006, Thm 4.2].

two times the log reciprocal of the probability mass function of $\hat{\theta}$. This expectation is the Shannon entropy H of the penalized MLE's distribution (i.e. the image measure of P under the Θ-valued deterministic transformation $\hat{\theta}$).

Corollary 2.1.7. *Let $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be a penalized MLE over all $\theta \in \Theta$ indexing a countable iid model. Then*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P) + \frac{2H(\hat{\theta}) - \mathbb{E}\mathcal{L}(\hat{\theta})}{n}.$$

It is known that the risk of the MLE is bounded by the log-cardinality of the model (e.g. Li [1999]); Corollary 2.1.7 implies a generalization of this fact for penalized MLEs:

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P) + \frac{2\log|\Theta| - \mathbb{E}\mathcal{L}(\hat{\theta})}{n}.$$

Importantly, Corollary 2.1.7 also applies to models of infinite cardinality.

Quadratic form as pseudo-penalty

Other simple corollaries come from using a quadratic pseudo-penalty $L(\theta) = (\theta - \mathbb{E}\hat{\theta})' M (\theta - \mathbb{E}\hat{\theta})$ for some positive definite matrix M. The expected pseudo-penalty is then

$$\mathbb{E}L(\hat{\theta}) = \operatorname{tr} M \mathbb{C}\hat{\theta}$$

where $\mathbb{C}\hat{\theta}$ denotes the covariance matrix of the random vector $\hat{\theta}(X^n)$ with $X^n \overset{iid}{\sim} P$. For the log summation term, we note that

$$\begin{aligned}\sum_{\theta_\epsilon \in \Theta_\epsilon} e^{-(\theta_\epsilon - \mathbb{E}\hat{\theta})' M (\theta_\epsilon - \mathbb{E}\hat{\theta})} &\leq \sum_{\theta_\epsilon \in \Theta_\epsilon} e^{-\lambda_d(M)\|\theta_\epsilon - \mathbb{E}\hat{\theta}\|^2} \\ &\leq \left(1 + \frac{2\sqrt{\pi}}{\epsilon\sqrt{\lambda_d(M)}}\right)^d\end{aligned}$$

by Lemma 2.3.6. Using αI_d as M gives us Corollary 2.1.8.

Corollary 2.1.8. *Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over an ϵ-discretization*

$\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$ *indexing an iid model with penalty* $\mathcal{L}$*. Then for any* $\alpha \geq 0$*,*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta_\epsilon, \mathcal{L}}(P) + \frac{2d \log(1 + \frac{2\sqrt{\pi}}{\epsilon\sqrt{\alpha}}) + \alpha \mathbb{V}\hat{\theta} - \mathbb{E}\mathcal{L}(\hat{\theta})}{n}.$$

As described in Section 2.2, one gets desirable order $1/n$ behavior from $\mathcal{R}^{(n)}_{\Theta_\epsilon, \mathcal{L}}(P)$ by using ϵ proportional to $1/\sqrt{n}$. For either of these two corollaries above to have order $1/n$ bounds, the numerator of the second term should be stable in n. In Corollary 2.1.8, one sets α proportional to $1/\epsilon^2$ and thus needs $\mathbb{V}\hat{\theta}$ to have order $1/n$. In many cases, such as ordinary MLE with an exponential family, the covariance matrix of the optimizer over Θ is indeed bounded by a matrix divided by n. However, one still needs to handle the discrepancy in behavior between the continuous and discretized estimator.

In a sense, Corollary 2.1.8 shifts the bounding problem to another risk-related quantity, while the pseudo-penalties used in the Bhattacharyya pseudo-penalty and log reciprocal pmf pseudo-penalty provide more direct ways of deriving exact risk bounds of order $1/n$.

Penalty as pseudo-penalty

Another simple corollary to Theorem 2.1.1 uses $L = \alpha\mathcal{L}$.

Corollary 2.1.9. *Assume* $X^n \overset{iid}{\sim} P$*, and let* $\hat{\theta}$ *be the penalized MLE over* Θ *indexing a countable iid model with penalty* $\mathcal{L}$*. Then*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta, \mathcal{L}}(P) + \frac{2 \log \sum_{\theta \in \Theta} e^{-\frac{\alpha+1}{2}\mathcal{L}(\theta)} + \alpha \mathbb{E}\mathcal{L}(\hat{\theta})}{n}.$$

Bayesian MAP (maximum a posteriori) is a common penalized likelihood procedure that has insufficient penalty for the index of resolvability bound (2.2) to be valid. In that case, Corollary 2.1.4 applies (where $\mathcal{L}$ comprises the logs of the reciprocals of prior masses), but the sum of exponential terms may be infinite. An alternative approach comes from Corollary 2.1.9 by setting $\alpha = 1$.

Corollary 2.1.10. *Assume* $X^n \overset{iid}{\sim} P$*, and let* $\hat{\theta}$ *be the MAP estimate over* Θ *indexing a*

countable iid model with prior pmf q. Then

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta,\log 1/q}(P) + \frac{\mathbb{E}\log(1/q(\hat{\theta}))}{n}.$$

For ϵ-discretizations, realize that q has to change as the refinement increases; thus the second term in Corollary 2.1.10 should be considered to have order strictly larger than $1/n$ in that context.

2.1.2 Simple concrete examples

When Θ indexes an exponential family, for any P_θ in the family there is a "Pythagorean" information identity

$$D(P\|P_\theta) = D(P\|P_{\theta^*}) + D(P_{\theta^*}\|P_\theta)$$

where P_{θ^*} is the rI-projection of P onto Θ; if there is the distribution in the family that agrees with P about the expectation of the sufficient statistic, that distribution is the rI-projection — see [Csiszár and Matúš, 2003, Thm 3 and Cor 6]. In such cases,

$$\begin{aligned}\mathcal{R}_{\Theta,\mathcal{L}}(P) &:= \inf_{\theta\in\Theta}\{D(P\|P_\theta) + \mathcal{L}(\theta)\} \\ &= D(P\|P_{\theta^*}) + \inf_{\theta\in\Theta}\{D(P_{\theta^*}\|P_\theta) + \mathcal{L}(\theta)\} \\ &= D(P\|P_{\theta^*}) + \mathcal{R}_{\Theta,\mathcal{L}}(P_{\theta^*}).\end{aligned}$$

This also holds when the model is a submodel Θ_ϵ (e.g. a discretization) of an exponential family Θ.

$$\mathcal{R}_{\Theta_\epsilon,\mathcal{L}}(P) = D(P\|P_{\theta^*}) + \mathcal{R}_{\Theta_\epsilon,\mathcal{L}}(P_{\theta^*})$$

One consequence of this is that the rI-projection of P onto Θ_ϵ is the same as the rI-projection of P_{θ^*} onto Θ_ϵ; to see why, consider the above identity with $\mathcal{L} = 0$.

Suppose $X^n \overset{iid}{\sim} P$ are real-valued observations, and P is estimated by a penalized MLE

of the standard ϵ-discretized Gaussian location model

$$\{P_\theta = N(\theta, 1/s) : \theta \in \Theta_\epsilon = \epsilon\mathbb{Z}\} \subset \{P_\theta = N(\theta, 1/s) : \theta \in \Theta = \mathbb{R}\}.$$

Suppose a squared norm penalty[9] is used: $\mathcal{L}(\theta) = \frac{s_0}{2}\|\theta\|^2$ for $s_0 \geq 0$. If $s = 0$, the estimator is the ordinary MLE; otherwise, the resulting penalized MLE is the Bayesian MAP when using the discretized Gaussian $N(0, 1/s_0)$ as the prior for θ. According to the preceding paragraph,

$$\mathcal{R}^{(n)}_{\Theta_\epsilon,\mathcal{L}}(P) = D(P\|P_{\theta^*}) + \inf_{\theta\in\Theta_\epsilon}\{\tfrac{s}{2}(\theta - \mathbb{E}X)^2 + \tfrac{s_0}{2n}\theta^2\}.$$

The quantity inside the infimum is minimized at $\frac{s}{s+s_0/n}\mathbb{E}X \in \mathbb{R}$. Because the penalized likelihood is unimodal, the minimizer *on the grid* is a neighboring grid-point,[10] which is within ϵ of the true optimizer. Using this fact, it is straightforward to derive a bound on the infimum.

$$\begin{aligned}\inf_{\theta\in\Theta_\epsilon}\{\tfrac{s}{2}(\theta - \mathbb{E}X)^2 + \tfrac{s_0}{2n}\theta^2\} &\leq \frac{s}{2}\left[\left(1 - \frac{s}{s+s_0/n}\right)^2(\mathbb{E}X)^2 + 2\left(1 - \frac{s}{s+s_0/n}\right)\epsilon|\mathbb{E}X| + \epsilon^2\right] \\ &\quad + \frac{s_0}{2n}\left[\left(\frac{s}{s+s_0/n}\right)^2(\mathbb{E}X)^2 + 2\left(\frac{s}{s+s_0/n}\right)\epsilon|\mathbb{E}X| + \epsilon^2\right] \\ &\leq \left[\frac{s_0^2/s}{n^2} + \frac{s_0}{n}\right](\mathbb{E}X)^2 + \frac{2s_0\epsilon}{n}|\mathbb{E}X| + \frac{s_0\epsilon^2}{n} + s\epsilon^2\end{aligned} \tag{2.3}$$

We will compare corrective terms that arise from three of this section's corollaries for bounding risk. First, Corollary 2.1.4 implies that a bound can be acquired by adding to

9. We use the *squared norm* notation to suggest extensions to more general $\mathbb{R}^d$, even though this example is limited to $\mathbb{R}$.

10. Since the penalized likelihood is quadratic in this case, the grid's minimzer is *the nearest grid-point*, which is within $\epsilon/2$ of the true optimizer. However, we will content ourselves with ϵ since that represents a more typical scenario.

$\mathcal{R}^{(n)}_{\Theta_\epsilon,\mathcal{L}}(P)$ the corrective term

$$\frac{2}{n}\log\sum_{\theta\in\Theta_\epsilon} e^{-\frac{s_0}{4}\theta^2} \leq \frac{2}{n}\log(1+2\sqrt{\pi}/\epsilon\sqrt{s_0}).$$

The Gaussian summation was bounded using Lemma 2.3.5. Overall,

$$\begin{aligned}\mathbb{E}D_B(P,P_{\hat{\theta}}) \leq D(P\|\Theta) + \frac{s_0(\mathbb{E}X)^2 + 2s_0\epsilon|\mathbb{E}X| + s_0\epsilon^2 + 2\log(1+4/\epsilon\sqrt{s_0})}{n} \\ + \frac{s_0^2(\mathbb{E}X)^2/s}{n^2} + s\epsilon^2.\end{aligned}$$

Secondly, we use the quadratic pseudo-penalty approach to bounding risk after upper bounding the variance of the estimator and lower bounding the expected penalty.

When the penalized likelihood is unimodal over a one-dimensional parameter space, the variance of the *grid's* optimizer $\hat{\theta}$ can be conveniently bounded in terms of the variance of the continuous model's optimizer $\hat{\theta}'$. Define $\delta := \hat{\theta} - \hat{\theta}'$, which has absolute value no greater than ϵ because of unimodality.

$$\begin{aligned}\mathbb{V}\hat{\theta} &= \mathbb{E}\|(\hat{\theta}'+\delta) - \mathbb{E}(\hat{\theta}'+\delta)\|^2 \\ &\leq 2\mathbb{E}\|\hat{\theta}' - \mathbb{E}\hat{\theta}'\|^2 + 2\mathbb{E}\|\delta - \mathbb{E}\delta\|^2 \\ &\leq 2\mathbb{V}\hat{\theta}' + 2\epsilon^2\end{aligned}$$

using Lemma 2.3.7 and the fact that the variance of a bounded random variable is at most half its range.

In our present case, the continuous optimizer is $\frac{s}{s+s_0/n}\bar{X}_n$ (and the MLE $\hat{\theta}$ is the grid-point closest to $\bar{X}_n$), so its variance is $(\frac{s}{s+s_0/n})^2\mathbb{V}X/n$ for $X \sim P$. For simplicity, we use the upper bound $\mathbb{V}X/n$.

The expected penalty is lower bounded using Lemma 2.3.8, which will be stated after

this example.

$$\begin{aligned}
\mathbb{E}\mathcal{L}(\hat{\theta}) &= \frac{s_0}{2}\mathbb{E}\|\hat{\theta}' + \delta\|^2 \\
&\geq \frac{s_0}{2}\mathbb{E}(\|\hat{\theta}'\| - \epsilon)^2 \\
&\geq \frac{s_0}{2}\mathbb{E}[(\hat{\theta}')^2 - 2\epsilon^2(\hat{\theta}')^2 - 2\epsilon^4 - \epsilon^2 - 1] \\
&\geq \frac{s_0}{2}[(\mathbb{E}\hat{\theta}')^2 - 2\epsilon^2\mathbb{E}(\hat{\theta}')^2 - 2\epsilon^4 - \epsilon^2 - 1] \\
&= \frac{s_0}{2}[(\mathbb{E}\hat{\theta}')^2 - 2\epsilon^2[(\mathbb{E}\hat{\theta}')^2 + \mathbb{V}\hat{\theta}'] - 2\epsilon^4 - \epsilon^2 - 1]
\end{aligned}$$

Since $\mathbb{E}\mathcal{L}(\hat{\theta})/n$ gets subtracted in the risk bound, the first term can be used to eliminate the $(\mathbb{E}X)^2/n$ term in (2.3).

Corollary 2.1.8 with $\alpha = 1/\epsilon^2$ implies the exact risk bound[11]

$$\begin{aligned}
\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq D(P\|\Theta) + \frac{s_0\epsilon^2(\mathbb{E}X)^2 + \frac{1+s_0\epsilon^4}{\epsilon^2 n}\mathbb{V}X + 2s_0\epsilon|\mathbb{E}X| + 4 + s_0 + (1+s_0)\epsilon^2 + \epsilon^4}{n} \\
+ \frac{s_0^2(\mathbb{E}X)^2/s}{n^2} + s\epsilon^2.
\end{aligned}$$

Lastly, we will try Corollary 2.1.9 with $\alpha = 1$. There is a log summation term

$$2\log\sum_{\theta\in\Theta_\epsilon} e^{-\frac{s_0}{2}\|\theta\|^2} \leq 2\log(1 + \sqrt{2\pi}/\epsilon\sqrt{s_0}).$$

And we upper bound the expected penalty by

$$\begin{aligned}
\mathbb{E}\mathcal{L}(\theta) &= \frac{s_0}{2}\mathbb{E}(\hat{\theta}' + \delta)^2 \\
&\leq s_0[\mathbb{E}(\hat{\theta}')^2 + \epsilon^2] \\
&= s_0[(\mathbb{E}X)^2 + \mathbb{V}X/n + \epsilon^2]
\end{aligned}$$

11. This risk bound monotonically increases in s_0; it fails to capture any trade-off in the severity of the penalty.

Overall,

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq D(P\|\Theta) + \frac{2s_0(\mathbb{E}X)^2 + 2s_0\epsilon|\mathbb{E}X| + 2s_0\epsilon^2 + 2\log(1 + 3/\epsilon\sqrt{s_0})}{n} + \frac{s_0^2(\mathbb{E}X)^2/s + \mathbb{V}X}{n^2} + s\epsilon^2.$$

We demonstrate Corollary 2.1.5 with the exponential distributions $p_\theta(x) = \theta e^{-\theta x}\mathbb{I}\{x \geq 0\}$, defined for $\theta > 0$. If the data-generating distribution is some P_{θ^*} in the model, then the Hellinger affinity is the geometric average of the parameters divided by their arithmetic average.

$$\begin{aligned} A(P_{\theta^*}, P_\theta) &= \int_0^\infty \sqrt{\theta^*} e^{-\theta^* x/2} \sqrt{\theta} e^{-\theta x/2} dx \\ &= \frac{\sqrt{\theta^*}\sqrt{\theta}}{\frac{1}{2}\theta^* + \frac{1}{2}\theta} \end{aligned}$$

For simplicity, let us confine our eventual choice of ϵ to be no greater than 1, in which case, we can bound $A(P_{\theta^*}, P_\theta)^{1/\epsilon^2}$ by $A(P_{\theta^*}, P_\theta)$. (This approach will not allow us to avoid the $\log n$ in the numerator of the risk bound, but it does provide a clean demonstration of the usefulness of Corollary 2.1.5.)

Without a penalty, these Hellinger affinities will create an unbounded summation term over the grid $\{k\epsilon : k \geq 1\}$. The penalty $\mathcal{L}(\theta) = 2\log\theta$ gives a clean bound that works uniformly for $\theta^* > 0$.

$$\begin{aligned} \sum_{\theta\in\Theta} e^{-\frac{1}{2}\mathcal{L}(\theta)} A(P, P_\theta) &= \sum_{\theta\in\Theta} \left(\frac{1}{\theta}\right)\left(\frac{2\sqrt{\theta^*}\sqrt{\theta}}{\theta^* + \theta}\right) \\ &= 2\sqrt{\theta^*} \sum_{\theta\in\Theta} \frac{1}{\sqrt{\theta}(\theta^* + \theta)} \\ &= 2\sqrt{\theta^*} \sum_{k\geq 1} \frac{1}{\sqrt{k\epsilon}(\theta^* + k\epsilon)} \\ &\leq 2\sqrt{\theta^*} \int_0^\infty \frac{1}{\sqrt{t}(\theta^* + t)} d(t/\epsilon) \\ &= \frac{2\sqrt{\theta^*}}{\epsilon}\left(\frac{2\tan^{-1}(\sqrt{t}/\sqrt{\theta^*})|_0^\infty}{\sqrt{\theta^*}}\right) \\ &= 2\pi/\epsilon \end{aligned}$$

Thus, with $\mathcal{L}(\theta) = 2\log\theta$, the exponential distribution in the ϵ-discretized grid that maximizes penalized likelihood has the risk bound

$$\begin{aligned}\mathbb{E}_{X^n \overset{iid}{\sim} P_{\theta^*}} d(P_{\theta^*}, P_{\hat{\theta}}) &\leq \frac{1}{1-\epsilon^2 n}\left[\inf_{\theta\in\Theta}\left\{D(P_{\theta^*}\|P_\theta) + \frac{2\log\theta}{n}\right\} + \frac{2\log 2\pi/\epsilon}{n}\right] \\ &= \frac{1}{1-\epsilon^2 n}\left[\inf_{\theta\in\Theta}\left\{\log\frac{\theta^*}{\theta} + \frac{\theta-\theta^*}{\theta^*} + \frac{2}{n}\log\theta\right\} + \frac{2}{n}\log\frac{2\pi}{\epsilon}\right].\end{aligned}$$

A rough bound on the optimum value of the objective can be obtained from Theorem 2.3.3.[12] It is straight-forward to derive $I_{P_{\theta^*}}(\theta) = 1/\theta^2$, which implies

$$\mathbb{E}_{X^n \overset{iid}{\sim} P_{\theta^*}} d(P_{\theta^*}, P_{\hat{\theta}}) \leq \frac{1}{1-\epsilon^2 n}\left[\frac{\epsilon^2}{(\theta^*-\epsilon)^2} + \frac{2}{n}\log\theta^* + \frac{2}{n}\log\frac{2\pi}{\epsilon}\right]$$

as long as $\theta^* > \epsilon$ (which will eventually hold). Using $\epsilon = 1/\sqrt{2n}$ results in an order $(\log n)/n$ bound.

$$\mathbb{E}_{X^n \overset{iid}{\sim} P_{\theta^*}} d(P_{\theta^*}, P_{\hat{\theta}}) \leq \frac{1}{n}\left[\frac{1}{(\theta^* - 1/\sqrt{2n})^2} + 9 + 4\log\theta^* + 2\log n\right]$$

And because this penalized relative entropy is unimodal, the optimizer on the grid will be within ϵ of $\frac{n-2}{n}\theta^*$. The resulting bound is

$$\begin{aligned}\mathbb{E}_{X^n \overset{iid}{\sim} P_{\theta^*}} d(P_{\theta^*}, P_{\hat{\theta}}) &\leq \frac{1}{1-\epsilon^2 n}\left[\frac{\frac{n-2}{n}\theta^* + \epsilon}{\theta^*} - 1 - \log\frac{[\frac{n-2}{n}\theta^* - \epsilon]^{1-2/n}}{\theta^*} + \frac{2}{n}\log\frac{2\pi}{\epsilon}\right] \\ &\leq \frac{1}{1-\epsilon^2 n}\left[\frac{\epsilon}{\theta^*} - \log\frac{[\frac{n-2}{n}\theta^* - \epsilon]^{1-2/n}}{\theta^*} + \frac{2}{n}\log\frac{2\pi}{\epsilon}\right] \\ &\leq \frac{2}{n-1}\log 2\pi\theta^* + \frac{\epsilon}{\theta^*} + \frac{2}{n-1}\log\frac{1}{\epsilon} - \frac{n-2}{n}\log\left[\frac{n-2}{n} - \frac{\epsilon}{\theta^*}\right].\end{aligned}$$

Using ϵ of order $1/(n-1)$ produces $\frac{1}{n-1}\log(n-1)$ convergence. While this example does not provide order $1/n$ bounds, it does show how Corollary 2.1.5 can have advantages over our other corollaries. The procedure described here is exactly Bayesian MAP with prior proportional to $1/\theta^2$. However, Corollary 2.1.4 is no use, as half the penalty is not Kraft-summable, and Corollary 2.1.10 is unable to provide a bound on the part outside the

12. To be more exact, the optimal θ over the continuum is $\frac{n-2}{n}\theta^*$, when $n \geq 2$.

infimum that works uniformly over $\theta^* > 0$.

2.2 Continuous parameter spaces

An analogue of Theorem 2.1.1 holds when the optimization is taken over a continuous parameter space. As prescribed in Barron et al. [2008], one can construct a discrete grid $\Theta_\epsilon \subseteq \Theta$, then try to bound the discrepancy between the penalized MLE and a selection from the grid.

Section 2.1 dealt with discrete parameter spaces, so we did not have to worry about the measurability of $\hat{\theta}$ or functions thereof. We do not much concern ourselves with measurability in the present continuous context either. Zhang [2006] points to *outer measure* approaches in the empirical process literature [van der Vaart and Wellner, 1996, Ch 2] as a justification for side-stepping measurability questions, but we feel that a simpler and more powerful convention is possible. Chapter C describes a generalization of the concept of *probability measure* that affirms the reality of our inequalities and deems measurability a secondary concern.

Theorem 2.2.1. *Let $X \sim P$, and let $\hat{\theta}$ be an estimator over Θ indexing a model. Given any countable subset $\Theta_\epsilon \subseteq \Theta$,*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}_{\hat{\theta},\mathcal{L}}(P) + 2\log \sum_{\theta_\epsilon \in \Theta_\epsilon} e^{-\frac{1}{2}[\mathcal{L}(\theta_\epsilon)+L(\theta_\epsilon)]} + \mathbb{E}L(\hat{\theta}) \\ + 2\mathbb{E}\left[\log \frac{\sqrt{p_{\hat{\theta}}(X)}e^{-\frac{1}{2}[\mathcal{L}(\hat{\theta})+L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} - \sup_{\theta_\epsilon \in \Theta_\epsilon} \log \frac{\sqrt{p_{\theta_\epsilon}(X)}e^{-\frac{1}{2}[\mathcal{L}(\theta_\epsilon)+L(\theta_\epsilon)]}}{A(P, P_{\theta_\epsilon})}\right].$$

The bound is also true when expectation and infimum are interchanged in the discrepancy term. If there is a maximizing grid point, then it can be thought of as an optimal representer for P from Θ_ϵ.

As before, we can state a corollary that subtracts the expected penalty rather than involving it in the summation.

Corollary 2.2.2. *Let $X \sim P$, and let $\hat{\theta}$ be an estimator over Θ indexing a model. Given*

any countable subset $\Theta_\epsilon \subseteq \Theta$,

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}_{\hat{\theta},\mathcal{L}}(P) + 2\log \sum_{\theta_\epsilon \in \Theta_\epsilon} e^{-\frac{1}{2}L(\theta_\epsilon)} + \mathbb{E}L(\hat{\theta}) - \mathbb{E}\mathcal{L}(\hat{\theta}) \\ + 2\mathbb{E}\left[\log \frac{\sqrt{p_{\hat{\theta}}(X)}e^{-\frac{1}{2}L(\hat{\theta})}}{A(P, P_{\hat{\theta}})} - \sup_{\theta_\epsilon \in \Theta_\epsilon} \log \frac{\sqrt{p_{\theta_\epsilon}(X)}e^{-\frac{1}{2}L(\theta_\epsilon)}}{A(P, P_{\theta_\epsilon})}\right].$$

Also as before, if $X^n \overset{iid}{\sim} P$ and the data are modeled as iid, then we can divide both sides of the inequality by n to see that

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\hat{\theta},\mathcal{L}} + \frac{2\log \sum_{\theta_\epsilon \in \Theta_\epsilon} e^{-\frac{1}{2}[\mathcal{L}(\theta_\epsilon)+L(\theta_\epsilon)]} + \mathbb{E}L(\hat{\theta})}{n} \\ + \frac{2}{n}\mathbb{E}\left[\log \frac{\sqrt{p_{\hat{\theta}}(X^n)}e^{-\frac{1}{2}[\mathcal{L}(\hat{\theta})+L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})^n} - \sup_{\theta_\epsilon \in \Theta_\epsilon} \log \frac{\sqrt{p_{\theta_\epsilon}(X^n)}e^{-\frac{1}{2}[\mathcal{L}(\theta_\epsilon)+L(\theta_\epsilon)]}}{A(P, P_{\theta_\epsilon})^n}\right].$$

Likewise, an iid version of Corollary 2.2.2 can be stated. A probabilistic loss bound analogous to Theorem 2.1.3 holds as well; see Theorem 4.0.1 for an example.

A summation over grid points is usually harder to work out than the analogous integral would be. But notice that in this case, the grid plays no role in the estimation; it is only a constructed for the sake of the analysis. It does not need to exactly coincide with an ϵ-discretization. This provides us with an opportunity to design the grid such that the summation can be replaced by an integral; this trick works out most neatly when $\Theta = \mathbb{R}^d$. Suppose $f : \Theta \to \mathbb{R}$ is continuous. Then by the mean value theorem, any hypercube h in Θ of side-length ϵ has at least one point θ_h whose value at the function $f(\theta_h)$ equals the average value of the function over h.

$$f(\theta_h) = \frac{1}{\epsilon^d}\int_h f(\theta)d\theta$$

If one uses these θ_h as the grid points, then the summation is proportional to the integral.

$$\sum_h f(\theta_h) = \sum_h \frac{1}{\epsilon^d}\int_h f(\theta)d\theta \\ = \frac{1}{\epsilon^d}\int_\Theta f(\theta)d\theta$$

We will call any such grid an *integration grid* for f. The distance from any point to its farthest neighboring grid-point is at most $2\epsilon\sqrt{d}$.

One approach for handling the discrepancy is to bound the supremum over the grid by the expectation with respect to a distribution on neighboring grid-points such that $\hat{\theta}$ is the mean. The trick from Lemma 2.3.2 is an option if one can bound the expectation of the largest eigenvalue of the discrepancy's Hessian near $\hat{\theta}$.

Consider using a quadratic pseudo-penalty of the form $L(\theta) = (\theta - \mathbb{E}\hat{\theta})' M (\theta - \mathbb{E}\hat{\theta})$. The resulting $\mathbb{E}L(\hat{\theta})$ term is $\operatorname{tr} M\mathbb{C}\hat{\theta}$. With an integration grid, the relevant integral is

$$\int_{\mathbb{R}^d} e^{-\frac{1}{2}(\theta - \mathbb{E}\hat{\theta})' M (\theta - \mathbb{E}\hat{\theta})} = (2\pi)^{d/2} |M|^{-1/2}$$

The Hessian matrix $\frac{1}{2}\nabla\nabla' L(\theta)$ is simply M. The following result comes from using a quadratic pseudo-penalty with $M = \alpha I_d$. Note that adding αI_d to a matrix adds α to each of its eigenvalues and therefore adds α to its largest eigenvalue.

Corollary 2.2.3. *Let $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over $\Theta = \mathbb{R}^d$ indexing an iid model. If $\log A(P, P_\theta)/\sqrt{p_\theta}$ is twice continuously differentiable in θ, then for any $\alpha, \epsilon > 0$,*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^{(n)}_{\Theta,\mathcal{L}}(P) - \frac{\mathbb{E}\mathcal{L}(\hat{\theta})}{n} + \frac{d}{n}\left[\log\frac{2\pi}{\alpha\epsilon^2} + \alpha\frac{\mathbb{V}\hat{\theta}}{d} + \epsilon^2\alpha + 4\epsilon^2\mathbb{E}\sup_{\tilde{\theta}\in B(\hat{\theta},\epsilon\sqrt{d})} \lambda_1\left(\nabla\nabla' \log\frac{A(P, P_{\hat{\theta}})^n}{\sqrt{p_{\hat{\theta}}(X^n)}}\right)_+\right].$$

In exponential family models, a condition on the sufficient statistic can guarantee Corollary 2.2.3's smoothness conditions for both p_θ and $A(P, P_\theta)$. Additionally, the Hessian simplifies in an interesting way as seen in the following identity. (The reader can refer to Chapter A for the definition of "geometric mixture.")

Lemma 2.2.4. *Let θ be the natural parameter vector of an exponential family with an open and convex parameter space and a twice continuously differentiable sufficient statistic vector ϕ. Then p_θ and $A(P, P_\theta)$ are twice continuously differentiable in θ. Furthermore, if P is*

not singular with respect to the family, then

$$\nabla\nabla' \log \frac{A(P, P_\theta)}{\sqrt{p_\theta(X)}} = \tfrac{1}{4}\mathbb{C}\,\phi(Y)$$

where $X \sim P$ and the distribution of Y is the $\frac{1}{2}$-geometric mixture between P and P_θ.

This Hessian does not depend on X. The iid version of the identity has $\frac{n}{4}\mathbb{C}\,\phi(Y)$.

What we really need to bound is the largest eigenvalue. Because the eigenvalues are positive in this case, we can bound the largest eigenvalue of the covariance matrix $\mathbb{C}\,\phi(Y)$ by its trace $\mathbb{V}\,\phi(Y)$.

Lemma 2.2.5. *Let $\Theta = \mathbb{R}^d$ parameterize the iid Gaussian location family with covariance $\sigma^2 I_d$ by $\theta = \sigma\mathbb{E}_{X\sim P_\theta} X$. Assume $X^n \overset{iid}{\sim} P$ and let $\hat{\theta}$ denote the MLE. If the distribution of Y is the $\frac{1}{2}$-log-mixture between P and P_θ with $\theta \in B(\hat{\theta}, \delta)$, then*

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} \mathbb{V}\phi(Y) \le \frac{2}{\sigma^2}\left[\mathbb{V}X + \mathbb{V}\widetilde{X} + \|\mathbb{E}\widetilde{X} - \mathbb{E}X\|^2\right] + 2\delta^2$$

where $\phi(x) = x/\sigma$ is the sufficient statistic, $X \sim P$, and $\widetilde{X}$ has density proportional to $\sqrt{p}$.

The above observations can be brought together to state an exact risk bound for the unpenalized MLE of the Gaussian location $N(\theta, \sigma^2 I_d)$ family with $X^n \overset{iid}{\sim} P$.

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \le D(P\|\Theta) + \frac{d[\log d + 9\, v_P/\sigma^2] + 15}{n}$$

with $v_P := \frac{1}{d}[\mathbb{V}X + \mathbb{V}\widetilde{X} + \|\mathbb{E}\widetilde{X} - \mathbb{E}X\|^2]$ where $X \sim P$ and $\widetilde{X}$ has density proportional to $\sqrt{p}$. To see this, use Lemma 2.2.5 with $\delta = 2\epsilon\sqrt{d}$ and Corollary 2.2.3 with $\epsilon = 1/\sqrt{nd}$ and $\alpha = 2\pi n = 2\pi/d\epsilon^2$. Then, use the fact that $n\mathbb{V}\hat{\theta}/d = \mathbb{V}X/d\sigma^2 \le v_P/\sigma^2$. Finally, use $1/n \le 1$ and round numbers up to integers.

If $\mathbb{E}X$ is finite, then the rI-projection is $N(\mathbb{E}X, \sigma^2 I_d)$, and thus

$$D(P\|\Theta) = D(P\|N(\mathbb{E}X, \sigma^2 I_d)).$$

The purpose of this Gaussian location example is to demonstrate the continuous extension process without assumptions on the form of the data-generating distribution. It is

useful as a "warm-up" for Theorem 4.0.1. The result itself is not important; indeed, even without any assumptions on P, better risk bounds for Gaussian location are easy to show (Theorems 2.3.11 and 2.3.12).

2.3 Proofs

Proof of ***Theorem 2.1.1****.* We follow the pattern of Jonathan Li's version of the resolvability bound proof [Li, 1999].

$$\begin{aligned} D_B(P, P_{\hat{\theta}}) &:= 2 \log \frac{1}{A(P, P_{\hat{\theta}})} \\ &= 2 \log \frac{\sqrt{p_{\hat{\theta}}(X)/p(X)} e^{-\frac{1}{2}[\mathcal{L}(\hat{\theta}) + L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} + \log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta}) + L(\hat{\theta}) \\ &\leq 2 \log \sum_{\theta \in \Theta} \frac{\sqrt{p_{\theta}(X)/p(X)} e^{-\frac{1}{2}[\mathcal{L}(\theta) + L(\theta)]}}{A(P, P_{\theta})} + \log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta}) + L(\hat{\theta}) \end{aligned}$$

We were able to bound the random quantity by the sum over all $\theta \in \Theta$ because each of these terms is non-negative.

We will take the expectation of both sides for $X \sim P$. To deal with the first term, we use Jensen's inequality and the definition of Hellinger affinity.

$$\begin{aligned} 2\mathbb{E} \log \sum_{\theta \in \Theta} \frac{\sqrt{p_{\theta}(X)/p(X)} e^{-\frac{1}{2}[\mathcal{L}(\theta) + L(\theta)]}}{A(P, P_{\theta})} &\leq 2 \log \sum_{\theta \in \Theta} \mathbb{E} \frac{\sqrt{p_{\theta}(X)/p(X)} e^{-\frac{1}{2}[\mathcal{L}(\theta) + L(\theta)]}}{A(P, P_{\theta})} \\ &= 2 \log \sum_{\theta \in \Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta) + L(\theta)]} \end{aligned}$$

Returning to the overall inequality, we have

$$\begin{aligned}
\mathbb{E}D_B(P, P_{\hat{\theta}}) &\leq 2\log \sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]} + \mathbb{E}\left[\log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta})\right] + \mathbb{E}L(\hat{\theta}) \\
&= 2\log \sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]} + \mathbb{E}\min_{\theta\in\Theta}\left\{\log \frac{p(X)}{p_{\theta}(X)} + \mathcal{L}(\theta)\right\} + \mathbb{E}L(\hat{\theta}) \\
&\leq 2\log \sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]} + \inf_{\theta\in\Theta}\mathbb{E}\left\{\log \frac{p(X)}{p_{\theta}(X)} + \mathcal{L}(\theta)\right\} + \mathbb{E}L(\hat{\theta}) \\
&= 2\log \sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]} + \inf_{\theta\in\Theta}\{D(P\|P_{\theta}) + \mathcal{L}(\theta)\} + \mathbb{E}L(\hat{\theta}).
\end{aligned}$$

□

Proof of ***Theorem 2.1.3****.* Following the steps described in [Barron et al., 2008, Theorem 2.3], we use Markov's inequality then bound a non-negative random variable by the sum of its possible values.

$$\begin{aligned}
P\left\{D_B(P, P_{\hat{\theta}}) - \left[\log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta}) + L(\hat{\theta})\right] \geq t\right\} &= P\left\{2\log \frac{\sqrt{p_{\hat{\theta}}(X)/p(X)}e^{-\frac{1}{2}[\mathcal{L}(\hat{\theta})+L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} \geq t\right\} \\
&= P\left\{\frac{\sqrt{p_{\hat{\theta}}(X)/p(X)}e^{-\frac{1}{2}[\mathcal{L}(\hat{\theta})+L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} \geq e^{t/2}\right\} \\
&\leq e^{-t/2}\mathbb{E}\frac{\sqrt{p_{\hat{\theta}}(X)/p(X)}e^{-\frac{1}{2}[\mathcal{L}(\hat{\theta})+L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} \\
&\leq e^{-t/2}\sum_{\theta\in\Theta} e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]}
\end{aligned}$$

□

Jensen differences

For any random vector Y and any function f, we will call $\mathbb{E}f(Y) - f(\mathbb{E}Y)$ a *Jensen difference.*

Lemma 2.3.1. *Let Y be a random vector with convex support $S \subseteq \mathbb{R}^d$. If $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, then*

$$\inf_{y\in S} \lambda_d(\nabla\nabla' f(y)) \leq \frac{\mathbb{E}f(Y) - f(\mathbb{E}Y)}{\mathbb{V}Y/2} \leq \sup_{y\in S} \lambda_1(\nabla\nabla' f(y)).$$

Proof. We start with a second-order Taylor expansion with Lagrange remainder.

$$f(Y) = f(\mathbb{E}Y) + (Y - \mathbb{E}Y)'\nabla f(\mathbb{E}Y) + \tfrac{1}{2}(Y - \mathbb{E}Y)'\nabla\nabla' f(\tilde{Y})(Y - \mathbb{E}Y)$$

for some $\tilde{Y}$ on the segment from Y to $\mathbb{E}Y$. By Lemma 2.3.9, the quadratic form has the bounds

$$\|Y - \mathbb{E}Y\|^2 \lambda_d(\nabla\nabla' f(\tilde{Y})) \leq (Y - \mathbb{E}Y)'\nabla\nabla' f(\tilde{Y})(Y - \mathbb{E}Y) \leq \|Y - \mathbb{E}Y\|^2 \lambda_1(\nabla\nabla' f(\tilde{Y})).$$

The smallest and largest eigenvalues of the Hessian at $\tilde{Y}$ are bounded by the infimum of smallest eigenvalue and supremum of largest eigenvalue taken over the support of Y.

$$\|Y - \mathbb{E}Y\|^2 \inf_{y \in S} \lambda_d(\nabla\nabla' f(y)) \leq (Y - \mathbb{E}Y)'\nabla\nabla' f(\tilde{Y})(Y - \mathbb{E}Y) \leq \|Y - \mathbb{E}Y\|^2 \sup_{y \in S} \lambda_1(\nabla\nabla' f(y))$$

Substituting this second-order Taylor expansion into $\mathbb{E}f(Y) - f(\mathbb{E}Y)$ gives the desired result.

□

Infimum on a grid

In many cases we will need to ensure that the infimum of a function on a grid of its domain approaches the overall infimum as the grid becomes increasingly refined. Lemma 2.3.2 will prove to be remarkably useful for such tasks.

Lemma 2.3.2. *Let $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$, and assume $f : \Theta \to \mathbb{R}$ is twice continuously differentiable. If θ is in the convex hull of $\Theta_\epsilon \cap B(\theta, \delta)$, then*

$$\inf_{\theta_\epsilon \in \Theta_\epsilon} f(\theta_\epsilon) \leq f(\theta) + \frac{\delta^2}{2} \sup_{\tilde{\theta} \in B(\theta,\delta)} \lambda_1(\nabla\nabla' f(\tilde{\theta}))_+.$$

Proof. We first bound the infimum over Θ_ϵ by the infimum over $\Theta_\epsilon \cap B(\theta, \delta)$. Then that infimum is bounded by the expectation using any distribution Q on those grid-points. We have assumed that θ is some weighted average of nearby grid-points (the ones at most δ distance away), and we can use that same weighted averaging to define Q. Then the

expectation of the random selection is θ, and we apply Lemma 2.3.1.

$$\begin{aligned}
\inf_{\theta_\epsilon \in \Theta_\epsilon} f(\theta_\epsilon) &\leq \inf_{\theta_\epsilon \in \Theta_\epsilon \cap B(\theta,\delta)} f(\theta_\epsilon) \\
&\leq \mathbb{E}_{\theta_\epsilon \sim Q} f(\theta_\epsilon) \\
&\leq f(\mathbb{E}_{\theta_\epsilon \sim Q} \theta_\epsilon) + \tfrac{1}{2} \mathbb{E}_{\theta_\epsilon \sim Q} \|\theta_\epsilon - \mathbb{E}_{\theta_\epsilon \sim Q} \theta_\epsilon\|^2 \sup_{\tilde{\theta} \in B(\theta,\delta)} \lambda_1(\nabla \nabla' f(\tilde{\theta})) \\
&\leq f(\theta) + \tfrac{1}{2} \delta^2 \sup_{\tilde{\theta} \in B(\theta,\delta)} \lambda_1(\nabla \nabla' f(\tilde{\theta}))
\end{aligned}$$

assuming $\lambda_1(\nabla \nabla' f(\tilde{\theta}))$ is non-negative. If the maximum eigenvalue is negative, i.e. if f is strictly concave within the ball, then the second order term is upper bounded by zero. $\square$

Suppose $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$ is an ϵ-discretization, as defined in Section 2.1. If Θ is convex, then every θ in the convex hull of Θ_ϵ satisfies the conditions of Lemma 2.3.2 with $\epsilon\sqrt{d}$ as δ. In particular, if every dimension of Θ is either $\mathbb{R}$ or a closed half-line, then there is an obvious ϵ-discretization that makes Lemma 2.3.2 apply for every $\theta \in \Theta$. For less favorably shaped Θ, one can consider adding more grid-points "on top of" an ϵ-discretization.

Behavior of $\mathcal{R}^{(\mathbf{n})}_{\mathbf{\Theta}_\epsilon, \mathcal{L}}(\mathbf{P})$

One way to bound $\mathcal{R}^{(n)}_{\Theta_\epsilon, \mathcal{L}}(P)$ is to use an approach similar to that just described for the infimum on a grid. Suppose $p_\theta(x)$ is twice continuously differentiable in θ. We define a type of Fisher "cross-information" matrix

$$I_P(\tilde{\theta}) := \mathbb{E}_{X \sim P} \nabla \nabla' \left[\log \frac{1}{p_\theta(X)} \right]_{\theta = \tilde{\theta}}$$

where the Hessian is taken with respect to θ. Note that if p_θ represents an exponential family, then P does not play a role. In that case, $I_P(\tilde{\theta})$ reduces to the ordinary Fisher information matrix.

Let $B(\theta, \delta)$ denote the closed Euclidean ball centered at θ with radius δ, and let $\lambda_j(\cdot)$ denote the jth largest eigenvalue of its matrix argument.

Theorem 2.3.3. *Let $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$. Assume that $\mathcal{L} : \Theta \to \mathbb{R}$ is twice continuously differentiable and that $p_\theta(x)$ is twice continuously differentiable in θ for every fixed x in its*

domain. If $\theta \in \Theta$ is in the convex hull of $\Theta_\epsilon \cap B(\theta, \delta)$, then

$$\mathcal{R}^{(n)}_{\Theta_\epsilon,\mathcal{L}}(P) \leq D(P\|P_\theta) + \frac{\mathcal{L}(\theta)}{n} + \frac{\delta^2}{2} \sup_{\tilde{\theta} \in B(\theta,\delta)} \lambda_1(I_P(\tilde{\theta}) + \tfrac{1}{n}\nabla\nabla'\mathcal{L}(\tilde{\theta}))_+.$$

Proof. Define $f_X(\theta) := \log \frac{p(X)}{p_\theta(X)} + \frac{\mathcal{L}(\theta)}{n}$, and let $X \sim P$. We use a second-order Taylor expansion at θ with Lagrange remainder and reason similarly to the proofs of Lemmas 2.3.1 and 2.3.2.

$$\begin{aligned}
\mathcal{R}^{(n)}_{\Theta_\epsilon,\mathcal{L}}(P) &= \inf_{\theta_\epsilon \in \Theta_\epsilon} \mathbb{E} f_X(\theta_\epsilon) \\
&= \inf_{\theta_\epsilon \in \Theta_\epsilon} \mathbb{E}\left(f_X(\theta) + (\theta_\epsilon - \theta)'\nabla f_X(\theta) + \tfrac{1}{2}(\theta_\epsilon - \theta)'[\nabla\nabla' f_X(\tilde{\theta})](\theta_\epsilon - \theta)\right) \\
&= \inf_{\theta_\epsilon \in \Theta_\epsilon} \left(\mathbb{E} f_X(\theta) + (\theta_\epsilon - \theta)'\mathbb{E}\nabla f_X(\theta) + \tfrac{1}{2}(\theta_\epsilon - \theta)'[\mathbb{E}\nabla\nabla' f_X(\tilde{\theta})](\theta_\epsilon - \theta)\right)
\end{aligned}$$

for some $\tilde{\theta}$ between θ and θ_ϵ.

The infimum is bounded by the expectation for any random θ_ϵ on the grid-points. In particular, use the distribution on neighboring grid-points that makes θ_ϵ have expectation θ. The first-order term is eliminated, while the second-order term is bounded by half the expected squared length of the vector $\theta_\epsilon - \theta$ times the largest eigenvalue (if positive). □

When $\Theta_\epsilon \subseteq \Theta$ is an ϵ-discretization, we use $\epsilon\sqrt{d}$ as δ.

Corollary 2.3.4. *Let $\Theta \subseteq \mathbb{R}^d$ be a convex parameter space having densities twice continuously differentiable in θ. Let $\Theta_\epsilon \subseteq \Theta$ be an ϵ-discretization. For any θ in the convex hull of Θ_ϵ,*

$$\mathcal{R}^{(n)}_{\Theta_\epsilon,\mathcal{L}}(P) \leq D(P\|P_\theta) + \frac{\mathcal{L}(\theta)}{n} + \frac{\epsilon^2 d}{2} \sup_{\tilde{\theta} \in B(\theta,\epsilon\sqrt{d})} \lambda_1(I_P(\tilde{\theta}) + \tfrac{1}{n}\nabla\nabla'\mathcal{L}(\tilde{\theta}))_+.$$

If one uses discretization $\epsilon = a/\sqrt{n}$,

$$\mathcal{R}^{(n)}_{\Theta_\epsilon,\mathcal{L}}(P) \leq D(P\|P_\theta) + \frac{\mathcal{L}(\theta) + a^2 dz/2}{n}$$

with $z := \sup_{\tilde{\theta} \in B(\theta,\sqrt{ad})} \lambda_1(I_P(\tilde{\theta}) + \nabla\nabla'\mathcal{L}(\tilde{\theta}))_+$ which does not depend on n. Notice that this bound uses the $n = 1$ version of the supremum term, because they cannot increase with

n. Notice also that, in general, z will increase with d. One could set $a^2 = 1/d$ to cancel out all dimension dependence, but that has an undesirable overall effect on the risk bound results put forward in this paper.

One will most likely want to invoke these results with P_θ being the rI-projection of P onto Θ if it exists. In particular, if P is in the model, then we can let P_θ be P to get an exact bound of order $1/n$ for $\mathcal{R}^{(n)}_{\Theta_\epsilon,\mathcal{L}}(P)$.

Bounding summations over grid-points

Lemmas 2.3.5 and 2.3.6 provide bounds for summations of Gaussian-shaped functions over ϵ-discretizations of $\mathbb{R}^d$.

Lemma 2.3.5. *Let Θ_ϵ be an ϵ-discretization of $\mathbb{R}^d$. Then for any $c > 0$ and $v \in \Theta_\epsilon$,*

$$\sum_{\theta\in\Theta_\epsilon} e^{-c\|\theta-v\|^2} \leq \left(1 + \tfrac{\sqrt{\pi}}{\epsilon\sqrt{c}}\right)^d$$

Proof. We can assume without loss of generality that v is the zero vector and that Θ_ϵ includes zero. First, consider the one-dimensional problem. The "center" term equals 1 and the sum of the other terms is bounded by a Gaussian integral.

$$\begin{aligned}
\sum_{\theta\in\Theta_\epsilon} e^{-c\theta^2} &= \sum_{\theta\in\Theta_\epsilon} e^{-c\epsilon^2(\theta/\epsilon)^2} \\
&= \sum_{z\in\mathbb{Z}} e^{-c\epsilon^2 z^2} \\
&\leq 1 + \int_{\mathbb{R}} e^{-c\epsilon^2 z^2} dz \\
&= 1 + \frac{\sqrt{\pi}}{\epsilon\sqrt{c}}
\end{aligned}$$

The d-dimensional problem can be bounded in terms of d instances of the one-dimensional problem. Let $\Theta^{(1)}_\epsilon, \ldots, \Theta^{(d)}_\epsilon$ represent the underlying discretizations of $\mathbb{R}$, so that $\Theta_\epsilon =$

$\prod_j \Theta_\epsilon^{(j)}$.

$$\begin{aligned}
\sum_{\theta\in\Theta_\epsilon} e^{-c\|\theta\|^2} &= \sum_{\theta\in\Theta_\epsilon} e^{-c\sum_j \theta_j^2} \\
&= \sum_{\theta_1\in\Theta_\epsilon^{(1)}} \cdots \sum_{\theta_d\in\Theta_\epsilon^{(d)}} \prod_j e^{-c\theta_j^2} \\
&= \prod_j \sum_{\theta_j\in\Theta_\epsilon^{(j)}} e^{-c\theta_j^2} \\
&\leq \prod_j \left(1 + \tfrac{\sqrt{\pi}}{\epsilon\sqrt{c}}\right) \\
&= \left(1 + \tfrac{\sqrt{\pi}}{\epsilon\sqrt{c}}\right)^d
\end{aligned}$$

□

Similar reasoning provides a slightly larger bound if the peak of the Gaussian function is not necessarily in the discretization.

Lemma 2.3.6. *Let Θ_ϵ be an ϵ-discretization of $\mathbb{R}^d$. Then for any $c > 0$ and $v \in \mathbb{R}^d$,*

$$\sum_{\theta\in\Theta_\epsilon} e^{-c\|\theta-v\|^2} \leq \left(1 + \tfrac{2\sqrt{\pi}}{\epsilon\sqrt{c}}\right)^d$$

Proof. Again, we begin with the one-dimensional problem. The closest point to v contributes at most 1 to the sum. We reduce to Lemma 2.3.5 by comparison to $\Theta^*_{\epsilon/2}$, the $(\epsilon/2)$-grid that includes v. Each point on the original grid can be translated "inward" to a neighboring point on the new (more refined) grid. The sum over the new grid's points will be larger than the sum over the original grid's points.

$$\begin{aligned}
\sum_{\theta\in\Theta_\epsilon} e^{-c(\theta-v)^2} &\leq \sum_{\theta\in\Theta^*_{\epsilon/2}} e^{-c(\theta-v)^2} \\
&\leq 1 + \frac{\sqrt{\pi}}{(\epsilon/2)\sqrt{c}}
\end{aligned}$$

As before, the d-dimensional problem reduces to the one-dimensional problem.

$$\begin{aligned}\sum_{\theta\in\Theta_\epsilon} e^{-c\|\theta-v\|^2} &= \prod_j \sum_{\theta_j\in\Theta_\epsilon^{(j)}} e^{-c(\theta_j-v_j)^2} \\ &\le \prod_j \left(1+\tfrac{2\sqrt{\pi}}{\epsilon\sqrt{c}}\right) \\ &= \left(1+\tfrac{2\sqrt{\pi}}{\epsilon\sqrt{c}}\right)^d\end{aligned}$$

□

Miscellaneous facts

The following handy facts are known, but we provide brief proofs here nonetheless.

Lemma 2.3.7. *For any vectors u, v in a real inner product space,*

$$\|u-v\|^2 \le 2\|u\|^2 + 2\|v\|^2.$$

Proof. We apply the Cauchy-Schwarz inequality followed by the arithmetic-geometric mean inequality.

$$\begin{aligned}\|a-b\|^2 &= \|a\|^2 + \|b\|^2 - 2\langle a,b\rangle \\ &\le \|a\|^2 + \|b\|^2 + 2\|a\|\|b\| \\ &\le \|a\|^2 + \|b\|^2 + 2(\|a\|^2/2 + \|b\|^2/2)\end{aligned}$$

□

Lemma 2.3.8. *For vectors a and b in an inner product space,*

$$\|a-b\|^2 \ge \|a\|^2 - 2\|a\|^2\|b\|^2 - \|b\|^2 - 2\|b\|^4 - 1$$

Proof. We use the *Peter-Paul inequality* with parameter b^2, then Lemma 2.3.7.

$$\begin{aligned}
\|a\|^2 &= \|a - b + b\|^2 \\
&\leq \|a - b\|^2 + 2\|a - b\|\|b\| + \|b\|^2 \\
&\leq \|a - b\|^2 + [\|b\|^2\|a - b\|^2 + 1] + \|b\|^2 \\
&\leq \|a - b\|^2 + \|b\|^2(2\|a\|^2 + 2\|b\|^2) + 1 + \|b\|^2
\end{aligned}$$

□

Lemma 2.3.9. *Let $v \in \mathbb{R}^d$, and let M be a symmetric $d \times d$ matrix. Then*

$$\lambda_d(M) \leq \frac{v'Mv}{\|v\|^2} \leq \lambda_1(M).$$

Proof. Any symmetric matrix has an orthonormal eigenvector decomposition $M = Q\Lambda Q'$.

$$\begin{aligned}
v'Mv &= v'Q\Lambda Q'v \\
&= \sum_j \lambda_j (Q'v)_j^2 \\
&= \|v\|^2 \sum_j \lambda_j \left(Q' \frac{v}{\|v\|}\right)_j^2.
\end{aligned}$$

Realize that squared values in the summation are eigenvector-basis coordinates of the unit vector in the direction of v. As such, these squared coordinates must sum to 1. Thus the summation is a weighted average of the eigenvalues. It achieves its maximum λ_1 when v is in the direction of the first eigenvector, and it achieves its minimum λ_d when v is in the direction of the last eigenvector. □

*Proof of **Theorem 2.2.1**.* Compared to Theorem 2.1.1, this proof only requires adding and subtracting a discrepancy term. Define

$$g_X(\theta) := \frac{\sqrt{p_\theta(X)/p(X)}e^{-\frac{1}{2}[\mathcal{L}(\theta)+L(\theta)]}}{A(P, P_\theta)}.$$

Let θ_ϵ be any point in Θ_ϵ; it is allowed to depend on X.

$$\begin{aligned} D_B(P, P_{\hat{\theta}}) &:= 2 \log \frac{1}{A(P, P_{\hat{\theta}})} \\ &= 2 \log g_X(\hat{\theta}) + \log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta}) + L(\hat{\theta}) \\ &= 2 \log g_X(\theta_\epsilon) + \log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta}) + L(\hat{\theta}) + 2 \log \frac{g(\hat{\theta})}{g(\theta_\epsilon)} \\ &\leq 2 \log \sum_{\theta \in \Theta_\epsilon} g_X(\theta) + \log \frac{p(X)}{p_{\hat{\theta}}(X)} + \mathcal{L}(\hat{\theta}) + L(\hat{\theta}) + 2 \log \frac{g_X(\hat{\theta})}{g_X(\theta_\epsilon)} \end{aligned} \tag{2.4}$$

The inequality holds for every $\theta_\epsilon \in \Theta_\epsilon$, so it also holds for the (random) infimum.

The proof is completed by taking the expectation of (2.4) and proceeding analogously to Theorem 2.1.1. □

Proof of ***Lemma 2.2.4****.* First, $e^{-\psi(\theta)/2}$ cancels out of the numerator and denominator. What is left over in the log-likelihood is linear in θ, so it's Hessian is the zero matrix. Thus all we have left is

$$\nabla \nabla' \log \tilde{A}_\theta$$

where $\tilde{A}_\theta$ is short-hand for the [multiple] integral

$$\int_{\mathcal{X}} \sqrt{p(x) r(x)} e^{\frac{1}{2} \theta' \phi(x)} dx$$

and r is the exponential family's "carrier" function. By inspecting the form of $\tilde{A}_\theta$, we see that it is the partition function of a different exponential family with natural parameter θ, sufficient statistic $\phi/2$, and carrier $\sqrt{pr}$. This family's natural parameter space is at least as large as that of the original exponential family in question, because by Cauchy-Schwarz,

$$\begin{aligned} \int_{\mathcal{X}} \sqrt{p(x) r(x)} e^{\frac{1}{2} \theta' \phi(x)} dx &\leq \sqrt{\int_{\mathcal{X}} p(x) dx} \sqrt{\int_{\mathcal{X}} r(x) e^{\theta' \phi(x)} dx} \\ &= \sqrt{\int_{\mathcal{X}} r(x) e^{\theta' \phi(x)} dx}. \end{aligned}$$

This quantity is finite on the natural parameter space of the original exponential family. Therefore, all derivatives of $\tilde{A}_\theta$ with respect to θ can be taken through the integral [Lehmann and Romano, 2006, Theorem 2.7.1].

The Hessian of $\log \tilde{A}_\theta$ can be expressed as

$$\nabla\nabla' \log \tilde{A}_\theta = \frac{1}{\tilde{A}_\theta^2}\left[\tilde{A}_\theta(\nabla\nabla'\tilde{A}_\theta) - (\nabla\tilde{A}_\theta)(\nabla\tilde{A}_\theta)'\right]. \tag{2.5}$$

Using the derivative-integral interchange, we find the gradient of $\tilde{A}_\theta$ to be

$$\begin{aligned}\nabla\tilde{A}_\theta &= \nabla\int_{\mathcal{X}} \sqrt{p(x)r(x)}e^{\frac{1}{2}\theta'\phi(x)}dx \\ &= \tfrac{1}{2}\int_{\mathcal{X}} \phi(x)\sqrt{p(x)r(x)}e^{\frac{1}{2}\theta'\phi(x)}dx \\ &= \tfrac{1}{2}\tilde{A}_\theta\mathbb{E}\phi(Y)\end{aligned}$$

where the distribution of Y is the $\frac{1}{2}$-log-mixture between P and P_θ.

Likewise, the Hessian is

$$\begin{aligned}\nabla\nabla'\tilde{A}_\theta &= \tfrac{1}{4}\int_{\mathcal{X}} \phi(x)\phi(x)'\sqrt{p(x)r(x)}e^{\frac{1}{2}\theta'\phi(x)}dx \\ &= \tfrac{1}{4}\tilde{A}_\theta\mathbb{E}\phi(Y)\phi(Y)'.\end{aligned}$$

Referring back to (2.5), we conclude that the Hessian of $\log \tilde{A}_\theta$ is

$$\begin{aligned}\nabla\nabla' \log \tilde{A}_\theta &= \tfrac{1}{4}[\mathbb{E}\phi(Y)\phi(Y)' - (\mathbb{E}\phi(Y))(\mathbb{E}\phi(Y))'] \\ &= \tfrac{1}{4}\mathbb{C}\,\phi(Y).\end{aligned}$$

□

Lemma 2.3.10 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 4.0.1. It also implies a stochastic inequality (and therefore expectation inequality) for the squared norms, which is used for an example in Section 2.2.

Lemma 2.3.10. *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. □

*Proof of **Lemma 2.2.5**.* $\mathbb{E}Y$ is the minimizer of expected squared distance from Y, so

$$\begin{aligned}\mathbb{E}_{X^n \overset{iid}{\sim} P} \mathbb{V}\phi(Y) &= \tfrac{1}{\sigma^2}\mathbb{E}\|Y - \mathbb{E}Y\|^2 \\ &\leq \tfrac{1}{\sigma^2}\mathbb{E}\|Y - \sigma\theta\|^2\end{aligned}$$

Now, realize that $\sigma\theta$ is the location of the peak of the density p_θ. Likewise, the square root $\sqrt{p_\theta}$ also has its mode at $\sigma\theta$ and is spherically symmetric about that point. The density of Y is proportional to the product of $\sqrt{p}$ and $\sqrt{p_\theta}$. Making use of Lemma 2.3.10, we see that this multiplication results in a density that is more concentrated toward $\sigma\theta$ compared

to $\sqrt{p}$. Letting $\widetilde{X}$ have density proportional to $\sqrt{p}$,

$$\begin{aligned}
\mathbb{E}\|Y - \sigma\theta\|^2 &\leq \mathbb{E}\|\widetilde{X} - \sigma\theta\|^2 \\
&= \mathbb{V}\widetilde{X} + \|\mathbb{E}\widetilde{X} - \sigma\theta\|^2 \\
&\leq \mathbb{V}\widetilde{X} + [\|\mathbb{E}\widetilde{X} - \sigma\hat{\theta}\| + \|\sigma\hat{\theta} - \sigma\theta\|]^2 \\
&\leq \mathbb{V}\widetilde{X} + 2\|\mathbb{E}\widetilde{X} - \sigma\hat{\theta}\|^2 + 2\sigma^2\|\hat{\theta} - \theta\|^2 \\
&\leq \mathbb{V}\widetilde{X} + 2\|\mathbb{E}\widetilde{X} - \sigma\hat{\theta}\|^2 + 2\sigma^2\delta^2.
\end{aligned}$$

To get the overall result, we need to take an expectation of this with respect to the data; the random quantity is $\hat{\theta} \equiv \bar{X}_n/\sigma$. By the bias-variance decomposition,

$$\begin{aligned}
\mathbb{E}\|\mathbb{E}\widetilde{X} - \sigma\hat{\theta}\|^2 &= \mathbb{E}\|\mathbb{E}\widetilde{X} - \bar{X}_n\|^2 \\
&= \mathbb{V}\bar{X}_n + \|\mathbb{E}\widetilde{X} - \mathbb{E}\bar{X}_n\|^2 \\
&= \tfrac{1}{n}\mathbb{V}X + \|\mathbb{E}\widetilde{X} - \mathbb{E}X\|^2
\end{aligned}$$

□

Here we prove the claim that simple distribution-free risk bounds for the Gaussian location MLE are easy to obtain.

Theorem 2.3.11. *Assume $X^n \overset{iid}{\sim} P$. Then the estimator $P_{\hat{\theta}} := N(\bar{X}_n, \sigma^2 I)$ has the Bhattacharyya risk bound*

$$\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq D_B(P, P_{\theta^*}) + \frac{\mathbb{V}X/2\sigma^2}{n}$$

where $X \sim P$ and P_{θ^*} is $N(\mathbb{E}X, \sigma^2 I_d)$.

Proof. If P has an infinite second moment, then the inequality is trivially true. We proceed assuming the second moment is finite.

Let $\widetilde{P}$ denote the part of P that is absolutely continuous with respect to Lebesgue measure, and realize that the rest of P contributes nothing to its Hellinger affinity with another continuous distribution. The only inequality in our derivation comes from applying

Theorem B.0.3 (incorporating $\sqrt{\widetilde{p}}$ into the measure).

$$\begin{aligned}
\mathbb{E}D_B(P, P_{\hat{\theta}}) &= \mathbb{E} - 2\log \int_{\mathbb{R}^d} \sqrt{\widetilde{p}}\,(2\pi\sigma^2)^{-d/4} e^{-\frac{1}{4\sigma^2}\|y-\hat{\theta}\|^2} dy \\
&\leq -2\log \int_{\mathbb{R}^d} \sqrt{\widetilde{p}}\,(2\pi\sigma^2)^{-d/4} e^{-\frac{1}{4\sigma^2}\mathbb{E}\|y-\hat{\theta}\|^2} dy \\
&= -2\log \int_{\mathbb{R}^d} \sqrt{\widetilde{p}}\,(2\pi\sigma^2)^{-d/4} e^{-\frac{1}{4\sigma^2}(\|y-\mathbb{E}X\|^2 + \mathbb{E}\|\bar{X}_n - \mathbb{E}X\|^2)} dy \\
&= -2\log \int_{\mathbb{R}^d} \sqrt{\widetilde{p}}\,(2\pi\sigma^2)^{-d/4} e^{-\frac{1}{4\sigma^2}\|y-\mathbb{E}X\|^2} dy + \frac{1}{2\sigma^2}\mathbb{E}\|\bar{X}_n - \mathbb{E}X\|^2 \\
&= D_B(P, P_{\theta^*}) + \frac{\mathbb{V}X/2\sigma^2}{n}
\end{aligned}$$

□

An analogous result for Kullback risk holds as an identity.

Theorem 2.3.12. *Assume $X^n \overset{iid}{\sim} P$. Then the estimator $P_{\hat{\theta}} := N(\bar{X}_n, \sigma^2 I)$ has Kullback risk*

$$\mathbb{E}D(P\|P_{\hat{\theta}}) = D(P\|P_{\theta^*}) + \frac{\mathbb{V}X/2\sigma^2}{n}$$

where $X \sim P$ and P_{θ^*} is $N(\mathbb{E}X, \sigma^2 I_d)$.

Proof. P_{θ^*} is the rI-projection of P onto the log-convex set Θ, so we have the "Pythagorean" identity

$$D(P\|P_\theta) = D(P\|P_{\theta^*}) + D(P_{\theta^*}\|P_\theta)$$

for any P_θ in the model. The identity holds for the random $P_{\hat{\theta}}$ as well. In that case, the first term does not depend on the data. By the definition of relative entropy and Gaussian density, the expectation of the second term is

$$\begin{aligned}
\mathbb{E}_{X^n \overset{iid}{\sim} P} D(P_{\theta^*}\|P_{\hat{\theta}}) &= \mathbb{E}_{X^n \overset{iid}{\sim} P} \frac{1}{2\sigma^2} \mathbb{E}_{X\sim P}\left[\|X - \hat{\theta}\|^2 - \|X - \theta^*\|^2\right] \\
&= \frac{1}{2\sigma^2}\mathbb{E}_{X^n \overset{iid}{\sim} P}\|\hat{\theta} - \theta^*\|^2.
\end{aligned}$$

The Gaussian location MLE is $\hat{\theta} := \bar{X}_n$, so $\mathbb{E}_{X^n \overset{iid}{\sim} P}\|\hat{\theta} - \theta^*\|^2 = \frac{1}{n}\mathbb{V}X$. □

Chapter 3

Approximation error of mixtures

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. However, their inequality requires the family to have a uniformly bounded density ratio. As a result, it does not apply to familiar families, including GRBM models. 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 prove that $1/k$ rates can hold without a uniformly bounded density ratio; in particular, we prove such a result for expected redundancy rate of GRBMs.

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 product-measurable.[1] 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

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

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.[2] After

1. 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].

2. 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 3.0.1 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

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

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 3.0.1. *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 3.0.2 uses Theorem 3.0.1 to bound the approximation error of greedy k-component mixtures in terms of any specific mixture over the family.

Corollary 3.0.2. *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[3]

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

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 3.0.3. *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*

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

3. 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

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

The MDL method for bounding risk penalized likelihood estimation (which was the topic of Chapter 2) 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 3.0.4 bounds using Theorem 3.0.1. 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 3.0.4. *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

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

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 3.0.4 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 3.0.5 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 3.0.5. *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

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

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 3.0.5 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

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

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 3.0.6 states a bound that follows from Theorem 3.0.5.

Corollary 3.0.6. *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 3.0.6 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 3.0.4 with Theorem 3.0.7 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 Ł^p-norm $\|Y\|_p := (\mathbb{E}\|Y\|^p)^{1/p}$.

Theorem 3.0.7. *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*

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

Furthermore, if P *has the subgaussianity property that* $\mathbb{E}_{X\sim P}e^{t\|X-\mathbb{E}X\|}\le e^{\sigma_P^2t^2/2}$ *for all* $t\ge 0$*, 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*

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

3.1 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 3.1.1. *Let* (B_k) *be a non-negative and non-decreasing sequence of real numbers. If* (D_k) *is a sequence such that*

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

then

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

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

the fact that

$$D_k \le \frac{D_1 + B\sum_{j=2}^{k} 1/j}{k} \tag{3.1}$$

because the harmonic sum is bounded by the logarithm. We prove (3.1) by induction, assuming $B_k = B$ is fixed for all k. For $k = 2$,

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

as required. Next, assuming (3.1) holds for D_k,

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

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 3.1.2. *For any $t \ge 0$,*

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

Proof. It is easy to verify that if $t \ge 1$, 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 3.0.1**.* First, follow the proof of [Li, 1999, Lem 5.8] except use our Lemma 3.1.2 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.

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

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

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

The initial term is

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

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 3.1.1 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 3.0.5**.* 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 3.1.3. *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.

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

□

*Proof of **Theorem 3.0.7**.* 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,

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

By Lemma 3.1.3,

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

Lemmas 3.1.5 and 3.1.6 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 3.1.4. *If $X, X_1, \ldots, X_n \overset{iid}{\sim} P$, then for any convex, increasing, non-negative function f,*

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

Proof.

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

□

Lemma 3.1.5. *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 3.1.4 with $f(x) = x^z$. □

Lemma 3.1.6. *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 3.1.4 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. □

Chapter 4

Risk of Gaussian radial basis mixtures

Suppose $\Theta = \bigcup_{k \geq 1} \Theta^{(k)}$ is a model class and each $\Theta^{(k)}$ is a model of countable cardinality. Let us index the distributions in Θ by $\theta = (k, \mu)$ with $\mu \in \Theta^{(k)}$. Assume the penalty and pseudo-penalty have the form $Ł(\theta) = Ł_0(k) + Ł_k(\mu)$ and $L(\theta) = L_0(k) + L_k(\mu)$. Then Theorem 2.1.1 can be useful if the penalty plus pseudo-penalty on k is large enough to counteract the within-model summations.

$$\sum_{\theta \in \Theta} e^{-\frac{1}{2}[Ł(\theta)+L(\theta)]} = \sum_{k \geq 1} \left[e^{-\frac{1}{2}[Ł_0(k)+L_0(k)]} \sum_{\mu \in \Theta^{(k)}} e^{-\frac{1}{2}[Ł_k(\mu)+L_k(\mu)]} \right]$$

One can use $L_0(k) = 0$ to avoid having to worry about the behavior of $\hat{k}$. Then bounds on $\sum_{\theta \in \Theta^{(k)}} e^{-[Ł_k(\theta)+L_k(\theta)]}$ should be known so that one can devise a penalty on k that bounds the weighted sum of these summations. In particular, one can set $Ł_0(k)$ large enough that

$$\log \sum_{k \geq 1} \left[e^{-\frac{1}{2}Ł_0(k)} \sum_{\theta \in \Theta^{(k)}} e^{-\frac{1}{2}[Ł_k(\theta)+L_k(\theta)]} \right] \leq 0.$$

It remains to deal with $\mathbb{E}L_k(\hat{\theta})$, either by bounding it or by absorbing it into the risk as in Corollary 2.1.5. Importantly, this approach makes it possible to bound the risk when k is penalized and the parameter within the model is unpenalized, as is typical in model

selection.

The same ideas can be used if the $\Theta^{(k)}$ are continuous parameter spaces. One can apply Theorem 2.2.1 by devising grids $\Theta_\epsilon^{(k)} \subseteq \Theta^{(k)}$, and defining $\Theta_\epsilon := \bigcup_{k\geq 1} \Theta_\epsilon^{(k)}$. The discrepancy term in that bound can have its supremum restricted to $\Theta_\epsilon^{(\hat{k})}$ to cancel out $e^{-\frac{1}{2}[\text{Ł}_0(\hat{k})+L_0(\hat{k})]}$.

Using the new risk bound approach from Chapter 2 with an expected redundancy bound derived in Chapter 3, we derive a risk bound for GRBM estimation.[1] The greedy algorithm along with the notations $\mathcal{C}$, c_Q^2, and C_Q^2 were defined in Chapter 3.

Theorem 4.0.1. *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 $\text{Ł}(\theta) = 3dk \log 4nk$. 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 $\widetilde{X} \sim \sqrt{p}$ and

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

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 $2/n$.

4.1 Proofs

Lemma 4.1.1. *Let $\theta = (\mu_1, \dots, \mu_k)$ with each $\mu_j \in \mathbb{R}^d$ indexing a component mean of an equal-weighted k-component GRBM P_θ. Let $\delta = (\delta_1, \dots, \delta_k)$ where each $\delta_j \in \mathbb{R}^d$ has norm*

1. The proof of Theorem 4.0.1 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.

bounded by a. Then

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

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

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

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.

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

by Lemma 2.3.10. ($\widetilde{X} \sim \sqrt{p}$ should be understood to mean the normalized version of $\sqrt{p}$.)

For the second part, we use Corollary B.0.3, which is a form of Hölder's inequality.

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

□

Proof of ***Theorem 4.0.1****.* Invoke Theorem 2.2.1 with pseudo-penalty

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

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 3.1.6 implies

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

The summation part of Theorem 2.2.1 can be handled by using integration grids $\Theta_\epsilon^{(k)} \subseteq \Theta^{(k)} = \mathbb{R}^{dk}$, as described in Section 2.2.[2]

$$\sum_{k\geq 1} e^{-\frac{1}{2}\text{Ł}(k)} \sum_{\theta_\epsilon \in \Theta_\epsilon^{(k)}} e^{-\frac{\sqrt{n}}{2k}\|\mu_j - \mathbb{E}X\|^2} = \sum_{k\geq 1} e^{-\frac{1}{2}\text{Ł}(k)} \left(\frac{\sqrt{2\pi k}}{\epsilon n^{1/4}}\right)^{dk}. \tag{4.1}$$

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

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

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 Section 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 4.1.1. 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 3.1.6. 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

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

Finally, we bound the expected redundancy using Theorem 3.0.7 then bound the infimum over k by comparison to the particular choice $k = \lceil\sqrt{n}\rceil \leq \sqrt{2n}$.

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

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

]

Chapter 5

Computing Gaussian radial basis mixtures

A GRBM likelihood can be highly multimodal and therefore difficult to optimize. Indeed, the product of the n sums can be represented as a sum of k^n unimodal terms:

$$\begin{aligned}\prod_{i=1}^{n} p_\theta(X_i) &\propto \prod_{i=1}^{n}\sum_{j=1}^{k} e^{-\frac{1}{2\sigma^2}\|X_i-\mu_j\|^2} \\ &= \sum_{v\in\mathcal{V}}\prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}\|X_i-\mu_{v_i}\|^2} \qquad (5.1)\end{aligned}$$

where $\mathcal{V} = \{1,\ldots,k\}^n$ indexes the set of all k^n possible assignments of labels and $v = (v_1,\ldots,v_n)$ denotes a labeling by having each $v_i \in \{1,\ldots,k\}$.

A typical approach to optimizing a mixture model's likelihood is to use the expectation maximization (EM) algorithm. One introduces hidden variables $Z = [Z_{i,j}]$ where $Z_{i,j} = 1$ means that observation i gets label j; with this approach, the log likelihood has a summation form

$$\sum_{i=1}^{n} \log p_{\theta,Z}(X_i) = -\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{i,j}\frac{1}{2\sigma^2}\|X_i-\mu_j\|^2.$$

The original likelihood for $\theta = (\mu_1,\ldots,\mu_k)$ is the marginal version of this joint likelihood. The EM algorithm alternates between calculating the expectations $r_{i,j} := \mathbb{E}Z_{i,j}$ given a θ

and then finding the θ that optimizes the expected log likelihood. Given $[r_{i,j}]$, the optimal θ has

$$\begin{aligned}
\check{\mu}_j &:= -\operatorname*{argmax}_{\mu\in\mathbb{R}^d} \sum_{i=1}^{n} r_{i,j} \tfrac{1}{2\sigma^2} \|X_i - \mu\|^2 \\
&= \operatorname*{argmin}_{\mu\in\mathbb{R}^d} \sum_{i=1}^{n} r_{i,j} \|X_i - \mu\|^2 \\
&= \operatorname*{argmin}_{\mu\in\mathbb{R}^d} \left[n_j \|\mu - \bar{X}_j\|^2 + \sum_{i=1}^{n} r_{i,j} \|X_i - \bar{X}_j\|^2 \right] \\
&= \bar{X}_j
\end{aligned}$$

where $n_j := \sum_i r_{i,j}$ and $\bar{X}_j := \sum_i \frac{r_{i,j}}{n_j} X_i$. Given $\theta = (\mu_1, \ldots, \mu_k)$, the labels' expectations are

$$\begin{aligned}
\mathbb{E} Z_{i,j} &= \mathbb{P}[Z_{i,j} = 1] \\
&= \frac{e^{-\frac{1}{2\sigma^2}\|X_i - \mu_j\|^2}}{\sum_l e^{-\frac{1}{2\sigma^2}\|X_i - \mu_l\|^2}}.
\end{aligned}$$

The EM iterations converge to a local optimum of the likelihood, but the result may not actually be a good choice due to our high degree of multi-modality.

A recent line of work started by Balakrishnan et al. [2017] has side-stepped the question of convergence to the global optimizer, instead analyzing EM's iterative behavior to analyze it statistical risk. That paper established general conditions under which a ball centered at the true parameter value would be a basin of attraction for the population version of the EM operator. For a large enough sample size, the difference (in that ball) between the sample EM operator and the population EM operator can be bounded such that the EM estimate approaches the true parameter with high probability. That bound is the sum of two terms with distinct interpretations. There is an *algorithmic convergence* term $\kappa^t \|\theta^{(0)} - \theta^\star\|$ for initializer $\theta^{(0)}$, truth $\theta^\star$, and some modulus of contraction $\kappa \in (0, 1)$; this comes from the analysis of the population EM operator. The second term captures *statistical convergence* and is proportional to the supremum norm of $M - M_n$, the difference between the population and sample EM operators, over the ball. This result is also shown for a "sample-splitting"

version of EM, where the sample is partitioned into batches and each batch governs a single step of the algorithm. They show that their analysis is easily seen to apply for a two-component GRBM.

The performance of EM for two-component GRBMs has since received further attention. Klusowski and Brinda [2018] showed that the intersection of a suitable half-space and ball about the the midpoint between the component means is also a basin of attraction for the population EM in that model when the component means are separated well enough relative to the noise. Exact probabilistic bounds on the squared norm error of the EM parameter estimate were also derived when the initializer is in the region. We concluded the paper by describing a random initialization strategy that has a high probability of finding the basin of attraction when the component means are sufficiently well separated. Our work made use of symmetries inherent to two-component GRBMs. Extending the analysis to allow for more components will present new challenges.

In this chapter, we will explore a variety of approaches to initializing the EM algorithm and compare the likelihoods that they produce. Section 5.1 will describe the algorithms under consideration. Simulations in Section 5.2 will put these algorithms to the test to see which one tends to output the θ with the highest likelihood.

5.1 Algorithms for initializing EM

A naive option for initializing EM is to generate θ at random perhaps from a Normal distribution or by uniformly selecting from the data points. Another naive option is to randomly generate a right stochastic matrix $[r_{i,j}]$. A more sophisticated way to generate an initial θ is to use Markov chain Monte Carlo to draw it from a distribution resembling the likelihood; an exciting new algorithm for such draws is introduced in Section 5.1.1. There are also variational Bayesian algorithms for Gaussian mixtures that have garnered interest recently; in Section 5.1.2, we explain the mean field algorithm as a way to initialize EM. Finally, Section 5.1.3 describes a recently derived method of moments estimator that can be used as the initial θ.

5.1.1 Markov chain Monte Carlo

For a labeling v, let $n_{v,j}$ denote the number of observations assigned to label j, and let $\bar{X}_{v,j}$ be the mean of the observations with label j. Then we can almost rewrite the likelihood from (5.1) as

$$\begin{aligned}\prod_{i=1}^{n} p_\theta(X_i) &\propto \sum_{v\in\mathcal{V}} \prod_{j=1}^{k} e^{-\frac{1}{2\sigma^2}[n_{v,j}\|\mu_j-\bar{X}_{v,j}\|^2+\sum_{i:v_i=j}\|X_i-\bar{X}_{v,j}\|^2]} \\ &= \sum_{v\in\mathcal{V}} \left(\prod_{j=1}^{k} \frac{1}{n_{v,j}^{d/2}} e^{-\frac{1}{2\sigma^2}\sum_{i:v_i=j}\|X_i-\bar{X}_{v,j}\|^2}\right)\left(\prod_{j=1}^{k} n_{v,j}^{d/2} e^{-\frac{n_{v,j}}{2\sigma^2}\|\mu_j-\bar{X}_{v,j}\|^2}\right)\end{aligned}$$

a density for θ that is proportional to a mixture of k^n multivariate Gaussians. The expression is not legitimate, however, because many labelings have a component with zero observations. In a Bayesian context, on the other hand, it is possible to write the posterior as an actual Gaussian mixture. We use independent $N(\alpha, \frac{\sigma^2}{\beta} I_d)$ priors for $\mu_1, \ldots, \mu_k$.

$$\begin{aligned}\prod_{j=1}^{k} e^{-\frac{\beta}{2\sigma^2}\|\mu_j-\alpha\|^2} \prod_{i=1}^{n} p_\theta(X_i) &\propto \sum_{v\in\mathcal{V}} \prod_{j=1}^{k} e^{-\frac{1}{2\sigma^2}[n_{v,j}\|\mu_j-\bar{X}_{v,j}\|^2+\sum_{i:v_i=j}\|X_i-\bar{X}_{v,j}\|^2+\beta\|\mu_j-\alpha\|^2]} \\ &= \sum_{v\in\mathcal{V}} \underbrace{\left(\prod_{j=1}^{k} \frac{1}{(\beta+n_{v,j})^{d/2}} e^{-\frac{1}{2\sigma^2}[\frac{\beta n_{v,j}}{\beta+n_{v,j}}\|\bar{X}_{v,j}-\alpha\|^2+\sum_{i:v_i=j}\|X_i-\bar{X}_{v,j}\|^2]}\right)}_{w(v)} \\ &\quad \times \underbrace{\left(\prod_{j=1}^{k} (\beta+n_{v,j})^{d/2} e^{-\frac{(\beta+n_{v,j})}{2\sigma^2}\|\mu_j-\widetilde{\mu}_{v,j}\|^2}\right)}_{f_v(\theta)} \qquad (5.2)\end{aligned}$$

with $\widetilde{\mu}_{v,j} := \frac{\beta}{\beta+n_{v,j}}\alpha + \frac{n_{v,j}}{\beta+n_{v,j}}\bar{X}_{v,j}$. As the sample size grows, the posterior increasingly resembles the normalized likelihood. A draw from (5.2) would be achieved by drawing a labeling according to the weights proportional to $\{w(v) : v \in \mathcal{V}\}$ then drawing $\theta = (\mu_1, \ldots, \mu_k)$ from the Gaussian density proportional to f_v.

The usual MCMC algorithms have a target distribution as their steady state; they approach their steady states and thus produce approximate draws from the desired distribution *eventually*. The problem is that there are typically no guarantees on how long it will take before the process is approximately distributed according to the target. Here we introduce

a promising new annealing technique (with similarities to population annealing, parallel tempering, and evolutionary sampling) by which a Markov chain will be carefully guided to (5.2); an upcoming paper in collaboration with Andrew R. Barron and Jason M. Klusowski will describe the technique in more generality and will hopefully include statistical guarantees. A key insight for our application is the following observation by Barron.

Lemma 5.1.1. *Suppose V and V' are independent $\mathcal{V}$-valued random variables both drawn from probability density q, and let r be a probability density on $\mathcal{V}$. Given V and V', generate $B \sim Bern(a + \frac{r(V)-q(V)}{q(V)})$ assuming $a + \frac{r(V)-q(V)}{q(V)} \in [0,1]$. Then the random variable $\widetilde{V} := BV + (1-B)V'$ has marginal distribution R.*

We will use this trick to guide the weights of a Markov chain toward those desired in (5.2); we call the technique *teleport annealing*.

If the current and target distributions are in a smoothly time-parametrized family $\{Q_t\}$, we can approximate the crucial coin-flip quantity by

$$\frac{q_{t+h}(v) - q_t(v)}{q_t(v)} \approx h \underbrace{\frac{\partial}{\partial t} \log q_t(v)}_{\text{“}\delta_v(t)\text{”}} \tag{5.3}$$

for small enough h. Assume Q_t is a discrete distribution and let w_t be proportional to q_t.

$$\begin{aligned}
\delta_v(t) &:= \tfrac{\partial}{\partial t} \log q_t(v) \\
&= \tfrac{\partial}{\partial t} \log \frac{w_t(v)}{\sum_{v' \in \mathcal{V}} w_t(v')} \\
&= \tfrac{\partial}{\partial t} \log w_t(v) - \frac{1}{\sum_{v' \in \mathcal{V}} w_t(v')} \sum_{v'' \in \mathcal{V}} \tfrac{\partial}{\partial t} w_t(v'') \\
&= \tfrac{\partial}{\partial t} \log w_t(v) - \sum_{v'' \in \mathcal{V}} \frac{w_t(v'')}{\sum_{v' \in \mathcal{V}} w_t(v')} \tfrac{\partial}{\partial t} \log w_t(v'')
\end{aligned}$$

The second term is the expectation of the first according to the weights q_t. This reveals the advantage of the approximation (5.3): rather than calculating the normalizing factor, we will estimate the derivative of its logarithm.

In place of our original model, consider the time-parametrized family of Gaussian mix-

ture models with density proportional to

$$p_\theta^{(t)}(X_i) = \sum_{j=1}^{k} e^{-\frac{1}{2\sigma^2}[(1-t)\|X_i\|^2 + t\|X_i-\mu_j\|^2]}.$$

As in (5.2), we express the resulting posterior as a mixture

$$\prod_{j=1}^{k} e^{-\frac{\beta}{2\sigma^2}\|\mu_j-\alpha\|^2} \prod_{i=1}^{n} p_\theta^{(t)}(X_i) \propto \sum_{v\in\mathcal{V}} \prod_{j=1}^{k} e^{-\frac{1}{2\sigma^2}[tn_{v,j}\|\mu_j-\bar{X}_{v,j}\|^2 + t\sum_{i:v_i=j}\|X_i-\bar{X}_{v,j}\|^2 + \beta\|\mu_j-\alpha\|^2]}$$

$$= \sum_{v\in\mathcal{V}} \underbrace{\left(\prod_{j=1}^{k} \frac{1}{(\beta+tn_{v,j})^{d/2}} e^{-\frac{1}{2\sigma^2}[\frac{\beta tn_{v,j}}{\beta+tn_{v,j}}\|\bar{X}_{v,j}-\alpha\|^2 + t\sum_{i:v_i=j}\|X_i-\bar{X}_{v,j}\|^2]} \right)}_{w_t(v)}$$

$$\times \underbrace{\left(\prod_{j=1}^{k} (\beta+tn_{v,j})^{d/2} e^{-\frac{(\beta+tn_{v,j})}{2\sigma^2}\|\mu_j-\widetilde{\mu}_{v,j}^{(t)}\|^2} \right)}_{f_v^{(t)}(\theta)}$$

with $\widetilde{\mu}_{v,j}^{(t)} := \frac{\beta}{\beta+tn_{v,j}}\alpha + \frac{tn_{v,j}}{\beta+tn_{v,j}}\bar{X}_{v,j}$. At $t=1$, this is equal to the target distribution (5.2), while at $t=0$ it assigns equal probability to all labelings which makes it easy to initialize.

Our teleport annealing algorithm will begin by drawing N uniformly random labelings for the data, which will serve as initializations for parallel chains. For each chain, calculate the $t=0$ version of

$$\frac{\partial}{\partial t}\log w_t(v) = -\sum_{j=1}^{k}\left[\frac{d}{2}\frac{n_{v,j}}{\beta+tn_{v,j}} + \frac{1}{2\sigma^2}\frac{\beta n_{v,j}}{(\beta+tn_{v,j})^2}\|\bar{X}_{v,j}-\alpha\|^2 + \frac{1}{2\sigma^2}\sum_{i:v_i=j}\|X_i-\bar{X}_{v,j}\|^2\right]. \tag{5.4}$$

We then estimate the expectation of this quantity by averaging these values over the N chains. For each labeling v and time t, let $\hat{\delta}_v(t)$ denote (5.4) minus the average. Ideally, h times the largest absolute value of $\hat{\delta}_v(t)$ is no greater than .5, so that the coin-flips of Lemma 5.1.1 will be possible. For each chain, generate a coin-flip with heads probability $.5 + h\hat{\delta}_v(t)$ where v is the labeling of the chain in question. If heads, then leave the chain alone. If tails, then the chain teleports to the labeling of another chain chosen uniformly at random from the $N-1$ others.

Since $h\hat{\delta}_v(t)$ is only an estimate of an approximation of the required quantity from Lemma 5.1.1, we will follow each swapping step by running M steps of Gibbs sampling to move the distribution of labels and parameters closer to the true posterior for time t. Given a labeling v, the Gibbs sampling procedure draws independent Gaussian component means according to the density proportional to $f_v^{(t)}$. Then given component means, the label for the ith observation is assigned to label j with probability

$$\frac{e^{-\frac{1}{2\sigma^2}t\|X_i-\mu_j\|^2}}{\sum_{j'=1}^{k} e^{-\frac{1}{2\sigma^2}t\|X_i-\mu_{j'}\|^2}}.$$

These Gibbs sampling steps also help weaken the dependence among the parallel chains that arises from the teleportation step.

5.1.2 Variational Bayes

Calculus of variations is the study of optimization over a space of functionals. *Variational approximation* means identifying the functional in a set that is *closest* to a fixed target functional. When the target functional is a probability measure with a density only known up to a constant, the task of identifying the closest probability measure in a set is *variational Bayes*.

Mean field approximation

When relative entropy (with the target as the second argument) is used to quantify closeness and the search space comprises all probability measures with a specific product structure, the variational Bayes problem is called *mean field approximation*. The approximating distribution is the information projection of the target onto the set of all probability measures with the specified product structure. Inspired by the *mean field theory* of physics, Ghahramani introduced this technique for statistical learning.

Suppose some target distribution on $\mathcal{X} \times \mathcal{Y}$ can be represented as $P \otimes \{Q_x\}$ for some probability measure P on $\mathcal{X}$ and a probability kernel $\{Q_x : x \in \mathcal{X}\}$ of "conditional distributions" with densities $\{q_x\}$ relative to a σ-finite dominating measure. The relative entropy

from any product measure $\check{P} \otimes \check{Q}$ to the target is[1]

$$D(\check{P} \otimes \check{Q} \| P \otimes \{Q_x\}) = \mathbb{E}_{X \sim \check{P}} \mathbb{E}_{Y \sim \check{Q}} \log \frac{\check{p}(X)\check{q}(X)}{p(X) q_X(Y)}$$
$$= D(\check{P} \| P) + \mathbb{E}_{X \sim \check{P}} D(\check{Q} \| Q_X). \tag{5.5}$$

Section A discusses the reverse compensation identity (Theorem A.0.2) which implies that for any given $\check{P}$, the optimal choice of $\check{Q}$ is the $\check{P}$-geometric mixture of $\{Q_x\}$. Likewise, if the target distribution also has a representation as $Q \otimes \{P_y\}$ with the roles of marginal and conditional variable reversed, then the $\check{Q}$-geometric mixture of $\{P_y\}$ is the optimal choice of $\check{P}$ for fixed $\check{Q}$. The same logic continues to hold if the product structure has more than two components: any one component to be optimized plays the role of $\check{Q}$ in (5.5) while the rest of the components together play the role of $\check{P}$. The *mean field algorithm* constructs a product measure approximation by cycling through the components in this manner, updating each piece by setting it to the appropriate geometric mixture.[2] Equation (5.5) makes it clear that the algorithm is monotonic; each step can only decrease the relative entropy from the product approximation to the target; furthermore, it is guaranteed to converge to a local optimum [Bishop, 2006, Sec 10.1.1].

In Bayesian analysis, the posterior distribution represents an appropriate *belief* about the unknown parameter that arises from updating a prior belief based on observed data. However, posterior probabilities of parameter regions and posterior expectations of functions of the parameters are often challenging to calculate. If the parameters have a conjugate prior, then integrals can be calculated analytically; if the dimension of the parameter space is small, then integrals can be calculated numerically. Otherwise, practitioners turn to a variety of other approaches. Markov Chain Monte Carlo methods attempt to generate samples from the true posterior, but it is time-consuming and can do poorly when the posterior is badly multi-modal. Alternatively, the posterior's mean field approximation can

1. The freedom to choose the order of integration is justified by Tonelli's theorem because there is a an alternative representation of relative entropy with a non-negative integrand — see Lemma A.3.1.

2. Most sources explaining the mean field algorithm put the joint distributions in place of the conditional distributions in the expression that we call the "geometric mixture." Both definitions result in the same distribution, so one can use whichever is more convenient.

have an analytically tractable form. Most convenient is when each conditional family is an exponential family, in which case the geometric mixture is itself in that family as well[3]; one simply needs to update the hyper-parameters to identify the new distribution.

Variational Bayesian methods may be useful for calculating approximate posteriors, but they are also "statistically unsound" in a sense. Ideally, a statistical procedure should be eventually correct, if enough data is collected and the algorithm runs long enough. However, if for instance the correct posterior belief is that the variables are highly correlated, the product approximations will never indicate that belief regardless of the amount of data and run-time. Any change in the scale of a probability measure will have an exactly corresponding change in the scale of the mean field approximation approximation since relative entropy is scale-invariant. The resulting divergence between the probability measure and its approximation will remain unchanged in terms of relative entropy or any other f-divergence. Thus as a probability measure becomes more concentrated, these product approximations do not get closer to it, at least in terms of scale-invariant divergences.

Variational Bayesian methods trade correctness for convenience, but their popularity may be a sign that this trade-off is sometimes worth taking.

Mean field likelihood of GRBMs

The mean field algorithm applies to Bayesian estimation of Gaussian mixtures; the steps are explained in [Bishop, 2006, Sec 10.2]. In fact, the same approach works for approximating the likelihood of GRBMs as we now show. (Remember however that the likelihood cannot be normalized, so the theory of information projection does not apply exactly.)

In terms of a vector of component means $\theta = (\mu_1, \ldots, \mu_k)$ and label matrix Z, the log likelihood of X^n is

$$\log p_\theta(X^n|Z) = -\sum_{i=1}^{n}\sum_{j=1}^{k} Z_{i,j}\tfrac{1}{2\sigma^2}\|X_i - \mu_j\|^2 + \text{const.}$$

We will now apply mean field theory to consider approximating joint distributions over

3. The $\check{P}$-geometric mixture over an exponential family is the distribution corresponding to the expectation of the canonical parameter, so the problem is simplest when x is indexing a canonical parameterization.

(μ, Z) for which μ and Z are independent of each other. Let $r_{i,j}$ represent the probability that $Z_{i,j}$ equals 1; with that fixed distribution of Z, the optimal distribution for θ has a log density of the form

$$\begin{aligned}\mathbb{E}_{Z_{i,j}\sim\text{Bern}(r_{i,j})} \log p_\theta(X^n|Z) &= -\sum_{i=1}^{n}\sum_{j=1}^{k} r_{i,j}\tfrac{1}{2\sigma^2}\|X_i-\mu_j\|^2 + \text{const} \\ &= -\sum_{j=1}^{k}\tfrac{1}{2\sigma^2}\left[n_j\|\mu_j-\bar{\mu}_j\|^2 + \sum_{i=1}^{n} r_{i,j}\|X_i-\bar{\mu}_j\|^2\right] + \text{const}\end{aligned}$$

with $n_j := \sum_i r_{i,j}$ and $\bar{\mu}_j := \frac{1}{n_j}\sum_i r_{i,j}X_i$. Thus, we conclude that the optimal distribution has independent components and its distribution of μ_j is $N(\bar{\mu}_j, \sigma^2/n_j)$. Given a fixed distribution Q for θ, the optimal pmf of Z is proportional to

$$e^{\mathbb{E}_{\theta\sim Q}\log p_\theta(X^n|Z)} \propto e^{-\sum_i\sum_j Z_{i,j}\frac{1}{2\sigma^2}\mathbb{E}_{\theta\sim Q}\|X_i-\mu_j\|^2}.$$

The result is a product of independent multi-Bernoulli distributions, according to which $Z_{i,j}$ equals 1 with probability

$$\frac{e^{-\frac{1}{2\sigma^2}\mathbb{E}_{\theta\sim Q}\|X_i-\mu_j\|^2}}{\sum_l e^{-\frac{1}{2\sigma^2}\mathbb{E}_{\theta\sim Q}\|X_i-\mu_l\|^2}} = \frac{e^{-\frac{1}{2\sigma^2}[\|X_i-\mathbb{E}\mu_j\|^2+\mathbb{E}\|\mu_j-\mathbb{E}\mu_j\|^2]}}{\sum_l e^{-\frac{1}{2\sigma^2}[\|X_i-\mathbb{E}\mu_l\|^2+\mathbb{E}\|\mu_l-\mathbb{E}\mu_l\|^2]}}.$$

This provides a mean field algorithm for approximating the normalized likelihood. First, initialize a right stochastic matrix $[r_{i,j}]$ with distinct rows. Next, set $n_j := \sum_i r_{i,j}$ and $\bar{\mu}_j := \sum_i \frac{r_{i,j}}{n_j}X_i$. Update

$$r_{i,j} := \frac{e^{-\frac{1}{2\sigma^2}\|X_i-\bar{\mu}_j\|^2+1/2n_j}}{\sum_l e^{-\frac{1}{2\sigma^2}\|X_i-\bar{\mu}_l\|^2+1/2n_l}},$$

and repeat until convergence. (Comparing this to the discussion that began this chapter, we see that the mean field algorithm is nearly identical to the EM algorithm in this context; it only differs by down-weighting the responsibilities according to the component sample sizes.)

This algorithm can give undesirable behavior by entering a positive feedback loop that

sends some of the components' responsibilities to zero. Instead one may want to use the VB algorithm for the posterior when using independent $N(0, \sigma^2 I_d)$ priors on the component means; then the optimal distribution for component means given labels instead becomes $N(\frac{n_j}{n_j+1}\bar{\mu}_j, \frac{\sigma^2}{n_j+1})$ while the optimal responsibilities have $\frac{n_j}{n_j+1}\bar{\mu}_j$ in place of $\bar{\mu}_j$ and $n_j + 1$ in place of n_j.

Take EM steps from the resulting $(\bar{\mu}_1, \ldots, \bar{\mu}_k)$ to obtain a local optimizer of likelihood. We have yet to specify how the VB algorithm should be initialized; any of the other EM initializers described in this chapter could be used, for instance.

5.1.3 Method of third moments

Method of moments estimation procedures choose a model distribution whose moments match empirical moments of the data. For any distribution on $\mathbb{R}^d$, the first moments comprise a vector in $\mathbb{R}^d$. A particular value of the first moment may uniquely correspond to a model distribution if the model has fewer than d parameters. The second moments comprise a positive semi-definite matrix in $\mathbb{R}^{d \times d}$. Again, a particular combination of first and second moments may correspond to a unique model distribution if the model has few enough parameters. First and second moments are not sufficient to identify distributions within higher dimensional models, but one can then make reference to higher moments. Techniques have recently been developed to relate certain models' parameters to the generating distribution's tensor[4] of third moments and to efficiently find a model distribution corresponding approximately to a given set of first, second, and third moments.

The idea at the heart of the new tensor methods comes from Chang [1996] in the context of Markov models; the idea's generality and broader usefulness were not realized until Anandkumar et al. and Anandkumar et al. [2014]. Specifically, the tensor trick involves transforming a third-order tensor such that it becomes a sum of rank-one tensors built from orthonormal vectors that relate meaningfully to the model parameters. The method is best understood by example, and we now describe a tensor approach for estimating GRBMs adapted from explanations in Anandkumar et al. and [Hsu and Kakade, Sec 2].

4. The concept of *tensor* generalizes the concepts of vectors and matrices. It means an array that can have any specified number of dimensions.

Let P be a GRBM with component means $(\mu_1, \ldots, \mu_k)$ and component covariances $\sigma^2 I_d$ with the number of components k no greater than the dimension[5] d. Let μ denote the matrix comprising the component means as its column vectors. It will be convenient to give names to the sums of outer products of the component means:

$$\Psi := \mu\mu' = \sum_j \mu_j \mu_j' = \sum_j \mu_j \otimes \mu_j \qquad \text{and} \qquad \Gamma := \sum_j \mu_j \otimes \mu_j \otimes \mu_j.$$

It turns out that if Ψ and Γ are known, then it is possible to identify the component means within P, assuming they are linearly independent.[6] Let $Q\Lambda Q'$ be a spectral decomposition of Ψ with $\Lambda \in \mathbb{R}^{k \times k}$. Define the "whitening" matrix $W := \Lambda^{-1/2} Q'$; it transforms the component means into an orthonormal set $\{u_j := W\mu_j\}$. We verify orthonormality by checking that $W\mu$ is the inverse of its transpose.

$$\begin{aligned}(W\mu)(W\mu)' &= W(\mu\mu')W' \\ &= \Lambda^{-1/2} Q'(Q\Lambda Q')Q\lambda^{-1/2} \\ &= I_k\end{aligned}$$

Apply W' to each "side" of Γ to define

$$\begin{aligned}G :&= W\Gamma_{W'}W' \\ &= \sum_j (W\mu_j) \otimes (W\mu_j) \otimes (W\mu_j) \\ &= \sum_j u_j \otimes u_j \otimes u_j.\end{aligned}$$

5. Additional variables can be constructed from the data (e.g. second-order products) if one wants to be able to use more components.

6. If desired, one can ensure that the unknown component means are linearly independent with probability 1 by randomly translating the space.

Let vector subscripts denote application[7] into a tensor:

$$\begin{aligned} G_v &:= G \cdot v \\ &= \left[\sum_j u_j \otimes u_j \otimes u_j \right] \cdot v \\ &= \sum_j (v' u_j) u_j u_j'. \end{aligned}$$

As long as v is not orthogonal to any of the $\{u_j\}$, a spectral decomposition of G_v reveals the $\{u_j\}$ as its eigenvectors, up to sign.[8] To determine whether a sign should be reversed, compare the corresponding eigenvalue of the spectral decomposition's proposal to what the eigenvalue should be: the inner product of v with the proposed u_j. If they agree, the proposed eigenvector is correct, otherwise its negative is correct.

At last, the original component means are recovered by undoing the whitening transformation $\mu_j = Q\Lambda^{1/2} u_j$.

We have seen how knowledge of Ψ and Γ allows one to find the model GRBM. Next, we will learn how Ψ and Γ relate to the first three moments of P. The first moment is $\bar{\mu} := \mathbb{E}_{X \sim P} X = \frac{1}{k} \sum_j \mu_j$. Letting $Z \sim N(0, \sigma^2 I_d)$, the matrix of second moments is

$$\begin{aligned} \mathbb{E}_{X \sim P} X X' &= \sum_j \tfrac{1}{k} [\mathbb{E}(\mu_j + \sigma Z)(\mu_j + \sigma Z)'] \\ &= \tfrac{1}{k} \Psi + \sigma^2 I_d, \end{aligned}$$

and the tensor of third moments has as its (d_1, d_2, d_3)-entry (with subscripts of X and Z

7. In the context of this example, it is usually not important to keep track of which side of the tensor a vector is being multiplied into.

8. It is easy to ensure that no eigenvectors are missed by applying enough vectors into G; for instance, any orthonormal basis of $\mathbb{R}^d$ suffices.

denoting coordinates)

$$\begin{aligned}
\mathbb{E}_{X\sim P} X_{d_1} X_{d_2} X_{d_3} &= \sum_j \tfrac{1}{k} [\mathbb{E}(\mu_{j,d_1} + Z_{d_1})(\mu_{j,d_2} + Z_{d_2})(\mu_{j,d_3} + Z_{d_3})] \\
&= \sum_j \tfrac{1}{k} [\mu_{j,d_1}\mu_{j,d_2}\mu_{j,d_3} + \mathbb{E} Z_{d_1} Z_{d_2} \mu_{j,d_3} \mathbb{I}_{d_1=d_2} \\
&\quad + \mathbb{E} Z_{d_1} Z_{d_3} \mu_{j,d_2} \mathbb{I}_{d_1=d_3} + \mathbb{E} Z_{d_2} Z_{d_3} \mu_{j,d_1} \mathbb{I}_{d_2=d_3}] \\
&= \sum_j \tfrac{1}{k} \mu_{j,d_1}\mu_{j,d_2}\mu_{j,d_3} \\
&\quad + \sum_j \tfrac{1}{k} \sigma^2 (\mu_{j,d_3} \mathbb{I}_{d_1=d_2} + \mu_{j,d_2} \mathbb{I}_{d_1=d_3} + \mu_{j,d_1} \mathbb{I}_{d_2=d_3}) \\
&= \sum_j \tfrac{1}{k} \mu_{j,d_1}\mu_{j,d_2}\mu_{j,d_3} \\
&\quad + \sigma^2 (\bar{\mu}_{d_3} \mathbb{I}_{d_1=d_2} + \bar{\mu}_{d_2} \mathbb{I}_{d_1=d_3} + \bar{\mu}_{d_1} \mathbb{I}_{d_2=d_3}).
\end{aligned}$$

Notice that the first term is $1/k$ times the (d_1, d_2, d_3)-entry of Γ.

From these moment derivations, we see that with data $X_1 = (X_{1,1}, \ldots, X_{1,d}), \ldots, X_n = (X_{n,1}, \ldots, X_{n,d})$, an unbiased estimate for Ψ is

$$\hat{\Psi} := k \left[\tfrac{1}{n} \sum_i X_i X_i' - \sigma^2 I_d \right],$$

and an unbiased estimate for the (d_1, d_2, d_3)-entry of Γ is

$$\hat{\Gamma}_{d_1,d_2,d_3} := k \left[\tfrac{1}{n} \sum_i X_{i,d_1} X_{i,d_2} X_{i,d_3} - \sigma^2 (\bar{X}_{d_3} \mathbb{I}_{d_1=d_2} + \bar{X}_{d_2} \mathbb{I}_{d_1=d_3} + \bar{X}_{d_1} \mathbb{I}_{d_2=d_3}) \right]$$

with $\bar{X}$ denoting the sample mean.

One can apply the whitening and spectral decomposition procedures described above to the estimates $\hat{\Psi}$ and $\hat{\Gamma}$ to get estimates $\hat{\mu}_1, \ldots, \hat{\mu}_k$ of the component means. This is considered a method of moments estimator, as it corresponds to finding the model GRBM whose first three moments approximately match the empirical moments.

Given a tensor that is a sum of no more than d rank-one tensors built of *orthonormal*

vectors

$$G = \sum_j u_j \otimes u_j \otimes u_j,$$

we noted that the $\{u_j\}$ comprise the spectral decomposition of G_u for a uniformly random unit vector u. Anandkumar et al. [2014] also describes an alternative algorithm for finding $\{u_j\}$ called the *tensor power method*. It starts with a random u, then iterates

$$u^{(t+1)} \leftarrow \frac{G_{u^{(t)}} u^{(t)}}{\|G_{u^{(t)}} u^{(t)}\|}.$$

The $\{u_j\}$ are fixed points. They are also maximizers of $u' G_u u$. In fact, Hölder's identity (Corollary B.0.5) exactly characterizes the change in that objective function in terms of unnormalized Rényi divergences D_λ, defined in Section A.1. An immediate consequence of Theorem 5.1.2 is that the nonlinear power steps are monotonic in the orthonormal case.

Theorem 5.1.2. *Let $M_u = \sum_{k=1}^K c_k (u'\alpha_k)^p \alpha_k \alpha_k'$ with integer $p \geq 1$, reals $c_1, \ldots, c_K \geq 0$, and orthonormal vectors $\alpha_1, \ldots, \alpha_K$. Define $J(u) = u' M_u u$, and let R denote a nonlinear power iteration step. Then*

$$\frac{J(R^{t+1}(u))}{J(R^t(u))} = \exp\left(D_{1/(p+3)}(P_t \| Q_t) + \frac{p+2}{2} D_{2/(p+3)}(P_t \| Q_t) \right)$$

where P_t and Q_t are probability distributions on $\{1, \ldots, K\}$ with masses

$$P_t(\{k\}) \propto c_k^{\frac{(p+2)(p+1)^{t+1}-2}{p}} (u'\alpha_k)^{(p+2)(p+1)^{t+1}} \qquad \text{and} \qquad Q_t(\{k\}) \propto (u'\alpha_k)^{2(p+1)^t}.$$

Note that in practice, the estimated version of G does not have such an orthonormal internal structure.

5.2 Simulation

Next we will simulate data from complicated distributions to compare these algorithms' ability to find parameters with large likelihood. Our observations will not be exhaustive or

final. The purpose is simply to get a glimpse of a small piece the picture and to provide example code that others can use for running or comparing these algorithms.

Our simulation generates 100 Gaussian mixtures according to the following procedure: the number of components is Pois(100), their relative weights are independently Exp(1), their means are drawn independently from $N(0, I_d)$, and finally a number is drawn from Exp(1) and is used as the rate parameter in an exponential distribution that generates the components' standard deviations. Each of these 100 randomly generated distributions is used to generate an iid random sample with $n = 100$ and $d = 12$. The GRBM model for fitting the data uses $\sigma = 1$ and $k = 8$.

For each of the 100 datasets, the five algorithms described above are used to initialize EM. The simplest is a *Gaussian* initialization which simply draws initial component means independently from a Normal centered at the sample average and having standard deviations equal to 2 times the standard deviations the data in each dimension. The *mean field* algorithm from Section 5.1.2 is also performed using simple Gaussian initialization for itself. The *method of third moments* described in Section 5.1.3 is tried as well; repeated use tensor power iterations gives us $k = 8$ vectors, with each subsequent initializer randomly selected from the space orthogonal to the previously revealed vectors. Next, the Hartigan-Wong *k-means* algorithm generates initializers. The next method is three steps of Gibbs sampling starting with uniformly random labels. Finally, we perform the *teleport annealing* algorithm from Section 5.1.1 (using $\alpha = 0$, $\beta = 1$, and $h = .05$) that alternates between five teleportation steps and one Gibbs sampling step.[9]

Each algorithm is allowed roughly the same amount of computing time. Specifically, we perform a single run of the teleport annealing algorithm that generates 5000 initial points, runs EM from them, and calculates the resulting likelihoods; the other algorithms are tried repeatedly for about that same amount of time. Table 5.1 lists the number of initializers we will receive from each algorithm. After performing EM to reach local optima, the largest log likelihood achieved by each initialization algorithm is recorded.

The final result is a data frame of 100 rows (one for each random dataset) and 5 columns

9. Earlier simulations were performed to tune the settings for the algorithms under final consideration. The code is available at quantitations.com/research.

Gaussian	mean field	third moments	k-means	Gibbs sampling	teleport annealing
6250	4500	6000	20000	5000	5000

Table 5.1: The number of initializers each algorithm generated for the simulated datasets.

(one for each algorithm); entry (i, j) is the best log-likelihood achieved by the jth algorithm on the ith dataset. Each row is then "standardized" by subtracting the average of the six and dividing by their standard deviation plus one. The Gaussian initialization is considered a baseline, and its best log likelihood (standardized) is subtracted from that of the other algorithms. The grid of scatterplots in Figure 5.1 shows how well our algorithms did relative to each other. In this simulation, the k-means algorithm was remarkably dominant at providing initializers that converged to parameter values of large likelihood. This is likely explained in part by the fact that the algorithm works rapidly and thus gets many more tries.

To the extent that this simulation generalizes to other datasets, it suggests that a good way to find parameter values of large likelihood is to perform EM using a vast number of k-means solutions as initializers.

5.3 Proofs

*Proof of **Lemma 5.1.1**.* The proof is easiest to understand for discrete $\mathcal{V}$ with q and r as probability mass functions.

$$\begin{aligned}
\mathbb{P}\{\widetilde{V} = v\} &= \mathbb{P}\{V = v \cap B = 1\} + \mathbb{P}\{V' = v \cap B = 0\} \\
&= \mathbb{P}\{V = v\}\mathbb{P}\{B = 1 | V = v\} + \mathbb{P}\{V' = v\}\mathbb{P}\{B = 0\} \\
&= q(v)(a + \tfrac{r(v)-q(v)}{q(v)}) + q(v) \sum_{v' \in \mathcal{V}} q(v')(1 - a - \tfrac{r(v')-q(v')}{q(v')}) \\
&= aq(v) + r(v) - q(v) + (1-a)q(v) - q(v) \sum_{v' \in \mathcal{V}} (r(v') - q(v')) \\
&= r(v)
\end{aligned}$$

The logic extends beyond the case of discrete $\mathcal{V}$ if mathematical care is taken. □

*Proof of **Lemma 5.1.2**.* Andrew R. Barron realized that Hölder's inequality can be used

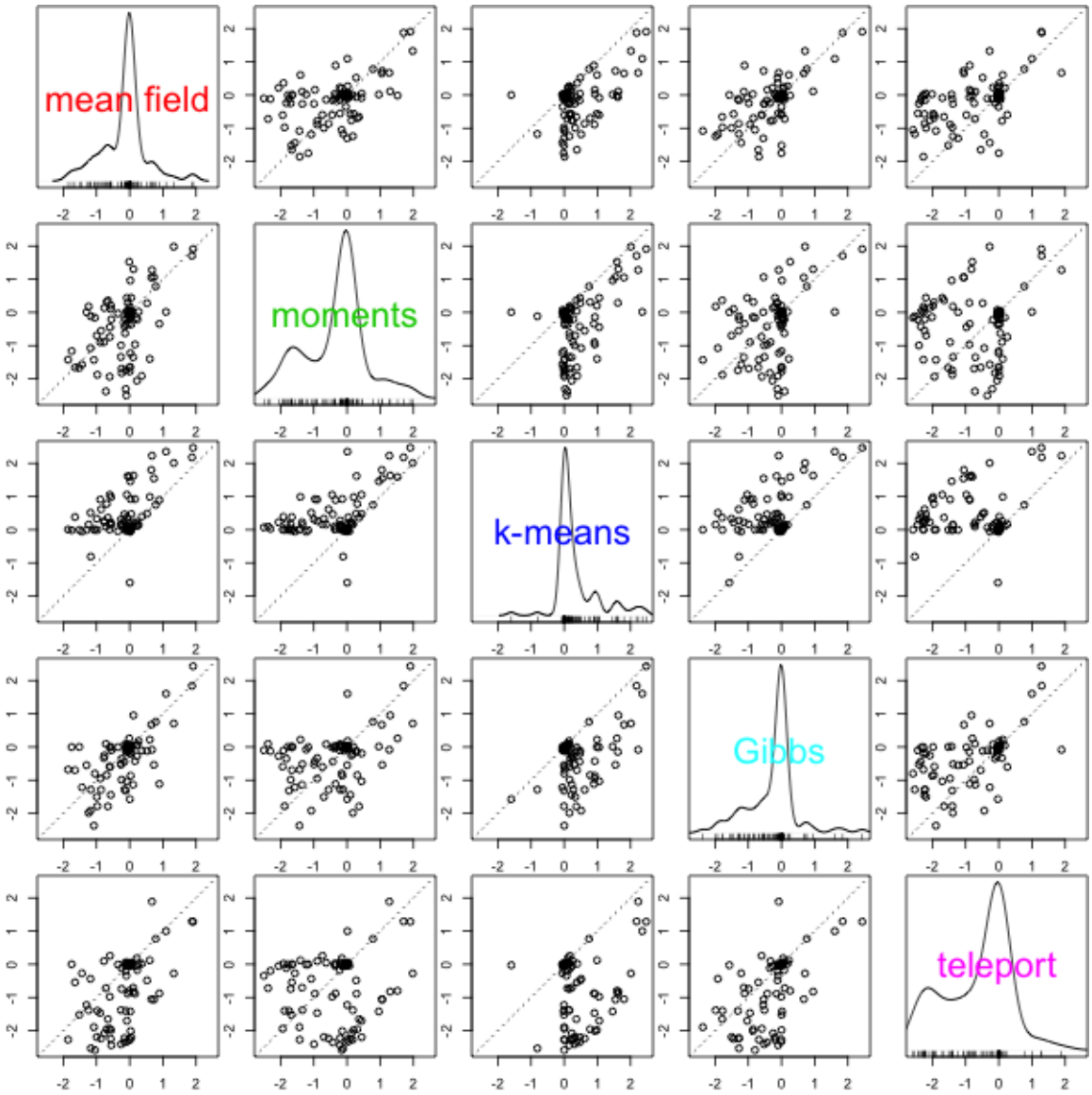

Figure 5.1: For each of 100 randomized datasets, the six algorithms under consideration were used to generate the number of initializers specified in Table 5.1, and each algorithm's best resulting log likelihood was recorded. For each of the 100 trials, the five algorithm's values were standardized, then the Gaussian initializations' best log likelihood value was subtracted from the others. The diagonal of our grid shows how the algorithm did relative to the simple Gaussian algorithm, while the off-diagonals show how they compared head-to-head. The $y = x$ dotted line splits the points according to which algorithm found an estimator of larger likelihood.

to show that the nonlinear power iterations monotonically increase the objective function in this context. A more precise result comes from instead using Hölder's *identity* (particularly Theorem B.0.6).

R^t has a convenient closed form that can be seen by looking at the first few iterates.

$$\begin{aligned} R(u) &\propto M_u u \\ &= \sum c_k u_k^{p+1} \alpha_k \end{aligned}$$

where $u_1, \ldots, u_K$ denote the "coordinates" of u in the α basis, i.e. the inner products of u with $\alpha_1, \ldots, \alpha_K$. The updated coordinates are proportional to $c_k(u_k)^{p+1}$. These can be substituted in to find the next iterate.

$$\begin{aligned} R^2(u) &:= R(R(u)) \\ &\propto \sum c_k (c_k u_k^{p+1})^{p+1} \alpha_k \\ &= \sum c_k c_k^{p+1} u_k^{(p+1)^2} \alpha_k \end{aligned}$$

The third step is

$$\begin{aligned} R^3(u) &:= R(R^2(u)) \\ &\propto \sum c_k (c_k c_k^{p+1} u_k^{(p+1)^2})^{p+1} \alpha_k \\ &= \sum c_k c_k^{p+1} c_k^{(p+1)^2} u_k^{(p+1)^3} \alpha_k. \end{aligned}$$

We can see the pattern: $R^k(u)$ will have coordinates proportional to $u_k^{(p+1)^t}$ times c_k taken to the following power:

$$\begin{aligned} \sum_{j=0}^{t-1} (p+1)^j &= \frac{(p+1)^t - 1}{(p+1) - 1} \\ &= \frac{(p+1)^t - 1}{p}. \end{aligned}$$

To be exact,

$$\begin{aligned} R^t(u) &= \frac{\sum c_k^{-1/p}(c_k^{1/p}u_k)^{(p+1)^t}\alpha_k}{\|\sum c_k^{-1/p}(c_k^{1/p}u_k)^{(p+1)^t}\alpha_k\|} \\ &= \frac{\sum c_k^{-1/p}(c_k^{1/p}u_k)^{(p+1)^t}\alpha_k}{\sqrt{\sum c_k^{-2/p}(c_k^{1/p}u_k)^{2(p+1)^t}}} \\ &= \frac{\sum c_k^{-1/p}r_{t,k}\alpha_k}{\sqrt{\sum c_k^{-2/p}r_{t,k}^2}} \end{aligned}$$

where $r_{t,k} := (c_k^{1/p}u_k)^{(p+1)^t}$. Notice the recursive relationship $r_{t+1,k} = r_{t,k}^{p+1}$. We can see how the algorithm evolves from u: it rapidly places an increasing proportion of its weight on whichever coordinate has the largest $c_k^{1/p}u_k$.

The objective function simplifies to

$$J(u) = \sum c_k(u'\alpha_k)^{p+2}.$$

After t iterations, it is

$$\begin{aligned} J(R^t(u)) &= \sum_k c_k(R^t(u)'\alpha_k)^{p+2} \\ &= \frac{\sum c_k(c_k^{-1/p}r_{t,k})^{p+2}}{(\sum c_k^{-2/p}r_{t,k}^2)^{(p+2)/2}} \\ &= \frac{\sum c_k^{-2/p}r_{t,k}^{(p+2)}}{(\sum c_k^{-2/p}r_{t,k}^2)^{(p+2)/2}}. \end{aligned}$$

The ratio of interest is

$$\begin{aligned} \frac{J(R^{t+1}(u))}{J(R^t(u))} &= \frac{\sum c_k^{-2/p}r_{t+1,k}^{(p+2)}/(\sum c_k^{-2/p}r_{t+1,k}^2)^{(p+2)/2}}{\sum c_k^{-2/p}r_{t,k}^{(p+2)}/(\sum c_k^{-2/p}r_{t,k}^2)^{(p+2)/2}} \\ &= \frac{\sum c_k^{-2/p}r_{t,k}^{(p+1)(p+2)}/(\sum c_k^{-2/p}r_{t,k}^{2(p+1)})^{(p+2)/2}}{\sum c_k^{-2/p}r_{t,k}^{(p+2)}/(\sum c_k^{-2/p}r_{t,k}^2)^{(p+2)/2}}. \end{aligned}$$

Considering the $c_k^{-2/p}$ as weights, we apply Hölder's identity to the numerator of $J(R^t(u))$.

$$\begin{aligned}\sum c_k^{-2/p} r_{t,k}^{(p+2)} &= \sum c_k^{-2/p} [r_{t,k}^{(p+1)(p+2)}]^{1/(p+3)} [r_{t,k}^2]^{(p+2)/(p+3)} \\ &= e^{-D_{1/(p+3)}(P_t \| Q_t)} \left[\sum c_k^{-2/p} r_{t,k}^{(p+1)(p+2)}\right]^{1/(p+3)} \left[\sum c_k^{-2/p} r_{t,k}^2\right]^{(p+2)/(p+3)}\end{aligned}$$

Making this substitution,

$$\begin{aligned}\frac{J(R^{t+1}(u))}{J(R^t(u))} &= e^{D_{1/(p+3)}(P_t \| Q_t)} \frac{(\sum c_k^{-2/p} r_{t,k}^{(p+1)(p+2)})^{1-1/(p+3)} (\sum c_k^{-2/p} r_{t,k}^2)^{(p+2)/2-(p+2)/(p+3)}}{(\sum c_k^{-2/p} r_{t,k}^{2(p+1)})^{(p+2)/2}} \\ &= e^{D_{1/(p+3)}(P_t \| Q_t)} \left[\frac{(\sum c_k^{-2/p} r_{t,k}^{(p+1)(p+2)})^{2/(p+3)} (\sum c_k^{-2/p} r_{t,k}^2)^{(p+1)/(p+3)}}{\sum c_k^{-2/p} r_{t,k}^{2(p+1)}}\right]^{(p+2)/2}.\end{aligned} \tag{5.6}$$

Finally, realize that the denominator of this fraction can be expressed as

$$\begin{aligned}\sum c_k^{-2/p} r_{t,k}^{2(p+1)} &= \sum c_k^{-2/p} [r_{t,k}^{(p+1)(p+2)}]^{2/(p+3)} [r_{t,k}^2]^{(p+1)/(p+3)} \\ &= e^{-D_{2/(p+3)}(P_t \| Q_t)} \left[\sum c_k^{-2/p} r_{t,k}^{(p+1)(p+2)}\right]^{2/(p+3)} \left[\sum c_k^{-2/p} r_{t,k}^2\right]^{(p+1)/(p+3)}\end{aligned}$$

where the last step used Hölder's identity once again. Substitute this into (5.6) to complete the proof. □

Appendix A

The compensation identities

Theorem A.0.1, called the *compensation identity* by [Topsøe, 2001, Thm 9.1], conveniently decomposes the expected relative entropy from a random probability measure to a fixed probability measure.[1]

Theorem A.0.1 (The compensation identity)**.** *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure γ, and suppose that $(x, y) \mapsto q_x(y)$ is product measurable. Let $X \sim P$ be an $\mathcal{X}$-valued random element. For any probability measure R defined on the same measurable space as γ,*

$$\mathbb{E}D(Q_X\|R) = D(\bar{Q}_P\|R) + \mathbb{E}D(Q_X\|\bar{Q}_P)$$

where $\bar{Q}_P$ represents the P-mixture over $\{q_x\}$.

A less familiar decomposition, which we will call the *reverse compensation identity*, holds when the expected relative entropy's *second* argument is random rather than its first. Instead of a mixture, it involves a *geometric-mixture*.[2] We define the *P-geometric mixture* of $\{q_x\}$ to be the probability measure with density

$$\widetilde{q}_P(y) := \frac{e^{\mathbb{E}_{X\sim P}\log q_X(y)}}{\int e^{\mathbb{E}_{X\sim P}\log q_X(y)} d\gamma(y)}.$$

1. This chapter's proofs are at the end.

2. What we call a "geometric mixture" is sometimes called a "log mixture" or "log-convex mixture," for instance by [Grünwald, 2007, Sec 19.6].

Jensen's inequality and Tonelli's theorem together provide an upper bound for the denominator.

$$\begin{aligned}\int e^{\mathbb{E}\log q_X(y)} d\gamma(y) &\leq \mathbb{E}\int e^{\log q_X(y)} d\gamma(y) \\ &= 1\end{aligned}$$

This integral can be zero, however, in which case the geometric-mixture is not well-defined.[3]

Theorem A.0.2 (The reverse compensation identity)**.** *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure γ, and suppose that $(x, y) \mapsto q_x(y)$ is product measurable. Let $X \sim P$ be an $\mathcal{X}$-valued random element. If $\int e^{\mathbb{E}\log q_X(y)} d\gamma(y) > 0$, then for any probability measure R defined on the same measurable space as γ,*

$$\mathbb{E}D(R\|Q_X) = D(R\|\widetilde{Q}_P) + \mathbb{E}D(\widetilde{Q}_P\|Q_X)$$

where $\widetilde{Q}_P$ represents the P-geometric mixture over $\{q_x\}$.

A two-point distribution version of Theorem A.0.2 is implied by [Csiszár and Matúš, 2003, Eq (3) with (4)] and similarly for any finite set of discrete distributions by [Veldhuis, 2002, Eq (9)].

A.1 Bias-variance decomposition

Theorems A.0.1 and A.0.2 are perfectly analogous to the bias-variance decomposition for Hilbert-space-valued random vectors.[4] The expected divergence from the a random element to a fixed element decomposes into the divergence from a "centroid" of the random element to that fixed element plus the internal variation of the random element from its centroid.[5]

3. An example of such a pathological case is when q_X has positive probabilities on two densities that are mutually singular.

4. In fact, the compensation identity and bias-variance decomposition are both instances of the same decomposition that works for all Bregman divergences — see [Telgarsky and Dasgupta, Lem 3.5] and Pfau [2013].

5. It follows that the centroid is the choice of fixed element that has the smallest possible expected divergence from the random element.

We suggest a notation that makes use of this intuition:

$$\begin{aligned}\bar{\mathbb{V}}Q_X :&= \inf_R \mathbb{E}D(Q_X\|R)\\ &= \mathbb{E}D(Q_X\|\bar{Q}_P)\end{aligned}$$

and[6]

$$\begin{aligned}\widetilde{\mathbb{V}}Q_X :&= \inf_R \mathbb{E}D(R\|Q_X)\\ &= \begin{cases}\mathbb{E}D(\widetilde{Q}_P\|Q_X), & \text{if } \int e^{\mathbb{E}\log q_X(y)}d\gamma(y) > 0\\ \infty, & \text{otherwise.}\end{cases}\end{aligned}$$

We also suggest the terminology *information risk* (*I-risk*), *information bias* (*I-bias*) squared, and *information variance* (*I-variance*) for the quantities in the compensation identity as well as the terminology *reverse information risk* (*rI-risk*), *reverse information bias* (*rI-bias*) squared, and *reverse information variance* (*rI-variance*) for the quantities in the reverse compensation identity. The language introduced here comports with that of information projections (I-projections) and reverse information projections (rI-projections).

There are straight-forward information-theoretic interpretations of the variance-like quantities. Roughly speaking, $\bar{\mathbb{V}}Q_X$ represents the smallest possible expected code-length redundancy one can achieve when the *coding* distribution is the random Q_X; to achieve it, one sets the decoding distribution to be $\bar{Q}_P$. On the other hand, $\widetilde{\mathbb{V}}Q_X$ represents the smallest possible expected code-length redundancy when the *decoding* distribution is the random Q_X; to achieve it, one sets the coding distribution to be $\widetilde{Q}_P$.

$\bar{\mathbb{V}}Q_X := \inf_R \mathbb{E}D(Q_X\|R)$ is equivalent to the familiar *mutual information* between X and Y considering $(X,Y) \sim P \otimes \{Q_x\}$ as a joint distribution. More generally, the term *f-informativity* has been used by Csiszár [1972] for $\inf_R \mathbb{E}D_f(Q_X\|R)$ in the context of an arbitrary f-divergence D_f.

Two-point distribution versions of these variance-like quantities are often used as diver-

6. The alternative representation of $\widetilde{\mathbb{V}}$ presented below is justified by Lemma A.3.4.

gences. The *Jensen-Shannon divergence* between probability measures Q and R is $\bar{\mathbb{V}}$ of the random probability measure that takes values Q and R each with probability $1/2$.

$$D_{\mathrm{JS}}(Q,R) := \tfrac{1}{2}D\left(Q\|\tfrac{Q+R}{2}\right) + \tfrac{1}{2}D\left(R\|\tfrac{Q+R}{2}\right)$$

Unnormalized Bhattacharyya divergence[7] is the $\widetilde{\mathbb{V}}$ analogue:

$$D_{\mathrm{UB}}(Q,R) = \tfrac{1}{2}D\left(\tfrac{\sqrt{qr}}{\gamma\sqrt{qr}}\|q\right) + \tfrac{1}{2}D\left(\tfrac{\sqrt{qr}}{\gamma\sqrt{qr}}\|r\right)$$

where q and r are densities of Q and R with respect to γ, and $\gamma\sqrt{qr}$ is short-hand for $\int\sqrt{q(y)r(y)}d\gamma(y)$ using de Finetti notation.[8] The derivation is straight-forward using the definition $D_{\mathrm{UB}}(Q,R) := \log\frac{1}{\gamma\sqrt{qr}}$, but it is more easily seen via Lemma A.3.2. *Unnormalized Renyi divergence* is a generalization $D_\lambda(Q\|R) := \log\frac{1}{\gamma\, q^\lambda r^{1-\lambda}}$, and a random distribution that takes values Q with probability λ and R with probability $1-\lambda$ has a $\widetilde{\mathbb{V}}$ of $D_\lambda(Q\|R)$.

The compensation identities can provide insights regarding regularization, and we conclude this subsection with one such observation. A simple way to regularize a point-estimator $\hat{\theta}$ is by shrinking it toward any constant point θ_0. The variance of $[1-\lambda]\hat{\theta}+\lambda\theta_0$ is $[1-\lambda]^2$ times the variance of the original estimator $\hat{\theta}$. Similarly, a density estimator's I-variance can always be decreased by *mixing with a fixed distribution.*

Theorem A.1.1. *Let $\{q_x : x\in\mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure γ, and suppose that $(x,y)\mapsto q_x(y)$ is product measurable. Let X be an $\mathcal{X}$-valued random element. For any fixed known probability measure $\check{Q}$, the I-variance of the mixture $\bar{\mathbb{V}}([1-\lambda]Q_X+\lambda\check{Q})$ is non-increasing as $\lambda\in[0,1]$ increases. The I-variance is strictly decreasing unless Q_X equals $\check{Q}$ with probability one.*

7. This terminology is based on [Grünwald, 2007, Eq (19.38).].

8. The de Finetti notation writes measures like ordinary functionals that can be applied to measurable functions; it is summarized (and advocated) in [Pollard, 2002, Sec 1.4].

A.2 Bayes rules

Suppose a random probability measure Q_X is known to have $X \sim P$. The compensation identity tells us that $\bar{Q}_P$ is the "best" fixed representative for Q_X in terms of having the smallest expected $D(Q_X \| \cdot)$. Likewise, the reverse compensation identity tells us that $\widetilde{Q}_P$ is the representative with the smallest expected $D(\cdot \| Q_X)$.

This situation arises in decision theory with regard to Bayesian statistics. A Bayes rule is a decision rule that minimizes the expected risk, where the expectation is taken with respect to some "prior" distribution on the parameter space. Let L be a non-negative product measurable loss function, let d denote a decision rule, and let P_0 and $P_n(Y)$ denote prior and posterior distributions on a parameter space $\mathcal{X}$ with "data" Y taking values in $\mathcal{Y}$. Tonelli's theorem justifies a change in the order of integration that verifies

$$\mathbb{E}_{X \sim P_0} \mathbb{E}_{Y \sim Q_X} L(X, d(Y)) = \mathbb{E}_{Y \sim \bar{Q}_{P_0}} \mathbb{E}_{X \sim P_n(Y)} L(X, d(Y)).$$

Any decision rule minimizing the posterior expected loss $\mathbb{E}_{X \sim P_n(y)} L(X, d(y))$ for every possible data value $y \in \mathcal{Y}$ is clearly a Bayes rule.

If the loss used is $L(X, \cdot) = D(Q_X \| \cdot)$, the compensation identity tells us that the *posterior mixture* $d(Y) = \bar{Q}_{P_n(Y)}$ minimizes the expected loss and is therefore the Bayes rule. From a Bayesian point of view, this observation is particularly satisfying since the posterior mixture is also the Bayesian's belief about what the next draw of datum will be; for this reason, it is also called the "predictive mixture." Furthermore, based on the information theoretic interpretation of relative entropy as coding redundancy, it makes sense that the true data-generating distribution Q_X is the first argument.

On the other hand, for the loss $L(X, \cdot) = D(\cdot \| Q_X)$, we know by the reverse compensation identity that $d(Y) = \widetilde{Q}_{P_n(Y)}$ must be the Bayes rule. We will call this distribution the posterior geometric mixture. It does not have the Bayesian interpretation of the ordinary posterior mixture, but from a purely decision theoretic point of view, the posterior geometric mixture is a perfectly sensible choice as well.

As the sample size increases, the distinction between the posterior mixture and the

posterior geometric mixture can become unimportant. This is because a highly concentrated P (with concentration quantified by $\widetilde{\mathbb{V}}$) has its centroids $\widetilde{Q}_P$ and $\bar{Q}_P$ close to each other. To see why, first note that $\widetilde{q}_P \leq \bar{q}_P / \int e^{\mathbb{E} \log q_X(y)} d\gamma(y)$. This results from applying Jensen's inequality to the numerator of $\widetilde{q}_P$:

$$\begin{aligned} e^{\mathbb{E} \log q_X(y)} &\leq e^{\log \mathbb{E} q_X(y)} \\ &= \mathbb{E} q_X(y) \end{aligned}$$

with equality if and only if P is a point-mass. Thus

$$\begin{aligned} D(\widetilde{Q}_P \| \bar{Q}_P) &= \mathbb{E}_{Y \sim \widetilde{Q}_P} \log \frac{\widetilde{q}_P(Y)}{\bar{q}_P(Y)} \\ &\leq \widetilde{Q}_P \log \frac{\bar{q}_P / \int e^{\mathbb{E} \log q_X(y)} d\gamma(y)}{\bar{q}_P} \\ &= \log \frac{1}{\int e^{\mathbb{E} \log q_X(y)} d\gamma(y)} \\ &= \widetilde{\mathbb{V}}_{X \sim P} Q_X. \end{aligned}$$

It would be nice if we could relate $\widetilde{\mathbb{V}}$ to more familiar ways of quantifying concentration. For now, we can at least relate it to more familiar conditions for asymptotic convergence. Given a sequence of probability measures (P_n) on $\mathcal{X}$ indexing a family $\{Q_x\}$, we will point to a few sufficient conditions for ensuring that the sequence has $\widetilde{\mathbb{V}}_{X \sim P_n} Q_X$ going to zero. A helpful observation is that by Fatou's Lemma,

$$\liminf_n \int e^{\mathbb{E}_{X \sim P_n} \log q_X(y))} d\gamma(y) \geq \int e^{\liminf_n \mathbb{E}_{X \sim P_n} \log q_X(y)} d\gamma(y)$$

If at every y, $\mathbb{E}_{X \sim P_n} \log q_X(y)$ converges to the log of some limiting probability density, then this bound becomes 1, which makes

$$\limsup_n \widetilde{\mathbb{V}}_{X \sim P_n} Q_X = \log \frac{1}{\liminf_n \int e^{\mathbb{E}_{X \sim P_n} \log q_X(y)}} d\gamma(y)$$

be bounded by 0. If (P_n) converges in total variation to a point-mass δ_{x_0}, then indeed $\mathbb{E}_{X \sim P_n} \log q_X(y) \to \log q_{x_0}(y)$ at every y. Weaker notions of convergence to δ_{x_0} paired

with appropriate regularity assumptions on $(x, y) \to q_x(y)$ can also result in point-wise convergence to q_{x_0}; the Portmanteau Theorem can be used, for example.

A.3 Proofs

It is known that relative entropy can be expressed in terms of a non-negative integrand. This fact enables us to use Tonelli's theorem to justify interchanges in the order of integration.[9]

Lemma A.3.1. *Let $\{q_x : x \in \mathcal{X}\}$ and $\{r_x : x \in \mathcal{X}\}$ be families of probability densities with respect to a σ-finite measure γ, and suppose that both $(x, y) \mapsto q_x(y)$ and $(x, y) \mapsto r_x(y)$ are product measurable. For any $\mathcal{X}$-valued random element X,*

$$\mathbb{E}\gamma\, q_X \log \frac{q_X}{r_X} = \gamma\, \mathbb{E} q_X \log \frac{q_X}{r_X}.$$

Proof. We use the fact that $\log z \leq z - 1$, then invoke Tonelli's theorem.

$$\begin{aligned}
\mathbb{E}\gamma\, q_X \log \frac{q_X}{r_X} &= \mathbb{E}\gamma\, q_X \left[\frac{r_X}{q_X} - 1 - \log \frac{r_X}{q_X}\right] \\
&= \gamma\, \mathbb{E} q_X \left[\frac{r_X}{q_X} - 1 - \log \frac{r_X}{q_X}\right] \\
&= \gamma\, \mathbb{E} r_X - \gamma\, \mathbb{E} q_X - \gamma\, \mathbb{E} q_X \log \frac{r_X}{q_X} \\
&= \mathbb{E} \underbrace{\gamma\, r_X}_{1} - \mathbb{E} \underbrace{\gamma\, q_X}_{1} + \gamma\, \mathbb{E} q_X \log \frac{q_X}{r_X}
\end{aligned}$$

□

9. The proofs in this section use a combination of integral notations: the expectation symbol ($\mathbb{E}$) for probability measures and de Finetti notation for more general integrals.

Proof of ***Theorem A.0.1****.* Lemma A.3.1 justifies changing the order of integration.

$$\begin{aligned}
\mathbb{E}D(Q_X\|R) &= \mathbb{E}\gamma\, q_X \log \frac{q_X}{r} \\
&= \mathbb{E}\gamma\, q_X \log \frac{\bar{q}}{r} + \mathbb{E}\gamma\, q_X \log \frac{q_X}{\bar{q}} \\
&= \gamma \underbrace{\mathbb{E}q_X}_{\bar{q}} \log \frac{\bar{q}}{r} + \mathbb{E}D(Q_X\|\bar{Q})
\end{aligned}$$

□

Lemma A.3.2. *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure γ, and suppose that $(x, y) \mapsto q_x(y)$ is product measurable. Let $X \sim P$ be an $\mathcal{X}$-valued random element. If $\gamma\, e^{\mathbb{E}\log q_X} > 0$, then for any probability measure R defined on the same measurable space as γ,*

$$\mathbb{E}D(R\|Q_X) = D(R\|\widetilde{Q}_P) + \log \frac{1}{\gamma\, e^{\mathbb{E}\log q_X}}.$$

Proof. Making use of the central trick from the explanations of the mean field approximation algorithm (e.g. Ormerod and Wand [2010]), we have

$$\begin{aligned}
\mathbb{E}D(R\|Q_X) &= \mathbb{E}R \log \frac{r}{q_X} \\
&= R\mathbb{E} \log \frac{r}{q_X} \\
&= R[\log r - \mathbb{E}\log q_X] \\
&= R[\log r - \log e^{\mathbb{E}\log q_X}] \\
&= R \log \frac{r}{e^{\mathbb{E}\log q_X}} \\
&= R \log \frac{r}{e^{\mathbb{E}\log q_X}/\gamma\, e^{\mathbb{E}\log q_X}} + \log \frac{1}{\gamma\, e^{\mathbb{E}\log q_X}}.
\end{aligned}$$

Again, Lemma A.3.1 justifies the order interchange. □

Lemma A.3.3. *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure γ, and suppose that $(x, y) \mapsto q_x(y)$ is product measurable. Let $X \sim P$ be*

an $\mathcal{X}$-valued random element. If $\gamma\, e^{\mathbb{E}\log q_X} > 0$, then

$$\mathbb{E}D(\widetilde{Q}_P\|Q_X) = \log\frac{1}{\gamma\, e^{\mathbb{E}\log q_X}}.$$

Proof. Use $\widetilde{Q}_P$ as R in Lemma A.3.2. □

*Proof of **Theorem A.0.2**.* Combine Lemmas A.3.2 and A.3.3. □

Lemma A.3.4. *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure γ, and suppose that $(x, y) \mapsto q_x(y)$ is product measurable. Let X be an $\mathcal{X}$-valued random element. If $\gamma\, e^{\mathbb{E}\log q_X} = 0$, then for any probability measure R, $\mathbb{E}D(R\|Q_X) = \infty$.*

Proof. The integrand of $\gamma\, e^{\mathbb{E}\log q_X}$ is non-negative, so the integral being zero implies that $\mathbb{E}\log q_X = -\infty$ γ-almost everywhere. Since γ dominates R, the condition also holds R-almost everywhere.

By Lemma A.3.1,

$$\begin{aligned}\mathbb{E}D(R\|Q_X) &= R\,\mathbb{E}\log\frac{r}{q_X}\\ &= R\,[\log r - \mathbb{E}\log q_X].\end{aligned}$$

Our previous observation tells us that $\gamma\, e^{\mathbb{E}\log q_X} = 0$ implies that the integrand $\log r - \mathbb{E}\log q_X$ equals ∞ with R-probability 1, so $\mathbb{E}D(R\|Q_X) = \infty$.

□

*Proof of **Theorem A.1.1**.* With P representing the distribution of X, the mixture's centroid is $[1-\lambda]\bar{Q}_P + \lambda\check{Q}$.

For any $\lambda_1 \in [\lambda, 1]$, a draw from $[1-\lambda_1]Q_x + \lambda_1\check{Q}$ can be achieved by "processing" a draw from $[1-\lambda]Q_x + \lambda\check{Q}$. One simply needs to switch it to a new draw from $\check{Q}$ with probability $\frac{\lambda_1-\lambda}{1-\lambda}$. The data processing inequality tells us that two processed distributions are no further in relative entropy than the unprocessed distributions were.

The same processing that transforms $[1-\lambda_1]Q_x+\lambda_1\check{Q}$ to $[1-\lambda]Q_x+\lambda\check{Q}$ also transforms

the centroids appropriately. Thus by the data processing inequality,

$$D([1-\lambda_1]Q_x + \lambda_1\check{Q} \,\|\, [1-\lambda_1]\bar{Q}_P + \lambda_1\check{Q}) \leq D([1-\lambda]Q_x + \lambda\check{Q} \,\|\, [1-\lambda]\bar{Q}_P + \lambda\check{Q}).$$

Since this holds for every $x \in \mathcal{X}$, it holds for any expectation over $\mathcal{X}$. □

Appendix B

Hölder's identity

Hölder's inequality is most commonly written

$$\int |f(y)g(y)|d\gamma(y) \leq \|f\|_p \|g\|_q \tag{B.1}$$

for conjugate exponents p and q. An alternative way of expressing this is to say that for any pair of non-negative functions f and g and any $\alpha \in [0,1]$,

$$\int f^{\alpha}(y)g^{1-\alpha}(y)d\gamma(y) \leq \left(\int f(y)d\gamma(y)\right)^{\alpha} \left(\int g(y)d\gamma(y)\right)^{1-\alpha}. \tag{B.2}$$

In other words, *the integral of the point-wise geometric average of two functions is bounded by the geometric average of their integrals.* In fact, this relationship holds for *arbitrary* geometric expectations over a random element indexing functions.[1]

Theorem B.0.1 (Hölder's inequality)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be product measurable. For any measure γ on $\mathcal{Y}$ and any $\mathcal{X}$-valued random element X such that $\mathbb{E} \log \int f(X,y)d\gamma(y) > -\infty$,*

$$\int e^{\mathbb{E}\log f(X,y)}d\gamma(y) \leq e^{\mathbb{E}\log \int f(X,y)d\gamma(y)}.$$

Inequalities (B.1) and (B.2) represent the two-point distribution version of Theorem B.0.1.

1. This chapter's proofs are at the end.

The generalization for an arbitrary finite measure on $\mathcal{X}$ is easy to derive by normalizing and then applying the result for probability measures.

Corollary B.0.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be product measurable. For any measure γ on $\mathcal{Y}$ and finite measure μ on $\mathcal{X}$,*

$$\int e^{\int \log f(x,y) d\mu(x)} d\gamma(y) \leq e^{\frac{1}{\mu(\mathcal{X})} \int [\log \int f(x,y)^{\mu(\mathcal{X})} d\gamma(y)] d\mu(x)}.$$

Using e^f as the function in Theorem B.0.1, and taking the log of both sides gives us an equivalent inequality that is also worth stating.

Corollary B.0.3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be product measurable. For any measure γ on $\mathcal{Y}$ and any $\mathcal{X}$-valued random element X,*

$$\log \int e^{\mathbb{E} f(X,y)} d\gamma(y) \leq \mathbb{E} \log \int e^{f(X,y)} d\gamma(y).$$

The fact that Hölder's inequality holds in this generality is perhaps not widely known. For example, Karakostas [2008] proved an extension of Hölder's inequality to *countable* products assuming γ is σ-finite; that result was improved by [Chen et al., 2016, Thm 2.11]. The inequalities they present are readily subsumed by Corollary B.0.2 by letting μ concentrate on a countable set.

[Haussler and Opper, 1997, Lemma 1] states our Corollary B.0.3, but the justification presented there is not quite adequate. They observe, using the two-point distribution version of Hölder's inequality, that the mapping $f \mapsto \log \gamma\, e^f$ is convex on the space of real-valued functions on a set. [Pettis] expectations commute with continuous affine functionals, and Jensen's inequality relies on the expectation commuting with a continuous affine functional tangent to the convex function. The existence of a tangent continuous affine functional is guaranteed for convex functions on finite-dimensional spaces, but not on infinite-dimensional spaces. As a simple example, consider any discontinuous linear functional; it is convex, but it has no continuous affine functional tangent to it. For a more concrete example, see [Perlman, 1974, Introduction].

Haussler and Opper [1997] reference Symanzik [1965] where the inequality in our Theo-

rem B.0.1 is stated and called *generalized Hölder's inequality*; he points to the classic text [Dunford and Schwartz, 1958, VI.11 Ex 36] where it is left as an exercise. Although that exercise does not say to assume σ-finiteness, the proof they hint at does require it. For σ-finite measures, at least, the proof can follow a different route from the one they hint at. We establish an identity that has an information-theoretic interpretation involving the non-negative "variance" functional $\widetilde{\mathbb{V}}$ for random probability measures defined and interpreted in Chapter A.

Theorem B.0.4. *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be product measurable. Let γ be a σ-finite measure on $\mathcal{Y}$, and let $X \sim P$ be an $\mathcal{X}$-valued random element. If $\int e^{f(x,y)} d\gamma(y)$ is in $(0, \infty)$ P-almost surely and $\mathbb{E} \log \int e^{f(X,y)} d\gamma(y) > -\infty$, then*

$$\mathbb{E} \log \int e^{f(X,y)} d\gamma(y) - \log \int e^{\mathbb{E} f(X,y)} d\gamma(y) = \widetilde{\mathbb{V}} Q_X$$

where Q_x has density $q_x(y) := \frac{e^{f(x,y)}}{\int e^{f(x,y)} d\gamma(y)}$ with respect to γ.

Corollary B.0.5 (Hölder's identity)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be product measurable. Let γ be a σ-finite measure on $\mathcal{Y}$, and let $X \sim P$ be an $\mathcal{X}$-valued random element. If $\int f(x,y) d\gamma(y)$ is in $(0, \infty)$ P-almost surely and $\mathbb{E} \log \int f(X,y) d\gamma(y) > -\infty$, then*

$$\frac{e^{\mathbb{E} \log \int f(X,y) d\gamma(y)}}{\int e^{\mathbb{E} \log f(X,y)} d\gamma(y)} = e^{\widetilde{\mathbb{V}} Q_X}$$

where Q_x has density $q_x(y) := \frac{f(x,y)}{\int f(x,y) d\gamma(y)}$ with respect to γ.

In the special case that X only takes two possible values, $\widetilde{\mathbb{V}} Q_X$ is an *unnormalized Renyi divergence* D_λ between the two possible distributions, as described in Section A.

Theorem B.0.6. *Let $\mathcal{Y}$ be a measurable space, and let $f : \mathcal{Y} \to \mathbb{R}^+$ and $g : \mathcal{Y} \to \mathbb{R}^+$ have finite positive γ-integrals. Then*

$$\frac{[\int f(y) d\gamma(y)]^\lambda [\int g(y) d\gamma(y)]^{1-\lambda}}{\int f^\lambda(y) g^{1-\lambda}(y) d\gamma(y)} = e^{D_\lambda(Q \| R)}$$

where Q has density $\frac{f(y)}{\int f(y) d\gamma(y)}$ and R has density $\frac{g(y)}{\int g(y) \gamma(y)}$ with respect to γ.

B.1 Proofs

Proof of ***Theorem B.0.4****.* We will write $f(X, \cdot)$ as f_X. The key is Lemma A.3.3.

$$\begin{aligned}\log \gamma\, e^{\mathbb{E} f_X} &= \log \gamma\, e^{\mathbb{E} \log[e^{f_X}/\gamma\, \exp f_X]} + \mathbb{E} \log \gamma\, e^{f_X} \\ &= -\mathbb{E} D(\widetilde{Q}_P \| Q_X) + \mathbb{E} \log \gamma\, e^{f_X}\end{aligned}$$

if the geometric mixture is well-defined.[2]

Next, assume that the geometric mixture is not well-defined; in other words, $\gamma\, e^{\mathbb{E} \log(e^{f_X}/\gamma\, e^{f_X})} = 0$. Because the integrand is non-negative, the integral can only be zero if the integrand is zero γ-almost everywhere. This requires the exponent, which simplifies to $\mathbb{E}[f_X - \log \gamma\, e^{f_X}]$, to be $-\infty$ almost everywhere. Assume that there exists a non-negligible set for which $\mathbb{E} f_X > -\infty$. Then on that set, $\mathbb{E}[f_X - \log \gamma\, e^{f_X}]$ can only be $-\infty$ if $\mathbb{E} \log \gamma\, e^{f_X}$ is ∞. Furthermore, the contribution of that non-negligible set ensures that $\log \gamma\, e^{\mathbb{E} f_X}$ is also strictly greater than $-\infty$, which tells us that the two sides of the proposed identity are both ∞.

In the one remaining case, the geometric mixture does not exist and $\mathbb{E} f_X = -\infty$ almost everywhere. These imply that $\mathbb{V} Q_X = \infty$ and $\log \gamma\, e^{\mathbb{E} f_X} = -\infty$, respectively. The theorem specifies that $\mathbb{E} \log \gamma\, e^{f_X} > -\infty$, so again the identity reduces to $\infty = \infty$.

An interesting observation is implicit in the above proof: $\mathbb{E} \log \gamma\, e^{f_X} = -\infty$ is only possible if $\mathbb{E} f_X = -\infty$ almost everywhere.

A closely related derivation in [Barron, 1988, Sec 4] was instructive; the accompanying discussion in that paper provides another interpretation of the quantities involved in Hölder's identity. □

Proof of ***Theorem B.0.1****.* When its conditions are met, Hölder's identity (Theorem B.0.5) implies the desired inequality result by non-negativity of $\widetilde{\mathbb{V}}$. □

Proof of ***Theorem B.0.6****.* Define the product measurable function $h_x(y)$ with x taking values in $\mathcal{X} = \{1, 2\}$ with $h_1(y) = f(y)$ and $h_2(y) = g(y)$. By the definition of unnormalized Renyi divergence, $\widetilde{\mathbb{V}}$ of the random distribution is equal to $D_\lambda(Q\|R)$ according to

2. The proofs in this section use a combination of integral notations: the expectation symbol ($\mathbb{E}$) for probability measures and de Finetti notation for more general integrals.

Lemma A.3.3. Therefore, the desired result is a direct consequence of Holder's identity, at least when γ is σ-finite. However, we deliberately omitted the σ-finiteness requirement. In fact, the reason we required σ-finiteness in previous Lemmas and Theorems was to justify interchanges in the order of integration. When one of the integrals concentrates on a finite set of atoms, then interchange is always valid by linearity of integration. Indeed, when $\mathcal{X}$ is finite, the Lemmas and Theorems of this paper are valid without the condition that γ is σ-finite. Alternatively, the sum of any finite collection of probability measures is itself a finite dominating measure for each of their densities. $\square$

Appendix C

Hypothetical measures

A central aspect of measure theory is the extension of non-negative countably additive set functions (known as *premeasures*) to larger domains. A σ-finite premeasure γ defined on a semi-algebra has a *unique* extension to a measure on the σ-algebra generated by its domain, by the Carathéodory construction [Bogachev, 2007, Prop 1.3.10].[1] An interpretation of this in terms of real-world modeling is that the premeasure is a state of knowledge of a substance's mass on certain sets; the substance's mass on some other sets can be inferred by the nature of *mass*, that is, non-negativity and countable additivity. What about a set A that is not in the completion of the σ-algebra generated by the original domain? We may not be able to infer a mass that it *must* have, but we can still exclude some values. Any value strictly less than its γ-induced inner measure (supremum of masses of its subsets) should be considered unreasonable, as should any value strictly larger than its outer measure (infimum of masses of its supersets). In fact, if the outer measure of A is finite, then given any value z between the inner and outer measure of A there exists an extension of γ to a measure on the σ-algebra generated by the original domain and A that assigns a measure of z to A [Bogachev, 2007, Thm 1.12.14]. It is sensible to conclude that any value in that range might be the "true" mass of A. This reasoning seems preferable to an insistence that conditions must be imposed to avoid the possibility of "unmeasurable" sets.

1. A measure has a unique extension to its completion, also via Carathéodory construction. And integration is a unique extension of the domain from indicator functions to measurable functions [Pollard, 2002, Sec 2.3].

This line of thinking can be implemented in a simpler and more powerful way than one might expect. Let γ be a premeasure on Ω with domain Σ, and let f be a function on Ω. Suppose $\mathcal{A} \subseteq 2^\Omega$ is at least fine enough that f is $\sigma(\Sigma \cup \mathcal{A})$-measurable. We define the *hypothetical measure* (we will say *hypomeasure*, for short) $\gamma^{\mathcal{A}} f$ to be an indexed family where the indices are extensions of γ to $\sigma(\Sigma \cup \mathcal{A})$ and each such extension indexes the integral of f according to that measure.[2] One might prefer to omit the superscript by letting $\mathcal{A}$ be σ-algebra generated by f by default, or more generally letting $\mathcal{A}$ be the union of the σ-algebras generated by the functions that are to be integrated in the statement at hand. It is appropriate that this results in the hypomeasure notation being indistinguishable from the ordinary integral, because the hypomeasure extends the concept of integral: when f is Σ-measurable, the hypomeasure $\gamma^{\sigma(f)} f$ is constant (equal to its ordinary integral's value) and can be treated as such.

The facts and operations that are valid for ordinary measures continue to hold pointwise for hypomeasures. By proceeding *as if* every function were measurable, one produces equations and inequalities that hold *point-wise*, where the points are the extensions of γ.

To see why this approach is more powerful than just calculating inner and outer measures, consider the following simple example. Let f_1 be the indicator function of an unmeasurable set A, and let f_2 be 2 times the indicator function of A. Suppose A has inner measure 0 and outer measure 1. Then one cannot compare the integrals of f_1 and f_2 by comparing by their inner/outer measure ranges, which are $[0, 1]$ and $[0, 2]$ respectively. However, the hypomeasures approach allows us to unhesitatingly assert that the "integral" of f_1 is no greater than the "integral" of f_2, regardless of what the masses of currently unknown sets turn out to be.

The hypomeasures approach greatly simplifies our work by allowing us to treat every function as measurable.[3] *Measurability only becomes relevant to follow-up questions* regarding the range of a hypomeasure. If the σ-field generated by f is a subset of the γ-completion of the domain of γ, then the γ-hypomeasure of f is constant, being everywhere equal to

2. Our notation is intended to resemble (and extend) the de Finetti notation for integrals.

3. The hypomeasure idea is fairly straight-forward and has likely been discussed before somewhere.

the ordinary measure of f according to the completion of γ. Measurability of f by the completion of γ is necessary and sufficient for this constancy when γ is a finite measure [Halmos, 1974, Thm 14.F].

Another follow-up question is perhaps concerning: do there exist *any* extensions of γ that can measure f? If so, we will call f *compatible* with γ. Incompatibility is possible; indeed, assuming Zorn's Lemma, no atomless measure can exist on a power set [Troitskii, 1994, Theorem 5]. Even a countable collection of sets has been devised that is incompatible with Lebesgue measure, assuming the continuum hypothesis [Bogachev, 2007, Cor 3.10.3].[4] However, such pathological functions are unusual in practice, so we suggest that a "presumption of innocence" is sensible. Furthermore, realize that there is nothing mathematically illegitimate about incompatible cases; they produce identities and inequalities that are vacuously true "point-wise" as there are not any points to check.[5]

When convenient, one can instruct the reader to *interpret "integrals" as hypomeasures*. In this way, identities and inequalities proven are mathematically legitimate regardless of measurability, and they are also realistically meaningful except in the pathological cases of incompatibility.

4. To clarify, the Vitali sets *are not* the example that we are referring to. It is easy to extend Lebesgue measure to include the Vitali sets, as they are disjoint [Bogachev, 2007, Thm 1.12.5]. The significance of the Vitali sets was that they demonstrated that there is no *translation-invariant* extension.

5. Careful not to be misled, though. Consider [Mattner, 1999, Sec 2.2] in which a non-negative integrand produces different results depending on the order of integration. Recall that Tonelli's Theorem requires product-measurability, which fails in Mattner's example. We can conclude that there is no extension of the measure for which the integrand is measurable; otherwise, Tonelli would apply and the iterated integrals would be valid. Thus, this is a case in which the hypomeasure has an empty domain.

Bibliography

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

Animashree Anandkumar, Daniel Hsu, and Sham M. Kakade. A Method of Moments for Mixture Models and Hidden Markov Models. In *Conference on Learning Theory*.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, 2017.

Andrew R. Barron. The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions. Report 7, University of Illinois, 1988.

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 and Nicolas Hengartner. Information Theory and Superefficiency. *The Annals of Statistics*, 26(5):1800–1825, 1998.

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.

Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

Vladimir I. Bogachev. *Measure Theory*, volume 1. Springer-Verlag, 2007.

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.

Joseph T. Chang. Full Reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency. *Mathematical Biosciences*, 137(1):51–73, 1996.

Sabyasachi Chatterjee. *Adaptation in Estimation and Annealing*. Thesis, 2014.

Sabyasachi Chatterjee and Andrew R. Barron. Information Theoretic validity of Penalized Likelihood. In *IEEE International Symposium on Information Theory*, pages 3027–3031. IEEE.

Wei Chen, Longbin Jia, and Yong Jiao. Hölder's inequalities involving the infinite product and their applications in martingale spaces. *Analysis Mathematica*, 42(2):121–141, 2016.

Imre Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4):191–213, 1972.

Imre Csiszár and František Matúš. Information Projections Revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.

Nelson Dunford and Jacob T. Schwartz. *Linear Operators, Part 1: General Theory*. Interscience Publishers, New York, 1958.

Zoubin Ghahramani. Factorial learning and the EM algorithm. In *Advances in Neural Information Processing Systems*, pages 617–624.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Paul R. Halmos. *Measure Theory*. Springer, 1974.

David Haussler and Manfred Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, pages 2451–2492, 1997.

Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science*, pages 11–20. ACM.

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.

G. L. Karakostas. An extension of Hölder's inequality and some results on infinite products. *Indian Journal of Mathematics*, 50:303–307, 2008.

Jason M Klusowski and W. D. Brinda. Statistical guarantees for estimating the centers of a two-component Gaussian mixture by EM. 2018.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer-Verlag, New York, 2006.

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.

Xi Luo. *Penalized Likelihoods: Fast Algorithms and Risk Bounds*. Thesis, 2009.

Lutz Mattner. Product measurability, parameter integrals, and a Fubini counterexample. 1999.

J. T. Ormerod and M. P. Wand. Explaining Variational Approximations. *The American Statistician*, 64(2):140–153, 2010.

Michael D. Perlman. Jensen's Inequality for a Convex Vector-Valued Function on an Infinite-Dimensional Space. *Journal of Multivariate Analysis*, 4(1):52–65, 1974.

David Pfau. A Generalized Bias-Variance Decomposition for Bregman Divergences. 2013.

David Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

Jorma Rissanen. Modeling by Shortest Data Description. *Automatica*, 14(5):465–471, 1978.

Jorma Rissanen. Stochastic Complexity and Modeling. *The Annals of Statistics*, 14(3): 1080–1100, 1986.

Omar Rivasplata. Subgaussian random variables: An expository note. 2012.

Alfréd Rényi. On Measures of Entropy and Information. In Jerzy Neyman, editor, *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1. University of California Press, 1961.

Kurt Symanzik. Proof and Refinements of an Inequality of Feynman. *Journal of Mathematical Physics*, 6(7):1155–1156, 1965.

Matus Telgarsky and Sanjoy Dasgupta. Agglomerative Bregman Clustering. In *International Coference on International Conference on Machine Learning*, pages 1011–1018. Omnipress.

Flemming Topsøe. Basic Concepts, Identities and Inequalities - the Toolkit of Information Theory. *Entropy*, 3(3):162–190, 2001.

V. G. Troitskii. Real Partitions of Measure Spaces. *Siberian Mathematical Journal*, 35(1): 189–191, 1994.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996.

Tim van Erven and Peter Harremoës. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Raymond Veldhuis. The Centroid of the Symmetrical Kullback-Leibler Distance. *IEEE Signal Processing Letters*, 9(3):96–99, 2002.

Tong Zhang. From epsilon-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.