Finite-Sample Risk Bounds for Maximum Likelihood Estimation With Arbitrary Penalties

W. D. Brinda and Jason M. Klusowski, Student Member, IEEE

Abstract—The minimum description length two-part coding index of resolvability provides a finite-sample upper bound on the statistical risk of penalized likelihood estimators over countable models. However, the bound does not apply to unpenalized maximum likelihood estimation or procedures with exceedingly small penalties. In this paper, we point out a more general inequality that holds for arbitrary penalties. 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 conclude by discussing implications for adaptive estimation.

Index Terms—Penalized likelihood estimation, minimum description length, codelength, statistical risk, redundancy.

I. INTRODUCTION

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 codelengths to a model (or model class) then selecting the distribution that provides the most efficient description of one’s data [1]. 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 codelengths 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 (θ, L) denote a discrete set indexing distributions along with a complexity function. With X ∼ P, the (pointwise) redundancy of any θ ∈ Θ is its two-part codelength minus log(1/p(x)), the codelength one gets by using P as the coding distribution.1 The expectation of redundancy is the relative entropy from P to Pθ plus L(θ). Let θ* ∈ Θ denote the minimizer of expected redundancy; it is the average-case optimal representative from (Θ, L) when the true distribution is P. Its expected redundancy will be denoted

\[ R_{Θ, L}(P) := \inf_{θ ∈ Θ} \{ D(P∥Pθ) + L(θ) \}, \]

or in the context of iid data Xn ∼ Pn and iid modeling {Pnθ : θ ∈ Θ}, its expected redundancy rate is denoted

\[ R^{(n)}_{Θ, L}(P) := \inf_{θ ∈ Θ} \left\{ D(P∥Pθ) + \frac{L(θ)}{n} \right\}. \]

Interestingly, [2] showed that if the complexity function is large enough, then the corresponding penalized likelihood estimator outperforms the best-case average representative. Specifically, the statistical risk is bounded by \( R_{Θ, L}(P) \); that result is stated for iid sampling in (2) below.2

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.3 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-Rényi divergence [4] with α ∈ (0, 1). For convenience, we frequently refer to the 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.

3Although the forthcoming resolvability bounds (i.e., as in (2) with \( L \) that is at least twice a codelength function) are valid under misspecification, they do not in general imply consistency in the sense that the corresponding penalized estimator eventually converges to the element \( θ^* \) of \( Θ \) that minimizes KL or Hellinger to the truth P. Indeed, there are various examples [3] in which the twice-codelength penalized estimator is inconsistent (i.e., provably never converges to \( θ^* \)).

1For 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 1. Specialization to iid data will be discussed thereafter.
we specialize our discussion and our present work to Bhattacharyya divergence [5] 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

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

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

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

Bhattacharyya divergence is bounded below by squared Hellinger distance (using $\log 1/x \geq 1 - x$) 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 [6].

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 [7]; 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, [8] worked out the continuous extension for Gaussian graphical models (building on [9]) 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 [10], 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 [11], whereas under regularity conditions (for which Bhattacharyya divergence is locally equivalent to one-half relative entropy, according to [7]) the optimal Bhattacharyya risk is of order $1/n$ [12]. Our variant on the resolvability method leads to the possibility of deriving exact bounds of order $1/n$.

Our bounds can be used for the penalized MLE over a discretization of an unbounded parameter space under a power decay condition on the Hellinger affinity, as in Theorems 11 and 12. We show that such a condition is satisfied by exponential families of distributions with a boundedness assumption on the largest eigenvalue of the covariance matrix of their sufficient statistics (see Lemma 9). For these models and others, we establish order $1/n$ bounds for the Bhattacharyya risk. The primary focus of this paper is to develop new tools towards this end.

One highly relevant line of work is [13], where 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 $R_{\Theta, L}(P)$ 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 [13, Th. 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 $L$; 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 II, we prove our more general version of the resolvability bound inequality using a derivation closely analogous to the one by [14]. We then explore corollaries that arise from various choices of pseudo-penalty. In Section III, we explain how our approach applies in the context of adaptive modeling. Additional work can be found in [15], including some simple concrete examples [15, Sec. 2.1.2], extension to continuous models [15, Sec. 2.2], and an application to Gaussian mixtures [15, Ch. 4].

Every result labeled a Theorem or Lemma has a formal proof, some of which are in the Appendix. 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 $CX$ means the covariance matrix, while $\nabla X$ represents its trace $\mathbb{E}\|X - \mathbb{E}X\|^2$. The notation $\lambda_j(\cdot)$ means the $j$th 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 $\Theta$. Multiplication and division take precedence over $\wedge$ and $\lor$; for instance, $ab \land c$ means $(ab) \land c$.

II. 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. This existence requirement will be implicit throughout our paper. Theorem 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 1:** Let \( X \sim P \), and let \( \hat{\theta} \) be the penalized MLE over \( \Theta \) indexing a countable model with penalty \( L \). Then for any \( L: \Theta \rightarrow \mathbb{R} \),

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

**Proof:** We follow the pattern of Jonathan Li’s version of the resolvability bound proof [14].

\[
D_B(P, P_{\hat{\theta}}) := 2 \log \frac{1}{A(P, P_{\hat{\theta}})} = 2 \log \frac{\sqrt{p_0(X)/p(X)} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} + \log \frac{p(X)}{p_0(X)} + L(\hat{\theta}) + L(\hat{\theta}) 
\]

\[
\leq 2 \log \sum_{\theta \in \Theta} \frac{\sqrt{p_0(X)/p(X)} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} + \log \frac{p(X)}{p_0(X)} + L(\hat{\theta}) + L(\hat{\theta}).
\]

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.

\[
2 \mathbb{E} \log \sum_{\theta \in \Theta} \frac{\sqrt{p_0(X)/p(X)} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} 
\]

\[
\leq 2 \log \sum_{\theta \in \Theta} \frac{\sqrt{p_0(X)/p(X)} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]}}{A(P, P_{\hat{\theta}})} - \frac{\sqrt{p_0(X)/p(X)} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]}}{A(P, P)} + \log \frac{p(X)}{p_0(X)} + L(\hat{\theta}) + L(\hat{\theta}).
\]

Returning to the overall inequality, we have

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

\[
+ \mathbb{E} \left[ \log \frac{p(X)}{p_0(X)} + L(\hat{\theta}) \right] + \mathbb{E} L(\hat{\theta}) 
\]

\[
= 2 \log \sum_{\theta \in \Theta} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]} 
\]

\[
+ \mathbb{E} \min_{\theta \in \Theta} \left[ \log \frac{p(X)}{p_0(X)} + L(\theta) \right] + \mathbb{E} L(\hat{\theta}) 
\]

\[
\leq 2 \log \sum_{\theta \in \Theta} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]} 
\]

\[
+ \mathbb{E} \left( \log \frac{p(X)}{p_0(X)} + L(\hat{\theta}) \right) + \mathbb{E} L(\hat{\theta}) 
\]

\[
\leq 2 \log \sum_{\theta \in \Theta} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]} + \mathbb{E} L(\hat{\theta}).
\]

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 \( \{P^n_{\theta}: \theta \in \Theta\} \). Because \( D_B(P^n, P_{\hat{\theta}}) = n D_B(P, P_{\hat{\theta}}) \) and \( D(P^n \| P^n_{\theta}) = n D(P \| P_{\theta}) \), we can divide both sides of Theorem 1 by \( n \) to reveal the role of sample size in this context:

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

We will see three major advantages to Theorem 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 \( L \) 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 1 enables 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 1 gives us a corollary that moves \( \mathcal{L} \) out of the summation.

**Corollary 2:** Let \( X \sim P \), and let \( \hat{\theta} \) be the penalized MLE over \( \Theta \) indexing a countable model with penalty \( L \). Then for any \( L: \Theta \rightarrow \mathbb{R} \),

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

The iid data and model version is

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

We will use the term pseudo-penalty for the function labeled \( L \) in either Theorem 1 or Corollary 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 3:** Let \( X \sim P \), and let \( \hat{\theta} \) be the penalized MLE over \( \Theta \) indexing a countable model with penalty \( L \). Then for any \( L: \Theta \rightarrow \mathbb{R} \),

\[
P \left\{ D_B(P, P_{\hat{\theta}}) - \left[ \log \frac{p(X)}{p_0(X)} + L(\hat{\theta}) + L(\hat{\theta}) \right] \geq t \right\} \leq e^{-t/2} \sum_{\theta \in \Theta} e^{-\frac{1}{2} [L(\theta) + L(\hat{\theta})]}.
\]

\cite{4}We will say “the” penalized MLE, even though we do not require uniqueness; any scheme can be used for breaking ties.
which we review in this section.

variable by the sum of its possible values.

Given any variable.

penalty is stochastically less than an exponential random

form

Several of our corollaries have \( \mathcal{L} \) and \( L \) designed to make \( \sum_{\theta \in \Theta} e^{-\frac{1}{2}L(\theta)+L(\theta)} \leq 1 \). 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 \( \epsilon \)-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 \( \epsilon \)-discretization of \( \Theta \subseteq \mathbb{R}^d \) is a set of the form \( \Theta \cap (v + \epsilon \mathbb{Z}^d) \). See Section III-D for a discussion of the behavior of \( \mathcal{R}^n_{\Theta, \mathcal{L}}(P) \) in that context.

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 document to highlight the fact that many of the exact risk bounds we derive are of order \( 1/n \) in that case.

A. Zero as Pseudo-Penalty

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

Corollary 4: Assume \( X^n \overset{iid}{\sim} P \), and let \( \hat{\theta} \) be the penalized MLE over \( \Theta \) indexing a countable iid model with penalty \( \mathcal{L} \). Then

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

The usual statement of the resolvability bound [7] 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,

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

implies

\[
\mathbb{E}D_B(P, P_{\hat{\theta}}) \leq \mathcal{R}^n_{\Theta, \mathcal{L}}(P). \tag{2}
\]

The quantity on the right-hand side of (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 \( \hat{\mathcal{L}} := \mathcal{L} + 2\log z \). Thus one has a resolvability bound of the form (2) with the equivalent penalty \( \hat{\mathcal{L}} \), which satisfies (1) with equality.

Additionally, the resolvability bounds give an exact upper bound on the minimax risk for any model \( \Theta \) that can be covered by finitely many relative entropy balls of radius \( \epsilon^2 \); 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 \( \Theta_\epsilon \). With data \( X^n \overset{iid}{\sim} P_{\theta^*} \) for any \( \theta^* \in \Theta \), the MLE restricted to \( \Theta_\epsilon \) has the resolvability risk bound

\[
\mathbb{E}D_B(P_{\theta^*}, P_{\hat{\theta}}) \leq \inf_{\theta \in \Theta_\epsilon} \left( D(P_{\theta^*} || P_{\hat{\theta}}) + \frac{2\mathcal{M}(\epsilon)}{n} \right)
\]

\[
= \inf_{\theta \in \Theta_\epsilon} D(P_{\theta^*} || P_{\hat{\theta}}) + \frac{2\mathcal{M}(\epsilon)}{n}
\]

\[
\leq \epsilon^2 + \frac{2\mathcal{M}(\epsilon)}{n}.
\]

If an explicit bound for \( \mathcal{M}(\epsilon) \) is known, then the overall risk bound can be optimized over the radius \( \epsilon \) — see for instance [7, Sec. 1.5].

Because this approach to upper bounding minimax risk requires twice-Kraft-valid codelengths, it only applies to models that can be covered by finitely many relative entropy balls. However, Corollary 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. For a minima result, one still needs this quantity to be uniformly bounded over all data-generating distribution \( \theta^* \in \Theta \). See Corollary 10 below as an example.

B. Bhattacharyya Divergence as Pseudo-Penalty

Important corollaries to Theorems 1 and 2 come from setting the pseudo-penalty equal to \( \alpha D_B(P, P_{\hat{\theta}}) \); 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: \( A(P^n, P_{\theta}) = A(P, P_{\theta})^n \).

The following corollaries serve as the starting point for the main bounds in Theorems 12 and 11, after which, more refined techniques are used in controlling the two terms in (3) and (4).
Corollary 5: Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over $\Theta$ indexing a countable iid model with penalty $L$. Then for any $\alpha \in [0, 1]$, 
\[
\mathbb{E}D_B(P, P_\hat{\theta}) \leq \frac{1}{1 - \alpha} \left[ R_{\hat{\theta}, L}^{(n)}(P) + \frac{2 \log \sum_{\theta \in \Theta} e^{-\frac{1}{2} L(\theta)} A(P, P_\theta)^{an}}{n} \right].
\] (3)

Corollary 6: Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over $\Theta$ indexing a countable iid model with penalty $L$. Then for any $\alpha \in [0, 1]$, 
\[
\mathbb{E}D_B(P, P_\hat{\theta}) \leq \frac{1}{1 - \alpha} \left[ R_{\hat{\theta}, L}^{(n)}(P) + \frac{2 \log \sum_{\theta \in \Theta} A(P, P_\theta)^{an} - \mathbb{E}L(\hat{\theta})}{n} \right].
\] (4)

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

Consider a penalized MLE selected from an $\epsilon$-discretization of a continuous parameter space; as the sample size increases, one typically wants to shrink $\epsilon$ to make the grid more refined (see Section III-D). Examining Corollaries 5 and 6, we see two opposing forces at work as $n$ increases: the grid-points themselves proliferate, while the $\alpha$th power depresses the terms in the summation. An easy case occurs when $A(P, P_\theta)$ is bounded by a Gaussian-shaped curve; we apply Corollary 6 and invoke Lemma 27.

Corollary 7: Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over an $\epsilon$-discretization $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$ indexing an iid model with penalty $L$. Assume $A(P, P_\theta) \leq e^{-c\|\theta - \theta^*\|^2}$ for some $c > 0$ and some $\theta^* \in \Theta$. Then 
\[
\mathbb{E}D_B(P, P_\hat{\theta}) \leq 2 \left[ R_{\hat{\theta}, L}^{(n)}(P) + \frac{2d \log (1 + \sqrt{\frac{\epsilon}{\epsilon n c}}) - \mathbb{E}L(\hat{\theta})}{n} \right].
\]

With $\epsilon$ proportional to $1/\sqrt{n}$, our bound on the summation of Hellinger affinities is stable. Corollary 8 sets $L = 0$ to demonstrate a more concrete instantiation of this result.

Corollary 8: Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the MLE over an $\epsilon$-discretization $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$ indexing an iid model using $\epsilon = \sqrt{2/n}$. Assume $A(P, P_\theta) \leq e^{-c\|\theta - \theta^*\|^2}$ for some $c > 0$ and some $\theta^* \in \Theta$. Then 
\[
\mathbb{E}D_B(P, P_\hat{\theta}) \leq 2 \frac{D(P)^{\epsilon}}{D(\Theta_\epsilon)} + \frac{4d \log (1 + \sqrt{2})}{n}. \tag{5}
\]

If $P$ is $P_{\theta^*}$ in an exponential family with natural parameter indexed by $\theta$, then Hellinger affinities do have a Gaussian-shaped bound as long as the minimum eigenvalue of the sufficient statistic’s covariance matrix is uniformly bounded below by a positive number. We use the notation $\hat{\lambda}_j(\cdot)$ for the $j$th largest eigenvalue of the matrix argument.

Lemma 9: Let $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ be an exponential family with natural parameter $\theta$ and sufficient statistic $\phi$. Then 
\[
A(P_{\theta^*}, P_\theta) \leq e^{-c\|\theta - \theta^*\|^2},
\]
where $c := \frac{1}{3} \inf_{\hat{\theta} \in \Theta} \hat{\lambda}_d(CX - P_{\hat{\theta}}\phi(X))$.

In Lemma 9, $c$ does not depend on $\theta^*$. In addition the $\epsilon$-discretization is also a KL-net, then the risk of the estimator described in Corollary 8 is uniformly bounded over data-generating distributions in $\Theta$. The minimax risk is no greater than the supremum risk of this particular estimator.

Corollary 10: Let $\Theta \subseteq \mathbb{R}^d$ index a set of distributions. Assume that for some $c > 0$, every $\theta^* \in \Theta$ has the property that $A(P_{\theta^*}, P_\theta) \leq e^{-c\|\theta - \theta^*\|^2}$. Assume further that there exists $\beta > 0$ such that for all $\epsilon > 0$, every $\epsilon$-discretization $\Theta_\epsilon \subseteq \Theta$ is also a KL-net with balls of radius $\beta \epsilon^2$. Then the minimax Bhattacharya risk of $\Theta$ has the upper bound 
\[
\min_{\theta^*} \mathbb{E}_{X^n \overset{iid}{\sim} P_{\theta^*}} D_B(P_{\theta^*}, P_\theta) \leq \frac{4\beta + d \log(1 + 4/\sqrt{\epsilon})}{n}. \tag{6}
\]

In general, however, Hellinger affinity being uniformly bounded by a Gaussian curve may be too severe of a requirement. A weaker condition is to require only a power decay for $\theta$ far from some $\theta^*$.

Theorem 11: Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over an $\epsilon$-discretization $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$ indexing an iid model with penalty $L$. Assume that for some $\theta^*, \hat{\theta} \in \Theta$, radius $R$ and $a, c > 0$, the Hellinger affinity $A(P, P_\theta)$ is bounded by $a/\|\theta - \theta^*\|^b$ outside the ball $B(\theta^*, R)$ and bounded by $e^{-c\|\theta - \theta^*\|^2}$ inside the ball. If $R \geq 11a^{1/b} \vee 3\epsilon$, and $n \geq 2(d + 1)/\beta$, then,
\[
\mathbb{E}D_B(P, P_\hat{\theta}) \leq 2 \left[ R_{\hat{\theta}, L}^{(n)}(P) + \frac{d(2 \log (1 + \sqrt{\frac{\epsilon}{\epsilon n c}}) + 2 \log (1 + \sqrt{\frac{2R}{\epsilon \sqrt{\epsilon n c}}} + 3 - \mathbb{E}L(\hat{\theta})}{n} \right]. \tag{7}
\]

Proof: The part of the summation where Hellinger affinity is bounded by a Gaussian curve has the same bound as in Corollary 7, which is a direct consequence of Lemma 27.

\[
\sum_{\theta \in \Theta \cap B(\theta^*, R)} A(P, P_\theta)^{an} \leq \sum_{\theta \in \Theta \cap B(\theta^*, R)} e^{-c\|\theta - \theta^*\|^2} \leq \sum_{\theta \in \Theta} e^{-c\|\theta - \theta^*\|^2} \leq (1 + \frac{2\sqrt{\pi}}{\epsilon \sqrt{nac}})^d. \tag{8}
\]

Notice that the “center” point for this Gaussian curve $\hat{\theta}^*$ can be different from the center of the ball $\theta^*$.

The summation of the remaining terms is handled by Lemma 31, assuming $n \geq (d + 1)/\beta a b$. 
\[
\sum_{\theta \in \Theta \cap B(\theta^*, R)} A(P, P_\theta)^{an} \leq \sum_{\theta \in \Theta \cap B(\theta^*, R)} \left( \frac{\alpha}{\epsilon \sqrt{nab \log(R/\alpha^{1/b})}} \right)^d \leq \left( \frac{4R}{\epsilon \sqrt{nab \log(R/\alpha^{1/b})}} \right)^d. \tag{9}
\]
The assumption that $R \geq 11 a^{1/b}$ assures us that $\log(R/4a^{1/b}) \geq 1$, simplifying the bound.

Each of (5) and (6) are at least 1, so by Lemma 20, the sum of their logs is bounded by the log of their sum plus $2 \log 2$. Finally, substitute $a = 1/2$.

The sample size requirement in Theorem 11 can be avoided by using a squared norm penalty. The bound we derive has superlinear order in the dimension.

**Theorem 12:** Assume $X^n \overset{iid}{\sim} P$, and let $\tilde{\theta}$ be the penalized MLE over an $\epsilon$-discretization $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$ indexing an iid model with penalty $L(\theta) = ||\theta||^2$. Assume that for some $\theta^*, \tilde{\theta}^* \in \Theta$, radius $R$ and $a, c > 0$, the Hellinger affinity $A(P, P_0)$ is bounded by $a/||\theta - \theta^*||^b$ outside the ball $B(\theta^*, R)$ and bounded by $e^{-c||\theta - \theta^*||^2}$ inside the ball. If $R \geq 11a^{1/b}/3e$, then

$$\mathbb{E}D_B(P, P_\theta) \leq 2R_{\Theta_\epsilon}^{(n)} \|P\|_2(P) + \frac{4d}{n} \left( \log \left( 1 + \frac{2\sqrt{2\pi}c}{\epsilon \sqrt{e}} \right) + \log \left( 1 + \frac{2\sqrt{d+6}}{\epsilon \sqrt{e}} \right) \right),$$

$$+ \frac{4 \log(2 + \frac{22}{R^2}) + 2||\theta^*||^2 + 8}{n}.$$

**Proof:** This time we use Corollary 5 rather than Corollary 6. The challenge is to bound the summation

$$\sum_{\theta \in \Theta \cap B(\theta^*, R)} e^{-||\theta||^2} \left( \frac{a}{||\theta - \theta^*||^b} \right)^{an}.$$

Assuming $n \geq 2(d+1)/b$, we can bound that term as in Theorem 11. With smaller $n$, we invoke Lemmas 32 and 33. In each case, the bound is no greater than the one we have claimed.

As in Corollary 7, the bounds in Theorems 11 and 12 remain stable if $\epsilon$ is proportional to $1/\sqrt{n}$.

As an example, we will see how these bound apply in a location family parametrized by the mean in $\Theta \subseteq \mathbb{R}^d$. First, we establish the power decay, assuming $P$ has a finite first moment. By Lemma 35,

$$A(P, P_0) \leq \frac{2(sp + s\theta)}{||\theta - \theta^*||^b},$$

where $\theta^* := \mathbb{E}_{X \sim P}X$, and the other constants are the first central moments $s_P := \mathbb{E}_{X \sim P}X - \theta^*$ and $s\theta := \mathbb{E}_{X \sim P}||X - \theta||$. Therefore, Theorems 11 and 12 apply if we can find a Gaussian-shaped Hellinger affinity bound that holds inside the ball centered at $\theta^*$ with radius $R = 22(s_P + s\theta) \vee 3e$.

In particular, let us assume the model comprises distributions that are continuous with respect to Lebesgue measure. Then we will also assume that $P$ is continuous; otherwise, the risk bound is infinite anyway. These assumptions ensure the existence of exact medians, enabling us to use Lemma 37.

Let $v$ be the vector of marginal medians of the model distribution with mean $\theta = 0$. The marginal median vector of any model distribution $P_\theta$ is then $\theta + v$. Let $m_P$ be the marginal median vector of $P$. By Lemma 37, for any $r \geq 0$, the inequality

$$A(P, P_\theta) \leq e^{-c||\theta + v - m_P||^2}$$

holds for $\theta$ within $B(m_P - v, r)$, where $c$ is $\frac{1}{2d}$ times the minimum squared marginal density of $P_\theta$ within $r$ of its median. It remains to identify an $r$ large enough that $B(m_P - v, r)$ contains $B(\theta^*, R)$. Using the triangle inequality and then Lemma 21 to bound the distance between means and medians,

$$||\theta - (m_P - v)|| = ||\theta - \theta^* + v - (m_P - \theta^*)|| \leq ||\theta - \theta^*|| + ||v|| + ||m_P - \theta^*|| \leq ||\theta - \theta^*|| + s\theta \sqrt{d} + s_P \sqrt{d}.$$

For $\theta$ in the ball $B(\theta^*, R)$, the first term is bounded by $R$. This tells us that the ball $B(m_P - v, R + \sqrt{d}(s\theta + s_P))$ contains $B(\theta^*, R)$.

Thus if all the marginal densities of $P_\theta$ are positive within $R + \sqrt{d}(s\theta + s_P)$ of their medians, then there is a positive $c$ for which

$$A(P, P_\theta) \leq e^{-c||\theta - (m_P - v)||^2}$$

in $B(\theta^*, R)$, confirming that Theorems 11 and 12 hold.

If the data-generating distribution is itself in the location family, then $P = P_{\theta^*}$ and $s_P = s\theta$. Thus the bound holds uniformly over $\theta^* \in \Theta$. If there exists $\beta > 0$ such that every $\epsilon$-discretization of the family is a KL-net with radius $\beta \epsilon^2$, then a minimax risk bound can be derived in the same manner as Corollary 10.

**C. Log Reciprocal pmf of $\hat{\theta}$ as Pseudo-Penalty**

In Section II-B, we chose a pseudo-penalty to have an expectation that 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's log summation term by letting $L$ be twice a code-length function. The smallest resulting $EL(\hat{\theta})$ comes from setting $L$ to be twice 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 $\Theta$-valued deterministic transformation $\hat{\theta}$).

**Corollary 13:** 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

$$\mathbb{E}D_B(P, P_\theta) \leq R_{\Theta_\epsilon(L)}^{(n)}(P) + \frac{2H(\hat{\theta}) - EL(\hat{\theta})}{n}.$$

It is known that the risk of the MLE is bounded by the log-cardinality of the model (e.g. [14]); Corollary 13 implies a generalization of this fact for penalized MLEs:

$$\mathbb{E}D_B(P, P_\theta) \leq R_{\Theta_\epsilon(L)}^{(n)}(P) + \frac{2log(||\theta|| - EL(\hat{\theta}))}{n}.$$

Importantly, Corollary 13 also applies to models of infinite cardinality.

**Lemma 14:** Let $\Theta_\epsilon \subseteq \mathbb{R}^d$ be an $\epsilon$-discretization, and let $\hat{\theta}$ be a $\Theta_\epsilon$-valued random vector. Suppose that for some $\theta^* \in \mathbb{R}^d$ and some radius $R \geq 0$, every $\theta \in \Theta_\epsilon$ outside of $B(\theta^*, R)$ has probability bounded by $e^{-c||\theta - \theta^*||^2}$. Then the entropy of $\hat{\theta}$ has the bound

$$H(\hat{\theta}) \leq \frac{d}{2} \left( \frac{4\sqrt{\pi}}{\epsilon \sqrt{c}} \right)^d + dlog(1 + \frac{2(c^{-1/2} \vee R \vee 3e)}{\epsilon}).$$
If $\epsilon \sqrt{n} \leq 4\sqrt{n}$, then this bound grows exponentially in $d$. However, if $c$ and $R$ are known, then one can set $\epsilon \geq 4\sqrt{n}/(\sqrt{c} \vee R/3)$ and find that $H(\hat{\theta})$ is guaranteed to be bounded by $3d$. Of course, one needs to take the behavior of the index of resolvability into account as well; good overall behavior will typically require that $c$ has order $n$.

In certain models satisfying $D(P_{\|P}) \geq a/\|\theta - \theta^*\|^2$ for some $a > 0$, we surmise that it may be possible to establish the applicability of Lemma 14 (with $c$ having order $n$) by using information theoretic large deviation techniques along the lines of [16, Th. 19.2].

**D. Quadratic Form as Pseudo-Penalty**

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

$$\sum_{\theta \in \Theta} e^{-L(\theta)} \leq \sum_{\theta \in \Theta} e^{-\lambda_d(M)\|\theta - \hat{\theta}\|^2} \leq \left(1 + \frac{2\sqrt{n}}{\epsilon\lambda_d(M)}\right)^d,$$

by Lemma 27. Using $aI_d$ as $M$ gives us Corollary 15.

**Corollary 15:** Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over an $\epsilon$-discretization $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$ indexing an iid model with penalty $L$. Then for any $\alpha \geq 0$,

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

As described in Section III-D, one gets desirable order $1/n$ behavior from $\mathcal{R}^{(n)}_{\epsilon, \theta}(P)$ by using $\epsilon$ 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 15, one sets $\alpha$ 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 $\Theta$ 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 15 shifts the problem to another risk-related quantity, while the pseudo-penalties used in Sections II-B and II-C provide more direct ways of deriving exact risk bounds of order $1/n$.

**E. Penalty as Pseudo-Penalty**

Another simple corollary to Theorem 1 uses $L = aL$.

**Corollary 16:** Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the penalized MLE over $\Theta$ indexing a countable iid model with penalty $L$. Then

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

Bayesian MAP (maximum a posteriori) is a common penalized likelihood procedure that has insufficient penalty for the index of resolvability bound (2) to be valid. In that case, Corollary 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 16 by setting $a = 1$.

**Corollary 17:** Assume $X^n \overset{iid}{\sim} P$, and let $\hat{\theta}$ be the MAP estimate over $\Theta$ 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 $\epsilon$-discretizations, realize that $q$ has to change as the refinement increases; thus the second term in Corollary 17 should be considered to have order strictly larger than $1/n$ in that context.

**III. Adaptive Modeling**

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 $\Theta$ by $\nu = (k, \theta)$, where with $\theta \in \Theta^{(k)}$. Assume the penalty and pseudo-penalty have the form $L(\nu) = \mathcal{L}_0(k) + \mathcal{L}_1(\theta)$ and $L(\nu) = L_0(k) + L_k(\theta)$. Then Theorem 1 can be useful if the penalty plus pseudo-penalty on $k$ is large enough to counteract the within-model summations.

$$\sum_{\nu \in \Theta} e^{-\frac{1}{2}[L_0(k) + L(k)]} \leq \sum_{k \geq 1} e^{-\frac{1}{2}[\mathcal{L}_0(k) + L_0(k)]} \sum_{\theta \in \Theta^{(k)}} e^{-\frac{1}{2}[\mathcal{L}_1(\theta) + L_k(\theta)]}.$$ 

One can use $L_0(k) = 0$ to avoid having to worry about the behavior of $k$. Then bounds on $\sum_{\theta \in \Theta^{(k)}} e^{-\frac{1}{2}[\mathcal{L}_1(\theta) + L_k(\theta)]}$ should be known so that one can devise a penalty on $k$ that bounds the weighted sum of these summations. In particular, if such bounds do not depend on any unknown quantities, then one can set $L_0(k) \geq k\sqrt{2 + 2\log\sum_{\theta \in \Theta^{(k)}} e^{-\frac{1}{2}[\mathcal{L}_1(\theta) + L_k(\theta)]}}$ and have

$$\log\sum_{k \geq 1} e^{-\frac{1}{2}[\mathcal{L}_0(k) + L(k)]} \leq 0.$$ 

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 5.

An important feature of the resolvability bound method is its generality; bounds can be derived that assume very little about the data-generating distribution. In non-adaptive models, however, the bound cannot become small if the data-generating distribution is far from Gaussian. Our hope is to derive similar exact risk bounds for penalized MLEs over flexible model classes as well, such as Gaussian mixtures, so that $D(P|\theta)$ can be made small (or possibly zero) for large classes of potential data-generating distributions.

**APPENDIX**

**A. Miscellaneous Facts**

The following handy facts are known, but we provide brief proofs here nonetheless.
Lemma 18: For any vectors $u, v$ in a real inner product space,
\[ \| u - v \|^2 \leq 2\| u \|^2 + 2\| v \|^2. \]

Proof: We apply the Cauchy-Schwarz inequality followed by the arithmetic-geometric mean inequality.
\[
\| a - b \|^2 = \| a \|^2 + \| b \|^2 - 2(a, b) \\
\leq \| a \|^2 + \| b \|^2 + 2\| a \|\| b \| \\
\leq \| a \|^2 + \| b \|^2 + 2(\| a \|^2/2 + \| b \|^2/2).
\]

Lemma 19: Let $v \in \mathbb{R}^d$, and let $M$ be a symmetric $d \times d$ matrix. Then
\[ \lambda_d(M) \leq \frac{v' M v}{\| v \|^2} \leq \lambda_1(M). \]

Proof: Any symmetric matrix has an orthonormal eigenvector decomposition $M = Q \Lambda Q'$.
\[
v' M v = v' Q \Lambda Q' v \\
= \sum_j \lambda_j (Q' v)_j^2 \\
= \| v \|^2 \sum_j \lambda_j \left( Q' v \frac{v}{\| v \|} \right)_j^2.
\]
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 $\lambda_1$ when $v$ is in the direction of the first eigenvector, and it achieves its minimum $\lambda_d$ when $v$ is in the direction of the last eigenvector.

Lemma 20: Let $a_1, \ldots, a_K \geq 1/K$. Then
\[ \log \sum_k a_k \leq \sum_k \log a_k + K \log K. \]

Proof: We apply the log-sum inequality and realize that it produces coefficients bounded by 1.
\[
\log \sum_k a_k = \frac{1}{\sum_k K a_k} \left[ \left( \sum_k K a_k \right) \log \sum_k \frac{K a_k}{1} \right] \\
\leq \frac{1}{\sum_k K a_k} \left[ \sum_k K a_k \log \frac{K a_k}{1} \right] \\
\leq \sum_k \log K a_k \\
= \sum_k \log a_k + K \log K.
\]

Lemma 21: Let $X \sim P$, where $P$ is a probability distribution on $\mathbb{R}^d$ with marginal median vector $m_P$. Then
\[ \| m_P - \mathbb{E} X \| \leq \sqrt{d} \| X - \mathbb{E} X \|. \]

Proof: Superscripts indicate coordinates. We use subadditivity of square root and the fact that the median minimizes expected absolute deviation.
\[
\| \mathbb{E} X - m_P \| \leq \| \mathbb{E} X - m_P \|_1 \\
= \| \sum_j [\mathbb{E} (X^{(j)}) - m_P^{(j)}] \| \\
= \| \sum_j [\mathbb{E} (X^{(j)}) - m_P^{(j)}] \| \\
\leq \sum_j \| \mathbb{E} (X^{(j)}) - m_P^{(j)} \| \\
\leq \sum_j \| \mathbb{E} (X^{(j)}) - \mathbb{E} (X^{(j)}) \| \\
\leq \sqrt{d} \| \mathbb{E} X - \mathbb{E} X \|.
\]

We used the fact that $l^1$ and $l^2$ satisfy $\| v \| \leq \| v \|_1 \leq \sqrt{d} \| v \|$.

B. 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 22: Let $Y$ be a random vector with convex support $S \subseteq \mathbb{R}^d$. If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is twice continuously differentiable, then
\[ \inf_{y \in S} \lambda_d(\nabla \nabla' f(y)) \leq \frac{\mathbb{E} f(Y) - f(\mathbb{E} Y)}{\mathbb{E} Y/2} \leq \sup_{y \in S} \lambda_1(\nabla \nabla' f(y)). \]

Proof: We start with a second-order Taylor expansion with Lagrange remainder.
\[ f(Y) = f(\mathbb{E} Y) + (Y - \mathbb{E} Y) \nabla f(\mathbb{E} Y) \\
+ \frac{1}{2} (Y - \mathbb{E} Y) \nabla \nabla' f(\tilde{Y}) (Y - \mathbb{E} Y), \]
for some $\tilde{Y}$ on the segment from $Y$ to $\mathbb{E} Y$. By Lemma 19, 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.
C. 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 23 will prove to be remarkably useful for such tasks.

Lemma 23: Let $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$, and assume $f : \Theta \to \mathbb{R}$ is twice continuously differentiable. If $\theta$ is in the convex hull of $\Theta_\epsilon \cap B(\theta, \delta)$, then

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

Proof: We first bound the infimum over $\Theta_\epsilon$ 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 $\theta$ is some weighted average of nearby grid-points (the ones at most $\delta$ distance away), and we can use that same weighted averaging to define $Q$. Then the expectation of the random selection is $\theta$, and we apply Lemma 22.

$$\inf_{\theta_\epsilon \in \Theta_\epsilon} f(\theta_\epsilon) \leq \inf_{\theta_\epsilon \in \Theta_\epsilon \cap B(\theta, \delta)} f(\theta_\epsilon) \leq \mathbb{E}_{\theta_\epsilon \sim Q} f(\theta_\epsilon) \leq f(\mathbb{E}_{\theta_\epsilon \sim Q} \theta_\epsilon) + \frac{1}{2} \mathbb{E}_{\theta_\epsilon \sim Q} \|\theta_\epsilon - \mathbb{E}_{\theta_\epsilon \sim Q} \theta_\epsilon\|^2 \sup_{\delta B(\theta, \delta)} \lambda_1(\nabla^2 f(\hat{\theta})) \leq f(\theta) + \frac{\delta^2}{2} \sup_{\delta B(\theta, \delta)} \lambda_1(\nabla^2 f(\hat{\theta})),$$

assuming $\lambda_1(\nabla^2 f(\hat{\theta}))$ 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. □

Suppose $\Theta_\epsilon \subseteq \Theta \subseteq \mathbb{R}^d$ is an $\epsilon$-discretization, as defined in Section II. If $\Theta$ is convex, then $\theta$ is in the convex hull of $\Theta_\epsilon$ satisfies the conditions of Lemma 23 with $\epsilon \sqrt{d}$ as $\delta$. In particular, if every dimension of $\Theta$ is either $\mathbb{R}$ or a closed half-line, then there is an obvious $\epsilon$-discretization that makes Lemma 23 apply for every $\theta \in \Theta$. For less favorably shaped $\Theta$, one can consider adding more grid-points “on top of” an $\epsilon$-discretization.

D. Behavior of $\mathcal{R}^{(n)}_{\Theta_\epsilon, \mathcal{L}}(P)$

One way to bound $\mathcal{R}^{(n)}_{\Theta_\epsilon, \mathcal{L}}(P)$ is to use an approach similar to Section III-C. Suppose $p_\theta(x)$ is twice continuously differentiable in $\theta$. We define a type of Fisher “cross-information” matrix

$$I_\theta(\hat{\theta}) := \mathbb{E}_{X \sim p} \nabla^2 \left[ \log \frac{1}{p_\theta(X)} \right]_{\theta = \hat{\theta}},$$

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

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

Theorem 24: 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 $\theta$ for every fixed $x$ in its domain. If $\theta \in \Theta$ is in the convex hull of $\Theta_\epsilon \cap B(\theta, \delta)$, then

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

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 $\theta$ with Lagrange remainder and reason similarly to the proofs of Lemmas 22 and 23.

$$\mathcal{R}^{(n)}_{\Theta_\epsilon, \mathcal{L}}(P) = \inf_{\theta_\epsilon \in \Theta_\epsilon} f_X(\theta_\epsilon) = \inf_{\theta_\epsilon \in \Theta_\epsilon} \left( f_X(\theta) + (\theta - \theta)^\top \nabla f_X(\theta) + \frac{1}{2} (\theta - \theta)^\top \nabla^2 f_X(\theta) (\theta - \theta) \right)$$

$$= \inf_{\theta_\epsilon \in \Theta_\epsilon} \left( f_X(\theta) + (\theta - \theta)^\top \nabla f_X(\theta) + \frac{1}{2} (\theta - \theta)^\top [\nabla^2 f_X(\theta) + \frac{1}{n} \nabla^2 \mathcal{L}(\hat{\theta})] (\theta - \theta) \right),$$

for some $\hat{\theta}$ between $\theta$ and $\theta_\epsilon$.

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

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

Corollary 25: Let $\Theta \subseteq \mathbb{R}^d$ be a convex parameter space having densities twice continuously differentiable in $\theta$. Let $\Theta_\epsilon \subseteq \Theta$ be an $\epsilon$-discretization. For any $\theta$ in the convex hull of $\Theta_\epsilon$,

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

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 $z := \sup_{\theta \in B(\theta, \epsilon \sqrt{d})} \lambda_1(I_\theta(\hat{\theta}) + \nabla^2 \mathcal{L}(\hat{\theta}))$ 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 $P_\theta$ being the r1-projection of $P$ onto $\Theta$ if it exists. In particular, if $P$ is in the model, then we can let $P_\theta$ be $P$ to get an exact bound of order $1/n$ for $\mathcal{R}^{(n)}_{\Theta_\epsilon, \mathcal{L}}(P)$.
E. Bounding Summations Over Grid-Points

Lemmas 26 and 27 provide bounds for summations of Gaussian-shaped functions over $\epsilon$-discretizations of $\mathbb{R}^d$.

**Lemma 26:** Let $\Theta_\epsilon$ be an $\epsilon$-discretization of $\mathbb{R}^d$. Then for any $c > 0$ and $v \in \Theta_\epsilon$,
\[
\sum_{\theta \in \Theta_\epsilon} e^{-c\|\theta - v\|^2} \leq \left(1 + \frac{\sqrt{\pi}}{\epsilon \sqrt{c}}\right)^d.
\]

**Proof:** We can assume without loss of generality that $v$ is the zero vector and that $\Theta_\epsilon$ 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.
\[
\sum_{\theta \in \Theta_\epsilon} e^{-c\theta^2} = \sum_{\theta \in \Theta_\epsilon} e^{-c\epsilon^2 \theta^2} = \sum_{\epsilon \in \mathbb{Z}} e^{-c\epsilon^2 \frac{\theta^2}{\epsilon^2}} \leq 1 + \int_{\mathbb{R}} e^{-c\epsilon^2 \frac{\theta^2}{\epsilon^2}} d\theta = 1 + \frac{\sqrt{\pi}}{\epsilon \sqrt{c}}.
\]

The $d$-dimensional problem can be bounded in terms of $d$ instances of the one-dimensional problem. Let $\Theta_\epsilon^{(1)}, \ldots, \Theta_\epsilon^{(d)}$ represent the underlying discretizations of $\mathbb{R}$, so that $\Theta_\epsilon = \prod_j \Theta_\epsilon^{(j)}$.
\[
\sum_{\theta \in \Theta_\epsilon} e^{-c\|\theta\|^2} = \sum_{\theta \in \Theta_\epsilon} \sum_{j} e^{-c\theta_j^2} = \sum_{\theta_1 \in \Theta_\epsilon^{(1)}} \cdots \sum_{\theta_d \in \Theta_\epsilon^{(d)}} \prod_j e^{-c\theta_j^2} \leq \prod_j \left(1 + \frac{\sqrt{\pi}}{\epsilon \sqrt{c}}\right) = \left(1 + \frac{\sqrt{\pi}}{\epsilon \sqrt{c}}\right)^d.
\]

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

**Lemma 27:** Let $\Theta_\epsilon$ be an $\epsilon$-discretization of $\mathbb{R}^d$. Then for any $c > 0$ and $v \in \mathbb{R}^d$,
\[
\sum_{\theta \in \Theta_\epsilon} e^{-c\|\theta - v\|^2} \leq \left(1 + \frac{2\sqrt{\pi}}{\epsilon \sqrt{c}}\right)^d.
\]

**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 26 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.
\[
\sum_{\theta \in \Theta_\epsilon} e^{-c\|\theta - v\|^2} \leq \sum_{\theta \in \Theta_{\epsilon/2}} e^{-c\|\theta - v\|^2} \leq 1 + \frac{\sqrt{\pi}}{(\epsilon/2) \sqrt{c}}.
\]

As before, the $d$-dimensional problem reduces to the one-dimensional problem.
\[
\sum_{\theta \in \Theta_\epsilon} e^{-c\|\theta - v\|^2} = \prod_j \sum_{\theta_j \in \Theta_{\epsilon/2}^{(j)}} e^{-c\theta_j^2} \leq \prod_j \left(1 + \frac{2\sqrt{\pi}}{\epsilon \sqrt{c}}\right) = \left(1 + \frac{2\sqrt{\pi}}{\epsilon \sqrt{c}}\right)^d.
\]

There are important situations in which a function can be bounded by one type of behavior near its peak and by another type of behavior further away. When the tail behavior has a spherically symmetric bound, Lemma 28 can help us convert the summation problem into a one-dimensional integral.

**Lemma 28:** Let $f$ be a real-valued function of the form $f(\theta) = g(|\theta - \theta^*|)$ for some non-increasing and non-negative function $g$. Let $\Theta_\epsilon$ be an $\epsilon$-discretization of $\Theta \subseteq \mathbb{R}^d$. For any radius $R \geq 3\epsilon$,
\[
\sum_{\theta \in \Theta_\epsilon \cap B(\theta^*, R)} f(\theta) \leq \frac{2\pi^{d/2}}{(\epsilon/4)^d \Gamma(d/2)} \int_{R/4}^{\infty} g(r)r^{d-1}dr,
\]
if the integral is well-defined.

**Proof:** First, we bound the summation outside the ball with diameter $2R$ by the summation outside the coordinate-axes-aligned hypercube inscribed by the ball, which has sides of length $\sqrt{2}R$. Next, we introduce a new more refined grid $\Theta_{\epsilon/2}$, which is the $\epsilon/2$-discretization that includes the point $\theta^*$. Consider the hyperplanes of $\Theta_{\epsilon}$ grid-points orthogonal to the first coordinate axis. One can “translate inward” each of these hyperplanes to a hyperplane in $\Theta_{\epsilon/2}$ that is closer to $\theta^*$. The argument can be repeated for each coordinate axis in turn. Because have assumed $R \geq 3\epsilon$, each translated point outside of the $\sqrt{2}R$ hypercube remains outside of the $R$ hypercube. Thus it suffices to sum over the points of $\Theta_{\epsilon/2}$ outside of the hypercube of side-length $R$. For the remainder of this proof, we will assume without loss of generality that $\theta^*$ is the zero vector.

We will complete our proof by bounding the function’s values at each point by its average value over a unique hypercube closer to zero. Given the standard $\epsilon$-discretization of $\mathbb{R}^d$, let $j$ index shells radiating outward from the origin. The $j$th shell comprises the grid-points on the boundary of the centered hypercube of side-length $2je$, along with the boundary hypercubes of volume $e^j$ demarcated by those grid-points. (We will consider the origin point itself to be the 0th shell.) The total number of points in the first $j$ shells is $(2j+1)^d$, so the number of points in the $(j+1)$st shell is $[2(j+1)+1]^d - [2j+1]^d$. Similarly, the number of hypercubes
in the $(j+1)$st shell is $[2(j+1)]^d - [2j]^d$. Because $t \mapsto t^d$ is convex and increasing on $\mathbb{R}^+$, $[2(j+1)]^d - [2j]^d \leq [2(j+1) + 2]^d - [2j + 2]^d = [2(j+2)]^d - [2(j+1)]^d$.

In other words, the number of points in the $(j+1)$st shell is no greater than the number of hypercubes in the $(j+2)$nd shell. Finally, we introduce yet another grid, $\Theta_{\epsilon/4}^*$. We know that the number of points in the $j$th shell of $\Theta_{\epsilon/4}^*$ is bounded by the number of hypercubes in the $(j+1)$st shell of $\Theta_{\epsilon/4}^*$. If we can establish that these hypercubes are closer to the origin than are the points in the $j$th shell of $\Theta_{\epsilon/4}^*$, then we can bound the sum of the points’ function values by the sum of the hypercubes’ average function values.

The points comprising the $j$th shell of $\Theta_{\epsilon/4}^*$ are inscribed by a sphere of radius $j\epsilon/2$; that sphere is inscribed by a hypercube of radius $j\epsilon/2\sqrt{d}$. As $j$ increases, the $(j+1)$st shell of $\Theta_{\epsilon/4}^*$ will be about half as far from the origin as the $j$th shell of $\Theta_{\epsilon/4}^*$. Because we have assumed $R \geq 3\epsilon$, the smallest $j$ we will need to worry about is $j = 3$. The third shell of $\Theta_{\epsilon/4}^*$ has distance $3\epsilon/2$ from the origin, while the 4th shell of $\Theta_{\epsilon/4}^*$ has distance $\epsilon$ from the origin. As $\epsilon < 3\epsilon/2$, every hypercube in the 4th shell of $\Theta_{\epsilon/4}^*$ is entirely closer to the origin than any point in the 3rd shell of $\Theta_{\epsilon/4}^*$. The comparison continues to hold for all $j \geq 3$.

The average value of $f$ on a hypercube within $\Theta_{\epsilon/4}^*$ is equal to the integral over that region divided by the hypervolume $(\epsilon/4)^d$. The inner-most hypercube that we need to consider is a distance of $2\epsilon/\sqrt{d}$ from the origin. We let $H_{0}(z)$ denote the coordinate-axes-aligned hypercube centered at the origin with side-length $2\epsilon$; we will need to integrate over the complement of this hypercube. Because $g/((\epsilon/4)^d)$ is non-negative, we can bound this integral by the integral over a larger region, the complement of a ball. We then use spherical symmetry to reduce the problem to a one-dimensional integral.

\[
\frac{1}{(\epsilon/4)^d} \int_{H_{0}(R/4)} g(\|\theta\|) d\theta \leq \frac{1}{(\epsilon/4)^d} \int_{B(0,R/4)} g(\|\theta\|) d\theta = \frac{1}{(\epsilon/4)^d} \int_{R/4}^{\infty} g(r) S_{r} dr \leq \frac{2\pi^{d/2}}{(\epsilon/4)^d \Gamma(d/2)} \int_{R/4}^{\infty} g(r) r^{d-1} dr,
\]

where $S_{r}$ is the “surface area” of any ball in $\mathbb{R}^d$ with radius $r$, which is $2\pi^{d/2}r^{d-1}/\Gamma(d/2)$.

A more manageable quantity for the right-hand-side of Lemma 28 can be derived.

**Lemma 29:** For any $\epsilon, d > 0$, $2\pi^{d/2}/(\epsilon/4)^d \Gamma(d/2) \leq \left( \frac{20\epsilon}{\sqrt{d}} \right)^d$.

**Proof:** [17, Th. 1] provides a Stirling lower bound for the gamma function:

\[
\Gamma(d/2) \geq \frac{\sqrt{2\pi} (d/2)^{d/2}}{e^{d/2} d^{d/2}}.
\]

We also upper bound $\sqrt{d}$ by $(2.9/2)^{d/2}$. The overall bound of $(20\epsilon\sqrt{d})^d$ comes from rounding numbers up to the nearest integer.

Next, we apply Lemma 29 to power decay functions.

**Lemma 30:** Let $\Theta_{\epsilon}^*$ be an $\epsilon$-discretization of $\mathbb{R}^d$. For any $R \geq 3\epsilon$ and $q > d$,

\[
\sum_{\theta \in \Theta_{\epsilon}^* \cap B(\theta^*; R)} \frac{1}{\|\theta - \theta^*\|^q} \leq \left( \frac{20}{\epsilon \sqrt{d}} \right)^d \frac{(4/R)^{q-d}}{q - d}.
\]

**Proof:** This is a straight-forward application of Lemmas 28 and 29.

\[
\sum_{\theta \in \Theta_{\epsilon}^* \cap B(\theta^*; R)} \frac{1}{\|\theta - \theta^*\|^q} \leq \left( \frac{20}{\epsilon \sqrt{d}} \right)^d \int_{R/4}^{\infty} \frac{r^{q-d}}{r^q dr} = \left( \frac{20}{\epsilon \sqrt{d}} \right)^d \left( \frac{R^{q-d}}{q - d} \right)_{R/4}^{\infty}.
\]

In our applications, the decaying functions will often be taken to the $an$ power for some $a \in [0, 1]$. For power decay in that case, Lemma 30 can be used to derive a bound that is exponential in dimension and is stable if $\epsilon$ is proportional to $1/\sqrt{n}$.

**Lemma 31:** Assume $n \geq (d + 1)/ab, a > 0$, and $R \geq 4a^{1/b} \sqrt{3\epsilon}$. Then

\[
\sum_{\theta \in \Theta_{\epsilon}^* \cap B(\theta^*; R)} \left( \frac{a}{\|\theta - \theta^*\|^b} \right)^{an} \leq \left( \frac{4R}{\epsilon \sqrt{nab} \log(R/4a^{1/b})} \right)^d.
\]

**Proof:** Start with Lemma 30, then apply the assumption that $ban - d \geq 1$.

\[
\sum_{\theta \in \Theta_{\epsilon}^* \cap B(\theta^*; R)} \left( \frac{a}{\|\theta - \theta^*\|^b} \right)^{an} = \sum_{\theta \in \Theta_{\epsilon}^* \cap B(\theta^*; R)} \left( \frac{a}{\|\theta - \theta^*\|^b} \right)^{ban} \leq \left( \frac{20}{\epsilon \sqrt{d}} \right)^d \left( \frac{4a^{1/b}}{R} \right)^{ban} \leq \left( \frac{20}{\epsilon \sqrt{d}} \right)^d \left( \frac{4a^{1/b}}{R} \right)^{ban}.
\]

Assuming $4a^{1/b} < R$, the quantity $n^{d/2}(4a^{1/b})^{ban}$ is maximized at $n = \frac{2\log(R/4a^{1/b})}{2\epsilon \sqrt{d} \log(R/4a^{1/b})}$. Substituting this critical value and rounding up gives us the desired bound.

Suppose the sample size is not large enough for Lemma 31 to be valid. If the summand is multiplied by a Gaussian-shaped function, then it is still possible to derive a bound that is stable if $\epsilon$ is proportional to $1/\sqrt{n}$, although the dependence on dimension becomes worse. Lemmas 32 and 33 splits the problem into two additional ranges for $n$.
Lemma 32: Assume \( n \leq (d - 1)/ab \), \( a, \kappa > 0 \), and \( R \geq 4a^{1/b} \sqrt{3\varepsilon} \). Then
\[
\sum_{\theta \in \Theta \cap B(\theta^*, R)} e^{-\kappa ||\theta||^2} \left( \frac{a}{\|\theta - \theta^*\|^b} \right)^{an} \leq 2 e^{\kappa ||\theta||^2} \left( \frac{4\sqrt{2\pi e \sqrt{d} \vee a^{2/b} \kappa}}{\varepsilon \sqrt{nabk}} \right)^d .
\]

Proof: By Lemma 18, we can upper bound \( ||\theta||^2 \) in terms of \( ||\theta - \theta^*||^2 \) and \( ||\theta^*||^2 \).
\[ e^{-\kappa ||\theta||^2} \leq e^{-\frac{2}{b} ||\theta - \theta^*||^2 + \kappa ||\theta^*||^2} . \]

Using this, we apply Lemma 28.
\[
\sum_{\theta \in \Theta \cap B(\theta^*, R)} e^{-\kappa ||\theta||^2} \left( \frac{a}{\|\theta - \theta^*\|^b} \right)^{an} \leq e^{\kappa ||\theta^*||^2} a^{an} \sum_{\theta \in \Theta \cap B(\theta^*, R)} e^{-\frac{2}{b} ||\theta - \theta^*||^2} \left( \frac{1}{\|\theta - \theta^*\|^b} \right)^{an} \leq e^{\kappa ||\theta^*||^2} a^{an} \frac{2\pi^{d/2}}{(\epsilon^4)^d \Gamma(d/2)} \int_{R/4}^{\infty} e^{-\kappa r^2/2} r^{d-ban-1} dr .
\]

The integral from \( R/4 \) to \( \infty \) can be upper bounded by the integral from 0 to \( \infty \). Then we change the variable to \( r^2 \) and compare the integrand to a Gamma distribution’s density.
\[
\int_{R/4}^{\infty} e^{-\kappa r^2/2} r^{d-ban-1} dr \leq \int_{R/4}^{\infty} e^{-\kappa r^2/2} r^{d-ban-1} dr = \frac{1}{2} \int_0^{\infty} e^{-\kappa r^2/2} (r^2)^{(d-ban-2)/2} \Gamma(d/2) (r^2) \frac{2\pi^{d/2}}{(\epsilon^4)^d \Gamma(d/2)} = \frac{\Gamma(d-ban)}{2(\kappa/2)^{(d-ban)/2}} .
\]

Applying the Stirling upper and lower bounds for the gamma function [17, Th. 1], we arrive at the bound
\[
\sum_{\theta \in \Theta \cap B(\theta^*, R)} e^{-\kappa ||\theta||^2} \left( \frac{a}{\|\theta - \theta^*\|^b} \right)^{an} \leq e^{\kappa ||\theta^*||^2} a^{an} \frac{2\pi^{d/2}}{(\epsilon^4)^d \Gamma(d/2)} \frac{\Gamma(d-ban)}{2(\kappa/2)^{(d-ban)/2}} \leq e^{\kappa ||\theta^*||^2} a^{an} \frac{4\sqrt{2\pi} e^{\sqrt{d}}}{\epsilon \sqrt{n\kappa}} \left( \frac{a^{2/b} \kappa e}{d} \right)^{ban/2} \times e^{1/6(d-ban)} \sqrt{1 - \frac{ban}{d}} .
\]

We have assumed that \( d - ban \geq 1 \), so we can upper bound the final two factors by \( e^{1/6} \) and 1. If \( d \leq a^{2/b} \kappa e \), then we substitute \( d/ba \) for \( n \) to upper bound \( n^{d/2}(2d/ba)^{ban/2} \) by \( (a^{2/b} \kappa e)^{d/2} \). Otherwise, the optimizer of the product of these two factors is \( n = \frac{d}{ba \log(d/a^{2/b} \kappa e)} \). Substituting this quantity into the second factor and \( d/ba \) into the first, we bound the product of the two factors by \( (\frac{d}{ba \log(d/a^{2/b} \kappa e)})^d \). Using \( e^{1/6} \leq 2 \) and \( d/e \leq ed \) simplifies the bound.

Lemma 33: Assume \( n \in \{(d - 1)/ab, (d + 1)/ab\} \), \( a, \kappa > 0 \), and \( R \geq 4a^{1/b} \sqrt{3\varepsilon} \). Then
\[
\sum_{\theta \in \Theta \cap B(\theta^*, R)} e^{-\kappa ||\theta||^2} \left( \frac{a}{\|\theta - \theta^*\|^b} \right)^{an} \leq e^{\kappa ||\theta^*||^2} \left( \frac{20}{\epsilon \sqrt{nab}} \right)^d \left( \frac{22}{R^3 + 2\sqrt{\kappa}} \right) .
\]

Proof: Begin as in Lemma 32. Our assumption about \( n \) ensures that the exponent of \( r \) is negative in the integral. First assume \( R/4 < 1 \).
\[
\int_{R/4}^{\infty} e^{-\kappa r^2/2} r^{d-ban-1} dr = \int_{R/4}^{\infty} e^{-\kappa r^2/2} r^{d-ban-1} dr + \int_{1}^{\infty} e^{-\kappa r^2/2} r^{d-ban-1} dr \leq \int_{R/4}^{\infty} r^{d-ban-1} dr + \int_{1}^{\infty} e^{-\kappa r^2/2} dr \leq \int_{R/4}^{\infty} r^{d-ban-1} dr + 2\pi \sqrt{\kappa} .
\]

If \( R/4 \geq 1 \), then the integral is bounded by \( \int_{R/4}^{\infty} e^{-\kappa r^2/2} dr \) which remains less than our bound.

We use Lemma 29 for the coefficient of the integral.
\[
\sum_{\theta \in \Theta \cap B(\theta^*, R)} e^{-\kappa ||\theta||^2} \left( \frac{a}{\|\theta - \theta^*\|^b} \right)^{an} \leq e^{\kappa ||\theta^*||^2} a^{an} \left( \frac{20}{\epsilon \sqrt{d}} \right)^{\frac{d}{2}} \left( \frac{22}{R^3 + 2\sqrt{\kappa}} \right) .
\]

If \( a < 1 \), the product \( n^{d/2} a^{an} \) is maximized at \( n = -d/2a \log a \). Substituting that into the second factor gives \( e^{-d/2} \); we bound the first factor using the fact that \( n \leq (d + 1)/ba \leq 2d/ba \). If \( a \geq 1 \), substitute \( 2d/ba \) in both factors. In either case, the product is bounded by \( (\frac{d}{\sqrt{ba}})^d \).

F. Hellinger Affinity

Lemma 34: Let \( P \) and \( Q \) be probability measures on \((X, A)\). For any event \( \mathcal{H} \in \mathcal{A} \),
\[
A(P, Q) \leq \sqrt{PH} \sqrt{QH} + \sqrt{P\mathcal{H}} \sqrt{Q\mathcal{H}} .
\]

Proof: Let \( p \) and \( q \) be densities of \( P \) and \( Q \) with respect to a common dominating measure \( \mu \). We use the
Cauchy-Schwarz inequality.\textsuperscript{[7]}  

\[ A(P, Q) = \int_H p(x)q(x)d\mu(x) \]

\[ = \int_H \sqrt{p(x)q(x)} dx + \int_{H^c} \sqrt{p(x)q(x)} d\mu(x) \]

\[ \leq \sqrt{\int_H p(x)d\mu(x)} \sqrt{\int_H q(x)d\mu(x)} \]

\[ + \sqrt{\int_{H^c} p(x)d\mu(x)} \sqrt{\int_{H^c} q(x)d\mu(x)}. \]

\textbf{Lemma 35:} Let \( P \) and \( Q \) be probability distributions on \( \mathbb{R}^d \). If they both have finite first moments, then

\[ A(P, Q) \leq \frac{2(\|X - E_X\| + \|Y - E_Y\|)}{\|E_X - E_Y\|}, \]

where \( X \sim P \) and \( Y \sim Q \).

\textbf{Proof:} Let \( H \) denote the halfspace containing \( E_X \) and demarcated by the perpendicular bisector of the path from \( E_X \) to \( E_Y \).

The \( P \)-probability of the complement of \( H \) is bounded by the probability of the complement of a ball within \( H \). We use Markov’s inequality to bound the probability that the deviation \( \|X - E_X\| \) is larger than \( \|E_X - E_Y\|/2 \).

\[ P(H^c) \leq P(\|X - E_X\| < \frac{\|E_X - E_Y\|}{2}) \]

\[ \leq \frac{\|E_X - E_Y\|}{\|E_X - E_Y\|/2}. \]

The same logic also allows us to bound \( QH \). Now invoke Lemma 34.

\[ A(P, Q) \leq \sqrt{P(H^c)Q(H^c)} \leq \sqrt{Q(H^c)P(H^c)} \]

\[ \leq \frac{\|E_Y - E_Y\|}{\|E_X - E_Y\|/2} + \frac{\|X - E_X\|}{\|E_X - E_Y\|/2}. \]

\textbf{Lemma 36:} Let \( P \) and \( Q \) be probability distributions on \( \mathbb{R} \). Suppose they have medians \( m_P \) and \( m_Q \), that is, \( P(\infty, m_P) = 1/2 \) and \( Q(\infty, m_Q) = 1/2 \). Then

\[ A(P, Q) \leq e^{-c z^2/2}, \]

where \( z \) is the \( P \) probability of the interval defined by an open endpoint at \( m_P \) and a closed endpoint at \( m_Q \).

\textbf{Proof:} Assume without loss of generality that \( m_P \leq m_Q \). Apply Lemma 34 using the interval \((\infty, m_P) \) for \( H \).

\[ A(P, Q) \leq \sqrt{P(H^c)Q(H^c)} \leq e^{-c z^2/2}, \]

where \( c := \frac{1}{2d} \min_{j \in \{1, \ldots, d\}} \min_{x \in [m_Q - R, m_Q + R]} q_j(x)^2. \)

\textbf{Proof of Lemma 9:} Let \( r \) denote the family’s carrier function and \( \psi \) denote the log-partition function.

\[ A(P_r, P_\theta) := \int_\chi r(x)e^{\theta^*\phi(x)}(x) = \int_\chi r(x)e^{\theta^*\phi(x)}dx \]

\[ = e^{-\frac{1}{2}(\psi(\theta^*) + \psi(\theta))} \int_\chi r(x)e^{\theta^*\phi(x)}dx \]

\[ = e^{-\frac{1}{2}(\psi(\theta^*) + \psi(\theta))} \langle x, (\theta^*)/2 \rangle \]

The exponent is a negative Jensen difference with a distribution that puts 1/2 mass on each of \( \theta \) and \( \theta^* \). Its expectation

\[ = \frac{1}{\sqrt{2}} \sqrt{1 + 1 - 4z^2} \]

\[ \leq 1 - z^2/2 \]

\[ \leq e^{-c^2/2}. \]

We recommend using mathematical software to algebraically verify the second-to-last step.

Regarding Lemma 36, note that \( A \) is symmetric in its arguments, so letting \( z \) be the \( P \) probability of the interval provides a valid bound as well.

\textbf{Lemma 37:} Let \( P \) and \( Q \) be probability distributions on \( \mathbb{R}^d \). Assume they have marginal medians \( m_P = (m_{P,1}, \ldots, m_{P,d}) \) and \( m_Q = (m_{Q,1}, \ldots, m_{Q,d}) \), and assume that \( Q \) has marginal densities \( q_1, \ldots, q_d \) with respect to Lebesgue measure. Let \( R \geq 0 \). Then for all \( Q \) with \( m_Q \in B(m_P, R) \),

\[ A(P, Q) \leq e^{-c (\|m_Q - m_P\|^2)}, \]

where

\[ c := \frac{1}{2d} \min_{j \in \{1, \ldots, d\}} \min_{x \in [m_Q - R, m_Q + R]} q_j(x)^2. \]

\textbf{Proof:} Let \( P^* \) and \( Q^* \) be the marginal distributions along the coordinate with the largest absolute difference between \( m_P \) and \( m_Q \); call the coordinates in this direction \( m_P^* \) and \( m_Q^* \).

By Lemma 36,

\[ A(P^*, Q^*) \leq e^{-c^2/2}, \]

where \( z \) is the \( Q^* \) probability of the interval from \( m_P^* \) to \( m_Q^* \).

It is at least as large as the absolute difference between the coordinates times the minimum value of the density \( q^* \) in the interval between them. The largest squared coordinate difference is at least as large as the average squared coordinate difference, that is

\[ |m_Q^* - m_P^*|^2 \geq \frac{1}{d} \|m_Q - m_P\|^2. \]

The 1/2 factor in \( c \) comes from Lemma 36.

The marginal distributions \( P^* \) and \( Q^* \) can be produced by “processing” draws from \( P \) and \( Q \). Because Hellinger affinity is a monotonically decreasing transformation of squared Hellinger divergence, the Data Processing Inequality implies that \( A(P, Q) \leq A(P^*, Q^*) \).

\textbf{Proof of Lemma 9:} Let \( r \) denote the family’s carrier function and \( \psi \) denote the log-partition function.

\[ A(P_r, P_\theta) := \int_\chi r(x)e^{\theta^*\phi(x)}(x) = \int_\chi r(x)e^{\theta^*\phi(x)}dx \]

\[ = e^{-\frac{1}{2}(\psi(\theta^*) + \psi(\theta))} \int_\chi r(x)e^{\theta^*\phi(x)}dx \]

\[ = e^{-\frac{1}{2}(\psi(\theta^*) + \psi(\theta))} \langle x, (\theta^*)/2 \rangle \]

The exponent is a negative Jensen difference with a distribution that puts 1/2 mass on each of \( \theta \) and \( \theta^* \). Its expectation

\[ = \frac{1}{\sqrt{2}} \sqrt{1 + 1 - 4z^2} \]

\[ \leq 1 - z^2/2 \]

\[ \leq e^{-c^2/2}. \]
is $(\theta^* + \theta)/2$, and its variance is $\|\theta - \theta^*\|^2/4$. Applying Lemma 22,

$$\frac{\psi(\theta^*) + \psi(\theta)}{2} - \psi\left(\frac{\theta^* + \theta}{2}\right) = \frac{\|\theta - \theta^*\|^2}{8} \inf_{\theta \in \Theta} \lambda_d(\nabla^2 \psi(\theta)).$$

It is a well-known fact about exponential families that $\nabla^2 \psi(\theta)$ is equal to the covariance matrix of the sufficient statistic vector $C_{\sim}^p$.\end{proof}

\section{Entropy of Subprobability Measures}

We extend the notion of entropy to more general measures. Let $Q$ be a measure on a countable set $X$. Then we define its entropy

$$H(Q) := \sum_{x \in X} q(x) \log \frac{1}{q(x)},$$

where $q(x) := Q(x)$ is the density of $Q$ with respect to counting measure.\footnote{One can likewise extend the notion of differential entropy $h$ for Borel measures on $\mathbb{R}^d$ by using Lebesgue measure rather than counting measure. Lemmas 38 and 39 also hold for differential entropy when $Q$ is a finite Borel measure on $X = \mathbb{R}^d$.}

\begin{lemma}
Let $Q$ be a finite measure on a countable set $X$, and let $\hat{Q} := \frac{Q}{\int Q}$ be the normalized version of $Q$. Then

$$H(Q) = (Q\lambda) H(\hat{Q}) + (Q \lambda) \log \frac{1}{Q \lambda}.$$\end{lemma}

\begin{proof}
$$H(Q) := \sum_{x \in X} q(x) \log \frac{1}{q(x)} = (Q\lambda) \sum_{x \in X} \frac{q(x)}{Q\lambda} \log \frac{1}{q(x)/Q\lambda} = (Q\lambda) \sum_{x \in X} \frac{q(x)}{Q\lambda} \log \frac{1}{q(x)/Q\lambda} + (Q\lambda) \sum_{x \in X} \frac{q(x)}{Q\lambda} \log \frac{1}{Q\lambda}.$$

\end{proof}

\begin{lemma}
Let $Q$ be a subprobability measure on a countable set $X$, and let $\hat{Q} := \frac{Q}{\int Q}$ be the normalized version of $Q$. Then

$$H(Q) \leq H(\hat{Q}) + 1/e.$$\end{lemma}

\begin{proof}
We apply Lemma 38, noting that $Q\lambda \leq 1$ and that the function $-z \log z$ has maximum 1/e.\end{proof}

In particular, for any subprobability distribution $Q$, $H(Q) \leq \log |X| + 1/e$. A cleaner inequality holds if $|X| \geq 3$.

\begin{lemma}
Let $Q$ be a subprobability measure on a countable set $X$. If $|X| \geq 3$, then

$$H(Q) \leq \log |X|.$$\end{lemma}

\begin{proof}
Consider the expression in Lemma 38, first bounding $H(\hat{Q})$ by $\log |X|$. The function $z \mapsto z \log |X| - z \log z$ (with domain $[0, 1]$) is maximized at $e^{\log |X| - 1}$ and 1. When $e^{\log |X| - 1} \geq 1$, then the function is bounded by $\log |X|$ which is no greater than $\log |X|$. This case applies when $|X| \geq e$. Otherwise, the function’s maximum value is $|X|/e$. Thus, the proposed inequality does not hold for sets of size 1 or 2.\end{proof}

\begin{proof}[Proof of Lemma 14] Let $q$ denote the pmf of $\hat{\theta}$.

$$H(\hat{\theta}) = \sum_{\theta \in \Theta \cap B} q(\theta) \log \frac{1}{q(\theta)}.$$

We will bound two parts of the summation separately: outside a ball centered at $\theta^*$ and then inside that ball.

The function $z \mapsto z \log(1/z)$ increases as $z$ goes from 0 to 1/e. If $|\theta - \theta^*| \geq 1/\sqrt{C}$, then the term of the entropy summation can only be increased by substituting the exponential probability bound

$$q(\theta) \log \frac{1}{q(\theta)} \leq e^{-c|\theta - \theta^*|^2} c |\theta - \theta^*|^2.$$\end{proof}

Thus, outside the ball $B := B(\theta^*, R / e^{-1/2} / \sqrt{C})$, we can bound the summation by an integral using Lemma 28 then compare the integral with a Gamma pdf.

$$\sum_{\theta \in \Theta \cap \overline{B}} q(\theta) \log \frac{1}{q(\theta)} \leq \sum_{\theta \in \Theta \cap \overline{B}} e^{-c|\theta - \theta^*|^2} c |\theta - \theta^*|^2 \leq \int_{\overline{B}} e^{-c|\theta - \theta^*|^2} c |\theta - \theta^*|^2 d\theta.$$

Next, we need to bound the terms coming from grid-points inside $B$. $Q$ restricted to this subset of grid-points can be considered a subprobability measure. Because the ball has radius at least 3e, it contains enough grid-points for Lemma 40 to apply; thus the entropy of this subprobability is bounded by the log-cardinality of the ball’s grid-points. The number of grid-points in the hypercube circumscribing the ball is no more than $(1 + 2(R_{4\sqrt{C} / \sqrt{C}} + 3e)^d)$.

$$\sum_{\theta \in \Theta \cap B} q(\theta) \log \frac{1}{q(\theta)} \leq \log \left(1 + \frac{2(R / e^{-1/2} / \sqrt{C} + 3e)}{e} \right)^d.$$\end{proof}

\section*{Acknowledgment}

The authors would like to thank their advisor, Andrew R. Barron, for his guidance in preparing this document, especially for driving them toward the most important aspects and consequences of the ideas therein.
The authors are also grateful to the anonymous reviewers whose detailed feedback led to dramatic improvements to this paper. Most importantly, they brought [13] to the authors’ attention, which inspired them to write Section II-B.

REFERENCES

[1] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.
[2] A. R. Barron and T. M. Cover, “Minimum complexity density estimation,” *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1034–1054, Jul. 1991.
[3] P. Grünwald and J. Langford, “Suboptimal behavior of Bayes and MDL in classification under misspecification,” *Mach. Learn.*, vol. 66, nos. 2–3, pp. 119–149, 2007.
[4] A. Rényi, “On measures of entropy and information,” in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1961, pp. 547–561.
[5] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bull. Calcutta Math. Soc.*, vol. 35, no. 1, pp. 99–109, 1943.
[6] T. van Erven and P. Harremoes, “Rényi divergence and Kullback–Leibler divergence,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.
[7] A. R. Barron, C. Huang, J. Li, and X. Luo, “The MDL principle, penalized likelihoods, and statistical risk,” in *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, I. Tabus, Ed. Tampere, Finland: Tampere Univ. Press, 2008.
[8] S. Chatterjee and A. Barron, “Information theoretic validity of penalized likelihood,” in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2014, pp. 3027–3031.
[9] X. Luo, “Penalized likelihoods: Fast algorithms and risk bounds,” Ph.D. dissertation, Dept. Statist. Data Sci., Yale Univ., New Haven, CT, USA, 2009.
[10] S. Chatterjee, “Adaptation in estimation and annealing,” Ph.D. dissertation, Dept. Statist. Data Sci., Yale Univ., New Haven, CT, USA, 2014.

W. D. Brinda received a B.S. in Neuroscience (cum laude) from Tulane University in 2007, an M.S. in Applied Statistics from Indiana University in 2011, and a Ph.D. in Statistics and Data Science from Yale University in 2017. His research interests include model selection, information theory, and causality.

Jason M. Klusowski (S’16) received a B.S. (hons.) in Statistics & Mathematics from the University of Manitoba in 2013 and an M.A. in Statistics from Yale University in 2017. He is currently working towards the Ph.D. degree in the Department of Statistics and Data Science at Yale University. His research interests are in high-dimensional density and function estimation, and network sampling.