MDL, Penalized Likelihood, and Statistical Risk

Andrew R. Barron
Statistics Department
Yale University
New Haven, CT 06520-8290
Andrew.Barron@yale.edu

Cong Huang
Statistics Dept.
Yale University
New Haven, CT 06520
Cong.Huang@aya.yale.edu

Johnathan Q. Li
Radar Networks, Inc.
410 Townsend St.
San Francisco, CA 94107
qiang.li@aya.yale.edu

Xi Luo
Statistics Dept.
Yale University
New Haven, CT 06520
Xi.Luo@yale.edu

***Abstract*— We determine, for both countable and uncountable collections of functions, information-theoretic conditions on a penalty pen(f) such that the optimizer $\hat{f}$ of the penalized log likelihood criterion $\log 1/$**likelihood**$(f)+$**pen**(f) has risk not more than the index of resolvability corresponding to the accuracy of the optimizer of the expected value of the criterion. If $\mathcal{F}$ is the linear span of a dictionary of functions, traditional description-length penalties are based on the number of non-zero terms (the ℓ_0 norm of the coefficients). We specialize our general conclusions to show the ℓ_1 norm of the coefficients times a suitable multiplier λ is also an information-theoretically valid penalty.**

I. Introduction

From work in information theory and statistics, there are connections between high-quality data compression and accurate statistical estimation. The original Shannon [63] code construction and the condition of Kraft characterizing valid codelengths show the correspondence between probability distributions $p(data)$ for data and optimal variable-length binary codes of length essentially $\log_2 1/p(data)$ bits (see, e.g., [28]). The development of universal data compression and, in particular, the minimum description-length (MDL) principle has built this correspondence further for the case of distributions $p_f(data)$ that depend on an unknown function f believed to belong to a family $\mathcal{F}$ which may be given parametrically (see [14] or [36] and work cited therein). The function f may provide a density or log-density function, or, in the case that the data consists of pairs of inputs X and outputs Y, the function $f(x)$ may refer to a regression function, classification function, Poisson intensity function, etc. that captures an essential aspect of the conditional distribution of Y given X. Starting from a discussion of coding redundancy, we analyze statistical risk of estimation, capturing its relationship to the accuracy of approximation and the level of complexity of functions f in $\mathcal{F}$, to contribute to a general theory of penalized likelihood.

This 2008 Information Theory Workshop plenary presentation is abridged from the Festschrift for Jorma Rissanen [13].

Ideal procedures adapt to the complexity revealed by the data. Results for mixture-based and prediction-based procedures are discussed and new results are presented for procedures that optimize penalized likelihood. Penalties $pen(f)$ are typically related to parameter dimension or to function irregularity. We develop means to determine when such penalties capture information-theoretic complexity to provide for quality compression and accurate function estimation.

An index of resolvability captures the performance of these procedures. It upper bounds the statistical risk as does a related expression involving an expected redundancy of data compression. These resolvability and redundancy bounds on risk have been developed for penalized likelihood restricted to a countable set of functions which discretizes $\mathcal{F}$, with complexity penalty $pen(f) = L(f)$ equal to an information-theoretic codelength for f, see [11], [4], [46], [43], [44],[36]. The present paper gives a simple and natural method to extend the previous information-theoretic bounds for penalized likelihood from the countable to uncountable $\mathcal{F}$ case.

Early advocacy of penalized likelihood estimation with penalty on the roughness of the density is in [35], [31], [65],[71]. Risk results for quadratic penalties in Hilbert space settings are developed in [29] based on functional analysis tools. Empirical process techniques built around metric entropy calculations yield rate results for penalties designed for a wide variety of function classes in [64]. Related theory for constrained maximum likelihood in nonparametric settings is in [53] and for minimum contrast estimators, sieves, and penalties is in [16], [17], [9].

The use of ℓ_1 penalization of log-likelihoods is a currently popular approach, see [54]. The penalty is applied to coefficients in linear models for f, coinciding with a generalized linear model $p_f(\underline{u})$ for the data, where the terms of the linear model are members of a dictionary of candidates. For special cases, see [42], [1], [34],[77]. That work has focussed on algorithmic development related to Lasso [70], basis pursuit [23], [24], LARS [32], coordinate algorithms [33] and relaxed greedy algorithms [40], [6], [48], [22], [78], [10]. A new algorithmic result is established at the end of this paper.

Recent work analyzes risk of ℓ_1 penalized procedures. Some of it, requiring restrictions on the correlation of dictionary members, addresses whether the procedure performs as well as a subset selection rule, as in [18], [19], [20], [50]. For general dictionaries without correlation conditions, we ask whether an ℓ_1 penalized criterion performs as well as the best tradeoff between approximation error and ℓ_1 norm of coefficients. This is examined for ℓ_1 penalized least squares in the manuscripts [80] and [39] and for ℓ_1 penalized likelihood in the present paper. Penalized likelihood risk bounds should capture the tradeoff of Kullback-Leibler approximation error and penalty. This motivates demonstration that the ℓ_1 penalty satisfies the information-theoretic requirements for the results we seek.

Extending information-theoretic risk results to uncountable families $\mathcal{F}$, the main tool developed in Section 3 is the notion of a variable-complexity cover to allow for variable penalty levels. Distortion is based on discrepancies between log-likelihood and its theoretical analog rather than based on the metrics of traditional metric entropy. In brief, a valid penalty $pen(f)$ is one for which for each f in $\mathcal{F}$ there is a representor in the cover for which $pen(f)$ is not less than its complexity plus distortion.

Often $\mathcal{F}$ is arranged as a union of families $\mathcal{F}_m$ of functions of similar characteristics, e.g., parametric families $\mathcal{F}_m = \{f_{\theta,m} : \theta \in R^{d_m}\}$ of given parameter dimension d_m. Consider linear combinations of a dictionary $\mathcal{H}$ of functions. Such $f_\theta(x) = \sum_{h\in\mathcal{H}} \theta_h h(x)$ are specified by the coefficients $\theta = (\theta_h : h \in \mathcal{H})$. The set of linear combinations $\mathcal{F}$ is the union of models $\mathcal{F}_m$ for subsets m of $\mathcal{H}$ in which the $f_{\theta,m}(x) = \sum_{h\in m} \theta_h h(x)$. These families have dimension $d_m = \text{card}(m)$ when the functions in m are linearly independent.

The data may come from a general sample space. It is traditional to think of finite length strings $\underline{U} = \underline{U}_n = (U_1, U_2, \ldots, U_n)$, consisting of a sequence of outcomes $X_1, X_2, \ldots, X_n$ or outcome pairs $(X_i, Y_i)_{i=1}^n$. We write $\underline{\mathcal{U}}$ for the sample space and $P_{\underline{U}|f}$ (or more briefly P_f) for the distributions on $\underline{U}$. Likewise $E_{\underline{U}|f}$ or more briefly E_f denotes the expected value. When being explicit about sample size, we index by n, as in $P_{\underline{U}_n|f}$ or $P_f^{(n)}$.

For lossless data compression, $\underline{\mathcal{U}}$ is countable, $p_f(\underline{u})$ is the probability mass function, and $q(\underline{u})$, satisfying Kraft's inequality $\sum_{\underline{u}\in\underline{\mathcal{U}}} q(\underline{u}) \le 1$, is a coding distribution with codelengths $\log_2 1/q(\underline{u})$ in bits. The pointwise coding redundancy is $\log 1/q(\underline{u}) - \log 1/p_f(\underline{u})$, the difference between the actual codelength and the codelength we would have had if f were given. Following past MDL work, we allow continuous sample spaces and density functions relative to a given reference measure, yet, we refer to the log density ratio as a redundancy. See [2] for a limiting code redundancy interpretation of the absolutely continuous case involving fine discretizations.

Thus our setting is that the distributions $P_{\underline{U}|f}$ have density functions $p(\underline{u}|f) = p_f(\underline{u})$, relative to a fixed reference measure, which provides the likelihood function of f at data $\underline{U}$. The reference measure is assumed to be a product of measures on the individual spaces. For the special case of i.i.d. modeling, there is a space $\mathcal{U}$ for the individual outcomes with distributions $P_f^{(1)} = P_f$ and then $\underline{\mathcal{U}}$ is taken to be the product space $\mathcal{U}^n$ and $P_{\underline{U}_n|f} = P_f^n$ is taken to be the product measure with joint density $p_f(\underline{u}_n) = \prod_{i=1}^n p_f(u_i)$.

The object of universal data compression and universal modeling is the choice of a distribution $q(\underline{u}), \underline{u} \in \underline{\mathcal{U}}$, such that the redundancy $\log 1/q(\underline{u}) - \log 1/p_f(\underline{u})$ is kept not larger than need be (measured either pointwise or in expectation over $\underline{u}$ and either on the average or in worst case over f) for functions in each class of interest.

As discussed in [61], [14], [36], minimum description-length methods choose q in one of several interrelated ways: by Bayes *mixtures*, by *predictive models*, by *two-stage* codes, or by *normalized maximum-likelihood* codes. We discuss some aspects of these with an eye toward redundancy and resolvability bounds on risk.

Our treatment of penalized likelihood gives general information-theoretic penalty formulation in sections 2 and 3, with risk bounds given for squared Hellinger and related distances, and then application to ℓ_1 penalties in sections 4 and 5. For information-theoretic context, we first review redundancy and resolvability bounds for mixture models and their implications for the risk of predictive estimators, with the stronger Kullback-Leiber loss. This material shows that tools are in place for dealing with uncountable families by mixture models and their associated predictive interpretations. Then penalized likelihood is studied because of its familiarity and comparative ease of computation.

A. Mixture models

These models for $\underline{U}$ use a prior distribution w on $\mathcal{F}$ leading to a mixture density $q(\underline{u}) = q_w(\underline{u}) = \int p_f(\underline{u}) w(df)$. For instance with $\mathcal{F}$ a union of families $\mathcal{F}_m$, the prior may be built from a probability $w(m)$ and a distribution on $\mathcal{F}_m$ for each m. If $\mathcal{F}_m$ is given parametrically the prior may originate on the parameters yielding $q(\underline{u}|m) = q_{w_m}(\underline{u}) = \int p_{f_{\theta,m}}(\underline{u}) w(d\theta|m)$, and an overall mixture $q(\underline{u}) = \sum_m w(m) q(\underline{u}|m)$. A mixture distribution has average case optimal redundancy, averaging over $\underline{u}$ according to $p_f(\underline{u})$ and averaging over functions according to the prior.

Here we discuss tools for redundancy and resolvability bounds for mixtures and the bounds they yield on risk.

The expected redundancy of the mixture q_{w_m} takes the form $E_f[\log p(\underline{U}|f)/q_{w_m}(\underline{U})]$ which is the Kullback-Leibler divergence between the mixture and the target. In each family, the minimax procedure with smallest worst case expected redundancy corresponds to a prior w_m yielding the largest minimum average redundancy. Suitable approximate forms for the optimal w_m, called the least favorable prior or capacity achieving prior, are available in an asymptotic setting [26], [27] and [73]. For smooth parametric families with a Fisher information $I(\theta|m)$, an asymptotically optimal prior is proportional to $|I(\theta|m)|^{1/2}$ and the resulting redundancy behaves asymptotically like $\frac{d_m}{2}\log n$ plus a specified constant determined by the logarithm of the integral of this root Fisher information. There are finite sample bounds of the same form but with slightly larger constants, from examination of the resolvability of mixtures we come to shortly.

Pointwise minimax redundancy theory identifies the smallest constant penalty to add to $\log 1/p(\underline{U}|\hat{f}_m)$, where $\hat{f}_m = f_{\hat{\theta},m}$ is the maximizer of the likelihood, such that the result retains a data-compression interpretation. This problem is studied in an asymptotic setting in [62], [14], [66], [67], [68], [69], [74], showing that the same value $\frac{d_m}{2}\log\frac{n}{2\pi} + \log\int |I(\theta|m)|^{1/2} d\theta$ characterizes this smallest constant penalty asymptotically. Such theory provides additional data compression justification for a penalty with main term proportional to the dimension d_m.

Choosing $w(m)$ is also addressed from an information-theoretic standpoint. The MDL parameter cost primarily de-

termined by the dimension d_m. Thinking of $\log 1/w(m)$ as a codelength, we set it using the log-cardinality of models of the same dimension (one can not do much better than that for most such models). For models which correspond to subsets m of size d chosen out of p candidate terms in a dictionary, the $\log 1/w(m)$ can be set to be $\log \binom{p}{d}$, plus a small additional description length for the dimension d. Often p is large compared to the sample size n, so this $\log 1/w(m)$ of order $d_m \log p/d_m$ substantially adds to $\frac{d_m}{2} \log n$ in the total description length.

B. Index of resolvability of mixtures

For mixtures an expected redundancy bound is developed in [8] and shown to bound an associated statistical risk. The Kullback divergence $D(P_{\underline{U}}||Q_{\underline{U}}) = E \log p(\underline{U})/q(\underline{U})$ is the total expected redundancy for data $\underline{U}$ described using $q(\underline{u})$ but governed by a density $p(\underline{u})$. Suppose this density has the form $p_{f^*}(\underline{u})$. A tool for examining redundancy is $D_n(f^*, f) = D(P_{\underline{U}|f^*}||P_{\underline{U}|f})$ measuring how well f approximates f^*. In the i.i.d. modeling case this divergence takes the form $D_n(f^*, f) = nD(f^*, f)$ where $D(f^*, f)$ is the divergence between the single observation distributions $D(P_{f^*}||P_f)$.

The resolvability bound on expected redundancy of mixtures is as follows. Let the distribution $Q_{\underline{U}}$ be a general mixture with density $q(\underline{u}) = \int p(\underline{u}|f)W(df)$ formed from a prior W. Let B be any measurable subset of functions in $\mathcal{F}$. Then by restriction of the integral to B followed by Jensen's inequality, the redundancy of the mixture Q is bounded by the sum of the maximum divergence of distributions in B from the target f^* and the log reciprocal prior probability of B, and thus,

$$D(P_{\underline{U}|f^*}||Q_{\underline{U}}) \;\le\; \min_B \left\{ \max_{f\in B} D_n(f^*, f) + \log \frac{1}{W(B)} \right\}.$$

In i.i.d. modeling, we divide by n to obtain a redundancy rate bound. When the $B = \{f\}$ are singleton sets, the right side is the same as the index of resolvability given in [11], used there for two-stage codes, as will be discussed further. The optimal sets for the resolvability bound for mixture codes take the form of Kullback balls $B_{r,f^*} = \{f : D(f^*, f) \le r^2\}$, yielding

$$(1/n)D(P_{\underline{U}_n|f^*}||Q_{\underline{U}_n}) \;\le\; \min_{r\ge 0} \left\{ r^2 + \frac{\log 1/W(B_{r,f^*})}{n} \right\}.$$

As in [8] with suitable choices of prior, it provides the usual $(d_m/2)(\log n)/n$ behavior of redundancy rate in finite-dimensional families, and rates of the form $(1/n)^\rho$ for positive $\rho < 1$ for various infinite-dimensional families of functions. Similar characterization arises from a stronger Bayes resolvability bound $D(P_{\underline{U}|f^*}||Q_{\underline{U}}) \le -\log \int e^{-D_n(f^*,f)} W(df)$ as developed in [3], [8], [37], and [79].

C. Implications for predictive risk

For predictive models of a sequence $\underline{U}_N = (U_n)_{n=1}^N$ the joint distribution $q(\underline{U}_N)$ (for universal modeling or coding) is formed by gluing together predictive distributions $q(u_n|\underline{u}_{n-1})$, that is, by multiplying together these conditional densities for $n = 1, 2, \ldots, N$. In the i.i.d. modeling case, given f, the density for U_n given the past is $p(u_n|f)$. Predictive distributions are often created in the form $p(u_n|\hat{f}_{n-1})$ by plugging in an estimate $\hat{f}_{n-1}$ based on the past $\underline{u}_{n-1}$. Predictive distribution need not be restricted to be of plug-in form. Indeed, the prior average of a predictive redundancy is optimized by a Bayes predictive density $q(u_n|\underline{u}_{n-1})$. The predictive redundancy is $E_f D(P_{U_n|f}||Q_{U_n|\underline{U}_{n-1}})$, which is the Kullback risk of the predictive density, based on a sample of size $n-1$. The model built by multiplying the Bayes predictive densities together is the mixture $q_w(\underline{u})$. Correspondingly, by the chain rule, the total codelength and its redundancy yield the same values, respectively, as the mixture codelength and redundancy discussed in (1) above. Indeed, the total redundancy of the predictive model is

$$D(P_{\underline{U}_N|f}||Q_{\underline{U}_N}) \;=\; \sum_{n=1}^N E_f D(P_{U_n|f}||Q_{U_n|\underline{U}_{n-1}}),$$

the cumulative Kullback risk. Dividing by N we see that the Cesàro average of the risks of the predictive distributions is bounded by the index of resolvability discussed above.

This chain rule property has been put to use for related conclusions. For example, it is the basis of the analysis of negligibility of superefficiency in [12] and it plays a critical role for non-finite dimensional families in identifying the minimax rates of estimation in [76] and [38].

D. Two-stage codes

We turn our attention to models based on *two-stage* codes. We recall some previous results here, and give in the next sections some simple generalizations to penalized likelihoods. Two-stage codes were used in the original formulation of the MDL principle [57], [58] and in the analysis of [11]. One works with a countable set $\tilde{\mathcal{F}}$ of possible functions, perhaps obtained by discretization of the underlying family $\mathcal{F}$. A key ingredient in building the total two-stage description length are assignments of complexities $L_n(f)$, for $f \in \mathcal{F}$, satisfying the Kraft inequality $\sum_{f\in\mathcal{F}} 2^{-L_n(f)} \le 1$.

These complexities typically have the form of a codelength for the model class m (of the form $L(m) = \log 1/w(m)$ as discussed above), plus a codelength $L(f|m)$ or $L(\theta|m)$ for the parameters that determine the functions in $\mathcal{F}_m$, which may be discretized to a grid of precision δ for each coordinate, each of which is described using about $\log 1/\delta$ bits. Under, respectively, first or second order smoothness conditions on how the likelihood depends on the parameters, the codelength for the parameters comes out best if the precision δ is of order $\frac{1}{n}$ or $\frac{1}{\sqrt{n}}$, leading to $L(f|m)$ of approximately $d_m \log n$ or $\frac{d_m}{2} \log n$, respectively, for functions in smooth families $\mathcal{F}_m$.

For each function f and data $\underline{U}$, one has a two-stage codelength $L_n(f) + \log 1/p_f(\underline{U})$ corresponding to the bits of description of f followed by the bits of the Shannon code for $\underline{U}$ given f. Then the minimum total two-stage codelength takes the form

$$\min_{f\in\tilde{\mathcal{F}}} \left\{ \log \frac{1}{p_f(\underline{U})} + L_n(f) \right\}.$$

A minimizer $\hat{f}$ is called the minimum complexity estimator for density estimation [11] and it is called the complexity regularization estimator for regression and classification problems [4].

Typical behavior of the minimal two stage codelength is revealed by investigating what happens when the data $\underline{U}_n$ are distributed according to $p_{f^*}(\underline{u}_n)$ for various possible f^*. Eventually exact discovery occurs when f^* is in $\tilde{F}$ as shown in [2], [11], but its complexity, as ultimately revealed by the data, may be too great for full specification of f^* to be a suitable description with moderate sample sizes. It is helpful to have a notion of a surrogate function f_n^* in the list $\tilde{F}$, appropriate to the current sample size n, in place of f^* which is not necessarily in the countable $\tilde{F}$. The appropriateness of such an f_n^* is judged by whether it captures expected compression and estimation properties of the target.

The redundancy rate of the two-stage description is shown in [11] to be not more than the index of resolvability defined by

$$R_n(f^*) \;=\; \min_{f\in\tilde{\mathcal{F}}} \left\{ \frac{1}{n} D(P_{\underline{U}_n|f^*}||P_{\underline{U}_n|f}) + \frac{1}{n} L_n(f) \right\}.$$

For i.i.d. modeling it takes the form

$$R_n(f^*) \;=\; \min_{f\in\tilde{\mathcal{F}}} \left\{ D(f^*, f) + \frac{L_n(f)}{n} \right\},$$

capturing the ideal tradeoff in error of approximation of f^* and the complexity relative to the sample size. The function f_n^* which achieves this minimum is the population counterpart to the sample-based $\hat{f}$. It best resolves the target for the given sample size. Since $\hat{f}$ is the sample-based minimizer, one has an inequality between the pointwise redundancy and a pointwise version of the resolvability

$$\log \frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} + L_n(\hat{f}) \;\;\leq\;\; \log \frac{p_{f^*}(\underline{U})}{p_{f_n^*}(\underline{U})} + L_n(f_n^*).$$

The resolvability bound on the expected redundancy is the result of taking the expectation of this pointwise inequality.

This $R_n(f^*)$ also bounds the statistical risk of $\hat{f}$, as we recall and develop further in Section 2, with a simplified proof and with extension in Section 3 to uncountable $\mathcal{F}$. The heart of our statistical analysis is the demonstration that the loss function we examine is smaller in expectation and stochastically not much more than the pointwise redundancy.

When $\tilde{\mathcal{F}}$ is a union of sets $\tilde{\mathcal{F}}_m$, the complexities may take the form $L(m, f) = L(m) + L(f|m)$ for the description of m followed by the description of f given m. Then there are actually three-stages with minimum total

$$\min_m \min_{f\in\tilde{\mathcal{F}}_m} \left\{ L(m) + L(f|m) + \log \frac{1}{p_f(\underline{U})} \right\}.$$

The associated minimizer $\hat{m}$ provides a model selection in accordance with the MDL principle. Likewise the resolvability takes the form

$$R_n(f^*) = \min_m \min_{f\in\tilde{\mathcal{F}}_m} \left\{ D(f^*, f) + \frac{L(m, f)}{n} \right\}.$$

The ideal model selection is the choice m_n^* achieving this minimum, and the performance of the sample based MDL selection $\hat{m}$ is captured by the resolvability at m_n^*.

Two-stage codes in parametric families are closely related to average-case optimal mixture-codes, and their codelengths achieve similar forms [2]. It is always possible to construct a mixture of shorter codelength than any two-stage code. Thus, when it is computationally feasible, it is preferable to use mixture codes.

Nevertheless, in many estimation settings, it is common to use a penalized likelihood criterion. We address information-theoretic and statistical properties of such procedures.

II. Risk and Resolvability for Countable $\tilde{\mathcal{F}}$

Here we recall risk bounds for penalized likelihood with a countable $\tilde{\mathcal{F}}$. Henceforth we use base e exponentials and logarithms to simplify the mathematics (the units for coding interpretations become nats rather than bits).

For our loss function, we need of another measure of divergence analogous to the Kullback-Leibler divergence. For pairs of probability distributions P and $\tilde{P}$ on a measurable space, the Bhattacharyya, Hellinger, Chernoff, Rényi divergence [15], [30], [25], [56] is given by $d(P, \tilde{P}) = 2\log 1/\int (p(u)\tilde{p}(u))^{1/2}$ where p and $\tilde{p}$, respectively, are the densities of P and $\tilde{P}$ with respect to a reference measure that dominates the distributions and with respect to which the integrals are taken. Writing $D(P||\tilde{P}) = -2E\log(\tilde{p}(U)/p(U))^{1/2}$ and employing Jensen's inequality shows that $D(P||\tilde{P}) \geq d(P, \tilde{P})$.

On a sequence space $\mathcal{U}^n$, if P^n and $\tilde{P}^n$ are n-fold products of the measures P and $\tilde{P}$, then $d(P^n, \tilde{P}^n) = nd(P, \tilde{P})$ and $D(P^n, \tilde{P}^n) = nD(P, \tilde{P})$. Analogous to notation used above, we use $d_n(f^*, f)$ to denote the divergence between the joint distributions $P_{\underline{U}|f^*}$ and $P_{\underline{U}|f}$, and likewise $d(f^*, f)$ to be the divergence between the distributions $P_{U_1|f^*}$ and $P_{U_1|f}$.

This divergence is closely connected to familiar distances such as the L_1 distance between the densities and the Hellinger distance. It upper bounds the square of the L_1 distance and the square of the Hellinger distance with which it is equivalent as explained below. This $d(P, \tilde{P})$, like the squared Hellinger distance, is locally equivalent to one-half the Kullback-Leibler divergence when $\log p(u)/\tilde{p}(u)$ is upper-bounded by a constant. Moreover, it evaluates to familiar quantities in special cases, e.g., for two normals of mean μ and $\tilde{\mu}$ and variance 1, this $d(P, \tilde{P})$ is $\frac{1}{4}(\mu - \tilde{\mu})^2$. The most important reason four our use of this loss function is that it allows clean examination of the risk, without putting any conditions on the density functions $p_f(\underline{u})$.

The integral used in the divergence is called the Hellinger affinity $A(P, \tilde{P}) = \int p^{1/2}\tilde{p}^{1/2}$. It is related to the squared Hellinger distance $H^2(P, \tilde{P}) = \int (p(u)^{1/2} - \tilde{p}(u)^{1/2})^2$ by $A = 1 - \frac{1}{2}H^2$ and hence the divergence $d(P, \tilde{P}) = -2\log A = -2\log(1 - \frac{1}{2}H^2)$ is not less than $H^2(P, \tilde{P})$. We let $A_n(f^*, f)$ denote the Hellinger affinity between the joint distributions $P_{\underline{U}|f^*}$ and $P_{\underline{U}|f}$. Its role in part of our analysis will be as a normalizer, equaling the expectation of $[p_f(\underline{U})/p_{f^*}(\underline{U})]^{1/2}$ for each fixed f.

The following result from Johnathan Li's thesis [46] is a simplification of a conclusion from [11]. It is also presented in [43] and in [36]. We repeat it here because it is a stepping stone for the extensions we give in this paper.

Theorem 2.1: Resolvability bound on risk ([46]). For a countable $\tilde{\mathcal{F}}$, and $\mathcal{L}_n(f) = 2L_n(f)$ satisfying $\sum e^{-L_n(f)} \leq 1$, let $\hat{f}$ be the estimator achieving

$$\min_{f\in\tilde{\mathcal{F}}} \left\{ \log \frac{1}{p_f(\underline{U}_n)} \, + \, \mathcal{L}_{\rm n}({\rm f}) \right\}.$$

Then, for any target function f^* and for all sample sizes, the expected divergence of $\hat{f}$ from f^* is bounded by the index of resolvability

$$Ed_n(f^*, \hat{f}) \, \leq \, \min_{f\in\tilde{\mathcal{F}}} \{ \, D_n(f^*, f) + \mathcal{L}_n(f) \, \}.$$

In particular with i.i.d. modeling, the risk satisfies

$$Ed(f^*, \hat{f}) \, \leq \, \min_{f\in\tilde{\mathcal{F}}} \left\{ D(f^*, f) + \frac{\mathcal{L}_n(f)}{n} \right\}.$$

Proof of Theorem 2.1: We have

$$2 \, \log \frac{1}{A_n(f^*, \hat{f})} \, =$$

$$2 \, \log \left[\frac{(p_{\hat{f}}(\underline{U})/p_{f^*}(\underline{U}))^{1/2} e^{-L(\hat{f})}}{A_n(f^*, \hat{f})} \right] \, + \, \log \frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} + \mathcal{L}_n(\hat{f}).$$

Inside the first part on the right side the ratio is evaluated at $\hat{f}$. We replace it by the sum of such ratios over all $f \in \tilde{\mathcal{F}}$ obtaining the upper bound

$$2 \log \sum_{f\in\tilde{\mathcal{F}}} \left[\frac{(p_f(\underline{U})/p_{f^*}(\underline{U}))^{1/2} e^{-L(f)}}{A_n(f^*, f)} \right] + \log \frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} + \mathcal{L}_n(\hat{f}).$$

Now we take the expected value for $\underline{U}$ distributed according to $P_{\underline{U}|f^*}$. For the expectation of the first part, by Jensen, obtaining a further upper bound, we may bring the expectation inside the log and then bring it also inside the sum. There we note for each fixed f that $E(p_f(\underline{U})/p_{f^*}(\underline{U}))^{1/2} = A_n(P_{f^*}, P_f)$, so there is a cancelation of the ratio. Then all that is left inside the log is $\sum e^{-L(f)}$ which by assumption is not more than 1. Thus the expected value of the first part is bounded by 0. What then remains is the expectation of the pointwise redundancy, which being less than the value at f_n^*, is bounded by the index of resolvability, which completes the proof for the general case. Dividing through by n gives the conclusion for the i.i.d. case.

If $\log p_{f*}(u)/p_f(u) \leq B$ for all u in $\mathcal{U}$, then by [75], Lemma 4, we have $d(f^*, f) \leq D(f^*||f) \leq C_B \, d(f^*, f)$, for a constant C_B given there that is less than $2 + B$. Consequently, we have the following.

Corollary 2.2: If, in the i.i.d. case, the log density ratios are bounded by a constant B, that is, if $|\log p_{f*}(u)/p_f(u)| \leq B$ for all $f \in \tilde{\mathcal{F}}$, then there is a constant $C_B \leq 2 + B$ such that the Kullback risk satisfies

$$ED(f^*, \hat{f}) \leq C_B \min_{f\in\tilde{\mathcal{F}}} \left\{ D(f^*, f) + \frac{\mathcal{L}_n(f)}{n} \right\}.$$

Remarks.

The factor 2 in the penalty is a byproduct of using the Chernoff-Rényi divergence with parameter $1/2$. As in [11], other multipliers may be used, though the best bound there occurs with the factor 2. See [79], Thm. 4.1 or [36], Ch. 15, for analogous bounds for Chernoff-Rényi divergences with parameter λ between 0 and 1.

Refinements in Li's thesis deal with the case that the distribution of the data is not near any of the P_f. He extends the result to bound the distance of the estimate from a reversed information projection onto a convex hull of the P_f.

Applications of resolvability bounds on risk for penalized likelihood are developed in [11], [46], [47], [55],[43], [44], [72]. Corresponding results for complexity penalized least squares are developed in [4] with applications and further extensions in [5], [7], [51], [52], [39].

The proof of Theorem 2.1 here is the same as in Li's Thesis. One slight difference is we have pointed out that the expected redundancy of the two-stage code is also a bound on the risk, as also noted in [36]. It even more closely relates the risk and coding notions.

The above proof compares the loss $d_n(f^*, \hat{f})$ with the pointwise redundancy $r_n = \log p_{f^*}(\underline{U})/p_{\hat{f}}(\underline{U}) + \mathcal{L}_n(\hat{f})$ and shows that the difference has mean bounded by 0. In like manner we obtain a measure of concentration of this difference.

Theorem 2.3: Tightness of the relationship between loss and redundancy: The difference between the loss $d_n(f^*, \hat{f})$ and the pointwise redundancy r_n is stochastically less than an exponential random variable of mean 2.

Proof of Theorem 2.3: As shown in the proof of Theorem 2.1 the difference in question is bounded by

$$2 \log \sum_{f\in\tilde{\mathcal{F}}} \left[\frac{(p_f(\underline{U})/p_{f^*}(\underline{U}))^{1/2} e^{-L(f)}}{A_n(f^*, f)} \right].$$

The probability that this exceeds any positive τ is bounded first by dividing through by 2, then exponentiating and using Markov's inequality, yielding $e^{-\tau/2}$ times an expectation shown in the proof of Theorem 2.1 to be not more than 1. This completes the proof of Theorem 2.3.

Further remarks:

In the i.i.d. case we measure the loss by the individual divergence obtained by dividing through by n. Consequently, in this case the difference between the loss $d(f^*, \hat{f})$ and pointwise redundancy rate is stochastically less than an exponential of mean $2/n$. It is exponentially unlikely (with probability not more than $e^{-n\tau/2}$) to be greater than any positive τ.

The original bound in [11] also proceeded by a tail probability calculation, though noticeably more elaborate than given here. An advantage of that proof is its change of measure from the one at f^* to the one at f_n^*, showing that the behavior when f^* is true can indeed be addressed by the behavior one would have if one thought of the distribution as being governed by the f_n^* which best resolves f^* at the given sample size.

In this section we assumed the space $\tilde{\mathcal{F}}$ of candidate fits is countable. From statistics and engineering standpoints, it

is awkward to force a user of this theory to construct a discretization of his space of functions in order to use our result. We overcome this difficulty in the next section.

III. Risk and Resolvability for uncountable $\mathcal{F}$

We come to the main new contributions of the paper. We consider estimators $\hat{f}$ that maximize $p_f(\underline{U})e^{-pen(f)}$ or, equivalently, that achieve the following minimum:

$$\min_{f\in\mathcal{F}} \left\{ \log \frac{1}{p_f(\underline{U})} + pen(f) \right\}.$$

Since the log ratio separates, for any target p_*, this sample minimization is equivalent to the following,

$$\min_{f\in\mathcal{F}} \left\{ \log \frac{p_*(\underline{U})}{p_f(\underline{U})} + pen(f) \right\}.$$

We want to know for proposed penalties $pen(f), f \in \mathcal{F}$, when it will be the case that $\hat{f}$ has risk controlled by the population-based counterpart:

$$\min_{f\in\mathcal{F}} \left\{ E \log \frac{p_*(\underline{U})}{p_f(\underline{U})} + pen(f) \right\},$$

where the expectation is with respect to $p_*(\underline{U})$. One may specialize to $p_* = p_{f^*}$ in the family. In general, it need not be a member of the family $\{p_f : f \in \mathcal{F}\}$, though such a bound is only useful when the target is approximated by such densities.

There are two related aspects to the question of whether such a bound holds. One concerns whether the optimal sample quantities suitably mirror the population quantities even for such possibly larger $\mathcal{F}$, and the other is to capture what is essential for the penalty.

A quantity that may be considered in examining this matter is the discrepancy between sample and population values, defined by,

$$\log \frac{p_*(\underline{U})}{p_f(\underline{U})} - E \log \frac{p_*(\underline{U})}{p_f(\underline{U})}.$$

Perhaps it is ideally centered, yielding mean 0 when defined in this way, with subtraction of the Kullback divergence. However, control of this discrepancy to produce bounds on Kullback risk (using, e.g., Bernstein-type bounds), would require conditions relating the variance of the log density ratios to the expected log ratio. Though such development is possible, e.g., if the log densities ratios are bounded, it is not as clean an approach as what follows.

Instead, we use the following discrepancy which is of similar spirit to the above and easier to control in the desired manner,

$$\log \frac{p_*(\underline{U})}{p_f(\underline{U})} - 2 \log \frac{1}{E(p_f(\underline{U})/p_*(\underline{U}))^{1/2}}.$$

This discrepancy does not subtract off as large a value, so it is not mean centered, but that is not an obstacle if we are willing to use the Hellinger risk, as the control needed of the discrepancy is one-sided in character. No moment condition is needed working with the expected square-roots that give the Hellinger affinities, which are automatically bounded by 1.

In Theorem 2.1, the penalty $\mathcal{L}(f)$ is used to show that if added to the discrepancy, then uniformly for f in the countable $\tilde{\mathcal{F}}$ (i.e., even with a data-based $\hat{f}$ in place of a fixed f) we have that the expectation of the penalized discrepancy is positive.

This leads us to consider, in the uncountable case, penalties which exhibit a similar discrepancy control. We say that a collection $\mathcal{F}$ with a penalty $pen(f)$ for $f \in \mathcal{F}$ has a *variable–complexity variable–discrepancy cover* suitable for p_* if there exists a countable $\tilde{\mathcal{F}}$ and $\mathcal{L}(\tilde{f}) = 2L(\tilde{f})$ satisfying $\sum_{\tilde{f}} e^{-L(\tilde{f})} \leq 1$, such that the following condition $(*)$ holds for all $\underline{U}$:

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}} \left\{ \log \frac{p_*(\underline{U})}{p_{\tilde{f}}(\underline{U})} - 2 \log \frac{1}{E(p_{\tilde{f}}(\underline{U})/p_*(\underline{U}))^{1/2}} \; + \; \mathcal{L}(\tilde{f}) \right\} \leq$$

$$\inf_{f\in\mathcal{F}} \left\{ \log \frac{p_*(\underline{U})}{p_f(\underline{U})} - 2 \log \frac{1}{E(p_f(\underline{U})/p_*(\underline{U}))^{1/2}} \; + \; pen(f) \right\}. \qquad (*)$$

This condition captures the aim that the penalty in the uncountable case mirrors an information-theoretically valid penalty in the countable case. The above condition will give what we want because the minimum over the countable $\tilde{f}$ is shown to have non-negative expectation and so the minimum over all f in $\mathcal{F}$ will also.

Equivalent to condition $(*)$ is that there be a $\tilde{\mathcal{F}}$ and $L(\tilde{f})$ with $\sum e^{-L(\tilde{f})} \leq 1$ such that for f in $\mathcal{F}$ the penalty satisfies

$$pen(f) \; \geq$$

$$\min_{\tilde{f}\in\tilde{\mathcal{F}}} \left\{ \log \frac{p_f(\underline{U})}{p_{\tilde{f}}(\underline{U})} - 2 \log \frac{E(p_f(\underline{U})/p_*(\underline{U}))^{1/2}}{E(p_{\tilde{f}}(\underline{U})/p_*(\underline{U}))^{1/2}} \; + \; 2L(\tilde{f}) \right\}.$$

That is, the penalty exceeds the minimum complexity plus discrepancy difference. The log ratios separate so the minimizing $\tilde{f}$ does not depend on f. Nevertheless, the following characterization $(**)$ is convenient. For each f in $\mathcal{F}$ there is an associated representor $\tilde{f}$ in $\tilde{\mathcal{F}}$ for which

$$pen(f) \; \geq \; \{ \log \frac{p_f(\underline{U})}{p_{\tilde{f}}(\underline{U})} - 2 \log \frac{E(p_f(\underline{U})/p_*(\underline{U}))^{1/2}}{E(p_{\tilde{f}}(\underline{U})/p_*(\underline{U}))^{1/2}} + 2L(\tilde{f}) \}. \qquad (**)$$

The idea is that if $\tilde{f}$ is close to f then the discrepancy difference is small. Then the complexity of such $\tilde{f}$ along with the discrepancy difference assesses whether a penalty $pen(f)$ is suitable. The minimizer in $\tilde{\mathcal{F}}$ depends on the data and accordingly we allow the representor $\tilde{f}$ of f to also have such dependence. With this freedom, in cases of interest, the variable complexity cover condition indeed holds for all $\underline{U}$, though it would suffice for our purposes that (*) hold in expectation.

One strategy to verify the condition would be to create a metric-based cover of $\mathcal{F}$ with a metric chosen such that for each f and its representor $\tilde{f}$ one has $|\log p_f(\underline{U})/p_{\tilde{f}}(\underline{U})|$ plus the difference in the divergences arranged if possible to be less than a distance between f and $\tilde{f}$. Some examples where this can be done are in [11]. Such covers give a metric entropy flavor, though the $L(\tilde{f})$ provides variable complexity rather than the fixed log-cardinality of metric entropy. The present

theory and applications show such covering by metric balls is not an essential ingredient.

Condition $(**)$ specifies that there be a cover with variable distortion plus complexity rather than a fixed distance and fixed cardinality. This is analogous to the distortion plus rate tradeoff in Shannon's rate-distortion theory. In our treatment, the distortion is the discrepancy difference (which does not need to be a metric), the codebook is the cover $\tilde{\mathcal{F}}$, the codelengths are the complexities $L(\tilde{f})$. Valid penalties $pen(f)$ exceed the minimal sum of distortion plus complexity.

Our main theorem, generalizing Theorem 2.1 to the case of uncountable $\mathcal{F}$, is the following.

Theorem 3.1: Consider $\mathcal{F}$ and $pen_n(f)$ satisfying the discrepancy plus complexity requirement $(*)$ and the estimator $\hat{f}$ achieving the optimum penalized likelihood

$$\min_{f\in\mathcal{F}} \left\{ \log \frac{1}{p_f(\underline{U})} + pen_n(f) \right\}.$$

If the data $\underline{U}$ are distributed according to $P_{\underline{U}|f^*}$, then

$$Ed_n(f^*,\hat{f}) \;\le\; \min_{f\in\mathcal{F}} \left\{ E\log \frac{p_{f^*}(\underline{U})}{p_f(\underline{U})} \;+\; pen_n(f) \right\}.$$

In particular, for i.i.d. modeling,

$$Ed(f^*,\hat{f}) \;\le\; \min_{f\in\mathcal{F}} \left\{ D(f^*,f) + \frac{pen_n(f)}{n} \right\}.$$

Proof of Theorem 3.1. From the characterization $(**)$, at $f=\hat{f}$ in $\mathcal{F}$ there is an associated $\tilde{f}$ in $\tilde{F}$ for which

$$2\log \frac{1}{A_n(P_{f^*},P_{\hat{f}})} \;\le$$

$$2\log\left[\frac{(p_{\tilde{f}}(\underline{U})/p_{f^*}(\underline{U}))^{1/2}e^{-L(\tilde{f})}}{A_n(P_{f^*},P_{\tilde{f}})}\right] + \left[\log\frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} + pen(\hat{f})\right].$$

The first part of the right side has expectation not more than 0 by the same analysis as in Theorem 2.1 (replacing the ratio inside the log, which is there evaluated at a random $\tilde{f}$, by its sum over all of $\tilde{\mathcal{F}}$ and bringing the expectation inside the log by Jensen's inequality). The expectation of the second part is an expected minimum which is bounded by the minimum expectation. This completes the proof.

In like manner we have the following.

Corollary 3.2: For $\mathcal{F}$ and $pen_n(f)$ satisfying the discrepancy plus complexity requirement, the difference between the loss $d_n(f^*,\hat{f})$ and the pointwise redundancy $r_n = \log p_{f^*}(\underline{U})/p_{\hat{f}}(\underline{U}) + pen_n(\hat{f})$ is stochastically less than an exponential random variable of mean 2.

Proof of Corollary 3.2. An interpretation of this assertion is that at a particular $f=\hat{f}$ the penalized discrepancy $\log p_{f^*}(\underline{U})/p_f(\underline{U}) - 2\log 1/A_n(f^*,f) + pen_n(f)$ is stochastically greater than $-Z$ where Z is an exponential random variable of mean 2. The requirement on the penalty enforces that uniformly in $\mathcal{F}$ this penalized discrepancy exceeds a minimum complexity penalized discrepancy from the countable class case, which as in the proof of Theorem 2.2 is already seen to be stochastically greater than such a random variable. This completes the proof.

Remark: Consider the case that f models the log density function of independent random variables $X_1,\ldots,X_n$, in the sense that for some reference density $p_0(x)$ we have

$$p_f(x) = \frac{p_0(x)\, e^{\,f(x)}}{c_f}$$

where c_f is the normalizing constant. Examining the difference in discrepancies at f and a representing $\tilde{f}$ we see that both $p_0(x)$ and c_f cancel out. What remains for our penalty requirement is that for each f in $\mathcal{F}$ there is a $\tilde{f}$ in a countable $\tilde{\mathcal{F}}$ with complexities $L(\tilde{f})$ for which

$$pen(f) \;\ge$$

$$2L(\tilde{f}) + \sum_{i=1}^{n}(f(X_i)-\tilde{f}(X_i)) + 2n\log E\exp\{\tfrac{1}{2}(\tilde{f}(X)-f(X))\}$$

where the expectation is with respect to a distribution for X constructed to have density which is the normalized pointwise affinity $p_a(x) = [p_{f^*}(x)p_f(x)]^{1/2}/A(f^*,f)$.

In the final section we illustrate how to demonstrate the existence of such representors $\tilde{f}$ using an ℓ_1 penalty on coefficients in representation of f in the linear span of a dictionary of candidate basis functions.

IV. Information-theoretic validity of ℓ_1 penalty

Let $\mathcal{F}$ be the linear span of a dictionary $\mathcal{H}$ of functions. Thus any f in $\mathcal{F}$ is of the form $f(x) = f_\theta(x) = \sum_h \theta_h h(x)$ where the coefficients are denoted $\theta = (\theta_h : h\in\mathcal{H})$. We assume that the functions in the dictionary are bounded. We want to show that a weighted ℓ_1 norm of the coefficients $||\theta||_1 = \sum_h |\theta_h| a_h$ can be used to formulate a valid penalty. Here we use the weights $a_h = \|h\|_\infty$. For f in $\mathcal{F}$ we denote $V_f = \min\{\|\theta\|_1 : f_\theta = f\}$. With the definition of V_f further extended to a closure of $\mathcal{F}$, this V_f is called the variation of f with respect to $\mathcal{H}$. We will show that certain multiples of V_f are valid penalties.

The dictionary $\mathcal{H}$ is a finite set of p candidate terms, typically much larger than the sample size. As we shall see, the codelengths of our representors will arise via a variable number of terms times the log cardinality of the dictionary. Accordingly, for sensible risk bounds, it is only the logarithm of p, and not p itself, that we need to be small compared to the sample size n.

A valid penalty will be seen to be a multiple of V_f, by arranging the number of terms in the representor to be proportional to V_f and by showing that a representor with that many terms suitably controls the discrepancy difference. We proceed now to give the specifics.

The countable set $\tilde{\mathcal{F}}$ of representors is taken to be the set of all functions of the form $\tilde{f}(x) = V\frac{1}{K}\sum_{k=1}^{K} h_k(x)/a_{h_k}$ for terms h_k in $\mathcal{H}\cup -\mathcal{H}\cup\{0\}$, where the number of terms K is in $\{1,2,\ldots\}$ and the nonnegative multipliers V will be determined from K in a manner we will specify later. We let

p be the cardinality of $\mathcal{H} \cup -\mathcal{H} \cup \{0\}$, allowing for h or $-h$ or 0 to be a term in $\tilde{f}$ for each h in $\mathcal{H}$.

The main part of the codelength $L(\tilde{f})$ is $K \log p$ nats to describe the choices of $h_1, \ldots, h_K$. The other part is for the description of K and it is negligible in comparison, but to include it simply, we may use a possibly crude codelength for the integer K such as $K \log 2$. Adding these contributions of $K \log 2$ for the description of K and of $K \log p$ for the description of $\tilde{f}$ given K, we have

$$L(\tilde{f}) = K \log(2p).$$

To establish existence of a representor $\tilde{f}$ of f with the desired properties, we put a distribution on choices of $h_1, \ldots, h_K$ in which each is selected independently, where h_k is h with probability $|\theta_h| a_h / V$ (with a sign flip if θ_h is negative). Here $K = K_f = \lceil V_f/\delta \rceil$ is set to equal V_f/δ rounded up to the nearest integer, where $V_f = \sum_h |\theta_h| a_h$, where a small value for δ will be specified later. Moreover, we set $V = K\delta$, which is V_f rounded up to the nearest point in a grid of spacings δ. When V_f is strictly less than V there is leftover an event of probability $1 - V_f/V$ in which h_k is set to 0.

As f varies, so does the complexity of its representors. Yet for any one f, with $K = K_f$, each of the possibilities for the terms h_k produces a possible representor $\tilde{f}$ with the same complexity $K_f \log 2p$.

The key property of our random choice of $\tilde{f}(x)$ representing $f(x)$ is that, for each x, it is a sample average of i.i.d. choices $V h_k(x)/a_{h_k}$. Each of these terms has expectation $f(x)$ and variance $V \sum_h |\theta_h| h^2(x)/a_h - f^2(x)$ not more than V^2.

As the sample average of K such independent terms, $\tilde{f}(x)$ has expectation $f(x)$ and variance $(1/K)$ times the variance given for a single draw. We will also need expectations of exponentials of $\tilde{f}(x)$ which is made possible by the representation of such an exponential of sums as the product of the exponentials of the independent summands.

The existence argument proceeds as follows. The quantity we need to bound to set a valid penalty is the minimum over $\tilde{\mathcal{F}}$ of the complexity-penalized discrepancy difference:

$$2L(\tilde{f}) + \sum_{i=1}^{n} (f(X_i) - \tilde{f}(X_i)) + 2n \log \int p(x) e^{(\tilde{f}(x) - f(x))/2}$$

where $p(x) = p_a(x)$ is a probability density function as specified in the preceding section. The minimizing $\tilde{f}$ gives a value not more than the expectation over random $\tilde{f}$ obtained by the sample average of randomly selected h_k. We condition on the data $X_1, \ldots X_n$. The terms $f(X_i) - \tilde{f}(X_i)$ have expectation 0 so it remains to bound the expectation of the log term. It is less than or equal to the log of the expectation, so we bring that expectation inside the integral. Then at each x we examine the expectation of the exponential of $\frac{1}{2}[\tilde{f}(x) - f(x)]$. By the independence and identical distribution of the K summands that comprise the exponent, the expectation is equal to the Kth power of the expectation of $\exp\{\frac{1}{2K}[V h(x)/a_h - f(x)]\}$ for a randomly drawn h.

We now take advantage of classical bound of Hoeffding, easily verified by using the series expansion of the exponential. If T is a random variable with range bounded by B, then $E \exp\{\frac{1}{K}(T - \mu)\} \leq \exp\{\frac{B^2}{8K^2}\}$.

Let $R(x) = \max_h h(x)/a_h - \min_h h(x)/a_h$ be the range of $h(x)/a_h$ as h varies for the given x, which is uniformly bounded by 2. At x given, $T = \frac{1}{2} V h(x)/a_h$ is a random variable, induced by the random h, having range $\frac{V}{2} R(x)$. Then at the given x, using the Hoeffding inequality gives that the expectation of $\exp\{\frac{1}{2}(\tilde{f}(x) - f(x))\}$ is bounded by $\exp\{\frac{(V R(x))^2}{32K}$ which is not more than $\exp\{\frac{V^2}{8K}\}$.

The expectation of the log of the integral of this exponential is bounded by $\frac{V^2}{8K}$ or equivalently $\frac{1}{8} V \delta$. When multiplied by $2n$, it yields a discrepancy difference bound of

$$\tfrac{1}{4} n V \delta,$$

where V is not more than $V_f + \delta$.

Now twice the complexity plus the discrepancy bound has size $2K \log(2p) + \frac{1}{4} n V_f \delta + \frac{1}{4} n \delta^2$, which, with our choice of $K = \lceil V_f/\delta \rceil$ not more than $V_f/\delta + 1$, shows that a penalty of the form

$$pen_n(f) \;\geq\; \lambda V_f \;+\; C$$

is valid as long as λ is at least $\frac{2}{\delta} \log(2p) + \frac{1}{4} n \delta$ and $C = 2 \log(2p) + \frac{1}{4} n \delta^2$. We set $\delta = (\frac{8 \log 2p}{n})^{1/2}$ as it optimizes the bound on λ producing a critical value λ_n^* equal to $(2n \log 2p)^{1/2}$ and a value of $C = 4 \log(2p)$. The presence of the constant term C in the penalty does not affect the optimization that produces the penalized likelihood estimator, that is, the estimator is the same as if we used a pure ℓ_1 penalty equal to λV_f. Nevertheless, for application of our theory giving risk bounds, the C found here is part of our bound.

We summarize the conclusion with the following Theorem. The setting is as above with the density model $p_f(x)$ with exponent $f(x)$. The estimate is chosen with f in the linear span of the dictionary $\mathcal{H}$. The data are i.i.d. according to $p_{f^*}(x)$.

Theorem 4.1: The ℓ_1 penalized likelihood estimator $\hat{f} = f_{\hat{\theta}}$ achieving

$$\min_{\theta} \left\{ \log \frac{1}{p_{f_\theta}(\underline{X}_n)} + \lambda_n ||\theta||_1 \right\},$$

or, equivalently,

$$\min_{f} \left\{ \log \frac{1}{p_f(\underline{X}_n)} + \lambda_n V_f \right\},$$

has risk $Ed(f^*, \hat{f})$ bounded for every sample size by

$$R_n(f^*) \;\leq\; \inf_{f \in \mathcal{F}} \left\{ D(f^*, f) + \frac{\lambda_n V_f}{n} \right\} \;+\; \frac{4 \log 2p}{n}$$

provided $\frac{\lambda_n}{n} \geq \left[\frac{2 \log(2p)}{n}\right]^{1/2}$.

In particular, if f^* has finite variation V_{f^*} then for all n,

$$Ed(f^*, \hat{f}) \;\leq\; R_n(f^*) \leq \frac{\lambda_n V_{f^*}}{n} + \frac{4 \log 2p}{n}.$$

Note that the last term $\frac{4 \log 2p}{n}$, is typically negligible compared the main term, which is near

$$\left[\frac{2 \log 2p}{n}\right]^{1/2} V_{f^*}.$$

Not only does this result exhibit $[(\log p)/n]^{1/2}$ as the rate of convergence, but also it gives clean finite sample bounds.

Even if V_{f^*} is finite, the best resolvability can occur with simpler functions. In fact, until n is large compared to $V_{f^*}^2 \log p$, the index of resolvability will favor approximating functions f_n^* with smaller variation.

V. Refined resolvability for ℓ_1 penalization

Two directions of refinement of this risk conclusion for ℓ_1 penalized log likelihood are presented briefly. Details of these and other extensions are in the full paper [13].

First, consider infinite dictionaries with a finite metric dimension property. At a suitable precision of order $1/\sqrt{n}$, an L_∞ cover of the dictionary has size about $n^{d/2}$ where d is the metric dimension of the dictionary. Then analogous conclusions obtain with $\log p$ replaced by $(d/2)\log n$, so that if f^* has finite variation with respect to the dictionary then the risk is of order bounded by $[(d\log n)/n]^{1/2}$. Thus the performance of the ℓ_1 penalized log-likelihood estimator is in agreement with what was obtained previously for other estimators in [5], [7], [51], [52], [48], [41], [10]. A noteworthy feature is that unlike standard derivative-based regularity conditions which lead to rates that degrade with dimension, the variation condition with respect to a finite-dimensional dictionary has rate of statistical risk at least as good as the power $1/2$.

Secondly, an improved method of approximation with probabilistic proof originates in Makovoz [49], with a stratified sampling interpretation in [39]. It yields an improvement in which V^2/K is replaced by $\varepsilon_0^2 V^2/(K-K_0)$ where ε_0 is the distance attained by the best covering of the dictionary of size $K_0 < K$. It allows a somewhat smaller λ_n and improved risk bounds for ℓ_1 penalized log-likelihood estimators of order $\left[\frac{d}{n}\log\frac{n}{d}\right]^{\frac{1}{2}+\frac{1}{2d+2}}$, which remains near the rate $1/2$ when d is large. This conclusion is in agreement with what is achieved by other estimators in [76] and it is close to the lower bound on optimal rates given there. Similar implications for classification problems using convex hulls of a dictionary are in [45] and for ℓ_1 penalized least squares in [39].

This completes our story of the risk of penalized log likelihood. Common penalties for functions in uncountable sets $\mathcal{F}$ may be used, such as the ℓ_1 norm of the coefficients of f, which may, at first glance, not look like a complexity penalty. Nevertheless, variable cover arguments show that the ℓ_1 penalty does have the property we require. For suitable multipliers λ, the ℓ_1 penalized discrepancy exceeds the complexity penalized discrepancy, and hence inherits its clean risk properties.

VI. A Note on Computation

Consider a relaxed greedy algorithm in which we successively optimize the ℓ_1 penalized likelihood one term at a time, optimizing choices of α, β and h in the update

$$\hat{f}_k(x) = (1-\alpha)\hat{f}_{k-1}(x) + \beta h(x)$$

for each $k = 1, 2, \ldots$. Our result is that it solves the ℓ_1 penalized likelihood optimization, with a guarantee that after k steps we have a k component mixture within order $1/k$ of the optimum. Indeed, one initializes with $\hat{f}_0(x) = 0$ and $v_0 = 0$. Then for each step k, ones optimizes α, β, and h to provide the the kth term $h_k(x)$. At each iteration one loops through the dictionary trying each $h \in \mathcal{H}$, solving for the best associated scalars $0 \le \alpha \le 1$ and $\beta \in R$, and picks the h that best improves the ℓ_1 penalized log-likelihood, using $v_k = (1-\alpha)v_{k-1} + |\beta|\, a_{h_k}$ as the updated bound on the variation of $\hat{f}_k$. This is a case of what we call an ℓ_1 *penalized greedy pursuit*. This algorithm solves the penalized log-likelihood problem, with an explicit guarantee on how close we are to the optimum after k steps. Indeed, for any given data set $\underline{X}$ and for all $k \ge 1$,

$$\frac{1}{n}\left[\log\frac{1}{p_{\hat{f}_k}(\underline{X})} + \lambda v_k\right] \le \inf_f \left\{ \frac{1}{n}\left[\log\frac{1}{p_f(\underline{X})} + \lambda V_f\right] + \frac{2V_f^2}{k+1} \right\},$$

where the infimum is over functions in the linear span of the dictionary, and the variation corresponds to the weighted ℓ_1 norm $\|\theta\|_1 = \sum_{h\in\mathcal{H}} |\theta_h| a_h$, with a_h set to be not less than $\|h\|_\infty$. This inequality shows that $\hat{f}_k$ has penalized log-likelihood within order $1/k$ of the optimum.

This computation bound for ℓ_1 penalized log-likelihood is developed in the Yale thesis research of one of us, Xi Luo, adapting some ideas from the corresponding algorithmic theory for ℓ_1 penalized least squares from [39]. The proof of this computation bound and the risk analysis given above have aspects in common. So it is insightful to give the proof here.

It is equivalent to show that for each f in the linear span that

$$\frac{1}{n}\left[\log\frac{p_f(\underline{X}_n)}{p_{\hat{f}_k}(\underline{X}_n)} + \lambda(v_k - V_f)\right] \le \frac{2V_f^2}{k+1}.$$

The left side of this desired inequality which we shall call e_k is built from the difference in the criterion values at $\hat{f}_k$ and an arbitrary f. It can be expressed as

$$e_k = \frac{1}{n}\sum_{i=1}^n [f(X_i) - \hat{f}_k(X_i)] + \log \int p_f(x) e^{\hat{f}_k(x) - f(x)} + \lambda[v_k - V_f],$$

where the integral arising from the ratio of the normalizers for $p_{\hat{f}_k}$ and p_f. Without loss of generality, making $\mathcal{H}$ closed under sign change, we restrict to positive β. This e_k is evaluated with $\hat{f}_k(x) = (1-\alpha)\hat{f}_{k-1}(x) + \beta h(x)$ and $v_k = (1-\alpha)v_{k-1} + \beta a_h$, at the optimized α, β and h, so we have that it is as least as good as at an arbitrary h with $\beta = \alpha v/a_h$ where $v = V_f$. Thus for any h we have that e_k is not more than

$$\frac{1}{n}\sum_{i=1}^n [f(X_i) - \bar{\alpha}\hat{f}_{k-1}(X_i) - \alpha h(X_i)/a_h] +$$

$$\log \int p_f(x) e^{[\bar{\alpha}\hat{f}_{k-1}(x) + \alpha v h(x)/a_h - f(x)]} + \bar{\alpha}\lambda[v_{k-1} - v],$$

where $\bar{\alpha} = (1-\alpha)$. Now reinterpret the integral using the expectation of $e^{\alpha[vh(x)/a_h - f(x)]}$ with respect to $p(x) = e^{\bar{\alpha}[f_{k-1}(x)-f(x)]}p_f(x)/c$, where c is its normalizing constant. Accordingly, we add and subtract $\log c = \log \int e^{\bar{\alpha}[f_{k-1}(x)-f(x)]}p_f(x)$ which, by Jensen's inequality using $\bar{\alpha} \leq 1$, is not more than $\bar{\alpha} \log \int e^{[f_{k-1}(x)-f(x)]}p_f(x)$. Recognizing that this last integral is what arises in e_{k-1} and distributing f between the terms with coefficients $\bar{\alpha}$ and α, we obtain that e_k is not more than

$$\bar{\alpha}e_k + \alpha \frac{1}{n}\sum_{i=1}^{n}[f(X_i) - vh(X_i)/a_h] + \log \int e^{\alpha[vh(x)/a_h - f(x)]}p(x).$$

This inequality holds for all h so it holds in expectation with a random selection in which each h is drawn with probability $a_h|\theta_h|/v$ where the θ_h are the coefficients in the representation $f(x) = \sum_{h\in\mathcal{H}} \theta_h h(x)$ with $v = \sum_h |\theta_h| a_h = V_f$. We bring this expectation for random h inside the logarithm, and then inside the integral, obtaining an upper bound by Jensen's inequality. For each x and random h the quantities $[vh(x)/a_h - f(x)]$ have mean zero and have range of length not more than $2v$ since $a_h \geq \|h\|_\infty$. So by Hoeffding's moment generating function bound, the expectation for random h of $e^{\alpha[vh(x)/a_h - f(x)]}$ is not more than $e^{\alpha^2 v^2/2}$. Thus

$$e_k \leq (1-\alpha)e_{k-1} + \alpha^2 V_f^2$$

for all $0 \leq \alpha \leq 1$, and so in particular with $\alpha = 2/(k+1)$. Also $e_0 \leq 2V_f^2$, so by induction

$$e_k \leq \frac{2V_f^2}{k+1},$$

which is the desired result.

This computation bound and its regression counterpart in [39] is related to past relaxed greedy algorithm work (with $\lambda = 0$ in [40], [6], [48], [21], [22], [47], [78], [10]. These previous results control the number of terms k rather than their ℓ_1 norm. The result stated here for ℓ_1 penalized log-likelihood and in [39] for regression, takes the matter a step further to show that with suitable positive λ the greedy pursuit algorithm solves the ℓ_1 penalized problem.

This computation analysis fits with our risk results. In the proof of Theorem 3.1, instead of the exact penalized likelihood estimator $\hat{f}$, substitute its k term greedy fit $\hat{f}_k$, The computation bound shows that this penalized likelihood ratio is not more than its corresponding value at any f, with addition of $2V_f^2/(k+1)$. Accordingly, its risk is not more than

$$Ed(f^*, \hat{f}_k) \leq \min_{f\in\mathcal{F}} \left\{ D(f^*, f) + \frac{\lambda_n V_f}{n} + \frac{2V_f^2}{k+1} \right\} + \frac{C}{n}.$$

The key step in our results is demonstration of approximation, computation, or covering properties, by showing that they hold on the average for certain distributions on the dictionary of possibilities.

REFERENCES

[1] O. Banerjee, L.E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation," *J. Machine Learning Res.*, 2007.

[2] A.R. Barron, *Logically Smooth Density Estimation*, Ph.D. Thesis, Electrical Engineering Dept., Stanford Univ., 1985.

[3] A.R. Barron, "The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions," Univ. Illinois Statistics Dept. Technical Report #7. 1998. Available at www.stat.yale.edu/∼arb4/publications.htm

[4] A.R. Barron, "Complexity Regularization with application to artificial neural networks," In G. Roussas (Ed.) *Nonparametric Functional Estimation and Related Topics*. pp.561–576. Dordrecht, the Netherlands, Kluwer Academic Publishers. 1990.

[5] A.R. Barron, "Approximation and estimation bounds for artificial neural networks," *Computational Learning Theory: Proc. Fourth Annual ACM Workshop*, L. Valiant (ed.). San Mateo, California, Morgan Kaufmann Publ. pp.243–249. 1991.

[6] A.R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*. Vol. 39, pp.930–945. 1993.

[7] A.R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*. Vol.14, pp.113–143. 1994.

[8] A.R. Barron, "Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems," In A. Dawid, J.M. Bernardo, J.O. Berger and A. Smith (Eds.), *Bayesian Statistics*. Vol.6, pp.27–52. Oxford Univ. Press. 1998.

[9] A.R. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection by penalization," *Probab. Theory Rel. Fields*. Vol.113, pp.301–413. 1998.

[10] A.R. Barron, A. Cohen, W. Dahmen, and R. DeVore, "Approximation and learning by greedy algorithms," *Ann. Statist.* Vol.36, pp.64–94. 2008.

[11] A.R. Barron and T.M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*. Vol.37, No.4, pp.1034–1054. 1991.

[12] A.R. Barron and N. Hengartner, "Information theory and superefficiency," *Ann. Statist.* Vol.26, No.5, pp.1800–1825. 1998

[13] A.R. Barron, C. Huang, J.Q. Li, X. Luo, "The MDL principle, penalized likelihoods, and statistical risk," In *Festschrift for Jorma Rissanen*. 2008.

[14] A.R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*. Vol.44, No.6, pp.2743–2760. Special Commemorative Issue: Information Theory: 1948-1998.

[15] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions," *Bull. Calcutta Math. Soc.* Vol.35, pp.99–109. 1943.

[16] L. Birgé and P. Massart, "Rates of convergence for minimum contrast estimators," *Probab. Theory Rel. Fields*. Vol.97, 113–150. 1993.

[17] L. Birgé and P. Massart, "Minimum contrast estimators on sieves: exponential bounds and rates of convergence," *Bernoulli*. Vol.4, 329–375. 1998.

[18] F. Bunea, A.B. Tsybakov and M.H. Wegkamp. "Aggregation and sparsity via ℓ_1 penalized least squares," In G. Lugosi and H.U. Simon (Eds.), *Proc. 19th Ann. Conf. on Learning Theory: COLT 2006*. pp.379–391. Springer-Verlag, Heidelberg. 2006.

[19] F. Bunea, A.B. Tsybakov and M.H. Wegkamp, "Aggregation for Gaussian regression," *Ann. Statist.* Vol.35, pp.1674–1697. 2007.

[20] F. Bunea and A.B. Tsybakov and M.H. Wegkamp, "Sparse density estimation with ℓ_1 penalties," In N. Behouty and C. Gentile (Eds.), *Proc. 20th Ann. Conf. on Learning Theory: COLT 2007*. pp.530–543. Springer-Verlag, Heidelberg. 2007.

[21] G.H.L. Cheang *Neural Network Approximation and Estimation of Functions*. Ph.D. Thesis, Statistics Dept., Yale University. 1998.

[22] G.H.L. Cheang and A.R. Barron, "Penalized least squares, model selection, convex hull classes, and neural nets," In M. Verleysen (Ed.). *Proc. 9th ESANN*, pp.371–376. Brugge, Belgium, De-Facto Press. 2001.

[23] S.S. Chen, D.L. Donoho, "Basis pursuit," *Proc. Asilomar conference*. www-stat.stanford.edu/∼donoho/Reports/1994/asilomar.pdf

[24] S.S. Chen, D.L. Donoho and M.A. Saunders. "Atomic decompositions by basis pursuit," *SIAM J. Scientific Computing*. Vol.20. pp.33–61. 1999.

[25] H. Chernoff, "A measure of asymptotic efficiency of test of a hypothesis based on the sum of observations," *Ann. Math. Statist.* Vol.23, pp.493–507. 1952.

[26] B.S. Clarke and A.R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*. Vol.36, pp.453–471. 1990.

[27] B.S. Clarke and A.R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Planning and Inference*. Vol. 41, pp.37-60. 1994.

[28] T.M. Cover and J. Thomas. *Elements of Information Theory*. Second Edition. New York, Wiley-Interscience. 2006.

[29] D.D. Cox and F. O'Sullivan, "Asymptotic analysis of penalized likelihood and related estimators," *Ann. Statist.* Vol.18, pp.1676-1695. 1990.

[30] H. Cramér, *Mathematical Methods of Statistics*. Princeton Univ. Press. 1946.

[31] G.M. de Montricher, R.A. Tapia and J.R. Thompson, "Nonparametric maximum likelihood estimation of probability densities by penalty function methods," *Ann. Statist.* Vol.3, pp.1329–1348. 1975.

[32] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.* Vol.32, pp.407–499. 2004.

[33] J. Friedman, T. Hastie, and R. Tibshirani, "Pathwise coordinate optimization," *Ann. Applied Statist.* Vol.1, pp.302–332. 2007.

[34] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the lasso," *Biostatistics*. Dec. 12, 2007.

[35] I.J. Good and R.A. Gaskins, "Nonparametric Roughness Penalties for Probability Densities," *Biometrika*. Vol.58, pp.255–277. 1971.

[36] P. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, MIT Press. 2007.

[37] D. Haussler, and A.R. Barron, "How well do Bayes methods work for on-line prediction of + or -1 values?" In *Computational Learning and Cognition: Proc. Third NEC Research Symposium*, pp.74–100. SIAM, Philadelphia. 1993.

[38] D. Haussler and M. Opper, "Mutual Information, metric entropy, and cumulative relative entropy risk," *Ann. Statist.* Vol. 25, pp.2451–2492. 1997.

[39] C. Huang, G.H.L. Cheang, and A.R. Barron. "Risk of penalized least squares, greedy selection and ℓ_1-penalization from flexible function libraries," Submitted to *Ann. Statist.* 2008.

[40] K.L. Jones, "A simple lemma on greedy approximation in Hilbert spaces and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.* Vol.20, pp.608–613. 1992.

[41] A. Juditsky and A. Nemirovski, "Functional aggregation for nonparametric regression," *Annals of Statistics*. Vol.28, pp.681–712. 2000.

[42] K. Koh, S.J. Kim and S. Boyd, "An interior-point method for large-scale ℓ_1 regularized logistic regression," *J. Machine Learning Res.* Vol. 8, pp.1519–1555. 2007.

[43] E.D. Kolaczyk and R.D. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," *Ann. Statist.* Vol.32, pp.500–527. 2004.

[44] E.D. Kolaczyk and R.D. Nowak, "Multiscale generalized linear models for nonparametric function estimation," *Biometrika*. Vol.92, pp.119–133. 2005.

[45] V. Koltchinskii and D. Panchenko, "Complexities of convex combinations and bounding the generalization error in classification," *Ann. Statist.* Vol.33, pp.1455–1496. 2005.

[46] J.Q. Li, *Estimation of Mixture Models*. Ph.D. Thesis, Statistics Dept., Yale University, 1999.

[47] J.Q. Li and A.R. Barron, "Mixture density estimation," In S. Solla, T. Leen, and K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems*, Vol.12, pp.279–285.

[48] W.S. Lee, P. Bartlett, and R.C. Williamson, "Efficient agnostic learning of neural networks with bounded fan-in," *IEEE Trans. Inform. Theory*. Vol.42, pp.2118–2132. 1996.

[49] Y. Makovoz, Random approximates and neural networks. *J. Approximation Theory*, Vol.85, pp.98–109. 1996.

[50] L. Meier and S. van de Geer, and P. Bühlmann, "The Group Lasso for logistic regression," *J. Royal Statist. Soc, Ser. B*. Vol.70. 2008.

[51] D. Modha and E. Masry, "Rates of convergence in density estimation using neural networks," *Neural Computation*. Vol.8, pp.1107–1122. 1996.

[52] D. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," *IEEE Trans. Inform. Theory*. Vol.42, pp.2133–2145. 1996.

[53] A.S. Nemirovskii, B.T. Polyak, and A.B. Tsybakov, "Rate of convergence of nonparametric estimates of maximum likelihood type," *Probl. Inform. Transmission*. Vol.21, pp.258–272. 1985

[54] M.Y. Park and T. Hastie, "L_1-regularization path algorithm for generalized linear models," *J. Royal Statist. Soc: Ser. B*. Vol.69, pp.659–677.

[55] A. Rakhlin, D. Panchenko, and S. Mukherjee, "Risk bounds for mixture density estimation," *ESAIM: Probab. and Statist.*, Vol.9, pp.220–229.

[56] A. Rényi, "On measures of entropy and information," In *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Vol.1, pp.547–561. 1960.

[57] J. Rissanen, "Modeling by the shortest data description," *Automatica*. Vol.14, pp.465–471. 1978

[58] J. Rissanen, "A universal prior on integers and estimation by minimum description length," *Ann. Statist.* Vol. 11, pp.416–431. 1983.

[59] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*. Vol.30, pp.629–636. 1984.

[60] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.* Vol.14, pp.1080-1100. 1986.

[61] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. Hackensack, NJ, World Scientific. 1989.

[62] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*. Vol.42, pp.40–47. 1996.

[63] C.E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.* Vol.27, pp.379–423,623–656. 1948.

[64] X. Shen, "On the method of penalization," *Statistica Sinica*. Vol.8, pp. 337–357. 1998.

[65] B. Silverman, "On the estimation of probability function by the maximum penalized likelihood method," *Ann. Statist.* Vol.10, pp.795–810. 1982.

[66] J. Takeuchi and A.R. Barron, "Asymptotically minimax regret for exponential families," *Proc. Symp. Inform. Theory and its Appl.*, pp.665-668. 1997.

[67] J. Takeuchi and A.R. Barron, "Asymptotically minimax regret for exponential and curved exponential families," Summary at www.stat.yale.edu/∼arb4/publications.htm for the presention at *1998 Internat. Symp. Inform. Theory*.

[68] J. Takeuchi and A.R. Barron, "Asymptotically minimax regret by Bayes mixtures," In *Proc. 1998 Internat. Symp. Inform. Theory*.

[69] J. Takeuchi, T. Kawabata, and A.R. Barron, "Properties of Jeffreys' mixture for Markov Sources," To appear in the *IEEE Trans. Inform. Theory*.

[70] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal Statist. Soc., Ser. B*. Vol. 58, pp.267-288. 1996.

[71] G. Wahba, *Spline Models for Observational Data*. Philadelphia, SIAM. 1990

[72] R. Willett and R. Nowak, "Multiscale Poisson intensity and density estimation," www.ece.wisc.edu/ %7Enowak/multiscale_poisson.pdf 2005

[73] Q. Xie and A.R. Barron (1997), "Minimax redundancy for the class of memoryless sources, *IEEE Trans. Inform. Theory*. Vol.43, pp.646–657. 1997.

[74] Q. Xie and A.R. Barron, "Asymptotically minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inform. Theory*. Vol.46, pp.431–445. 2000.

[75] Y. Yang and A.R. Barron, "An asymptotic property of model selection criteria," *IEEE Trans. Inform. Theory*. Vol.44, pp.117–133. 1998.

[76] Y. Yang and A.R. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.* Vol.27, pp.1564–1599. 1999.

[77] H. Zhang, G. Wahba, Y. Lin, M. Voelker, R.K. Ferris, B. Klein, "Variable selection and model building via likelihood basis pursuit," *J. Amer. Statist. Assoc.* Vol.99, pp.659–672. 2005.

[78] T. Zhang, "Sequential greedy approximation for certain convex optimization problems," *IEEE Trans. Inform. Theory*. Vol. 49, pp.682-691. 2003.

[79] T. Zhang, "From epsilon-entropy to KL-entropy: analysis of minimum information complexity density estimation," *Ann. Statist.* Vol.34, pp.2180–2210. 2006.

[80] T. Zhang, "Some sharp performance bounds for least squares regression with ℓ_1 regularization," Manuscript at http://stat.rutgers.edu/∼tzhang/pubs.html