THE MDL PRINCIPLE, PENALIZED LIKELIHOODS, AND STATISTICAL RISK

ANDREW R. BARRON, CONG HUANG, JONATHAN Q. LI, AND XI LUO

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 $\log 1/\text{likelihood}(f) + \text{pen}(f)$ has statistical 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 of candidate fits (the ℓ_0 norm of the coefficients) as we review. 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.

1. INTRODUCTION

From work in the information theory and statistics communities, there are close connections between high-quality data compression and accurate statistical estimation. The original Shannon (1948) 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., Cover and Thomas 2007). The development of universal data compression and, in particular, the minimum description-length (MDL) principle has built this correspondence further to deal with 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, Barron, Rissanen and Yu 1998 or Grünwald 2007 and work cited therein). The function f may provide a density or log-density function (for instance we may have $p_f(x) = p_0(x)e^{f(x)}/c_f$ where p_0 is a reference distribution and c_f is a normalizing constant), 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.

Andrew Barron, Cong Huang, and Xi Rossi Luo are with the Department of Statistics, Yale University, P.O. Box 208290, New Haven, CT 06520-8290; Andrew.Barron@yale.edu, Cong.Huang@yale.edu and Xi.Luo@yale.edu; Jonathan Qiang Li is with Radar Networks, Inc., 410 Townsend St., San Francisco, CA 94107: qiang.li@aya.yale.edu.

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, the optimum sum of relative entropy approximation error and penalty relative to the sample size, is used to capture 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 (Barron and Cover 1991, Barron 1990, Li 1999, Kolaczyk and Nowak 2004,2005, and Grünwald 2007). The estimator is interpretable as a maximizing posterior probability with $L(f)$ equal to the log reciprocal prior probability of f. Without restriction to a countable class, resolvability bounds on risk of penalized likelihood estimators have been developed for functions in finite-dimensional families with penalty proportional to the dimension (Yang and Barron 1998, Barron, Birgé and Massart 1999). Moreover, resolvability bounds on risk of Bayes predictive density estimators have been developed as will be discussed below. 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 advocates of penalized likelihood estimation with penalty on the roughness (or irregularity) of the density include Good and Gaskins (1971), de Montricher, Tapia and Thompson (1975), and Silverman (1982). Reproducing kernel Hilbert space penalties are championed in Wahba (1990). Statistical rate results for quadratic penalties in Hilbert space settings corresponding to weighted ℓ_2 norms on coefficients in function expansions (including Sobolev-type penalties equal to squared L_2 norms of derivatives) are developed in Cox and O'Sullivan (1990) based on functional analysis tools. Later developments in this direction are in Cucker and Smale (2001). Empirical process techniques built around metric entropy calculations yield rate results for penalties designed for a wide variety of function classes in Shen (1998). Related theory for constrained maximum likelihood in nonparametric settings is in Nemirovski, Polyak and Tysbakov (1985) and for minimum contrast estimators and sieves in Birgé and Massart (1993,1998).

The use of ℓ_1 penalization of log-likelihoods is a currently popular approach, see Park and Hastie (2007). 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 Koh, Kim and Boyd (2007), Banerjee, Ghaoui and d'Aspermont (2007), Friedman, Hastie and Tibshirani (2007b), or Zhang, Wahba et al (2005). That work has focussed on algorithmic development, related to work for penalized least squares in Tibshirani's 1996 Lasso, Chen and Donoho's 1994,1999 basis pursuit, the LARS algorithm (Efron et al 2004), coordinate algorithms (Friedman et al 2007a) and relaxed greedy algorithms (Jones 1992, Barron 1993, Lee, Bartlett and Williamson 1996, Barron and Cheang 2001, Zhang 2003, and Barron, Cohen, et al 2008). A new algorithmic result is established at the end of this paper.

Recently there is activity to analyze risk of ℓ_1 penalized procedures. Some of it, requiring restrictions on the correlation of dictionary members, focusses on whether the procedure performs as well as the best subset selection rule, as in the work on ℓ_1 penalized least squares regression in Bunea, Tsybakov and Wegkamp (2006,2007a), on ℓ_1 penalized empirical L_2 criteria for density estimation in Bunea, Tsybakov and Wegkamp (2007b), and ℓ_1 penalized logistic regression in Meier, van de Geer and Bühlmann (2008). For general dictionaries without correlation conditions, it is natural to 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 manuscripts by Zhang (2007) and by Huang, Cheang and Barron (2008) and for ℓ_1 penalized likelihood in the present paper. Risk bounds for penalized likelihood should capture the corresponding tradeoff of Kullback-Leibler approximation error and the penalty, as is available for Bayes predictive estimators. This motivates our analysis of the risk of penalized likelihood estimators and demonstration that the ℓ_1 penalty satisfies the information-theoretic requirements for the results we seek.

Extending information-theoretic risk results to penalized likelihood with an uncountable family $\mathcal{F}$, the main tool developed in Section 3 is that of a variable-complexity cover. Such covers allow for variable penalty levels. The distortion used in measuring closeness to the cover 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 representer in the cover for which $pen(f)$ is not less than its complexity plus distortion.

The theory is simplified compared to alternatives that would glue together bounds for subclasses with their separate metric entropy (fixed complexity) covering properties. Indeed, it is not necessary to organize $\mathcal{F}$ to come from a list of function subclasses. Nevertheless, to relate to past work, various subclasses $\mathcal{F}_s$ may arise, corresponding to functions of various regularity s, quantified by number of derivatives or by weighted norms of coefficients in function expansions.

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. For instance, 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 are assumed to come from a sample space over which distributions indexed by f are provided. For our most general statements, other than a measure space, no particular structure need be assumed for this space. It is traditional to think of data in the form of a finite length string $\underline{U} = \underline{U}_n = (U_1, U_2, \ldots, U_n)$, 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 sometimes more briefly P_f if clear from the context) for the distributions on $\underline{U}$. Likewise $E_{\underline{U}|f}$ or sometimes 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, the space $\underline{\mathcal{U}}$ is countable, such as a discretization of an underlying continuous space, $p_f(\underline{u})$ 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. Then 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 Barron (1985) 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 on $\underline{\mathcal{U}}$. The likelihood function $likelihood(f)$ is $p_f(\underline{U})$ at specified data $\underline{U}$. When the sample space is a sequence space 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 in general is the choice of a single 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 Rissanen (1989), Barron, Rissanen and Yu (1998) and Grünwald (2007), 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. To put these results into an information-theoretic context, we first review below redundancy and resolvability bounds for mixture models and their implications for the risk of predictive estimators. These risk bounds are for the stronger Kullback-Leiber loss. This material shows that tools are already 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.

1.1. **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. These mixture densities are the same objects used in Bayesian model selection and Bayesian prediction. However, a difference is that with MDL we use data compression thinking to guide the choice of the prior weights to achieve operationally desirable properties.

We discuss tools for redundancy and resolvability bounds for mixtures and the bounds they yield on risk. First we recall results for parametric families in which the aim is to uniformly control the redundancy.

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 we recognize as the Kullback-Leibler divergence between the mixture and the target. In a well-studied problem, initiated in the characterization of communication channel capacity and extended to minimax redundancy of universal data compression (Gallager 1968,1974, Davisson 1973, Davisson and Leon-Garcia 1980, Haussler 1997, Clarke and Barron 1990,1994 and Xie and Barron 1997) the minimax procedure yielding the smallest worst case expected redundancy in each $\mathcal{F}_m$ corresponds to a

choice of prior w_m yielding the largest minimum average redundancy, interpretable as a maximum Shannon mutual information $I(f;\underline{U})$, and suitable approximate forms for the optimal w_m, called the least favorable prior or capacity achieving prior, are available in an asymptotic setting. Indeed, for smooth parametric families with a Fisher information $I(\theta|m)$, an asymptotically optimal prior is proportional to $|I(\theta|m)|^{1/2}$ with a sequence of boundary modifications, 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.

Building on the work of Shtarkov (1987), the theory of pointwise minimax redundancy identifies what is the smallest constant penalty that can be added 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 has been studied in an asymptotic setting in Rissanen (1996), Barron, Rissanen and Yu (1998), Takeuchi et al (1997a,1997b,1998,2007), and Xie and Barron (2000). One of the conclusions, in the cases studied there, is 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. That theory provides data compression justification for a penalty with main term proportional to the dimension d_m. Certain mixture procedures have asymptotically minimax pointwise redundancy and are shown to be close to the exact optimal normalized maximum likelihood. These mixtures use the same Fisher information based prior with boundary modification, with an additional modification required for non-exponential family cases, that puts some small mass on an enlargement of the family. That there are solutions of mixture form is of interest for our subsequent discussion of predictive distributions.

Choosing weights $w(m)$ to assign to the families can also be addressed from an information-theoretic standpoint, thinking of $\log 1/w(m)$ as a codelength. Indeed, since the MDL parameter cost, approximately $\frac{d_m}{2}\log n$, is determined by the dimension d_m, it is customary to set $\log 1/w(m)$ using the log-cardinality of models of the same dimension (one can not do much better than that for most such models). For example, 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 comparatively small additional description length for the dimension d. Often p is large compared to the sample size n, while the critical dimensions d which lead to the best resolvability are small compared to 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. Use of $\frac{d_m}{2}\log n$ alone is not in accord with the total minimum description length principle in such cases in which the contribution from $\log 1/w(m)$ is comparable or larger.

1.2. **Index of resolvability of mixtures.** We now come to a bound on expected redundancy of mixtures developed in Barron (1998), which is shown to bound an associated statistical risk. Recall the Kullback divergence $D(P_{\underline{U}}||Q_{\underline{U}}) = E \log p(\underline{U})/q(\underline{U})$ is the total expected redundancy if data $\underline{U}$ are described using $q(\underline{u})$ but the governing measure has a density $p(\underline{u})$. Suppose this density has the form $p_{f^*}(\underline{u})$, sometimes abbreviated $p_*(\underline{u})$. A tool in the examination of the redundancy is $D_n(f^*, f) = D(P_{\underline{U}|f^*}||P_{\underline{U}|f})$ which measures how well f approximates a hypothetical 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 $D(P_{f^*}||P_f)$. It is an important characteristic of mixtures that the divergence of a mixture from a product measure is considerably smaller than the order n divergence between pairs of distributions in the family.

Indeed, the resolvability bound on expected redundancy of mixtures is given 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, as in Barron (1998), 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, minimizing over any collection of such subsets B,

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

In i.i.d. modeling, we divide by n to obtain the following redundancy rate bound. This shows the redundancy of mixture codes controlled by an *index of resolvability*, expressing the tradeoff between the accuracy of approximating sets and their log prior probability relative to the sample size,

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

When the $B = \{f\}$ are singleton sets, the right side is the same as the index of resolvability given in Barron and Cover (1991), 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 illustrated in Barron (1998) 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 Barron (1988,1998), Haussler and Barron (1993), and Zhang (2006).

1.3. **Implications for predictive risk.** For predictive models the data are presumed to arise in a sequence $\underline{U}_N = (U_n)_{n=1}^N$ and 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 $\underline{u}_{n-1} = (u_i)_{i=1}^{n-1}$. Nevertheless, predictive distribution need not be restricted to be of such a plug-in form. Indeed, averaging with respect to a prior w, a one-step-ahead predictive redundancy is optimized by a Bayes predictive density $q(u_n|\underline{u}_{n-1})$. The one-step-ahead predictive redundancy is $E_f D(P_{U_n|f}||Q_{U_n|\underline{U}_{n-1}})$, which we recognize to be the Kullback risk of the predictive density, based on a sample of size $n-1$, as an estimate of the target density $p(u_n|f)$. Here and in what follows, it is to be understood that if the variables are not i.i.d. given f, the target becomes the conditional density $p(u_n|\underline{u}_{n-1}, f)$. 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

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

which is the cumulative Kullback risk. Dividing by N we see in particular 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 Barron and Hengartner (1998). That work shows for d-dimensional families that $\frac{d}{2n}$ is the asymptotically efficient level of individual Kullback risk based on samples of size n. Indeed, summing across sample sizes $n = 1, 2, \ldots, N$, it corresponds to a total Kullback risk (total redundancy) of $\frac{d}{2} \log N$, which cannot be improved upon asymptotically (except in a negligible set of parameters) according to Rissanen's (1984) award winning result. The predictive interpretation also plays a critical role for non-finite dimensional families $\mathcal{F}_s$ in identifying the efficient rates of estimation (also in Barron and Hengartner 1998) and in establishing the minimax rates of estimation (in Yang and Barron 1999 and Haussler and Opper 1997). For these cases typical individual risk rates are of the form some constant times $(1/n)^\rho$ for some positive rate $\rho \leq 1$. At the heart of that analysis, one observes that taking the Cesàro average Kullback risk across sample sizes up to N recovers the same form $(1/N)^\rho$ (albeit with a different constant multiplier). The idea is that minimax rates for total expected redundancy is somewhat easier to directly analyze than individual Kullback risk, though they are related by the chain rule given above.

1.4. **Two-stage codes.** We turn our attention to models based on $two\text{-}stage$ codes, also called two-part 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 by Rissanen (1978,1983) and in the analysis of Barron and Cover (1991). 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 $\sum_{f\in\mathcal{F}} 2^{-L_n(f)} \leq 1$, given the size n of the sample.

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 $\mathcal{F}_m$.

We are not forced to always have such growing parameter complexities. Indeed, as suggested by Cover and developed in Barron (1985) and Barron and Cover (1991), one may consider a more general notion of parameter complexity inspired by Kolmogorov. That work shows when any computable parameter value govern the data, ultimately a shorter total codelength obtains with it than for all other competitors and the true parameter value is discovered with probability one. Nevertheless, for any coding scheme, in second order smooth families with parameters in R^{d_m}, except for a null set of Lebesgue measure 0 as shown by Rissanen (1984,1986), the redundancy will not be of smaller order than $\frac{d_m}{2} \log n$. The implication for parameter coding is that for most parameters the representor in the code will need to have complexity of order not smaller than $\frac{d_m}{2} \log n$.

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

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

The minimizer $\hat{f}$ (breaking ties by choosing one of minimal $L_n(f)$) is called the minimum complexity estimator in the density estimation setting of Barron and Cover (1991) and it is called the complexity regularization estimator for regression and classification problems in Barron (1990).

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^*. As we have noted, eventually exact discovery is possible when f^* is in $\tilde{F}$, but its complexity, as will be ultimately revealed by the data, may be too great for full specification of f^* to be the suitable description with moderate sample sizes. It is helpful to have the 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 (defined as $\frac{1}{n}$ times the expected difference between the total codelength and the target $\log 1/p_{f^*}(\underline{U}_n)$) is shown in Barron and Cover (1991) 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 recognized as 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 will be the demonstration that the loss function we examine is smaller in expectation and stochastically not much more than the pointwise redundancy.

Returning to the form of the estimator, we note that 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. Thus, in fact, though it is customary to refer to two-stage coding, 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 $\hat{m}$ (again breaking ties by choosing the simplest such m) provides a model selection in accordance with the MDL principle. Likewise the resolvability takes the form

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

Again the ideal model selection, best resolving the target, is the choice m_n^* achieving the minimum, and the performance of the sample based MDL selection $\hat{m}$ is captured by the resolvability provided by m_n^*.

Two-stage codes in parametric families are closely related to average-case optimal mixture-codes. Indeed, in second order smooth families of dimension d, Laplace approximation, as in Barron (1985) or the references to pointwise redundancy given above, shows that log mixture likelihood is approximately the maximum log-likelihood minus the log ratio between the square root of the determinant of empirical total Fisher information and the prior density, plus $\frac{d}{2} \log 2\pi$. Two-stage codes can achieve the same form (although with a slightly suboptimal constant) provided one uses more elaborate parameter quantizations based on local diagonalization of the Fisher information with a rectangular grid in the locally transformed parameter space, as explained in Barron (1985), rather than merely using a rectangular grid in the original parameters. To avoid such complications and to have exact average-case optimality, when it is computationally feasible, it is preferable to use mixture models in such smooth families rather than two-stage codes.

Nevertheless, in many estimation settings, it is common to proceed by a penalized likelihood (or penalized squared error) criterion, and it is the intent of the present paper to address associated information-theoretic and statistical properties of such procedures.

To recap, we have seen in the minimum description-length principle that there are close connections between compression and statistical estimation.

The connections of information theory and statistics have additional foundations. While it is well-known that information-theoretic quantities determine fundamental limits of what is possible in communications, it is also true that corresponding information-theoretic quantities determine fundamental limits of what is possible in statistical estimation, as we recall in the next subsection.

1.5. **Information-theoretic determination of minimax rates.** In Haussler and Opper (1997) and Yang and Barron (1999) the problem of minimax rates of function estimation are shown to have an information-theoretic characterization. Suppose we have a loss function $\ell(f^*, f)$ which is a squared metric locally equivalent to Kullback divergence $D(f^*, f)$ (i.e., they agree to within a constant factor in a suitable subset of the function space), and assume that the data $\underline{U}_n = (U_1, \ldots, U_n)$ are i.i.d. from p_{f^*} with an f^* in a given function class $\mathcal{F}_s$. Here we use the subscript s to remind ourselves that we are referring to function subclasses that permit control on the quantities of interest that characterize minimax rates (that is, finite metric entropy or finite capacity). In the language of information theory the family of distributions $(P_{\underline{U}_n|f}, f \in \mathcal{F}_s)$ is a channel with inputs f and outputs $\underline{U}_n$.

Three quantities are shown to be important in the study of the statistical procedures: these are the Kolmogorov *metric entropy*, the Shannon *channel capacity*, and the *minimax risk* of Wald's statistical decision theory. The *metric entropy* $H_\epsilon = H_\epsilon(\mathcal{F}_s)$ is defined by

$$H_\epsilon(\mathcal{F}_s) \;=\; \inf_{\tilde{\mathcal{F}}} \left\{ \log \text{card}(\tilde{\mathcal{F}}) \;:\; \inf_{\tilde{f} \in \tilde{\mathcal{F}}} \ell(f, \tilde{f}) \le \epsilon^2, \quad \forall f \in \mathcal{F}_s \right\},$$

for which a critical $\epsilon_n = \epsilon_n(\mathcal{F}_s)$ is one for which ϵ_n^2 is of the same order as H_{ϵ_n}/n. Shannon's *channel capacity* $C_n = C_n(\mathcal{F}_s)$ is

$$C_n \;=\; \max_W I(f; \underline{U}_n)/n$$

where the maximum is over distributions W restricted to $\mathcal{F}_s$ and $I(f; \underline{U})$ is Shannon's mutual information for the channel, equal also to the Bayes average (w.r.t. W) of the redundancy of the mixture code. Finally, the *minimax risk* $r_n = r_n(\mathcal{F}_s)$ is

$$r_n \;=\; \inf_{\hat{f}} \sup_{f \in \mathcal{F}_s} E_f \, \ell(f, \hat{f}),$$

where the infimum is over all estimators based on the sample of size n. Suppose also that we are not in a finite-dimensional setting (where the metric entropy is order of a multiple of $\log 1/\epsilon$), but rather we are in an infinite-dimensional setting, where the metric entropy is of order at least $(1/\epsilon)^\gamma$ for some positive γ. Then from Haussler and Opper (1997) and Yang and Barron (1999), as also presented by Lafferty (2007), we have the equivalence of these quantities.

Theorem 1.1. *The minimax estimation rate equals the channel capacity rate equals the metric entropy rate at the critical precision. That is,*

$$r_n \sim C_n \sim \frac{H_{\epsilon_n}}{n} \sim \epsilon_n^2,$$

where $\sim$ means that the two sides agree to within constant factors.

In modern statistical practice it is rarely the case that one designs an estimator solely around one function class of bounded metric entropy. Indeed, even if one knew, in advance of seeing the data, how one wants to characterize regularity of the function (e.g. through a certain norm on coefficients), one usually does not have advance knowledge of an appropriate size of that norm, though such knowledge would be required for such metric entropy control. Instead, one creates an estimate that adapts, that is, it simultaneously gives the right levels of risk for various function subclasses. Penalized likelihood estimators provide a means by which to achieve such aims.

We say that the countable set $\tilde{\mathcal{F}}$ together with its variable complexities provides an *adaptive cover* of each of several function subclasses $\mathcal{F}_s$, if for each of the subclasses there is a subset of functions in $\tilde{\mathcal{F}}$ that have complexity bounded by a multiple of H_{ϵ_n} and that cover the subclass to within precision ϵ_n. Then application of the index of resolvability shows that the minimum complexity estimator is simultaneously minimax rate optimal for each such subclass. In this setting the role of Theorem 1.1 is to give the lower bounds showing that the achieved rates are indeed best possible.

As discussed in Yang and Barron (1998,1999), Barron, Birgé and Massart (1999), and Barron, Cohen, Dahmen and DeVore (2008), such adaptation (sometimes to within log-factors of the right rates) is shown to come for free for a variety of function classes when the models consist of subsets of basis functions from suitable dictionaries and the penalties are given by the dimension (times a log-factor).

The resolvability bounds go beyond such asymptotic rate considerations to give finite sample performance characterization specific to properties of the target f^*, not required to be tied to the worst case for functions in various classes.

2. 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).

In setting our loss function, we will have need of another measure of divergence. Analogous to the Kullback-Leibler divergence we have already discussed, for pairs of probability distributions

P and $\tilde{P}$ on a measurable space, we consider the Bhattacharyya, Hellinger, Chernoff, Rényi divergence (Bhattacharyya 1943, Cramér 1946, Chernoff 1952, Rényi 1960) 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 $P_{U_1|f}$.

We take this divergence to be our loss function in examination of the accuracy of penalized likelihood estimators. One reason is its close connection 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). Another is that $d(P,\tilde{P})$, 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. Thirdly, 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 $\frac{1}{4}(\mu-\tilde{\mu})^2$. Most important though for our present purposes is the cleanness with which it allows us to examine 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 $H^2(P,\tilde{P})$. In thinking about the affinity note that it is less than or equal to 1 with equality only when $P = \tilde{P}$. 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 Jonathan Li's 1999 Yale thesis is a simplification of a conclusion from Barron and Cover (1991). It is also presented in Kolaczyk and Nowak (2004) and in Grünwald (2007). 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 (Li 1999). For a countable $\tilde{\mathcal{F}}$, and $\mathcal{L}_{\rm n}(f) = 2L_n(f)$ 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 bound

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

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 Yang and Barron (1998), 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.

We comment that the presence of the factor 2 in the penalty $\mathcal{L}(f) = 2L(f)$ is a byproduct of using the Chernoff-Rényi divergence with parameter $1/2$. As in the original Barron and Cover (1991) bound, one may replace the 2 with any multiplier strictly bigger than 1, though the best bound there occurs with the factor 2. See Zhang (2006, Thm. 4.1) or Grünwald (2007, Ch. 15) for analogous risk bounds for Chernoff-Rényi divergences with parameter λ between 0 and 1.

Producing an exact minimizer of the complexity penalized estimator can be computationally difficult, but an approximate minimizer is still amenable to analysis by the above method. For instance in Li (1999) and Li and Barron (2000) a version of a greedy algorithm is given for estimating densities by sums of m components from a given dictionary of possible component densities (e.g. Gaussian mixtures). Analysis there shows that with m steps the complexity penalized likelihood is within order $1/m$ of the optimum.

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

Some implications of resolvability bounds on risk are discussed in Barron and Cover (1991). Corresponding results for complexity penalized least squares and other bounded loss functions were developed in Barron (1990). Applications to neural nets were developed in Barron (1991,1994), providing risk bounds for estimation of linear combinations of a dictionary, by penalized least squares with a penalty that incorporates aspects of the ℓ_0 and ℓ_1 norms of the coefficients, but restricted to a countable set (that restriction is lifted by Huang, Cheang and Barron (2008) and the developments we give in the next section). Analogous resolvability bounds for regression and log-density estimation by neural nets in a weakly dependent setting were given in Modha and Masry (1996a,b). For mixture density estimation (including Gaussian mixtures), direct implications of Theorem 2.1 using resolvability calculations are given in Li (1999) and Li and Barron (2000) and, building in part on those developments, Rakhlin, Panchenko, and Murherjee (2005) give related risk results using bounds for Rademacher averages of convex hulls.

Kolaczyk and Nowak (2004,2005), and Willett and Nowak (2005) give implications of Li's theorem for multiscale wavelet image estimation and Poisson intensity function estimation. In some of their investigations the data are functions (e.g. of continuous time or location) but the theory nevertheless applies as they make clear in their settings. Indeed, as we have indicated, the structure of the data $\underline{U}$ (other than that there be a dominating measure for the candidate distributions) is not essential for the validity of the general bounds.

The proof of Theorem 2.1 given here is essentially the same as in Li's Thesis. One slight difference is that along the way we have pointed out that the expected redundancy of the two-stage code is also a bound on the risk. This is also noted by Grünwald (2007) and, as he emphasizes, it even more closely relates the risk and coding notions. The resolvability form is more useful in obtaining bounds that exhibit the tradeoff between approximation accuracy and dimension or complexity.

To be specific, the proof of Theorem 2.1 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 is a random variable of mean bounded by 0. In a similar manner one can 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 (probability not more than $e^{-n\tau/2}$) to be greater than any positive τ.

The original bound of Barron and Cover (1991) also proceeded by a tail probability calculation, though it was noticeably more elaborate than given here. An advantage of that original proof is its change of measure from the one at f^* to the one at f_n^*, showing that questions about the behavior when f^* is true can indeed be resolved 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.

Remember that in this section we assumed that the space $\tilde{\mathcal{F}}$ of candidate fits is countable. From both statistics and engineering standpoints, it is awkward to have to force a user of this theory to construct a discretization of his space of functions in order to use this penalized likelihood result. We overcome this difficulty in the next section.

3. 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 when such a bound holds, it 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, at least by the techniques of which we are aware, would require control of higher order moments, particularly the variance, which, in order 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. Furthermore Bernstein-type bounds would entail a finite moment generating function of the log-likelihood ratio for generating function parameters in an open neighborhood of 0. 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 necessarily 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 conditions will be needed in this analysis other than working with the expected square-roots that give the Hellinger affinities, which are automatically bounded by 1. Note that this expected square root is a value of the moment generating function of the log-likelihood ratio $\log p_f(\underline{U})/p_*(\underline{U})$ and that its logarithm is a value of its cumulant generating function, but only evaluated at the specific positive value $1/2$.

In Theorem 2.1, the penalty $\mathcal{L}(f) = 2L(f)$ is used to show that if it is 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. We drop reference to dependence of the penalty on the sample size, but since the bounds we develop hold for any size data, there is no harm in allowing any of the quantities involved to change with n. In brief, 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 every 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 we use the complexity of such $\tilde{f}$ along with the discrepancy difference to assess whether a penalty $pen(f)$ is suitable. The countable set $\tilde{\mathcal{F}}$ of possible representors is taken to be non-stochastic. Nevertheless, the minimizer in $\tilde{\mathcal{F}}$ will depend 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 Barron and Cover (1991). 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.

An alternative perspective on formulation of conditions for penalized likelihood is in Shen (1998). He begins with an argument that the estimator is likely to have a penalty value in the set $\{f : pen(f) \leq Cpen(f^*)\}$, with C near 1. Under certain conditions, this set is compact with metric entropy properties using the Hellinger metric with a bracketing condition, permitting appeal to uniform large deviation properties of log likelihood ratios from Wong and Shen (1995) (which do require relationship between the variance and mean of the log-likelihood ratios) and to results on constrained maximum likelihood in nonparametric classes from Nemirovskii, Polyak and Tsybakov (1985). Presumably, one could also appeal then to results in Birgé and Massart (1993,1998) on what they call minimum contrast estimation. In these papers one can indeed see application to various function classes, including some that go beyond the traditional Sobolev type. However, for that machinery one inevitably has a number of typically unspecified, possibly large constants that arise giving a certain asymptotic rate flavor to the conclusions. Unlike Shen's method, we don't assume that the target f^* must have a finite penalty. What matters is that there be functions f close to f^* that do. Moreover, in seeking risk bounds of the form $\inf\{D(f^*, f) + pen(f)/n\}$ with constants equal to 1, we are striving to make the results of practical interest in non-asymptotic settings.

The ideas we develop here have parallels with other empirical measures of loss, such as the average squared error in regression problems, explored in the concurrently developed paper for which some of us are coauthors (Huang, Cheang and Barron 2008), building on earlier work with Cheang originating with his 1998 Yale thesis. In particular, that work does center by subtracting the expected loss in defining the discrepancies and does force uniform boundedness of the fits so that variances of the squared errors are proportional to the mean squared errors. The idea of bridging from the countable to the uncountable classes by the assumption that the penalty exceeds a complexity penalized discrepancy difference originates with Cong Huang in this regression work. Its use is simpler here in dealing with densities, because we use a milder loss function that allow arbitrary densities.

Our main theorem, generalizing Theorem 2.1 to 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

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

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

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

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

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

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-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: We complete this section with a further comment on the tool for verification of the requirement on the penalty. 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

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

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.

4. Information-theoretic validity of the 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. In the discussion below we leave the weights a_h free for us to determine what might be the most appropriate. With a little compromise we will settle upon $a_h = \|h\|_\infty$. With $f = f_\theta$ we denote $V_f = \|\theta\|_1$, with the understanding that when it happens that there are multiple θ representing the same f one takes $V_f = \min\{\|\theta\|_1 : f_\theta = f\}$. As suggested by the notion of the total variation which corresponds to the case that $\mathcal{H}$ consists of indicators of half-spaces, with the definition of V_f 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. (One can also work with an infinite $\mathcal{H}$ together with an empirical cover as explored in Section 5.) As we shall see, the codelengths of our representors will arise via a variable number of terms times the log cardinality of the dictionary (one could allow variable complexity of members h, but for simplicity that too will not be explored at this time). 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 $K \log 2$ (or more standard codelengths for integers may be used, e.g, of size slightly larger than $\log K$). 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

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

Some shortening of this codelength is possible, taking advantage of the fact that the order of the terms $h_1, \ldots, h_K$ does not matter and that repeats are allowed, as will be briefly addressed in Section 5. For simplicity we take advantage of the present form linear in K in the current section.

To establish the existence of a representor $\tilde{f}$ of f with the properties we want, we consider a distribution on choices of $h_1, h_2, \ldots, h_K$ 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 $K_f \log 2p$.

Now the critical 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 $v(x) = v_f(x)$ given by

$$V \sum_h |\theta_h| h^2(x)/a_h - f^2(x)$$

which is not more than $V \sum_h |\theta_h| ||h||_\infty^2 / a_h$.

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:

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

where $p(x) = p_a(x)$ is a probability density function as specified in the preceding section. The minimizing $\tilde{f}$ gives a value that is 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. The expected log is less than or equal to the log of the expectation and we bring that expectation inside the integral. So, indeed, at each x we are to 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 $\exp\{\frac{1}{2K}[Vh(x)/a_h - f(x)]\}$ for a randomly drawn h.

We now take advantage of classical lemmas of Bernstein and Hoeffding, easily verified by using the series expansion of the exponential. If T is a random variable satisfying the Bernstein moment condition $E|T-\mu|^m \leq \frac{m!}{2}\sigma^2(\text{BERN})^{m-2}$ for $m \geq 2$, where $\mu = ET$, and $\sigma^2 = E|T-\mu|^2$, then for $K > \text{BERN}$, we have $E \exp\{\frac{1}{K}(T-\mu)\} \leq \exp\{\frac{\sigma^2}{2K^2} \frac{1}{1-\text{BERN}/K}\}$. Here BERN is called the Bernstein constant. (Classically, the symbol h is used for the Bernstein constant, but we can't use that here since h means our random function.) If T has range bounded by B, then the Bernstein constant is bounded by B. Actually, in the case of bounded range, we have the stronger moment conditions $E|T-\mu|^m \leq \sigma^2 B^{m-2}$ for $m \geq 2$ and $E \exp\{\frac{1}{K}(T-\mu)\} \leq \exp\{\frac{\sigma^2}{2K^2} e^{B/K}\}$ holding for all positive K. Likewise in this case one has the Hoeffding bound $E \exp\{\frac{1}{K}(T-\mu)\} \leq \exp\{\frac{B^2}{8K^2}\}$.

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. Strictly speaking we only need the max and min for h that appear in the linear combination $f(x) = \sum_h \theta_h h(x)$ we are currently working to represent. For a uniform bound on the range use $2 \max_h ||h||_\infty / a_h$. Then at the given x, we may use Bernstein's Lemma with the $T = \frac{1}{2} V h(x)/a_h$, a random variable, induced by the random h, having mean $\frac{1}{2} f(x)$, variance $\frac{1}{4} v_f(x)$ and Bernstein constant bounded by the range $\frac{V}{2} R(x)$. This range divided by K is equal to $\delta R(x)/2$ in our setting. Then at the given x, using the Bernstein-like inequality or the Hoeffding inequality gives that the expectation of $\exp\{\frac{1}{2}(\tilde{f}(x) - f(x))\}$ is bounded by $\exp\{\frac{v(x)}{8K} e^{\delta R(x)/2}\}$ or by $\exp\{\frac{(VR(x))^2}{32K}\}$.

The next step is to bound $2n$ times the log of the integral of either of these exponential bounds with respect to a probability density $p(x)$. This involves the cumulant generating function of the exponent. In the first case, that is approximately the cumulant generating function for $v(x)/8K$, which should be near the mean of $v(x)/8K$ with respect to $p_a(x)$. Bringing the integral with respect to this density inside the sum defining $v(x)$, it would give $\frac{n}{4K}[V \sum_h |\theta_h| ||h||_2^2/a_h - ||f||_2^2]$ as the approximation to the remaining part of the discrepancy difference, where here $||h||_2^2 = \int p_a(x) h^2(x)$ is the squared $L_2(P_a)$ norm. Recalling that V is near $V_f = \sum_h |\theta_h| a_h$, it suggests via the Cauchy-Schwartz inequality that the best weights a_h would take the form $a_h = ||h||_2$. But with p_a unknown, use of $||h||_2$ in forming the penalty is not feasible. An empirical version with $||h||_2^2$ replaced by $(1/n) \sum_{i=1}^n h^2(X_i)$ may be considered for practical use, though the present theory does not yet provide means with which to support it.

Instead, we may bound $v(x)$ with $V \sum_h |\theta_h| ||h||_\infty^2 / a_h$. This is near $V_f \sum_h |\theta_h| ||h||_\infty^2 / a_h$, with $V_f = \sum_h |\theta_h| a_h$, which is optimized with the weights $a_h = ||h||_\infty$. With this choice of weights our bound on the variance is $V V_f$. Moreover, conveniently, this choice of a_h makes the range $R(x)$ of $h(x)/||h||_\infty$ bounded by 2. We are left then with an upper bound on $(v(x)/8K) e^{\delta R(x)/2}$, and hence an upper bound on its cumulant generating function, given by $(V_f V/8K) e^\delta$ which is $\frac{1}{8} V_f \delta e^\delta$. Now multiplying by $2n$, our bound on the discrepancy difference is $\frac{1}{4} n V_f \, \delta e^\delta$. Here one may alternatively use the Hoeffding bound, noting that the quantity $\frac{(VR(x))^2}{32K}$, now equal to $\frac{V^2}{8K}$ or equivalently $\frac{1}{8} V \, \delta$, when multiplied by $2n$ yields a discrepancy difference bound of

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

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

The form of the discrepancy difference bound above is to our liking, because it is proportional to V_f as desired. Now twice the complexity plus the discrepancy bound has size $2K \log(2p) +$

$\frac{1}{4}nV_f\delta + \frac{1}{4}n\delta^2$, 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 $C = 4\log(2p)$. We note that the presence of the constant term $C = 4\log(2p)$ 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 $\left[\frac{\log p}{n}\right]^{1/2}$ 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.

5. Refined resolvability for ℓ_1 penalized log likelihood

Three directions of refinement of this risk conclusion for ℓ_1 penalized log likelihood are presented briefly here, using the techniques introduced above. These parallel corresponding refinements for ℓ_1 penalized least squares in Huang et al (2008). This section may be skipped by those who only want the overview and who want to move to the computation results of Section 6. The present material is for readers who want to see some of the nuances of statistical rates of density estimation using ℓ_1 controls.

One refinement is that a valid codelength bound for $\tilde{f}$ can take the form $K \log(4e \max\{p/K, 1\})$ which is smaller when $K > 2e$. This leads to an improvement in the risk conclusion in which λ_n^* is as above but with $4e \max\{p/\sqrt{n}, 1\}$ in place of $2p$ inside the log factor so that the log factor may be replaced by a constant when p is small, not more than a multiple of $\sqrt{n}$. The idea of this improvement originates in the setting of Bunea et al (2007a). This improved codelength and risk conclusion follows directly from the above argument using Huang et al (2008), Lemmas 8.5 and 8.6 so we omit the detail. This refinement does not improve the order of the bound when the dictionary size p is a larger order power of n.

Secondly, we consider infinite dictionaries with a finite metric dimension property, and show that a suitable 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 $\left[\frac{d \log n}{n}\right]^{1/2}$. Thus the performance of the ℓ_1 penalized log-likelihood estimator is in agreement with what was obtained previously for other estimators in Barron (1991,1994), Modha and Masry (1996a,b), Lee, Bartlett, and Williamson (1996), Juditsky and Nemirovski (2000), Barron, Cohen, Dahmann and Devore (2008); where 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 $1/2$.

To explain, suppose a library $\mathcal{H}$ has the properties that $\|h\|_\infty \leq b$ and there are positive constants c and d such that for each positive $\varepsilon \leq b$ there is a finite L_∞ cover $\tilde{\mathcal{H}}$ of size $M_\varepsilon \leq (c/\varepsilon)^d$. Here the cover property is that for each h in $\mathcal{H}$ there is a representor $\tilde{h}$ in $\tilde{\mathcal{H}}$ with $\|h - \tilde{h}\|_\infty \leq \varepsilon$. Then d is called the metric dimension of $\mathcal{H}$. (As shown in Barron (1991,1994) this property holds for the dictionary of Lipshitz sigmoidal functions in d variables used in single hidden layer neural networks and related classes of sinusoidal functions; see Barron, Birgé and Massart (1999) for other examples.) Again with $\mathcal{F}$ the linear span of $\mathcal{H}$, functions take the form $f(x) = \sum_{h\in\mathcal{H}} \theta_h h(x)$ in $\mathcal{F}$ and we consider optimization of the ℓ_1 penalized log likelihood.

To adapt the above proof, we use the unweighted ℓ_1 norm $\|\theta\|_1 = \sum_{h\in\mathcal{H}} |\theta_h|$, multiplying by b^2 in the bound on the $v(x)$ to account a_h equal to 1 rather than $\|h\|_\infty$, and let V_f is the infimum of such $\|\theta\|_1$ among representations satisfying $f_\theta = f$. To obtain a representor $\tilde{f}$, we again draw $h_1, \ldots, h_K$ independently, with distribution that yields h with probability $|\theta_h|/V$, with $K = K_f$ and V as before. The new step is to replace each such h_j with its representor $\tilde{h}_j$ in $\tilde{\mathcal{H}}$, which changes the value of each $\tilde{f}(x)$ by at most $V\varepsilon$. Thus the discrepancy studied above

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

is increased by at most $2nV\varepsilon$, while the complexity is the same as before with the cardinality p replaced by M_ε. This yields a complexity penalized discrepancy bound of $2K_f \log(2M_\varepsilon) + \frac{1}{4}nb^2(V_f + \delta)\delta + 2n(V_f + \delta)\varepsilon$, where the three terms correspond to the three parts of the above analysis: namely, the complexity, the discrepancy, and the contribution of the cover of the dictionary.

Consequently, we have validity of the penalty $pen_n(f) = \lambda_n V_f + C$, with λ_n at least $\lambda_n^* = \frac{2}{\delta}\log(2M_\varepsilon) + \frac{1}{4}nb^2\delta + 2n\varepsilon$ and $C = 2\log(2M_\varepsilon) + \frac{1}{4}nb^2\delta^2 + 2n\delta\varepsilon$. Setting $\delta = \frac{1}{b}\left[\frac{8}{n}\log(2M_\varepsilon)\right]^{1/2}$ produces the best such λ_n^* equal to $b\left[2n\log(2M_\varepsilon)\right]^{1/2} + 2n\varepsilon$. With M_ε replaced by the bound $(c/\varepsilon)^d$, to balance the two terms in λ_n^* we set $\varepsilon = b\sqrt{d/n}$, valid for $d \leq n$. Then M_ε is within a constant factor of $(n/d)^{d/2}$ and we have the desired risk conclusion in a slightly improved form. Indeed, for any sequence of dictionaries and sample sizes with d/n small, $\frac{\lambda_n^*}{n}$ is near $b\left[\frac{d}{n}\log\frac{n}{d}\right]^{1/2}$ and $\frac{C}{n}$ is near $2\left[\frac{d}{n}\log\frac{n}{d}\right]$. To summarize, with λ_n not less than this λ_n^* for dictionaries of finite metric dimension, we have the resolvability bound on risk of the ℓ_1 penalized likelihood estimator:

$$Ed(f^*, \hat{f}) \;\leq\; R_n(f^*) \;\leq\; \inf_{f\in\mathcal{F}} \left\{ D(f^*, f) + \frac{\lambda_n V_f}{n} \right\} \;+\; \frac{C}{n}.$$

A feature of this analysis of resolvability of densities is that the constructed variable-complexity cover $\tilde{\mathcal{F}}$ is not data-dependent. This necessitated our appeal to L_∞ covering properties of the

dictionary in constructing the set of representors $\tilde{\mathcal{F}}$. Results for least squares in Lee, Bartlett, and Williamson (1996) and for penalized least squares in Huang et al (2008) allow for data-dependent covers (depending on observed and hypothetical input data), and accordingly allow for empirical L_1 or L_2 covering properties of the dictionary, thus allowing traditional step sigmoids in the neural net case. It is not clear whether there is a method to allow data-dependent covers in risk analysis for density estimation by penalized likelihood.

Thirdly, an improved method of approximation with probabilistic proof originates in the L_2 case in Makovoz (1996), with a stratified sampling interpretation in Huang et al (2008). 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 $K_0 < K$. We find 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 the dimension d is large. This conclusion is in agreement with what is achieved by other estimators in Yang and Barron (1999) and close to the lower bound on optimal rates given there. Similar implications for classification problems using convex hulls of a dictionary are in Koltchinskii and Panchenko (2005). The refined conclusion for ℓ_1 penalized least squares is given in Huang et al (2008) using empirical L_2 covering properties based on Makovoz's result.

Adapting the stratified sampling argument to ℓ_1 penalized log likelihood and the use of L_∞ covering properties proceeds as follows. Partition $\mathcal{H}$ into K_0 disjoint cells c. Let $v(c) \geq \sum_{h\in c}|\theta_h|$ and $f_c(x) = \frac{1}{v(c)}\sum_{h\in c}\theta_h h(x)$ which decomposes $f(x) = \sum_{h\in\mathcal{H}}\theta_h h(x)$ as $f(x) = \sum_c v(c) f_c(x)$. Consider positive integers $K(c)$. A convenient choice is $v(c) = \eta K(c)$ with $K(c) = \lceil \sum_{h\in c}|\theta_h|/\eta \rceil$. For each cell c draw $h_{c,k}$, for $k = 1, 2, \ldots, K(c)$, independently with outcome h with probability $\theta_h/v(c)$ for h in c (and outcome 0 with any leftover probability due to $v(c)$ possibly larger than $\sum_{h\in c}|\theta_c|$). Form the within cell sample averages $f_{c,K}(x) = \frac{1}{K(c)}\sum_{k=1}^{K(c)} h_{c,k}(x)$ and the random representor $\tilde{f}(x) = \sum_c v(c) f_{c,K}(x)$, which is seen to be an equally weighted average when $v(c)$ is proportional to $K(c)$. Now with $a_h = 1$ we proceed as in the analysis in the previous section, with the following exception.

For each x, the expectation, with respect to the distribution of the random terms, of $\exp\{\frac{1}{2}[\tilde{f}(x) - f(x)]\}$ is again straightforward by the independence of the terms $h_{c,k}$, but now they are not all identically distributed. This expectation becomes the product across the cells c of the $K(c)$ power of the expectation of $\exp\{\frac{1}{2}\frac{v(c)}{K(c)}[h_{c,1}(x) - f_c(x)]\}$. By the Hoeffding bound each of these expectations is not more than $\exp\{\frac{1}{32}\left(\frac{v(c)}{K(c)}\right)^2 R_c(x)\}$, where $R_c(x) = \max_{h\in c} h(x) - \min_{h\in c} h(x)$ is the

range of $h(x)$ for $h \in c$ for each x. With $mid_c(x) = [\min_{h\in c}(x) + \max_{h\in c}(x)]/2$ equal to the midrange function we recognize that it is the choice of function representing cell c optimizing $\max_{h\in c} |h(x) - mid_c(x)|$, equal to the half-range $R_c(x)/2$. Then we bound $R_c(x)$ by $\|R_c\|_\infty = 2\max_{h\in c} \|h - mid_c\|_\infty$, which is not more than $2\varepsilon_0$ if the partition is arranged to correspond to the best L_∞ cover of $\mathcal{H}$ of size K_0. Accordingly, the expectation of $2n \log \int p_a(x) \exp\{\frac{1}{2}[\tilde{f}(x) - f(x)]\}$ is not more than $\frac{1}{4}n \sum_c \frac{v(c)^2}{K(c)}\varepsilon_0^2$. Choosing $v(c)/K(c) = \eta$ to equal δ/ε_0 and $V = \sum_c v(c)$, this is $\frac{1}{4}nV\delta\varepsilon_0$, improving on the previous bound by the presence of the factor ε_0.

The other difference with the previous analysis is that with $K(c)$ equal to $\sum_{h\in c} |\theta_h|/\eta$ rounded up to an integer, the sum of these counts over the K_0 cells is a total count of $K = K_f$ between V_f/η and $V_f/\eta + K_0$. Likewise, $V = K\eta$ is between V_f and $V_f + K_0\eta$, with $\eta = \delta/\epsilon_0$.

So the complexity penalized discrepancy bound is now $2K_f \log(2M_\varepsilon) + \frac{1}{4}nV\delta\varepsilon_0 + 2nV\varepsilon$. Using the indicated bounds on K and V, and setting $\delta = \left[\frac{8}{n}\log(2M_\varepsilon)\right]^{1/2}$, it is not more than $\lambda_n^* V_f + C$, with $\lambda_n^* = \varepsilon_0 \left[2n \log(2M_\varepsilon)\right]^{1/2} + 2n\varepsilon$ the same as before but with the smaller ε_0 in place of b, which is the source of the improved rate. One sees that a good choice for the relationship between the precisions is $\varepsilon = \varepsilon_0/\sqrt{n}$, with which $C = K_0 \left[4\log(2M_\varepsilon) + 2n\delta\varepsilon/\varepsilon_0\right]$ becomes $C = K_0 \left[4\log(2M_\varepsilon) + (8\log(2M_\varepsilon))^{1/2}\right]$, the same order as before but with the multiplication by $K_0 \geq 1$. Again the resolvability is $\inf_{f\in\mathcal{F}} \left\{ D(f^*, f) + \frac{\lambda_n^* V_f}{n} + \frac{C}{n} \right\}$, with the improved λ_n^* and the inflated C. In particular, in the finite metric dimension case with K_0 of order $(1/\varepsilon_0)^d$, setting ε_0 of order $\left[\frac{d}{n}\log\frac{n}{d}\right]^{\frac{1}{2d+2}}$ one finds that both λ_n^*/n and C/n are of order $\left[\frac{d}{n}\log\frac{n}{d}\right]^{\frac{1}{2}+\frac{1}{2d+2}}$, providing the claimed improvement in rate.

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.

6. A Note on Computation

Building on past work on relaxed greedy algorithms, we consider successively optimizing 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 $k = 1, 2, \ldots$. The 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 $0 \leq \alpha \leq 1$ 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 $k \geq 1$,

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

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 Huang et al (2008). The proof of this computation bound and the risk analysis given above have many aspects in common. So it is insightful to give the proof here.

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

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

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

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

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

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

where $\bar{\alpha} = (1-\alpha)$. 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

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

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 may bring this expectation for random h inside the logarithm, and then inside the integral, obtaining an upper bound by Jensen's inequality. Now 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 $0 \leq \alpha \leq 1$, in particular with $\alpha = 2/(k+1)$, and $e_0 \leq 2V_f^2$, so by induction the result holds

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

This computation bound as well as its regression counterpart in Huang, Cheang and Barron (2008) holds even for $\lambda = 0$, which shows its relationship to past relaxed greedy algorithm work (by Jones 1992, Barron 1993, Lee, Bartlett and Williamson 1996, Cheang 1998, Cheang and Barron 2001, Li and Barron 2000, Zhang 2003 and Barron, Cohen, Dahmen, and DeVore 2008). These previous results remind us that explicit control on the ℓ_1 norm of the estimator is not necessary for similar conclusions. Instead, one can incorporate a penalty on the the number of terms k rather than their ℓ_1 norm and have fast computations by traditional relaxed greedy pursuit algorithms with $\lambda = 0$. The conclusion in the cited work is that it yields estimators which perform well as captured by risk bounds based on the best tradeoff between the accuracy of functions in the linear span and their ℓ_1 norm of coefficients. The result stated here for ℓ_1 penalized log-likelihood and in Huang et al (2008) 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 comfortably fits with our risk results. Indeed, the proof of our main risk conclusion (Theorem 3.1) involves the penalized likelihood ratio $\log \frac{p_{f^*}(X)}{p_{\hat{f}}(X)} + pen(\hat{f})$. Instead

of the exact penalized likelihood estimator $\hat{f}$, substitute its k term greedy fit $\hat{f}_k$, Then the computation bound of the current section 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

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

Finally, we note an intrinsic connection between the computation analysis and the information-theoretic validity of the penalty for statistical risk. Indeed, inspecting the proof of the computation bound we see that it can be adapted to show that $V^2/(K+1)$ bounds the discrepancy divided by n of an associated greedily obtained f_K, which may be used as a representor of f, rather than the sample average $\tilde{f}$ used in Section 4. Moreover with prescription of α_k and β_k, one again can describe such f_K using $K \log(2p)$ bits. Accordingly, the same analysis used to demonstrate the computation bound also demonstrates the information-theoretic validity of the ℓ_1 penalty.

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. As a reviewer notes, as information-theorists we are predisposed to look for opportunity to provide such an argument by Shannon's pioneering work. One can see other specific precursors for the probabilistic proof argument used here. For the purposes of demonstrating information-theoretically valid penalties for log-likelihood for Rissanen's MDL criterion, the idea for the probabilistic argument came in part from its use in the least squares setting, showing approximation bounds by greedy algorithms, in the line of research initiated by Jones.

References

Banerjee, O., L.E. Ghaoui, and A. d'Aspremont (2007). Model selection through sparse maximum likelihood estimation. To appear in the *Journal of Machine Learning Research*. Available at http://arxiv.org/abs/0707.0704

Barron, A. R. (1985). *Logically Smooth Density Estimation*. Ph. D. Thesis, Department of Electrical Engineering, Stanford University, Stanford, CA.

Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. University of Illinois Department of Statistics Technical Report #7. Available at www.stat.yale.edu/∼arb4/publications.htm

Barron, A. R. (1990). 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.

Barron, A. R. (1991). Approximation and estimation bounds for artificial neural networks. *Computational Learning Theory: Proceedings of the Fourth Annual ACM Workshop*, L. Valiant (ed.). San Mateo, California, Morgan Kaufmann Publ. pp. 243-249.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*. Vol. 39, pp. 930-945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*. Vol. 14, pp. 113-143.

Barron, A. R. (1998). 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 University Press.

Barron, A.R., L. Birgé, and P. Massart (1999). Risk bounds for model selection by penalization. *Probability Theory and Related Fields*. Vol. 113, pp. 301-413.

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

Barron, A.R., A. Cohen, W. Dahmen, and R. DeVore (2008). Approximation and learning by greedy algorithms. *Annals of Statistics*. Vol. 36, No.1, pp. 64-94.

Barron, A. R., and T. M. Cover (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*. Vol. 37, No.4, pp. 1034-1054.

Barron, A. R., and N. Hengartner (1998). Information theory and superefficiency. *Annals of Statistics*. Vol. 26, No.5, pp. 1800-1825.

Barron, A.R., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*. Vol. 44, No.6, pp. 2743-2760. Special Commemorative Issue: Information Theory: 1948-1998.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by probability distributions. *Bulletin of the Calcutta Mathematics Society*. Vol. 35, pp. 99-109.

Birgé, L., and P. Massart. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*. Vol. 97, 113-150.

Birgé, L., and P. Massart. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*. Vol. 4, 329-375.

Bunea, F., and A.B. Tsybakov and M.H. Wegkamp (2006). Aggregation and sparsity via ℓ_1 penalized least squares. In G. Lugosi and H.U. Simon (Eds.), *Proceedings of the 19th Annual Conference on Learning Theory: COLT 2006*. pp. 379-391. Springer-Verlag, Heidelberg.

Bunea, F., and A.B. Tsybakov and M.H. Wegkamp (2007a). Aggregation for Gaussian regression. *Annals of Statistics*. Vol. 35, pp. 1674-1697.

Bunea, F., and A.B. Tsybakov and M.H. Wegkamp (2007b). Sparse density estimation with ℓ_1 penalties. In N. Behouty and C. Gentile (Eds.), *Proceedings of the 20th Annual Conference on Learning Theory: COLT 2007*. pp. 530-543. Springer-Verlag, Heidelberg.

Cheang, G. H. L. (1998). *Neural Network Approximation and Estimation of Functions*. Ph.D. Thesis, Department of Statistics, Yale University.

Chen, S.S. and D.L. Donoho (1994). Basis pursuit. *Proceedings of the Asilomar Conference*. www-stat.stanford.edu/∼donoho/Reports/1994/asilomar.pdf

Chen, S.S, D.L. Donoho and M. A. Saunders (1999). Atomic decompositions by basis pursuit. *SIAM Journal on Scientific Computing*. Vol. 20. pp. 33-61.

Chernoff, H. (1952). A measure of asymptotic efficiency of test of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*. Vol. 23, pp. 493-507.

Clarke, B. S., and A. R. Barron (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*. Vol. 36, No.3, pp. 453-471.

Clarke, B. S., and A. R. Barron (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*. Vol. 41, pp. 37-60.

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

Cox, D. D., and F. O'Sullivan (1990). Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*. Vol. 18, pp. 1676-1695.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Cucker, F., and S. Smale (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematics Society*. Vol. 39. pp. 1-49.

Davisson, L. (1973). Universal noiseless coding. *IEEE Transactions on Information Theory*. Vol. 19, pp. 783-795.

Davisson, L., and Leon-Garcia (1980). A source matching approach to finding minimax codes. *IEEE Transactions on Information Theory*. Vol. 26, pp. 166-174.

de Montricher, G.M., R.A. Tapia and J. R. Thompson (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Annals of Statistics*. Vol. 3, pp. 1329-1348.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics*. Vol. 32, pp. 407-499.

Friedman, J., T. Hastie, and R. Tibshirani (2007a). Pathwise coordinate optimization. *Annals of Applied Statistics*. Vol. 1, pp. 302-332.

Friedman, J., T. Hastie, and R. Tibshirani (2007b). Sparse inverse covariance estimation with the lasso. *Biostatistics*. Dec. 12.

Gallager, R. G. (1968). *Information Theory and Reliable Communication*. New York, Wiley.

Gallager, R. G. (1974). Notes on Universal Coding, Supplement #3. MIT course 6.441.

Good, I.J., and R. A. Gaskins (1971). Nonparametric Roughness Penalties for Probability Densities. *Biometrika*. Vol. 58, pp. 255-277.

Grünwald, P. (2007). *The Minimum Description Length Principle*. Cambridge, MA, MIT Press.

Haussler, D. (1997). A general minimax result for relative entropy. *IEEE Transactions on Information Theory*. Vol. 43, No.4, pp. 1276-1280.

Haussler, D., and A.R. Barron (1993). 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.

Haussler, D., and M. Opper (1997). Mutual Information, metric entropy, and cumulative relative entropy risk. *Annals of Statistics*. Vol. 25, pp. 2451-2492.

Huang, C., and G.H.L. Cheang, and A.R. Barron (2008). Risk of penalized least squares, greedy selection and ℓ_1-penalization from flexible function libraries. Submitted to *Annals of Statistics*.

Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert spaces and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, Vol. 20, p. 608-613.

Juditsky, A. and A. Nemirovski (2000). Functional aggregation for nonparametric regression. *Annals of Statistics*. Vol. 28, pp. 681-712.

Koh, K., and S.J. Kim and S. Boyd (2007). An interior-point method for large-scale ℓ_1 regularized logistic regression. *Journal of Machine Learning Research*. Vol. 8, pp. 1519-1555.

Kolaczyk, E.D., and R. D. Nowak, (2004). Multiscale likelihood analysis and complexity penalized estimation. *Annals of Statistics*. Vol. 32, pp. 500-527.

Kolaczyk, E.D., and R. D. Nowak, (2005). Multiscale generalized linear models for nonparametric function estimation. *Biometrika*. Vol. 92, No. 1, pp. 119-133.

Koltchinskii, V. and D. Panchenko (2005). Complexities of convex combinations and bounding the generalization error in classification. *Annals of Statistics*. Vol. 33, pp. 1455-1496.

Lafferty, J. (2007). Challenges in statistical machine learning. Presented at *Information Theory and Applications*, University of California, San Diego, Feb. 2007. Video http://ita.ucsd.edu/workshop/07/talks

Lee, W.S., P. Bartlett, and R. C. Williamson (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*. Vol. 42, pp. 2118-2132.

Li, J.Q. (1999). *Estimation of Mixture Models*. Ph.D. Thesis, Department of Statistics, Yale University, New Haven, CT.

Li, J.Q., and A. R. Barron (2000). Mixture density estimation. In S. Solla, T. Leen, and K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems*, Vol. 12, pp. 279-285.

Makovoz, Y. (1996). Random approximates and neural networks. *Journal of Approximation Theory*, Vol. 85, pp. 98-109.

Meier, L., and S. van de Geer, and P. Bühlmann (2008). The Group Lasso for logistic regression. *Journal of the Royal Statistics Society, Series B*. Vol. 70.

Modha, D., and E. Masry (1996a). Rates of convergence in density estimation using neural networks. *Neural Computation*. Vol. 8, pp. 1107-1122.

Modha, D., and E. Masry (1996b). Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*. Vol. 42, pp. 2133-2145.

Nemirovskii, A.S., B.T. Polyak, and A. B. Tsybakov (1985). Rate of convergence of nonparametric estimates of maximum likelihood type. *Problems in Information Transmission*. Vol. 21, pp. 258-272.

Park, M.Y. and T. Hastie (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B*. Vol. 69, pp.659-677.

Rakhlin, A., D. Panchenko, and S. Mukherjee (2005). Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, Vol. 9, pp. 220-229.

Rényi, A. (1960). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1, pp. 547-561.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*. Vol. 14, pp. 465-471.

Rissanen, J. (1983). A universal prior on integers and estimation by minimum description length. *Annals of Statistics*. Vol. 11, pp. 416-431.

Rissanen, J. (1984). Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory*. Vol. 30, pp. 629-636.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*. Vol. 14, pp. 1080-1100.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Hackensack, NJ, World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*. Vol. 42, No. 1, pp. 40-47.

Shannon, C. (1948). The mathematical theory of communication. *Bell System Technical Journal*. Vol. 27, pp. 379-423,623-656.

Shen, X. (1998). On the method of penalization. *Statistica Sinica*. Vol. 8, pp. 337-357.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*. Vol. 23, No. 3, pp. 3-17.

Silverman, B. (1982). On the estimation of probability function by the maximum penalized likelihood method. *Annals of Statistics*. Vol. 10, pp. 795-810.

Takeuchi, J., and A. R. Barron (1997a). Asymptotically minimax regret for exponential families. In *Proceedings of the Symposium on Information Theory and its Applications*, pp. 665-668.

Takeuchi, J., and A. R. Barron (1997b). Asymptotically minimax regret for exponential and curved exponential families. Fourteen page summary at www.stat.yale.edu/∼arb4/publications.htm for the presentation at the *1998 International Symposium on Information Theory*.

Takeuchi, J., and A. R. Barron (1998). Asymptotically minimax regret by Bayes mixtures. In *Proceedings of the 1998 International Symposium on Information Theory*.

Takeuchi, J., T. Kawabata, and A. R. Barron (2007). Properties of Jeffreys' mixture for Markov Sources. Accepted to appear in the *IEEE Transactions on Information Theory*.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO, *Journal of Royal Statistical Society, Series B*. Vol. 58, pp. 267-288.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, SIAM.

Willett, R., and R. Nowak (2005). Multiscale Poisson intensity and density estimation. www.ece.wisc.edu/%7Enowak/multiscale_poisson.pdf

Wong, W. H., and X. Shen (1995). Probability inequalities for likelihood ratios and convergence rate of sieve estimates. *Annals of Statistics*. Vol. 23, pp. 339-362.

Xie, Q., and A. R. Barron (1997). Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*. Vol. 43, pp. 646-657.

Xie, Q., and A. R. Barron (2000). Asymptotically minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*. Vol. 46, pp. 431-445.

Yang, Y., and A. R. Barron (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*. Vol. 44, pp. 117-133.

Yang, Y., and A. R. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*. Vol. 27, pp. 1564-1599.

Zhang, H., G. Wahba, Y. Lin, M. Voelker, R.K. Ferris, B. Klein (2005). Variable selection and model building via likelihood basis pursuit. *Journal of American Statistical Association*. Vol. 99, pp. 659-672.

Zhang, T. (2003). Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*. Vol. 49, pp. 682-691.

Zhang, T. (2006). From epsilon-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Annals of Statistics*. Vol. 34, pp. 2180-2210.

Zhang, T. (2007). Some sharp performance bounds for least squares regression with ℓ_1 regularization. Manuscript at http://stat.rutgers.edu/∼tzhang/pubs.html