COMPLEXITY REGULARIZATION WITH APPLICATION TO ARTIFICIAL NEURAL NETWORKS

ANDREW R. BARRON
University of Illinois
Departments of Statistics and
Electrical & Computer Engineering
725 South Wright Street
Champaign, Illinois 61820 USA

ABSTRACT. Concepts of universal data compression lead to minimum description-length criteria for parsimonious statistical model selection for the estimation of functions. In this paper we define general complexity regularization criteria and establish bounds on the statistical risk of the estimated functions. These bounds establish consistency, yield rates of convergence, and demonstrate the near asymptotic optimality of the model selection criterion in both parametric and nonparametric cases. A fundamental role is played by an index of resolvability that quantifies the tradeoff between complexity and accuracy of candidate models. Applications are given to polynomial regression and artificial neural networks.

1. Introduction

In the context of statistical estimation of functions, a list of parametric models is provided from which one is to be automatically selected from the data. These lists of parametric models have the property that for essentially any target function there exist a convergent sequence of approximating functions on the list. In the absence of knowledge of the true function, the problem is to estimate an appropriate size model that provides a good tradeoff between the approximation error, which is best for larger models, and the parameter estimation error, which is best for smaller models.

In this paper we present an index of resolvability that defines the best tradeoff between approximation error and the complexity of models. A complexity regularization criterion is defined which adds an information-theoretic complexity penalty to the empirical loss, and the estimator is taken which achieves the optimum value of this criterion. It is shown that the statistical risk of the estimated function is bounded by the index of resolvability.

The methods and theory we present here are closely related to minimum complexity density estimation as developed in Barron and Cover (1990). The difference is that here we do not necessarily stick to the minimum description-length paradigm from Rissanen (1983,1984) and Barron (1985). In that paradigm the model selection criterion is required to correspond to a total description length for the sample. That paradigm works quite well for density estimation, using the fact that the best codes

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 561–576.
© 1991 *Kluwer Academic Publishers. Printed in the Netherlands.*

for a given distribution have length equal to minus the logarithm of the density. The complexity penalty then serves as the length of a preample that describes what estimated distribution is used in the description of that sample. However, for other curve fitting problems we may be interested in estimating, say, a regression function using a penalized squared-error criterion. If the distribution of the errors is unknown (and in particular not Gaussian), then the squared error term in the criterion does not correspond to the length of an accurate code for the data. Nevertheless, regularization of the estimator by the addition of a complexity penalty is still fruitful in this context, despite the lack of a complete justification on grounds of optimal data compression.

A Bayesian formulation of the model selection problem as in Schwarz (1978) yields test statistics that are similar to criteria motivated by the minimum description-length principle. Although closely related, the complexity regularization criterion for the estimation of functions does not necessarily correspond to a Bayesian solution. Again, the reason is that the choice of an empirical loss function is not constrained to correspond to the log-likelihood for a model of the error distribution.

Other statistical performance criteria have been analyzed for their asymptotic statistical properties. In particular, Shibata (1981) has demonstrated asymptotic optimality properties of the Akaike's final prediction error and Akaike's AIC criterion in the case of Gaussian errors. For more general error distributions, Li (1987) demonstrates asymptotic optimality properties of Mallows' C_p criterion, cross-validation, and generalized cross-validation. Li's analysis applies only to linear models and he requires a condition that indirectly restrict the number of candidate models that may be considered. The analysis that we give here is similar in spirit to the work of Shibata and Li.

An advantage of the complexity regularization criterion is that it is amenable to statistical analysis for essentially any list of candidate functions. In particular, linearity of the parametric models need not be assumed. Also, no restriction need be placed on the number of candidate models that is considered for each sample size. In essence the effect of the number of candidate models is automatically accounted for in the complexity penalty. To permit this increased generality, our bounds do not yield asymptotic optimality in the sense of Shibata for the nonparametric case. Nevertheless, the bounds do show that the statistical risk is within a logarithmic factor of the optimum. Moreover, our results are seen to be applicable both for parametrically represented functions and for nonparametric functions, without prior knowledge of whether the true function is in the list of parametric models.

Below we give two examples to which the convergence theory applies. The theory is not restricted to these cases. These examples provide lists of parametric families of functions $f_M(x,\theta)$ for $x \in R^d$ and $\theta \in R^m$, where the dimension m depends on the model M. As a concession to computational feasibility, the size of the list of models may be restricted to depend on the sample size n, as long as the restrictions are relaxed as $n \to \infty$. Statistical and information-theoretical considerations show that there is no essential loss to restrict the coordinates of θ to a grid of points spaced at width equal to a small multiple of $1/\sqrt{n}$, where n is the sample size. Because of the discretization of the parameters, the list of models leads to a countable list Γ_n of candidate functions.

For each candidate function f, a complexity $C_n(f)$ is assigned. In general, these numbers $C_n(f)$ are arbitrary subject to a summability restriction on $2^{-C_n(f)}$. This restriction means that $C_n(f)$ may be interpreted as a codelength for the description of f, or, what turns out to be operationally equivalent but motivationally quite distinct, $2^{-C_n(f)}$ may be interpreted as a prior probability of f. For the examples we give, the complexities may be chosen to take on the following form

$$C_n(f_M(\cdot,\theta)) = \frac{m}{2}\log n + c_\theta + c_M, \qquad (1)$$

where the logarithm is taken with base 2. Here $(m/2)\log n + c_\theta$ may be interpreted as a codelength for the parameter θ, or 2^{-c_θ} as a prior density function for the parameter. Likewise c_M may be interpreted as a codelength for the model M, or 2^{-c_M} as a prior probability, where $\sum_M 2^{-c_M} \le 1$. The term c_M indirectly accounts for the effect of the number of candidate models. When there are many competing models, this term help to avoid overfit problems due to selection bias.

The examples we study have the property that essentially every function f^* is well approximated by functions in the list Γ_n. In particular, for every L^2 function with compact domain in R^d, there is a sequence of functions f_n in Γ_n, such that the L^2 approximation error converges to zero as $n \to \infty$. A consequence is that the index of resolvability $R_n(f^*) = \min_{f \in \Gamma_n}(\| f - f^* \|^2 + C_n(f)/n)$, which quantifies the tradeoff between the accuracy and complexity of the best approximations, is seen to converge to zero. The speed at which it converges depends on how well tailored the list Γ_n is to the function f^*. If f^* is a member of one of the candidate families then $R_n(f^*) = O(\log n)/n$. In other smooth cases, $R_n(f^*)$ converges at rate given by a fractional power of $(\log n)/n$. The theory we develop here gives conditions such that the statistical risk of estimated functions converges at rate bounded by $R_n(f^*)$.

EXAMPLE 1.1: MULTIVARIATE BASIS FUNCTION SELECTION

Let $B_k(x)$, $x \in R^d$, $k \in \{0,1,2,...\}^d$ denote a collection of basis functions for multivariate functional expansion. In particular, we may take basis functions for polynomials, splines, or trigonometric series. For example, $B_k(x) = x_1^{k_1} \cdots x_d^{k_d}$ is the traditional basis for polynomials. Finite sets M of multi-indices k yield candidate models

$$f_M(x,\theta) = \textstyle\sum_{k \in M} \theta_k B_k(x). \qquad (2)$$

The lists of candidate models consist of collections of such f_M. In theory one may take the collection of all finite subsets M, and use the complexity regularization criterion to choose a statistically suitable subset. However, in practice it is more feasible to impose sequences of bounds on the maximal degree and the interaction order of the basis functions, and perhaps to restrict to hierarchical sets of terms, thereby yielding a sequence of lists Γ_n of candidate functions. Theoretically, the sequence of bounds should be fairly large (and tend to infinity as $n \to \infty$), so that we do not interfere with the desired approximation properties. The complexity regularization criterion makes the final choice of model within the possibly restricted class.

There are various statistical contexts in which such basis function expansions may be used, including least squares regression (see Cox (1988), Friedman (1990)), maximum likelihood estimation of the log-density (see Stone (1990), Barron and Sheu (1988)), and logistic regression for nonparametric discrimination. As we shall see, complexity regularization and indices of resolvability are defined for each of these contexts.

Suppose the target function f^* is in the Sobolev space of functions on $[0,1]^d$ for which all partial derivatives up to order r are square integrable. In the contexts of regression and log-density estimation using polynomials, splines, or trigonometric series, the index of resolvability converges at rate

$$R_n(f^*) = O(\frac{\log n}{n})^{2r/(2r+d)}. \tag{3}$$

This is shown for $d=1$ in Barron and Cover (1990) and extends by the same method to $d>1$ using multivariate approximation results in Sheu (1989).

These bounds on the index of resolvability then apply, using the main results from Section 4, to show that the statistical risk of the estimated function converges to zero at rate $O(\log n)/n$ in the case that the true function is in one of the candidate families and at rate $O(\log n/n)^{2r/(2r+d)}$ when the function is in the Sobolev space. Again these rates are obtained without apriori knowledge of the family or of the smoothness class.

EXAMPLE 1.2: ARTIFICIAL NEURAL NETWORKS

Feedforward artificial neural networks are a relatively flexible parametric family of functions $f(x,\theta)$ on R^d, which are expressed as the mathematical composition of fixed non-linear functions of one variable and parameterized linear functions of several variables. The most popular non-linear function in this context is the logistic $h(u) = 1/(1+e^{-u})$ for $-\infty < u < \infty$, although polynomial nonlinearities have also had some practical success. The references Farlow (1984), Rumelhart, et. al. (1986), Anderson and Rosenfeld (1988), Barron and Barron (1988), Lippmann (1987), and Lee and Lippmann (1990) provide a starting point on network methods of empirical modeling. Despite the name "neural," these networks are intended for general purpose approximation and estimation of functions; they are connected to biological networks only through a loose analogy. Artificial neural networks are intended as competitors to other methods of nonparametric function estimation, particularly in high-dimensional contexts in which the networks have an advantage of dimensionality reduction by composition. That is, if the true response is a function of many variables, that depends on these variables through the composition of lower-dimensional smooth functions, then network methods may be better suited than traditional series expansions which suffer from an exponentially large the number of candidate terms as a function of the dimension.

The architecture of an artificial neural network is the specification of the mathematical composition scheme. The layers of the network refer to the depth of composition. A basic network architecture is the multiple layer networks in which the output of the functions on one layer are inputs to every function on the next layer. Thus

$$f(x,\theta) = \sum_{m=1}^{M} \alpha_m g_m(x,\beta_m), \tag{4}$$

where each parameterized function $g(x,\beta)$ takes the form

$$g(x,\beta) = h(\beta^T z), \tag{5}$$

with $z = (1,z_1,...,z_K)$ for some K, and $\beta \in R^{K+1}$. These elemental functions (5) are called units, nodes, or artificial neurons. In the case of one layer of non-linear nodes, the $z_k = x_k$ are original input variables. In the case of two-layer networks, each intermediate input z_k is itself the output of a another parameterized function of the form (5). This process is repeated for any additional layers of the network. The parameter vector θ consists of the linear parameters α and the parameters β from all of the nodes in the network.

Cybenko (1989) demonstrated an essential approximation-theoretic property of families of networks with the logistic non-linearity (or other bounded non-linearities with distinct left and right limits). Namely, the set of such networks with one layer of non-linear nodes is dense in the space of all bounded continuous functions with compact domain in R_d. It follows that classes of multiple layer networks are also dense. Similar results show the density of networks in the space of L^2 functions with compact domain and show that the density of networks holds for unbounded non-linearities such as polynomials, see Barron (1989). Very little is known about the rates of approximation of networks. One result in this direction, based on the work of Jones (1990), is that if a sequences of of network functions exist for which $\| f - f^* \|$ tends to zero and if the sum of the absolute values of the α parameters is bounded, then there exists network functions f_M of the form (4) with integrated squared error bounded by $\| f_M - f^* \|^2 \le O(1/M)$.

Without appropriate statistical control on the size of an artificial neural network and a control on the number of estimated parameters, artificial neural networks are subject to the same overfit problems (inadequacy of the generalization of the predictions to new observation) that are associated with other parameter intensive models, such as the models in Example 1.1 above. To realize the potential benefits of neural network approximations, it is important to have a statistically accurate model selection criterion. Moreover, the criterion should be demonstrably applicable to network models that can be highly nonlinear functions of the parameters. These considerations were motivating factors in the development and analysis of the complexity regularization criterion.

The statistical convergence theory we present shows that artificial neural network estimated from data achieve the best rates of approximation, as measured by the index of resolvability for the given class of networks. This convergence theory does not necessarily hold for arbitrary neural network training algorithms; it applies to network estimators that optimize the complexity regularization criterion.

Results on consistency of statistically estimated networks were announced in Barron and Barron (1988). See also, White (1990) for consistency results for networks based on the method of sieves. The bounds on the statistical risk of functions estimated by complexity regularization that are proved in the present paper were announced in the neural network context in Barron (1989).

2. Loss Functions for Functional Estimation

Before giving our convergence result for functions estimated by complexity regularization, we address the issue of the choice of the loss function and the measure of statistical risk.

Let $(X_i,Y_i)_{i=1}^{n}$ be independent observations drawn from the unknown joint distribution of random variables X,Y, where the support of X is in $\mathbf{R}^d$. Here X is the vector of explanatory variables and Y is the response variable. Functions $f(X)$ are used in predicting the response. The error incurred by a prediction is measured by a distortion function $d(Y,f(X))$. (Informally d is sometimes referred to as a loss function, although in a strict decision-theoretic sense it is not, since it depends on the random variables X,Y). The empirical loss is $(1/n)\sum_{i=1}^{n} d(Y_i,f(X_i))$. Functions f_n are typically estimated by minimizing the empirical loss over a class of candidate functions, or by minimizing the empirical loss with a penalty added for the complexity of the functions.

We let f^* be a function which minimizes $E(d(Y,f(X)))$ over all measurable functions on $\mathbf{R}^d$. For the distortion functions we investigate, such a optimum function $f^*(x)$ exists and is related explicitly to the conditional distribution of Y given $X = x$. When a function f is used in place of the optimum function f^* the regret is measured by the difference between the expected distortions

$$r(f,f^*) = E(d(Y,f(X))) - E(d(Y,f^*(X))). \tag{6}$$

In our formulation, the quantity $r(f,f^*)$ is the theoretical loss incurred by the function f when the true best function is f^*. This loss function $r(f,f^*)$ depends on the function f over its whole domain (the support of the distribution of X) and not just on its values at one or several points. It quantifies the ability of an estimate to generalize, on the average, to new data from the distribution of X,Y. The statistical risk is the expected value of the loss for a given estimator f_n,

$$\text{risk} = E(r(f_n,f^*)). \tag{7}$$

More generally, we define $d^* = \inf_f E(d(Y,f(X)))$, $r(f) = E(d(Y,f(X))) - d^*$, and risk $= E(r(f_n))$, all of which are defined even if there does not exist an optimal function f^*. This $r(f)$ is the same as the loss function $r(f,f^*)$ in the case that an optimal function $f^*(x)$ exists.

The most common choice for the distortion function $d(Y,f(X))$ is the squared error $(Y-f(X))^2$, although other choices can also be handled in the theory, including the absolute error $|Y-f(X))|$. In the case of a dichotomous random variable Y, assumed to take values of either $+1$ or -1, reasonable choices include, in addition to the squared error, the zero-one distortion function $1_{\{Y \neq \mathrm{sgn}(f(X))\}}$, and the logistic distortion function $-Yf(X) + \log(e^{f(X)} + e^{-f(X)})$, which is the same, except for a linear rescaling as the distortion function used in traditional logistic regression.

A general class of distortion functions are those which take the form $d(Y,f(X)) = -\log p(Y \mid f(X))$ where $p(y \mid f(x))$ models the conditional density of Y given X. In particular, the squared error distortion corresponds to a Gaussian conditional density and the logistic distortion corresponds to the Bernoulli model written in exponential form $p(y \mid x) = e^{yf(x)}/(e^{f(x)} + e^{-f(x)})$, $y = \pm 1$, where $f(x)$ models one-half the log-odds ratio in favor of class $+1$ versus class -1. For the estimation of the density function for the random vector X, we use $d(Y,f(X)) = -\log f(X)$ and require that $f(x)$ is non-negative and integrates to one with respect to a specified dominating measure.

For the squared error distortion, the loss function reduces to an integrated squared error

$$r(f,f^*) = E((f(X)-f^*(X))^2) \tag{8}$$

and $f^*(x) = E[Y \mid X=x]$ is the conditional mean of Y given $X = x$. In particular, for the dichotomous case with squared error loss, $f^*(x)$ is the difference between the conditional probabilities of class $+1$ and class -1. For the zero-one loss, $r(f,f^*)$ is the difference between the probability of error based on f and the Bayes optimal probability of error.

For distortion functions based on a family of conditional densities, f^* is the choice which makes the conditional density $p(y \mid f^*(x))$ closest to the true conditional density in the relative entropy sense. If the family of conditional densities is correctly specified, i.e. if it includes the true conditional density, then $r(f,f^*)$ is the average relative entropy distance between $p(\cdot \mid f^*(x))$ and $p(\cdot \mid f(x))$.

$$r(f,f^*) = E \log \frac{p(Y \mid f^*(X))}{p(Y \mid f(X))}. \tag{9}$$

In this context, a role will also be played by the squared Hellinger distance,

$$d_H^2(f(x),f^*(x)) = \int (\sqrt{p(y \mid f(x))}-\sqrt{p(y \mid f^*(x))}\,)^2 \,\mu(dy), \tag{10}$$

where μ is the measure assumed to dominate the family of conditional distributions for Y given X.

3. Complexity Regularization and the Index of Resolvability

We have already indicated that the complexity regularization criterion is motivated in certain cases by the minimum description-length principle. Indeed, these criteria are seen to coincide when the distortion function d is taken to be equal to minus the logarithm of likelihood and when the parameter λ in the definition of the criterion is set to equal one. Here we motivate the criterion in the context of uniform bounds on the statistical risk of estimators.

There are two contributions to the risk $E(r(f_n,f^*))$ of estimators based on the optimization of an empirical loss: namely, the approximation error $r(f,f^*)$ achieved by functions f in the given class as an approximation to the desired function f^*, and the estimation error, which is due to the discrepancy between the empirical and theoretical loss. By techniques in Vapnik (1982), Devroye (1988), or Haussler (1989), this discrepancy between empirical and theoretical averages can be shown to be uniformly bounded by $O(\sqrt{C_n/n}\,)$, in probability, for families of functions of complexity bounded by C_n. (Essentially, C_n is taken there as the logarithm of the number of functions required to approximate functions in the class to within a prescribed accuracy.) By generalizations of these techniques, it is shown that the estimation error can be bounded in probability by a multiple of $\sqrt{C_n(f)/n}$ for arbitrary bounded distortion functions and by $r(f,f^*) + C_n(f)/n$ for the squared error and for the likelihood-based distortion functions, uniformly for all candidate functions f. Here, instead of requiring a uniform complexity bound, we permit unbounded complexities $C_n(f)$ that may depend on the candidate functions. (These "complexities" $C_n(f)$ are arbitrary numbers satisfying a summability requirement, as given in equation (15) below, in accordance with an information-theoretic interpretation.) In the absence of constraints on the form or number of candidate functions, the empirical loss is not necessarily uniformly accurate for all candidate functions. The complexity penalties used here is the smallest penalty we are able to impose to force the criterion to be accurate at least for the functions which achieve the minimum criterion value.

Thus we are led to complexity regularization criteria and to corresponding indices of approximation. Depending on whether bounds of order $\sqrt{C_n(f)/n}$ or $C_n(f)/n$ arise in controlling the estimation error, we add the appropriate complexity penalty to the empirical loss to define the criterion for function estimation, for in this way it seen that the minimizer of the empirical criterion has a performance essentially as good as that achievable by the theoretical analog of the criterion.

Definition: Given a collection Γ_n of functions, numbers $C_n(f)$, $f \in \Gamma_n$, satisfying the summability condition (15), and a positive constant λ, the method of *complexity regularization* selects the function to minimize one of the following two criteria,

$$\frac{1}{n}\sum_{i=1}^{n} d(Y_i,f(X_i)) + \lambda(\frac{1}{n}C_n(f))^{1/2} \tag{11}$$

or

$$\frac{1}{n}\sum_{i=1}^{n} d(Y_i,f(X_i)) + \lambda\frac{1}{n}C_n(f). \tag{12}$$

The theoretical analog of these criteria leads to the following indices of resolvability, which quantify the tradeoff between the complexity and accuracy of approximations to f^*,

$$R_n^{(1)}(f^*) = \min_{f\in\Gamma_n}(r(f,f^*) + \lambda(\frac{1}{n}C_n(f))^{1/2}), \tag{13}$$

and

$$R_n^{(2)}(f^*) = \min_{f\in\Gamma_n}(r(f,f^*) + \lambda\frac{1}{n}C_n(f)). \tag{14}$$

The latter quantity is the index of resolvability introduced in an information-theoretic context in Barron and Cover (1990).

The requirement that is imposed on the numbers $C_n(f)$ is the following summability condition:

$$\sum_{f\in\Gamma_n} 2^{-C_n(f)} \le s, \tag{15}$$

for some finite constant s. There is an information-theoretic interpretation of the summability condition with $s = 1$: this is the Kraft-McMillan inequality which is necessary and sufficient for the existence of uniquely decodable binary codes, with codelengths $C_n(f)$, for $f \in \Gamma_n$. We shall also require that $C_n(f) \ge l$ for all f and all n, for some positive constant l. This also is automatically satisfied, with $l = 1$ in the case of binary codes for a set Γ_n with more than one function. A trivial consequence of the latter requirement is that $R_n^{(1)}$ is not of order smaller than $1/\sqrt{n}$ and $R_n^{(2)}$ is not of order smaller than $1/n$.

4. Bounds on the Statistical Risk

Now we present the main results on statistical convergence properties of functions using a complexity regularization criterion. Let

$$f_n^{(1)} = \underset{f\in\Gamma_n}{\operatorname{argmin}}(\frac{1}{n}\sum_{i=1}^{n} d(Y_i,f(X_i)) + \lambda(\frac{1}{n}C_n(f))^{1/2}), \tag{16}$$

and

$$f_n^{(2)} = \underset{f\in\Gamma_n}{\operatorname{argmin}}(\frac{1}{n}\sum_{i=1}^{n} d(Y_i,f(X_i)) + \lambda\frac{1}{n}C_n(f)). \tag{17}$$

The first estimator is used for distortion functions such as the zero-one distortion and the absolute error for which the ideal rate of convergence of the risk would be close to $1/\sqrt{n}$. The second estimator is used for squared error and log-likelihood based distortions for which close to $1/n$ would be the ideal rate.

Our main result, which we now give, shows that the statistical rate of convergence of functions estimated by complexity regularization is bounded by the index of resolvability.

The result requires in some cases that $d(Y,f(X))$ be almost surely bounded. This is forced by a constraint on the support of Y and by clipping the functions $f(X)$, or by explicit choice of a bounded distortion function. For distortions based on the log-likelihood, with a correctly specified family of conditional densities, no boundedness of the distortion is required.

Convergence Theorem for Complexity Regularization: *Assume that the indices of approximation $R_n^{(1)}(f^*)$ and $R_n^{(2)}(f^*)$ tend to zero as $n\to\infty$. If the range of $d(Y,f(X))$ for every f in Γ_n is in a fixed interval of length b, and if $\lambda > b/\sqrt{2\log e}$ in the definition of the complexity regularized estimator $f_n^{(1)}$, then the statistical risk of the estimator converges to zero at rate bounded by $R_n^{(1)}(f^*)$, i.e.,*

$$E(r(f_n^{(1)},f^*)) \le O(R_n^{(1)}(f^*)). \tag{18}$$

Indeed, for all $n \ge 1$,

$$E(r(f_n^{(1)},f^*)) \le R_n^{(1)}(f^*) + \frac{c_0}{\sqrt{n}}, \tag{19}$$

where $c_0 = (s+1)b\sqrt{\pi/2}$.

For the squared error distortion function, if the support of Y and the range of each function $f(X)$ is in a known interval of length b, then with $\lambda > 5b^2/3\log e$ in the definition of the estimator $f_n^{(2)}$, the mean squared error converges to zero at rate bounded *by $R_n^{(2)}(f^*)$, i.e.,*

$$E((f_n^{(2)}(X) - f^*(X))^2) \leq O(R_n^{(2)}(f^*)). \tag{20}$$

If the distortion function is $d(Y,f(X)) = -\log p(Y \mid f(X))$ where the true conditional density is $p(Y \mid f^*(X))$, then for all $\lambda > 1$ in the definition of the estimator $f_n^{(2)}$, the expected squared Hellinger distance between the conditional densities converges at rate bounded by the index of resolvability $R_n^{(2)}(f^*)$, i.e.,

$$E(d_H^2(f_n^{(2)}(X),f^*(X))) \leq O(R_n^{(2)}(f^*)). \tag{21}$$

The L^1 distance, which takes the form $\int \mid p(y \mid f(x)) - p(y \mid f^*(x)) \mid \mu(dy)$, is known to not be greater than twice the Hellinger distance. Therefore, a consequence of (21) is that the expected square of the L^1 distance also converges at rate bounded by $R_n^{(2)}(f^*)$.

For the Gaussian error case, the Hellinger distance can be evaluated and lower bounded as in Barron and Cover (1990). It is seen that for any $c > 0$, the risk $E(\min((f_n(X)-f^*(X))^2,c))$ converges to zero at rate bounded by $R_n^{(2)}(f^*)$.

5. Proof of the Convergence Theorem for Complexity Regularization

For simplicity in the proof we assume, without loss of generality, that the complexities are converted to base e. This means that the summability condition becomes $\sum_{f \in \Gamma_n} e^{-C_n(f)} \leq s$. Accordingly, for the purpose of the proof, the logarithms in the theorem are now interpreted as base e instead of base 2.

For the first two conclusions of the theorem, the proof uses a bound on the probability of the event that $r(f_n,f^*) > t$ in terms of a sum of probabilities of related events for each $f \in \Gamma_n$, to which inequalities of Hoeffding and Bernstein can be applied. The bounds on the probabilities are then integrated for $t > 0$ to obtain the indicated bounds on the risk. The proof of the third conclusion is based in part on results in Barron and Cover (1990) which uses inequalities of Chernoff.

The inequality of Hoeffding (1963, Theorem 2) states that if $(U_i)_{i=1}^n$ are independent random variables taking values in intervals of length b, then the distribution of the sample average $\overline{U} = (1/n)\sum_{i=1}^n U_i$ has the following exponential bound for all $\varepsilon > 0$,

$$P\{\overline{U} - E\overline{U} \geq \varepsilon\} \leq e^{-2n\varepsilon^2/b^2}. \tag{22}$$

The Bernstein-type inequalities make direct use of the variance as well as the expected values of the random variables. To state the Bernstein inequality, let U_i be independent random variables that satisfy the moment condition that for some $h > 0$,

$$E \mid U_i - EU_i \mid^k \leq \frac{\operatorname{var}(U_i)}{2} k!\, h^{k-2} \tag{23}$$

for $k \geq 2$, $i=1,\ldots,n$. This is satisfied in particular if $|U_i - EU_i| \leq M$ with $h = M/3$. Bernstein's inequality states that for $t > 0$,

$$P\{\overline{U} - E\overline{U} \geq t\sigma_{\overline{U}}\} \leq \exp\{-t^2/(2 + 2ht/n\sigma_{\overline{U}})\} \tag{24}$$

see Craig (1933), Bennett (1962).

5.1. PROOF OF THE THEOREM: BOUNDED LOSS FUNCTION CASE

We examine the theoretical and empirical loss,

$$r(f) = Ed(Y,f(X)) - d^* \tag{25}$$

and

$$\hat{r}_n(f) = \frac{1}{n}\sum_{i=1}^n d(Y_i,f(X_i) - d^*, \tag{26}$$

where we have subtracted the constant $d^* = \inf_f Ed(Y,f(X))$. By Hoeffding's inequality and the union of events bound, for any $\varepsilon_n(f) > 0$,

$$r(f) - \hat{r}_n(f) < \varepsilon_n(f) \quad \text{for all } f \in \Gamma_n, \tag{27}$$

except in an event of probability not greater than $\sum_{f \in \Gamma_n} e^{-2n(\varepsilon_n(f)/b)^2}$. Given $\delta > 0$, we choose $\varepsilon_n(f)$ such that $2n(\varepsilon_n(f)/b)^2 = C_n(f) + \ln 1/\delta$, to obtain that

$$r(f) - \hat{r}_n(f) < \frac{b}{\sqrt{2}}\left(\frac{C_n(f)}{n} + \frac{\ln 1/\delta}{n}\right)^{1/2} \quad \text{for all } f \in \Gamma_n \tag{28}$$

except in a set of probability not greater $\delta \sum_{f \in \Gamma_n} e^{-C_n(f)}$, which by the assumption on the complexities $C_n(f)$ is not greater than $s\delta$.

For the estimator f_n defined to achieve the minimum value of $\hat{r}_n(f) + \lambda(C_n(f)/n)^{1/2}$ and for $\lambda > b/\sqrt{2}$, we have that the following bounds hold on the loss $r(f_n)$, except in a set of probability not greater than $s\delta$,

$$\begin{aligned} r(f_n) &< \hat{r}_n(f_n) + \lambda\left(\frac{C_n(f_n)}{n}\right)^{1/2} + \frac{b}{\sqrt{2}}\left(\frac{\ln 1/\delta}{n}\right)^{1/2} \\ &\leq \hat{r}_n(f_n^*) + \lambda\left(\frac{C_n(f_n^*)}{n}\right)^{1/2} + \frac{b}{\sqrt{2}}\left(\frac{\ln 1/\delta}{n}\right)^{1/2}. \end{aligned} \tag{29}$$

Taking f_n^* to be a function that achieves the best resolvability, that is, a function minimizing $r(f) + \lambda(C_n(f)/n)^{1/2}$, and applying Hoeffding's inequality once more, to get that $\hat{r}_n(f_n^*) \leq r(f_n^*) + \lambda((\ln 1/\delta)/n)^{1/2}$ except in a set of probability not greater than δ, we obtain that

$$r(f_n) < r(f_n^*) + \lambda\left(\frac{C_n(f_n^*)}{n}\right)^{1/2} + 2\frac{b}{\sqrt{2}}\left(\frac{\ln 1/\delta}{n}\right)^{1/2}, \tag{30}$$

except in a set of probability not greater than $(s+1)\delta$. This shows that the loss of the estimator is bounded in terms of the index of resolvability,

$$r(f_n) \leq R_n^* + O\left(\frac{1}{n}\right)^{1/2}, \tag{31}$$

in probability, where the index of resolvability is

$$R_n^* = \min_{f \in \Gamma_n}(r(f) + \lambda(C_n(f)/n)^{1/2}). \tag{32}$$

In particular, setting $\delta = e^{-(1/2)nt^2/b^2}$ for $t > 0$, we have

$$P\{r(f_n) \geq R_n^* + t\} \leq (s+1)e^{-(1/2)nt^2/b^2}. \tag{33}$$

Integrating for $0 < t < \infty$, yields,

$$\begin{aligned} E\, r(f_n) - R_n^* &\leq \int_0^\infty P\{r(f_n) - R_n^* \geq t\}\, dt \\ &\leq (s+1)\int_0^\infty e^{-(1/2)nt^2/b^2} dt \\ &= (s+1)\frac{b\sqrt{\pi/2}}{\sqrt{n}} \\ &= \frac{c_0}{\sqrt{n}}. \end{aligned} \tag{34}$$

Thus for all $n \geq 1$, the risk of the estimator is bounded by

$$E\, r(f_n) \leq R_n^* + \frac{c_0}{\sqrt{n}}. \tag{35}$$

This completes the proof of the bounds on the statistical risk of the complexity regularization estimator for bounded loss functions.

5.2. PROOF OF THE THEOREM: SQUARED ERROR CASE

For the squared error loss function, the ideal rate of convergence is close to $1/n$ instead of $1/\sqrt{n}$. For this case we use the criterion and the index of resolvability with penalty $\lambda C_n(f)/n$ instead of $\lambda(C_n(f)/n)^{1/2}$ and instead of Hoeffding's inequality we use a variant of Bernstein's inequality.

Direct use of Bernstein's inequality for our purposes is workable but cumbersome. We find it easier to use the following inequality, that Craig (1933) develops in his proof of Bernstein's inequality. If U_i are independent random variables satisfying Bernstein's moment condition, then

$$P\{\overline{U} - E\overline{U} \geq \frac{\tau}{n\varepsilon} + \frac{n\varepsilon\,\mathrm{var}(\overline{U})}{2(1-c)}\} \leq \exp\{-\tau\}, \tag{36}$$

for $0 < \varepsilon h \leq c < 1$ and $\tau > 0$.

Now to treat the complexity regularization estimator with the squared error distortion function, denote the difference in empirical loss at f and f^* by

$$\begin{aligned} \hat{r}_n(f, f^*) &= \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f^*(X_i))^2 \\ &= -\frac{1}{n}\sum_{i=1}^n U_i, \end{aligned} \tag{37}$$

where

$$U_i = -(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2.$$

Under the assumption that $f(X_i)$ and Y_i take values in a fixed interval of length b, we have that Bernstein's condition is satisfied with $h = 2b^2/3$. Next we bound the variance of U_i. To this end, we expand the square in the definition of U_i to get

$$U_i = 2(Y_i - f^*(X_i))(f(X_i) - f^*(X_i)) - (f(X_i) - f^*(X_i))^2. \tag{38}$$

The covariance between these two terms is zero. The expected square of the first term is $4E((\sigma^2_{Y|X})(f(X) - f^*(X))^2)$ which is not greater than $b^2E(f(X) - f^*(X))^2$ (since the variance of a distribution concentrated on an interval of length b is not greater than the variance $b^2/4$ achieved by the distribution that places mass 1/2 at each endpoint). The expected square of the second term is also bounded by $b^2E(f(X) - f^*(X))^2$. Together these bounds yield

$$\mathrm{var}(U_i) \leq 2b^2 r(f, f^*). \tag{39}$$

It follows that $n\ \mathrm{var}(\overline{U}) \leq 2b^2 r(f, f^*)$. Now apply the union of events bound and the Bernstein-type inequality with $\tau = C_n(f) + \ln 1/\delta$ and $\varepsilon = 1/\lambda$, to obtain that

$$r(f, f^*) - \hat{r}_n(f, f^*) < \lambda\frac{C_n(f)}{n} + \frac{b^2}{\lambda(1-c)} r(f, f^*) + \lambda\frac{\ln 1/\delta}{n} \tag{40}$$

for all f in Γ_n, except in an event of probability not greater than $s\delta$. Set $c = \varepsilon h = 2b^2/(3\lambda)$. The assumption that $\lambda > 5b^2/3$ implies that $\alpha = b^2/(\lambda(1-c)) < 1$. We collect the terms involving $r(f, f^*)$ on the left side, and evaluate at the complexity regularization estimator f_n to obtain

$$\begin{aligned} (1-\alpha) r(f_n, f^*) &\leq \hat{r}_n(f_n, f^*) + \lambda\frac{C_n(f_n)}{n} + \lambda\frac{\ln 1/\delta}{n} \\ &\leq \hat{r}_n(f_n^*, f^*) + \lambda\frac{C_n(f_n^*)}{n} + \lambda\frac{\ln 1/\delta}{n}, \end{aligned} \tag{41}$$

where f_n^* is a function that achieves the best resolvability. Applying the Bernstein-type inequality once more, but now with $\tau = \ln 1/\delta$, to get that $\hat{r}_n(f_n^*, f^*) \leq r(f_n^*, f^*) + \alpha r(f_n^*, f^*) + \lambda(1/n)\ln 1/\delta$ except in an event of probability not greater that δ, we obtain that

$$(1-\alpha) r(f_n, f^*) \leq (1+\alpha) r(f_n^*, f^*) + \lambda\frac{C_n(f_n^*)}{n} + 2\lambda\frac{\ln 1/\delta}{n}, \tag{42}$$

except in an event of probability not greater than $(s+1)\delta$. Dividing through by $(1-\alpha)$, and using the definition of the index of resolvability, we have that

$$r(f_n, f^*) \leq \frac{1+\alpha}{1-\alpha} R_n(f^*) + 2\lambda\frac{\ln 1/\delta}{n}\frac{1}{1-\alpha}, \tag{43}$$

except in an event of probability not greater than $(s+1)\delta$. Setting $\delta = e^{-nt/(2\lambda)}$ and integrating the probability for $0 < t < \infty$ as in the previous case, we conclude that for all $n \geq 1$,

$$E(r(f_n, f^*)) \leq \frac{1+\alpha}{1-\alpha} R_n(f^*) + \frac{2\lambda}{n}\frac{s+1}{1-\alpha}. \tag{44}$$

This completes the proof of the bound on the mean squared error of the complexity regularization estimator.

5.3. PROOF OF THE THEOREM: LIKELIHOOD CASE

The third conclusion of the Theorem is essentially given in Barron and Cover (1990). There it is shown that the squared Hellinger distance converges in probability at rate given by the index of resolvability. Here we complete the reasoning to show that the expected squared Hellinger distance also converges at the same rate.

Let $P = P_{Y \mid f^*(X)} P_X$ be the true distribution of X, Y, let $P_n = P_{Y \mid f_n(X)} P_X$ be the joint distribution obtained by plugging in the minimum complexity estimator f_n, and let $P_n^* = P_{Y \mid f_n^*(X)} P_X$ be the joint distribution obtained by plugging in a function f_n^* that achieves the best resolvability. Using inequalities of Chernoff, exponential bounds are derived in Barron and Cover (1990) for the P_n^* probability that the ratio of the squared Hellinger distance (between f_n and f_n^*) and the index of resolvability is greater than an arbitrary positive constant. In particular,

$$P_n^*\{(1-\alpha)\frac{d_H^2(f_n, f_n^*)}{R_n(f^*)} > c\} \leq s\, e^{-cnR_n(f^*)} e^{\lambda C_n(f_n^*)}, \tag{45}$$

where $\alpha = 1/\lambda$. Applying the Lemma with $U = (1-\alpha) d_H^2(f_n, f_n^*)/R_n(f^*)$, with $r_1 = nR_n(f^*)/2$ and $r_2 = \lambda C_n(f_n^*)$, it follows that

$$\begin{aligned}(1-\alpha)\frac{E(d_H^2(f_n, f_n^*))}{R_n(f^*)} &\leq s\int_0^\infty e^{-cnR_n(f^*)/2} dc + \frac{nr(f_n^*, f^*) + \lambda C_n(f_n^*)}{nR_n(f^*)/2} + \frac{e^{-1}}{nR_n(f^*)/2} \\ &\leq \frac{s + e^{-1}}{nR_n(f^*)/2} + 2 \\ &\leq 2\frac{s + e^{-1}}{l} + 2,\end{aligned} \tag{46}$$

where we have used the fact that, in the present context, $r(f_n^*, f^*)$ is the relative entropy distance between P and P_n^*. Also, in the last line we used $nR_n(f^*) \geq C_n(f_n^*) \geq l$. We conclude that the expected squared Hellinger distance between f_n and f_n^* is bounded by a constant times the index of resolvability, for all $n \geq 1$,

$$E(d_H^2(f_n, f_n^*)) \leq c_1 R_n(f^*), \tag{47}$$

where the constant is $c_1 = 2(1+(s+e^{-1})/l)/(1-\alpha)$. Using the triangle inequality to get $d_H^2(f_n, f^*) \leq 2(d_H^2(f_n, f_n^*) + d_H^2(f^*, f_n^*))$ and then using the fact that the squared Hellinger distance $d_H^2(f^*, f_n^*)$ is bounded by the relative entropy $r(f_n^*, f^*)$ which in turn is less than $R_n(f^*)$, we obtain

$$E(d_H^2(f_n, f^*)) \leq 2(c_1 + 1)R_n(f^*). \tag{48}$$

This completes the proof of the Theorem.

Lemma: For any nonnegative random variable U, any pair of distributions P and Q, and constants $r_1 > 0$, and $r_2 \geq 0$,

$$E(U) \leq e^{-r_2}\int_0^\infty Q\{U > c\} e^{r_1 c} dc + \frac{D(P \| Q) + r_2 + e^{-1}}{r_1}, \tag{49}$$

where the expectation is with respect to P, and $D(P \| Q)$ is the relative entropy distance between P and Q.

Proof of the Lemma: The inequality is trivial if the relative entropy distance is infinite. Now suppose $D(P \| Q)$ is finite, so there is a density ratio dP/dQ. By a simple calculation as in Barron and Cover (1990, Lemma 2),

$$P\{U > c\} \leq Q\{U > c\} e^{r_1 c - r_2} + P\{\frac{1}{r_1}(r_2 + \log\frac{dP}{dQ}) > c\}. \tag{50}$$

Integrating for $c > 0$ then gives

$$E(U) \leq 2^{-r_2}\int Q\{U > c\} e^{cr_1} dc + \frac{1}{r_1}E(r_2 + \ln\frac{dP}{dQ})^+. \tag{51}$$

Using the fact that $E(\ln dP/dQ)^+ \leq D(P \| Q) + e^{-1}$ completes the proof of the Lemma.

6. Acknowledgement

This work is supported by an Office of Naval Research Contract N00014-89-J-1811.

7. References

Anderson, James A. and Rosenfeld, Edward (1988) *Neurocomputing: Foundations of Research,* MIT press.

Barron, Andrew R. (1985) "Logically smooth density estimation," Ph.D. dissertation, Department of Electrical Engineering, Stanford University.

Barron, Andrew R. (1989) "Statistical properties of artificial neural networks," *Proc. 28th Conference on Decision and Control,* IEEE, New York.

Barron, Andrew R. and Barron, Roger L. (1988) "Statistical learning networks: a unifying view," *Computing Science and Statistics: Proc. 20th Symp. Interface.*, Edward Wegman, editor, Amer. Statist. Assoc., Washington, DC., 192-203.

Barron, Andrew R. and Cover, Thomas M. (1990) "Minimum complexity density estimation," To appear in *IEEE Trans. Inform. Theory.*

Barron, Andrew R. and Sheu, Chyong-Hwa (1988) "Approximation of density functions by sequences of exponential families," To appear in *Ann. Statist.*

Bennett, George (1962) "Probability inequalities for the sum of independent random variables," *J. Amer. Statist. Assoc.*, **57**, 33-45.

Cox, Dennis D. (1988) "Approximation of least squares regression on nested subspaces," *Ann. Statist.*, **18**, 713-732.

Craig, Cecil C. (1933) "On the Tchebychef inequality of Bernstein" *Ann. Math. Statist.*, **4**, 94-102.

Cybenko, George (1989) "Approximations by superpositions of sigmoidal functions," *Math. Control, Signals, Systems*, **2**, 303-314.

Devroye, Luc (1988) "Automatic pattern recognition: a study of the probability of error," *IEEE Trans. Pattern Anal. Mach. Intelligence*, **10**, 530-543.

Farlow, Stanley J. (1984) "Self Organizing Methods in Modeling: GMDH Type Algorithms," Marcel Dekker, New York.

Friedman, Jerome H. (1990) "Multivariate adaptive regression splines (with discussion." To appear in the *Ann. Statist.*

Haussler, David (1989) "Generalizing the PAC model for neural net and other learning applications," Computer Research Laboratory, Technical Report 89-30, University of California, Santa Cruz.

Hoeffding, W. (1963) "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, **58**, 13-30.

Jones, Lee (1990). "A simple lemma on greedy approximations in Hilbert space and convergence rates for projection pursuit regression and neural network training," To appear in IAnn. Statist..

Lee, Yuchun and Lippmann, Richard P. (1990) "Practical characteristics of neural network and conventional pattern classifiers on artificial and speech problems," *Advances in Neural Information Processing Systems 2*, David S. Touretzky, editor, Morgan Kauffmann Publishers, San Mateo, CA.

Li, Ker-Chau (1987) "Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: discrete index set. *Ann. Statist.*, **15**, 958-975.

Lippmann, Richard P. (1987) "An introduction to computing with neural nets," *IEEE Communications Magazine*, **4**, 4-22.

Rissanen, Jorma (1983) "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, **11**, 416-431.

Rissanen, Jorma (1984) "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, **30**, 629-636.

Rumelhart, David E., McClelland, James L., et. al. (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press.

Schwarz, Gideon (1978) "Estimating the dimension of a model," *Ann. Statist.*, **6**, 461-464.

Sheu, Chyong-Hwa (1989) "Density estimation with Kullback-Leibler loss," Ph.D. Thesis, Department of Statistics, University of Illinois, Champaign, Illinois.

Shibata, Ritei (1981) "An optimal selection of regression variables," *Biometrika*, **68**, 45-54.

Vapnik, V. N. (1982) *Estimation of Dependences Based on Empirical Data*, Springer Verlag, New York.

White, Halbert (1990) "Connectionists nonparametric regression: multilayer feedforward networks can learn arbitrary mappings," To appear in *Neural Networks*.

DESIGNING PREDICTION BANDS

RUDOLF BERAN
University of California
Berkeley, California 94720 USA

ABSTRACT. This article develops four principles for the design of good prediction bands for a random process. The issues addressed include: appropriate asymptotic convergence of conditional and unconditional coverage probabilities; probability centering of prediction bands; controlling dispersion of conditional coverage probabilities; and increasing the rate of convergence of unconditional coverage probabilities. Examples illustrate the design issues and a proposed bootstrap construction for good prediction bands.

1. Introduction

Prediction bands for a random process will be discussed in the following setting. An observed learning sample Y_n and a potentially observable variable X have a joint distribution $P_{\theta,n}$. The unknown parameter θ lies in a parameter space Θ. To be predicted is the potentially observable random process $Z = \{Z(u,X): u \in U\}$, where the index set U may be *infinite* and the function $Z(\cdot,\cdot)$ is specified. In this article, both Θ and U are metric spaces.

The treatment will emphasize the case where the process Z is real-valued, the index set U is infinite, and the aim is to devise a good one-sided prediction band for Z. By careful extension of the index set U and of the function $Z(\cdot,\cdot)$, the analysis for this one-sided case also applies to two-sided prediction bands and to multivariate prediction regions. These possibilities will be illustrated through examples.

Let $z = \{z(u): u \in U\}$ denote a generic possible value of the random process Z. Define the one-sided prediction band for Z

$$D_n = \{z: z(u) \leq c_n(u) \text{ for every } u \in U\}, \tag{1.1}$$

where the critical values $c_n(u)$ depend on the learning sample Y_n. Clearly, D_n is equivalent to simultaneously asserting the one-sided prediction intervals

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 577–586.

© 1991 *Kluwer Academic Publishers. Printed in the Netherlands.*