ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 371-376

Penalized least squares, model selection, convex hull classes and neural nets

Gerald H. L. Cheang*
NIE, Nanyang Tech. Univ.
Singapore 637617

Andrew R. Barron†
Dept. of Stat., Yale Univ.
New Haven, CT 06520-8290

Abstract. We develop improved risk bounds for function estimation with models such as single hidden layer neural nets, using a penalized least squares criterion to select the size of the model. These results show the estimator achieves the best order of balance between approximation error and penalty relative to the sample size. Bounds are given both for the case that the target function is in the convex hull $\mathcal{C}$ of a class Φ of functions of dimension d (determined through empirical l_2 convering numbers) and for the case that the target is not in the convex hull.

1. Introduction

Penalized least squares model selection is commonly used to estimate a function from data $(X_i, Y_i)_{i=1}^n$, with a collection of models $\mathcal{F}_m$, with m in some index set $\mathcal{M}$. Let $\hat{f} = \hat{f}_{\hat{m}}$ be the estimate with model $\hat{m}$ selected by penalized least squares, achieving $\min_{m\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \hat{f}_m(X_i)\right)^2 + \frac{\text{pen}_n(m)}{n}$, where $\text{pen}_n(m)$ is a penalty term. The data $(X_i, Y_i)_{i=1}^n$ are assumed to be independently drawn according to the distribution of random variables (X, Y), with response variable Y and input vector X. Performance is measured by the squared $\mathcal{L}_2$ norm $\|\hat{f} - f\|^2 = \int |\hat{f}(x) - f(x)|^2 \mu(dx)$ where the target function is $f(x) = E[Y|X = x]$ and $\mu = P_X$ is the distribution of X. Bounds of the form

$$E\|f - \hat{f}\|^2 \leq C \min_{m\in\mathcal{M}} \left[\|f - f_m\|^2 + \frac{\text{pen}_n(m)}{n}\right] \tag{1}$$

characterize the risk of such functions estimates $\hat{f}$, where f_m is the function in $\mathcal{F}_m$ closest to f. Such bounds are available only for certain types of penalty and to the extent that (1) holds, one wants small values and radius $(\text{pen}_n(m))/n$ equal to $Kd_m(\log n)/n$ in accordance, roughly, with Schwartz's [14] BIC and Rissanen's [13] MDL criteria. This bound is superior to what can be obtained with the larger $(\text{pen}_n(m))/n = K\sqrt{d_m(\log n)/n}$ from Vapnik's [15] structural risk minimization (the derivation of which is similar, but here we more fully take into account what is possible with squared error criteria). With smaller

*geraldc@aya.yale.edu
†andrew.barron@yale.edu

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 371-376

$(\text{pen}_n(m))/n$ of order d_m/n (as in Akaike's [1] criterion) lacking the logarithmic factor, the bound (1) does not hold in general for estimation by nonlinearly parameterized functions (such as neural nets) that lack the homogeneous metric dimension property of Barron, Birge, and Massart [5].

Previously, bounds of the form (1) are in Barron [2],[4], and in [5], but there the least squares optimization was restricted to discrete parameter sets or to Lipshitz classes $\mathcal{F}_m$ (which permit $\mathcal{L}_\infty$ covering sets). For models with empirical l_2 or l_1 covering numbers (thereby allowing for more general neural net activation functions including step functions, which are not Lipshitz), bounds of the form (1) were announced by Cheang and Barron [6], building on the tools of Lee *et al* [11] concerning risk bounds for individual models without model selection. The approach we develop here does allow for model selection in a general setting which includes neural nets with step activation functions, and yields much better constants.

We allow for the case that the target function is not in the closure of the sequence of models $\mathcal{F}_m$, m in $\mathcal{M}$. Then having only a bound on $E\|\hat{f} - f\|$ of the form (1), with $C > 1$, would be undesirable since $\|f - f_m\|$ is not necessarily small. Thus, building on the work of Lee *et al* [11], we bound the additional squared error beyond that achieved by the best function in a convex set containing $\mathcal{F}_m$. Let $\mathcal{C}$ be the convex closure of $\mathcal{F}_m$ and let $f_\mathcal{C}$ be the projection of f onto $\mathcal{C}$. We bound the difference $E\|\hat{f} - f\|^2 - \|f_\mathcal{C} - f\|^2$ which by Pythagorean inequality upper bounds the error $E\|\hat{f} - f_\mathcal{C}\|^2$ between our estimate $\hat{f}$ and the ideal $f_\mathcal{C}$. An accuracy of approximation $a^2_{m,\mathcal{C}}$ quantifies the difference in squared error achievable by functions in $\mathcal{F}_m$ and functions in $\mathcal{C}$. We find that

$$E\|\hat{f} - f\|^2 - \|f_\mathcal{C} - f\|^2 \leq 4 \min_m \left\{ a^2_{m,\mathcal{C}} + \frac{\text{pen}_n(m)}{n} \right\} + \frac{4\bar{B}^2}{n}. \qquad (2)$$

Consequently $E\|\hat{f} - f_\mathcal{C}\|^2$ has the same bound. Now the presence of the constant 4 (greater than one) is not as much of a concern because $a^2_{m,\mathcal{C}}$ will be small if $f_\mathcal{C}$ is approximated well by members of $\mathcal{F}_m$. Applications are given for one hidden layer feed forward neural networks in section 4. A strategy of analysis familiar from empirical process theory plays a basic role in our analysis.

2. A Brief Look at Empirical Process Bounds

Let $\mathcal{G}$ be a class of real-valued functions g on a measurable space $\mathcal{D}$. Let $D_1, D_2, \ldots, D_n$ be independent random variables (data) with identical distribution. There is interest in the empirical process $P_n(g)$, $g \in \mathcal{G}$, where $P_n(g) = \frac{1}{n}\sum_{i=1}^n g(D_i)$, and the uniformity of its closeness to the expectation $P(g) = \int g(D)P(dD)$. In statistical learning theory (see Vapnik [15]) probability bounds on $\sup_{g\in\mathcal{G}} (P(g) - P_n(g))$ (Pollard [12], Dudley [7]), or on $\sup_{g\in\mathcal{G}} \frac{P(g)-P_n(g)}{\sigma(g)}$ or $\sup_{g\in\mathcal{G}} \frac{P(g)-P_n(g)}{\tau+\sigma^2(g)}$ (Vapnik [15], Haussler [8]), where $\sigma^2(g)$ is the variance of g, are useful in obtaining risk bounds for estimators that minimize empirical loss (or penalized empirical loss). We develop risk bounds in

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 371-376

section 3 with the following lemma.

Lemma 1. Let a function class $\mathcal{G}$ be given. For every positive τ, γ and small enough positive δ such that $\alpha^2 = \tau - \frac{4\delta^2}{\gamma^2} - \frac{4\delta}{\gamma}$ is positive, we have

$$\begin{aligned} & P\left\{\sup_{g\in\mathcal{G}} \frac{P_n'(g)-P_n(g)}{\frac{\tau\gamma}{2}+\frac{1}{2\gamma}\frac{1}{n}\sum_{i=1}^n\left(g(D_i')-g(D_i)\right)^2} \geq 1\right\} \\ \leq\ & P\left\{\sup_{g\in\mathcal{G}_{\underline{D},\underline{D}',\delta}} \frac{\frac{1}{n}\sum_{i=1}^n\left(g(D_i')-g(D_i)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n\left(g(D_i')-g(D_i)\right)^2}} \geq \alpha\right\} \leq EN_2\exp\left(-\frac{n\alpha^2}{2}\right), \end{aligned} \tag{3}$$

where $\mathcal{G}_{\underline{D},\underline{D}'}$ is the collection of functions $g \in \mathcal{G}$ restricted to the points $(\underline{D},\underline{D}') = (D_1,\ldots,D_n,D_1',\ldots,D_n')$ and $N_2 = \mathcal{N}_2(\delta,\mathcal{G}_{\underline{D},\underline{D}'})$ is its minimal cardinality of cover $\mathcal{G}_{\underline{D},\underline{D}',\delta}$ such that for each g in $\mathcal{G}_{\underline{D},\underline{D}'}$, there is a g_δ in $\mathcal{G}_{\underline{D},\underline{D}',\delta}$ with $d(g,g_\delta)\leq\delta$. The metric $d_{2n}(g,g_\delta)$ is defined using

$$d_{2n}(g,g_\delta) = \sqrt{\frac{1}{2n}\left(\sum_{i=1}^n (g(D_i)-g_\delta(D_i))^2 + \sum_{i=1}^n (g(D_i')-g_\delta(D_i'))^2\right)}. \tag{4}$$

Remark. Similar results, for example, in Lee *et al* [11], have the square root of the actual variance (instead of the empirical variance) in the denominator of (3). The advantage of our formulation is that in the application to function estimation problems better mean squared error bounds are obtained using the empirical variance in the lemma.

3. A Risk Bound

The target function is now denoted by $f^*(x) = E[Y|X=x]$. Let $f_{\mathcal{C}}$ be the projection of f^* onto $\mathcal{C}$, a closed convex class. To each f in $\mathcal{C}$, there corresponds $g : \mathcal{X}\times\mathcal{Y}\to\mathcal{R}$, a relative loss function which assigns to each $D=(X,Y)$, the regret

$$g(D) := (Y-f(X))^2 - (Y-f_{\mathcal{C}}(X))^2. \tag{5}$$

Let $\mathcal{G}_m$ be the class of functions g corresponding to f in $\mathcal{F}_m$, and let $\hat{f}_m$ be the least squares estimator, that is, $\hat{f}_m$ minimizes $\frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2$ over $f\in\mathcal{F}_m$. The corresponding $\hat{g}_m$ minimizes $P_n(g)$ over g in $\mathcal{G}_m$. Noting that $E(Y-f(X))^2 = E[Y-f^*(X)]^2 + \|f-f^*\|^2$, the expected regret of a function f is

$$Eg(D) = \|f-f^*\|^2 - \|f_{\mathcal{C}}-f^*\|^2. \tag{6}$$

The empirical loss with the training data $\{D_i\}_{i=1}^n$ is $P_n(g) = \frac{1}{n}\sum_{i=1}^n g(D_i)$ and with an independent copy $\{D_i'\}_{i=1}^n$ is $P_n'(g) = \frac{1}{n}\sum_{i=1}^n g(D_i')$. Note that

$$g(D) = (f(X)-f_{\mathcal{C}}(X))(f(X)+f_{\mathcal{C}}(X)-2Y). \tag{7}$$

Under the assumption that functions in $\mathcal{C}$ are bounded by B and that $|Y|\leq B_0$, then the right factor in (7) has magnitude $|f(X)+f_{\mathcal{C}}(X)-2Y|$ bounded, say, by $\bar{B} = 2(B_0+B)$ and $|g(D)|\leq 2B\bar{B}$.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 371-376

Let $\hat{f}_{\mathcal{C}}$ be the minimizer of $\sum_{i=1}^{n} (Y_i - f(X_i))^2$ over $f \in \mathcal{C}$. The relative empirical risk is

$$A^2_{m,\mathcal{C}} = \frac{1}{n}\sum_{i=1}\left(Y_i - \hat{f}_m(X_i)\right)^2 - \frac{1}{n}\sum_{i=1}\left(Y_i - \hat{f}_{\mathcal{C}}(X_i)\right)^2, \tag{8}$$

and its expected value is denoted by $a^2_{m,\mathcal{C}} = EA^2_{m,\mathcal{C}}$. If $\mathcal{F}_m$ consists of m-term approximations (convex combinations of m terms) with terms selected from a set Φ of bounded functions and $C = \overline{\text{conv}}(\Phi)$ with $|\phi(x)| \le c$ for $\phi \in \Phi$, then $A^2_{m,\mathcal{C}} \le \frac{4c^2}{m}$ for all possible $(X_i, Y_i)_{i=1}^n$ (see Lee *et al* [11] building on earlier work of Jones [10] and Barron [3]) and hence

$$a^2_{m,\mathcal{C}} \le \frac{4c^2}{m}. \tag{9}$$

The estimator $\hat{f}_m$ need not be an exact minimizer of $\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$ over f in $\mathcal{F}_m$. What matters for the bounds is that one has control of the relative empirical risk $A^2_{m,\mathcal{C}}$ or of its expected value $a^2_{m,\mathcal{C}}$. In particular, $A^2_{m,\mathcal{C}} \le \frac{4c^2}{m}$ also hold for estimators based on the Jones [10] relaxed greedy algorithm.

The use of Lemma 1 is crucial in the proof of Theorem 1, which bounds the risk when functions are estimated by penalized least squares criteria.

Theorem 1. Let the data be $(X_i, Y_i)_{i=1}^n$, independent, having the distributions of random variables (X, Y). Let the target function be $f^*(x) = E[Y|X = x]$ and let $f_{\mathcal{C}}$ be the projection of f^* onto $\mathcal{C}$, a convex class of functions which contains the subclasses $\mathcal{F}_m$, and let $\hat{f}_m$ be the least squares estimator minimizing $\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$ over $f \in \mathcal{F}_m$ for each m. Then let $\hat{f} = \hat{f}_{\hat{m}}$, with $\hat{m}$ the choice that minimizes the penalized least squares criterion

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_m(X_i))^2 + \frac{\text{pen}_n(m)}{n},$$

where the penalty $\text{pen}_n(m)$ satisfies

$$\sum_{m\in\mathcal{M}} \exp\left(-\frac{\text{pen}_n(m)}{\bar{B}^2} + \frac{n\delta_m^2}{2\bar{B}^4} + \frac{n\delta_m}{\bar{B}^2} + \log E\mathcal{N}_2\left(\frac{\delta_m}{\bar{B}}, \mathcal{F}_{m,\underline{D},\underline{D}'}\right)\right) \le 1. \tag{10}$$

Then the expected loss of the estimator $\hat{f}_{\hat{m}}$ compared to the best approximation $f_{\mathcal{C}}$ in the convex class $\mathcal{C}$ is

$$E\|\hat{f}_{\hat{m}} - f^*\|^2 \le \|f_{\mathcal{C}} - f^*\|^2 + 4R_n + \frac{4\bar{B}^2}{n} \tag{11}$$

where $R_n = \min_m\left\{a^2_{m,\mathcal{C}} + \frac{\text{pen}_n(m)}{n}\right\}$ and thus $E\|\hat{f}_{\hat{m}} - f_{\mathcal{C}}\|^2 \le 4R_n + \frac{4\bar{B}^2}{n}$.

In particular if the model $\mathcal{F}_m$ is of dimension d_m in the sense that the covering number satisfies $\mathcal{N}_2(\delta/2, F_{m,\underline{D},\underline{D}'}) \le n^{d_m}$ with $\delta = \frac{1}{n}$, then the risk bound holds with $\text{pen}_n(m) = Kd_m \log n$.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 371-376

4. Application to Neural Network Estimation

A single hidden layer feedforward sigmoidal network is a family of functions $f_m(x)$ of the form $f_m(x) = \sum_{i=1}^m \lambda_i \phi(a_i \cdot x - b_i), x \in \mathcal{R}^d$ parametrized by $(a_i, b_i, \lambda_i)_{i=1}^m$ with internal weight vectors a_i in $\mathcal{R}^d$, internal location parameter b_i in $\mathcal{R}$, external weights λ_i, and $\phi(z) = 1_{\{z>0\}}$. Thus $f_m(x)$ is a piecewise constant function. The network model can be used to approximate the target function $f(x)$ and to estimate it based on data $(X_i, Y_i)_{i=1}^n$, a random sample from a joint probability distribution $P_{X,Y}$ with $f^*(x) = E[Y_i|X_i = x]$. The model classes are chosen to be

$$\mathcal{F}_m = \left\{ f_m(x) = \sum_{i=1}^m \lambda_i \phi(a_i \cdot x - b_i), x \in \mathcal{R}^d : \sum_{i=1}^m |\lambda_i| \leq c, \|f\|_\infty < B \right\}. \quad (12)$$

Let

$$\Phi = \{\zeta \phi(a_i \cdot x - b_i), x \in \mathcal{R}^d : a_i \in \mathcal{R}^d, b_i \in \mathcal{R}, |\zeta| \leq c\}. \quad (13)$$

Thus we are taking convex combinations of functions from the class Φ consisting of indicators of half-spaces multiplied by constants of magnitude less than c.

Theorem 2. Let the data be $(X_i, Y_i)_{i=1}^n$, independently distributed with joint probability distribution $P_{X,Y}$ and $f^*(x) = E[Y_i|X_i = x]$, and $|Y| \leq B_0$. Let $\mathcal{F}_m$ be defined as in (12) and let $\mathcal{C} = \overline{\text{conv}}\Phi$. Let $f_\mathcal{C}$ be the projection of f^* on $\mathcal{C}$, and let $\hat{f}_m$ be either the least squares estimator minimizing $\frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2$ over each $f \in \mathcal{F}_m$ for each m, or more generally, any estimator achieving $A_{m,\mathcal{C}}^2 \leq \frac{4c^2}{m}$, with sample size $n \geq 8$. Let $\hat{f} = \hat{f}_{\hat{m}}$ with $\hat{m}$ as the choice that minimizes the penalized least squares criterion $\frac{1}{n}\sum_{i=1} \left(Y_i - \hat{f}_m(X_i)\right)^2 + \frac{\text{pen}_n(m)}{n}$. Let m_n^* be the particular value of m that minimizes $R_n = \min_m \left\{ a_{m,\mathcal{C}}^2 + \frac{\text{pen}_n(m)}{n} \right\}$, where the penalty $\text{pen}_n(m)$, of order $\bar{B}^2 m(d+1)\log n$, is any choice not smaller than

$$\bar{B}^2 \left[(d+1)m \log \frac{2en}{d+1} + m \log \left(\frac{en^2}{2} \right) + \log(m+1) + 4 + \log \frac{1}{q(m)} \right], \quad (14)$$

with $q(m)$, such as $q(m) = \frac{6}{\pi^2 m^2}$, chosen such that $\sum_m q(m) = 1$. Then R_n is of order $\min_m \left\{ \frac{c^2}{m} + \frac{\bar{B}^2 m d \log n}{n} \right\} = 2c\bar{B}\sqrt{\frac{d \log n}{n}}$ and the expected loss of the estimator $\hat{f}_{\hat{m}}$ compared to the best approximation $f_\mathcal{C}$ in the convex class $\mathcal{C}$ is

$$E\|\hat{f}_{\hat{m}} - f^*\|^2 \leq \|f_\mathcal{C} - f^*\|^2 + 4R_n + \frac{4\bar{B}^2}{n}, \quad (15)$$

and hence

$$E\|\hat{f}_{\hat{m}} - f_\mathcal{C}\|^2 \leq 4R_n + \frac{4\bar{B}^2}{n}.$$

Remark. The proof of Theorem 2 is a direct application of Theorem 1 with the appropriate calculation of the covering numbers of $\mathcal{F}_m$ which satisfies

$$\mathcal{N}_2\left(\delta_m/2, \mathcal{F}_{m,\underline{D},\underline{D}'}\right) \leq e\,(m+1)\left(\bar{B}^4 e/2\delta_m^2\right)^m (2en/d+1)^{(d+1)m}, \quad (16)$$

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 371-376

using the bound in Haussler [9, Corollary 3]. In order to achieve nearly the best tradeoff for the penalty in (10), we choose $\delta_m = \min\left(2\bar{B}^2/n, \bar{B}^2/2\right)$.

References

[1] H. Akaike (1973) Information theory and an extension of the maximum likelihood principle. In P. N. Petrov and F. Csaki (eds.) *Proc. 2nd International Symposium on Inform. Theory*, pp 267–281. Akademia Kiado, Budapest.
[2] A.R. Barron (1991) Complexity regularization with applications to artificial neural networks, in G. Rousas (ed.) *Nonparametric functional estimation*, pp 561–576. Kluwer, Dordrecht, the Netherlands and Boston.
[3] A.R. Barron (1993) Universal approximation bounds for superposition of a sigmoidal function. *IEEE Trans. Info. Thy.*, **39**:930–944.
[4] A.R. Barron (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning*, **1994**:
[5] A.R. Barron, L. Birgé and P. Massart (1999) Risk bounds for model selection via penalization. *Prob. Thy. Rel. Fields*, **113**:301–413.
[6] G.H.L. Cheang and A.R. Barron (1999) Estimation with two hidden layer neural nets. *Proc. 1999 IJCNN, July 1999, Washington, D.C.*
[7] R. Dudley (1974) Metric entropy of some classes of sets with differentiable boundaries. *J. Approximation Theory*, **10**:227–236; Correction (1979), *J. Approximation Theory*, **26**:192–193.
[8] D. Haussler (1992) Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. Comp.*, **100(1)**:78–150.
[9] D. Haussler (1995) Sphere packing numbers of subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Comb. Thy.*, Series A, **69**:217–232.
[10] L.K. Jones (1990) A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Stats.*, **20**:608–613.
[11] W.S. Lee, P.L. Bartlett and R.C. Williamson (1996) Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Info. Thy.*, **42**:2118 – 2132.
[12] D. Pollard (1984) *Convergence of stochastic processes.* Springer Verlag, Berlin.
[13] J. Rissanen (1978) Modeling by shortest data description. *Automatica*, **14**:465–471.
[14] G. Schwartz (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**:461–464.
[15] V.N. Vapnik (1996) *Statistical Learning Theory.* J. Wiley, New York.