Estimation with Two Hidden Layer Neural Nets

Gerald H. L. Cheang
National Institute of Educaton (Nanyang Technological University), Singapore

Andrew R. Barron
Yale Univeristy, New Haven, CT, USA

Abstract—Our presentation deals with function estimation by neural networks. Mean square error bounds are given for the case when the target function is in the convex hull of ellipsoids multiplied by a scalar constant. When the target function is not in this class but is bounded, we bound the difference between the mean square prediction error compared to the best approximation error of the target function (the expected regret). We also give a general theorem that gives the convergence rate of the expected regret when the functions are estimated by penalized least squares criteria.

I. Introduction

We consider function estimation by feedforward sigmoidal neural networks. A single hidden layer feedforward sigmoidal network is a familiy of functions $f_T(x)$ of the form

$$f_T(x,\theta) = \sum_{i=1}^{T} c_i \phi(a_i \cdot x - b_i), x \in \mathcal{R}^d \qquad (1)$$

parametrized by $\theta = (a_i, b_i, c_i)_{i=1}^T$. A two hidden layer sigmoidal network takes the form

$$f_{T_1,T_2}(x,\theta) = \sum_{i=1}^{T_1} c_i \phi(\sum_{j=1}^{T_2} a_{ji}\phi(\omega_{ji} + b_{ji}) - d_i), x \in \mathcal{R}^d. \qquad (2)$$

It is parametrized by $\theta = (a_i, d_i, b_{ji}, \omega_{ji}, c_{ji})_{i=1}^{T_1}{}_{j=1}^{T_2}$. We use the unit step sigmoid $\phi(z) = 1_{\{z>0\}}$ throughout this paper.

Gerald H. L. Cheang, +65-460-5690, fax +65-469-8952, cheanghlg@nie.edu.sg; Div. of Math, School of Science, Nat'l Inst. of Education (Nanyang Technological University), 469 Bukit Timah Rd, Singapore 259756, Singapore.

Andrew R. Barron, +1-203-432-0666, barron@stat.yale.edu; Dept. of Stats, Yale University, P. O. Box 208290, Yale Station, New Haven, CT 06520-8290, USA.

The target function f^* is estimated from data $(X_i, Y_i)_{i=1}^N$, an independent random sample of size N from a joint probability distribution $P_{X,Y}$ with $f^*(x) = E[Y_i|X_i = x]$ and f^* is in $\mathcal{L}_2(P_X)$. We are given a sequence of models $\mathcal{F}_M$ (consisting of a family of functions) indexed in a countable index set $\mathcal{M}$. For each model, we estimate $\hat{f}_{M,N}$ to minimize the empirical loss $\frac{1}{N}\sum_{i=1}^N (Y_i - f(X_i))^2$ over choices of $f \in \mathcal{F}_M$ and then we pick $\hat{M}$ and $\hat{f} = \hat{f}_{\hat{M},N}$ to minimize the penalized squared error criterion $\frac{1}{N}\sum_{i=1}^N (Y_i - \hat{f}_{M,N}(X_i))^2 + \text{pen}_N(M)$. We require convexity of the class $\bigcup_M \mathcal{F}_M$ consisting of the union of our models $\mathcal{F}_M$, for $M \in \mathcal{M}$. In our analysis we examine the risk compared to the best possible in $\mathcal{F} = \text{closure}(\bigcup_M \mathcal{F}_M)$ (where the closure is taken in $\mathcal{L}_2(P_X)$). We build classes of neural network estimators that satisfy the conditions on the sequence of models $\mathcal{F}_M$ and apply a general theorem bounding the risk of penalized least squares estimators under entropy conditions on the component models.

We extend the estimation bound results of Barron [1] to the case of unit step sigmoids and that of Lee *et al* [4] to include a penalty term. The result for two hidden layer network estimators is also new. We are able to obtain the risk bounds based on unexpectedly accurate and parsimonious neutal net approximations for balls and ellipsoids in $\mathcal{R}^d$ developed by the authors [3].

II. Main Results

A. Summary

Let $\mathcal{F}_V$ be the closure (in $\mathcal{L}_2(P_X)$) of the class of all single layer neural nets, with a given bound V on the

0-7803-5529-6/99/$10.00

sum of the absolute value of the outer weights. We give a penalized least squares estimator $\hat{f}_{\hat{T}}$ and show that if $f \in \mathcal{F}_V$ then the mean square prediction error $E\|f - \hat{f}_{\hat{T}}\|_2^2$ is bounded above by $KV^2\left(\frac{d\ln N}{N}\right)^{\frac{1}{2}}$. Let $V_{f,\mathcal{H}}$ denote the variation of f with respect to half-spaces, which is the smallest number such that f/V_f is in the closure of the convex hull of signed indicators of half-spaces. We show that the mean squared error between $\hat{f}_{\hat{T}}$ and f is bounded by $KV_f^2\left(\frac{d\ln N}{N}\right)^{\frac{1}{2}} + \frac{K'V_f^4}{N}$. When the target function has variation $V_{f,\mathcal{E}}$ with respect to a class $\mathcal{E}$ of ellipsoids, we show with a two hidden layer network estimator $\hat{f}_{\hat{T}_1,\hat{T}_2}$ that $E\|f - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2 \leq Kd^{3/2}V_{f,\mathcal{E}}^2\left(\frac{\ln N}{N}\right)^{\frac{1}{4}}$.

B. A Risk Bound

Consider a sequence of models $\mathcal{F}_M$ (consisting of a family of functions) indexed in a countable index set M. For each model, we estimate $\hat{f}_{M,N}$ to minimize the empirical loss $\frac{1}{N}\sum_{i=1}^N (Y_i - f(X_i))^2$ over choices of $f \in \mathcal{F}_M$ and then we pick $\hat{M}$ and $\hat{f} = \hat{f}_{\hat{M},N}$ to minimize the penalized squared error criterion

$$\frac{1}{N}\sum_{i=1}^N (Y_i - \hat{f}_{M,N}(X_i))^2 + \text{pen}_N(M). \qquad (3)$$

We also require convexity of the class $\bigcup_M \mathcal{F}_M$ consisting of the union of our models $\mathcal{F}_M$, for $M \in \mathcal{M}$.

In our analysis we examine the risk compared to the best possible in $\mathcal{F} = \text{closure}(\bigcup_M \mathcal{M})$ (where the closure is taken in $\mathcal{L}_2(P_X)$). Let $f_{\mathcal{F}}^*$ in $\mathcal{F}$ achieve $E(Y - f_{\mathcal{F}}^*(X))^2 = \inf_{f\in\mathcal{F}} E(Y - f(X))^2$. We define the loss function (regret)

$$\begin{aligned} r(f) &= r(f, f^*) \\ &:= E(Y - f(X))^2 - E(Y - f_{\mathcal{F}}^*(X))^2 \end{aligned} \qquad (4)$$

and the empirical loss function

$$\hat{r}(f) := \frac{1}{N}\sum_{i=1}^N (Y_i - f(X_i))^2 - \frac{1}{N}\sum_{i=1}^N (Y_i - f_{\mathcal{F}}^*(X_i))^2. \qquad (5)$$

The relative regret $r(f, f^*)$ measures the regret in $\mathcal{L}_2$ approximation of f^* by f compared to the best approximation in $\mathcal{F}$,

$$\begin{aligned} r(f, f^*) &= \|f^* - f\|_2^2 - \inf_{g\in\mathcal{F}} \|f^* - g\|_2^2 \\ &= \|f^* - f\|_2^2 - \|f^* - f_{\mathcal{F}}^*\|_2^2. \end{aligned} \qquad (6)$$

We define an index of resolvability

$$R_{N,M}(f^*) := \min_{f\in\mathcal{F}_M} \{r(f, f^*) + \text{pen}_N(M)\}. \qquad (7)$$

Let

$$R_N(f^*) := \min_{M\in\mathcal{M}} R_{N,M}(f^*)$$

be the minimum value of the resolvability and let a function that minimizes this resolvability be denoted by $f_{M^{**}}^*$.

For $N \in \{1, 2, \cdots\}$ and $x, y \in \mathcal{R}^N$, let

$$d_{l_1}(x, y) := \frac{1}{N}\sum_{i=1}^N |x_i - y_i|$$

For $U \subseteq \mathcal{R}^N$, $\epsilon > 0$, we say that $C \subseteq \mathcal{R}^N$ is an l_1 ϵ-cover of U if for all $x \in U$, there exists $y \in C$ such that $d_{l_1}(x, y) \leq \epsilon$. The l_1 covering number $\mathcal{N}(\epsilon, U)$ is the smallest number of l_1 balls that forms an l_1 ϵ-cover of U. Thus $\mathcal{N}(\epsilon, \mathcal{F}_{M|\underline{x}})$ is the l_1 ϵ-covering number of $\mathcal{F}_{M|\underline{x}}$ given the data $\underline{x} \in \mathcal{X}^N$. Suppose $\underline{x} = (x_1, \cdots, x_N) \in \mathcal{X}^N$ is given, then elements of $\mathcal{F}_{M|\underline{x}}$ will be functions in $\mathcal{F}_M$ evaluated at the points $\underline{x}$, for example $(f(x_1), \cdots, f(x_N))$. Define $\mathcal{N}_N(\epsilon, M) := \sup_{\underline{x}\in\mathcal{X}^N} \mathcal{N}_N(\epsilon, \mathcal{F}_{M|\underline{x}})$.

The following theorem bounds the expected regret under certain conditions. It relates the convergence rate of the expected regret to a multiple of the index of reslvability. First we cover the case that there is a fixed upper bound B to the values of B_0 and B_M for all $M \in \mathcal{M}$. Next we cover the case that B_M is unbounded for $M \in \mathcal{M}$.

Theorem 1

i. Let the data be $(X_i, Y_i)_{i=1}^N$, independent with probability distribution $P_{X,Y}$, $f^*(x) = E(Y_i|X_i = x)$, and $|Y_i| \leq B$, $|f| \leq B$ for all $f \in \mathcal{F}_M$, for $M \in \mathcal{M}$, and $|f_{\mathcal{F}}| \leq B$ and suppose that $\mathcal{F} = \text{closure}\bigcup_M \mathcal{F}_M$ is convex. Suppose $\delta_{M,N}$ and the penalty $\text{pen}_{M,N}$ are chosen to satisfy

$$\begin{aligned} &\sum_M 6\mathcal{N}_{2N}(\frac{\delta_{M,N}}{8}, M) \\ &\times \exp\left(-\frac{3(\text{pen}_N(M) - \delta_{M,N}/2)N}{5248B^2}\right) \\ &\leq 1, \end{aligned} \qquad (8)$$

then the estimator $\hat{f} = \hat{f}_{\hat{M},N}$ that minimizes the penalized squared error has expected regret compared to the best $f \in \mathcal{F}$ that is bounded by

$$E[r(\hat{f}_{\hat{M}})] \leq 2R_N(f^*) + \frac{c_1 B^2}{N}, \tag{9}$$

where $c_1 = 20992$.

ii. Let the data be $(X_i, Y_i)_{i=1}^N$, independent with probability distribution $P_{X,Y}$, $f^*(x) = E(Y_i|X_i = x)$, and $|Y_i| \leq B_o$, $|f| \leq B_M$ for all $f \in \mathcal{F}_M$, for $M \in \mathcal{M}$, and $|f^*_{\mathcal{F}}| \leq B_1$ and suppose that $\mathcal{F} = \text{closure} \bigcup_M \mathcal{F}_M$ is convex. Suppose for $\delta_{M,N}$ and the penalty $\text{pen}_{M,N}$ are chosen to satisfy

$$\begin{aligned} &\sum_{M \in \mathcal{M}} 6\mathcal{N}_{2N}(\frac{\delta_{M,N}}{8}, M) \\ &\times \quad \exp\left(- \frac{3\left(\text{pen}_N(M) - \frac{\delta_{M,N}}{2} - \frac{1312\bar{B}_M^4}{N}\right) N}{5248\bar{B}_M^2} \right) \\ &\leq \quad 1, \end{aligned} \tag{10}$$

then the estimator $\hat{f} = \hat{f}_{\hat{M},N}$ that minimizes the penalized squared error has expected regret compared to the best $f \in \mathcal{F}$ is bounded by

$$E[r(\hat{f}_{\hat{M}})] \leq 7R_N(f^*). \tag{11}$$

C. Remark

If each term in the summand (8) is a function of M, say $g(M)$, with $\sum_M g(M) \leq 1$ and if an upper bound $\bar{\mathcal{N}}$ is available for $\mathcal{N}$, then we can take the penalty to be

$$\text{pen}_N(M) = \frac{5248B^2}{3N} \ln \left[\frac{6\bar{\mathcal{N}}_{2N}(\frac{\delta_{M,N}}{8}, M)}{g(M)} \right] + \frac{\delta_{M,N}}{2}. \tag{12}$$

One can intepret $g(M)$ as a prior distribution on $\mathcal{M}$ and $1/\bar{\mathcal{N}}_{2N}(\frac{\delta_{M,N}}{8}, M)$ as a prior on the functions in $\mathcal{F}_M$.

D. Estimation with Single Hidden Layer Networks

In this section, we apply the result from Theorem 1 to estimation with single hidden layer neural networks with step activation functions. The range of the observed responses Y_i is assumed to be in $[-B_o, B_o]$ and the estimated single layer network takes the form (1).

Let

$$\mathcal{F}_T := \left\{ x \to \sum_{i=1}^T c_i \phi(a_i \cdot x - b_i) : a_i \in \mathcal{R}^d, b_i, c_i \in \mathcal{R} \right\}$$

be the class of single layer nets with T hidden units with no restrictions on the magnitude of the parameters. The subclass $\mathcal{F}_{B,T}$ of networks with a bound on the sum of absolute values of output weight is

$$\mathcal{F}_{B,T} := \left\{ x \to \sum_{i=1}^T c_i \phi(a_i \cdot x - b_i) : \sum_{i=1}^T |c_i| \leq B \right\}.$$

The closure of the class of single hidden layer neural networks $\mathcal{F}_B$ with sum of absolute values of output weights bounded by B is $\mathcal{F}_B := B\overline{\text{conv}}\{\phi(a \cdot x - b) : a \in \mathcal{R}^d, b \in \mathcal{R}\}$, which is the closure of $\bigcup_T \mathcal{F}_{B,T}$. When B is fixed, the convex target class $\mathcal{F}$ is $\mathcal{F}_B = \text{closure}\,(\bigcup_T \mathcal{F}_{B,T})$. Then the indices for application of Theorem 1 are integers $M = \{1, 2, ...\}$. The penalty takes the form $\text{pen}_{B,N}(T) = \frac{KB^2 m_T}{N} \ln N$, where K is a constant and m_T is the dimension of the parameter space.

Let $\mathcal{F}_V$ be the closure (in $\mathcal{L}_2(P_X)$) of the class of all single layer neural nets, with a given bound V on the sum of the absolute value of the outer weights. Denoting the penalized least squares estimator by $\hat{f}_{\hat{T}}$, we see that if $f \in \mathcal{F}_V$ then the mean square prediction error $E\|f - \hat{f}_{\hat{T}}\|_2^2$ is bounded above by $KV^2\left(\frac{d \ln N}{N}\right)^{\frac{1}{2}}$.

We also consider the case that B is not fixed but rather is part of the model specification and we allow the penalized criterion to make selection among indices $M = (B, T)$ in $\mathcal{M} = \{1, 2, ...\}^2$. In particular it will contain the target function f^* which we have assumed to be bounded by B_o. In this setting we obtain consistent estimation for all bounded functions with rate controlled by the index of resolvability which expresses the trade-off for each model $\mathcal{F}_{B,T}$ between its squared approximation error and the log l_1-covering number divided by sample-size. In particular, when f has finite variation V_f with respect to half-spaces, we get a trade-off of order $\frac{V_f^2}{T}$ plus $V_f^2\left(\frac{dT}{N}\right)\ln(N)$ as long as the candidate models include those with B at least V_f.

The mean squared error between $\hat{f}_{\hat{T}}$ and f is bounded by $KV_f^2\left(\frac{d\ln N}{N}\right)^{\frac{1}{2}} + \frac{K'V_f^4}{N}$.

The model selection allows such trade-off without prior knowledge of V_f. When the variation V_f is infinite the resolvability bound expresses the trade-off between the appoximation squared error $\|f - f_{T,B}\|^2$ and $B^2\left(\frac{dT}{N}\right)\ln(N) + \frac{B^4}{N}$. In this case ($V_f = \infty$) the criterion will determine from the data the value of B and T that achives a desirable trade-off. As N goes to infinity, the resulting B and T will diverge to infinity (to allow the approximation error to go to zero) while $\frac{BT}{N}$ and $\frac{B^4}{N}$ will tend to zero.

E. Estimation with Two Hidden Layer Networks

As before, the target function f is estimated from data $(X_i, Y_i)_{i=1}^N$, an independent with distribution $P_{X,Y}$ and $f^*(x) = E[Y_i|X_i = x]$. The range of the observed responses Y_i is assumed to be in $[-B_o, B_o]$ and the estimated two layer network takes the form (2).

A class of two hidden layer neural networks $\mathcal{F}_{B,T_1,T_2}$, with T_1 hidden units in the outer-layer and T_2 hidden units in the inner layer is defined to be

$$\mathcal{F}_{B,T_1,T_2} := \left\{ x \to \sum_{i=1}^{T_1} c_i \phi\left(\sum_{j=1}^{T_2} \omega_{ij}\phi(a_{ij}\cdot x - b_{ij}) - d_i\right) : \sum_{i=1}^{T} |c_i| \le B \right\}.$$

We may restrict $\sum_{j=1}^{T_2}|\omega_{ij}| \le 1$ since $\phi(z) = \phi(kz)$ for hard-limiter sigmoids (unit-step functions) when $k > 0$. Let $\mathcal{G}_B$ be the closure of $\bigcup_{T_1,T_2}\mathcal{F}_{B,T_1,T_2}$. Thus our candidate model classes are $\mathcal{F}_{B,M} = \{f : f \in \mathcal{F}_{B,T_1,T_2}\}$. The set $\mathcal{M}$ of indices M consists of all (T_1, T_2) and $\mathcal{F}_B = \text{closure}\bigcup_M \mathcal{F}_{B,M} = \text{closure}\bigcup_{T_1,T_2}\mathcal{F}_{B,T_1,T_2}$. Here we will focus for simplicity on the case that B is fixed. The penalty for application of takes the form $\text{pen}_{B,N}(T_1,T_2) = \frac{KB^2 m_{T_1,T_2}}{N}\ln N$, where m_{T_1,T_2} is the dimension of the parameter space.

Denote the penalized least squares estimator by $\hat{f}_{\hat{T}_1,\hat{T}_2}$. If $\frac{1}{B}f^*$ is in the class $\mathcal{H}$ (determined by convex combination of signed indicators of ellipsoids), then f^* is in $\mathcal{F}_B = \text{closure}\bigcup_{T_1,T_2}\mathcal{F}_{B,T_1,T_2}$ by [2], [3]. Using the bound in [3, Theorem 4] on the approximation error, we obtain

$$E\|f^* - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2 \le 2\min_{T_1,T_2}\left\{\frac{K_1B^2}{T_1} + \frac{K_2^2B^2d^2}{\sqrt{T_2}} + \frac{Km_{T_1,T_2}B^2}{N}\ln N\right\}. \quad (13)$$

Optimizing over T_1 and T_2 yields $K'd^{3/2}B^2\left(\frac{\ln N}{N}\right)^{\frac{1}{4}}$ as an upper bound to the mean squared error $E\|f^* - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2$. The bound tends to zero as $N \to \infty$. The optimal values of T_1 and T_2 are of order $\frac{1}{d}\left(\frac{N}{\ln N}\right)^{\frac{1}{4}}$ and $d\left(\frac{N}{\ln N}\right)^{\frac{1}{2}}$ respectively.

III. Conclusions

Risk bounds are given for neural network estimators for certain classes of functions : functions that are in the closure of the convex hull of signed indicators of half-spaces and functions in the convex hull of indicators of ellipsoids. As the case of convex combinations of indicators of ellipsoids illustrates, two hidden layer networks provide accurate estimators in cases where accurate one layer representations are not necessarily available.

References

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

[2] G. H. L. Cheang, "Neural network approximation and estimation of functions," *Proc. of the 1994 IEEE-IMS Workshop on Info. Thy. and Stat.*, 1994.

[3] G. H. L. Cheang and A. R. Barron, "A better approximation for balls," Preprint.

[4] W. S. Lee, P. L. Bartlett and R. C. Williamson, "Efficient agnostic learning of neural networks with bounded fan-in" *IEEE Trans. Info. Thy.*, vol. 42, pp. 2118–2132, 1996.