INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Neural Network Approximation and Estimation of Functions

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Gerald H. L. Cheang

Dissertation Director: Prof. Andrew R. Barron

May 1998

Abstract

Neural Network Approximation and Estimation of Functions

Gerald H. L. Cheang

May 1998

Approximation and estimation bounds for neural networks are obtained in this dissertation. Two hidden layer feedforward sigmoidal neural nets are used to estimate a target function of d variables. For example, if the target function f has finite total variation V_f with respect to a class of ellipsoids, then the $\mathcal{L}_2$ approximation error is bounded by $\frac{K_1 V_f}{T_1^{1/2}} + \frac{K_2 V_f d}{T_2^{1/4}}$, where K_1 and K_2 are constants, when such a function is approximated by a two layer neural net with T_1 nodes in the outer layer and T_2 nodes in the inner layer. When estimating the function using a random sample, the overall mean squared error in terms of the best approximation error, the dimension of the parameter space m_{T_1,T_2} and the sample size N is bounded by $K_1\|f - f_{T_1,T_2}\|_2^2 + \frac{K_2 m_{T_1,T_2}}{N} \log m_{T_1,T_2} N$. When this bound is optimized for T_1 and T_2, it is of order $d^{3/2} V_f^{5/4} (\frac{\log N}{N})^{1/4}$. It can be seen from our bounds that the number of nodes, and hence parameters, and the sample size are not required to be exponentially large in the dimension d to obtain accurate estimates. Probabilistic methods and approximation of the Gaussian play a special role in the derivation of the approximation bound. Minimum complexity regularization, and a calculation of an index of resolvability, are used in the derivation of our estimation bound. A heuristic algorithm for fitting single

hidden layer nets iteratively to a class of target functions is also given. Functions in this class (when normalized) lie in the closed convex hull of sigmoids. Finally, we suggest ways of extending some of these results.

Preface

This dissertation is dedicated to my late grandfather. Without his encouragement, Yale would not have been a possibility. I am also grateful to my parents for their support throughout this endeavour.

I would like to thank Prof. Andrew Barron for introducing me to the fascinating world of artificial neural networks. His numerous suggestions and insights into various research problems never cease to amaze me. His constant encouragement and advice are very much appreciated. Thanks are also due to Profs. John Hartigan and Nicolas Hengartner for their suggestions in improving the original version of this dissertation. I am also grateful to Dr. Catherine Macken and Prof. George Seber for their excellent teaching during my Auckland days, which aroused my interest in statistics.

Finally, thanks to Christoph Thiele, Michael Schmelzle, Sebastian Walter, Jason Wright, Armin Westerhof, Alexandra Thiry and many others here at Yale, whose companionship and sometimes insane antics have helped me preserve my sanity throughout these memorable years. Many thanks to Roland Schwaiger for his wonderful *gemütlich* hospitality at the Technozentrum, Universität Salzburg.

Gerald Cheang
New Haven, Connecticut
March 1998

Contents

List of Figures and Tables

Chapter 1

An Overview

1.1 Introduction

A single hidden layer feedforward sigmoidal network is a family of functions $f_T(x)$ of the form

$$f_T(x,\theta) = \sum_{i=1}^{T} c_i \phi(a_i \cdot x - b_i), \qquad x \in \mathcal{R}^d \tag{1.1}$$

parametrized by $\theta = (a_i, b_i, c_i)_{i=1}^T$ with internal weight vectors a_i in $\mathcal{R}^d$, internal location parameter b_i in $\mathcal{R}$, external weights c_i, and ϕ a fixed sigmoidal function. We use $a \cdot x$ to denote the inner product of vectors a and $x \in \mathcal{R}^d$. Such a network has d inputs, T hidden nodes and a linear output unit. A sigmoid is a bounded monotone function on $\mathcal{R}$. When $\phi(z) = 1_{\{z>0\}}$, the sigmoids $\phi(a_i \cdot x - b_i)$ provide indicators of half-spaces and $f_T(x,\theta)$ is a piecewise constant function. The network model can be used to approximate target functions $f(x)$ defined over bounded subsets of $\mathcal{R}^d$

and to estimate the function based on data $(X_i, Y_i)_{i=1}^N$, a random sample from a joint probability distribution $P_{X,Y}$ with $f(x) = E[Y_i|X_i = x]$. For the sake of brevity, such a network is sometimes referred to as a one-layer net or a single layer net. Similarly, we use the term "k-layer neural net" to mean a feedforward network with k "hidden" layers of sigmoidal units and one linear output unit. The probability distribution over the input space is P_X and the mean square distance between any two functions $f(x)$ and $g(x)$ is $\|f - g\|^2 = E_X|f(X) - g(X)|^2$.

In living organisms with a central nervous system, the neuron forms the basic building block of the central nervous system. The neuron is a signal generating cell that receives stimuli from the environment or from other neurons and generates an output to neighboring cells. The concept of artificial neural networks as a mathematical model first appeared in 1943 in McCullough and Pitts [40]. They regarded $c\phi(a \cdot x - b)$ as a simple neuron model with $\phi(z) = 1_{\{z>0\}}$, where the coordinates of x correspond to the voltages at the dentritic synapses and $a \cdot x$ corresponds to the accumulated voltage at the cell body; the neuron fires with output voltage c on the axon when $a \cdot x$ exceeds the threshold b. However, the aim of McCullough and Pitts [40] was not to model biological models, rather their aim was to show that arbitrary Boolean functions could be represented by a sufficiently large network composed of artificial neurons. These artificial neurons are called nodes or units interchangeably in the literature. Networks of such units are also called perceptrons. See, for example, Rosenblatt [49, 50]. There a loose analogy is drawn between a retinal perception

system (in which images falling on individual retinal nerve cells are processed and perceived as a whole image) and the way artificial neural networks receive and process input signals. However the similarity between physiological neural networks and artificial neural nets is only superficial. Though from time to time neuroscience attempts to bridge the gap, for the most part, artificial neural networks are not used to model their physiological counterparts. Indeed these networks and the parametrized functions they represent have been put to use in computer science, engineering, physics and statistics as tools for pattern recognition, signal processing and estimation of functions (see Cheng and Titterington [13], Hopfield [28], Buntine and Weigend [11], Rumelhart *et al* [51], Bishop [9], Ripley [48], Barron and Barron [5]).

This work will be concerned with seeking extensions for approximation and estimation bounds for two hidden layer sigmoidal networks. Such a network takes the form

$$f_{T_1,T_2}(x,\theta) = \sum_{i=1}^{T_1} c_i \phi(\sum_{j=1}^{T_2} a_{ji}\phi(\omega_{ji}\cdot x + b_{ji}) - d_i), x \in \mathcal{R}^d \tag{1.2}$$

There are T_1 nodes in the outer layer and T_2 nodes in the inner layer, for each node in the outer layer, giving a total of $T_1 + T_1T_2$ nodes. It is parametrized by $\theta = (a_i, d_i, b_{ji}, \omega_{ji}, c_{ji})_{i=1}^{T_1}{}_{j=1}^{T_2}$. The architecture of the two hidden layer net is as follows. Firstly, the individual co-ordinates of x form the input layer. These are fed into the first hidden layer (called inner layer here) with T_2 nodes. The output from these nodes are then fed into the next hidden layer (called outer layer here) consisting of T_1 nodes, which lastly goes into the output. The function $f_{T_1,T_2}(x,\theta)$ is parametrized

by $\theta = (a_i, d_i, b_{ji}, \omega_{ji}, c_{ji})_{i=1}^{T_1}{}_{j=1}^{T_2}$. The space spanned by families of single layer sigmoidal networks is dense in the space of all functions in $\mathcal{L}_2(P_X)$, for any probability measure P_X (Hornik *et al* [29]) and classes of functions have been identified that permit approximation bounds of reasonable accuracy in terms of the number of nodes (for example, Barron [3] and Makovoz [38]). Two hidden layer neural networks can approximate functions that can be approximated by a single layer network. Indeed, each single layer network has a two layer representation with certain trivial choices of second layer parameters. For example, consider any node $\phi(a_i \cdot x - b_i)$ in (1.1). This can be trivially extended to a node in the outer layer of a two layer neural net by noting that $\phi(\phi(a_i \cdot x - b_i) - d) = \phi(a_i \cdot x - b_i)$ when $0 < d < 1$. We identify apparantly broader classes of functions that permit reasonable approximation bounds using two layer networks. Approximation and estimation bounds for the two hidden layer case will be built up from existing results for the single layer case.

1.2 Historical Background and Motivation

In standard parametric function estimation, the target function to be estimated is assumed to take on a fixed parametric form. One can appeal to, for example, Searle [55], if the model is linear in all the parameters, that is $Y = \theta \cdot X + \epsilon$, where ϵ is the error. If it is non-linear in at least one parameter, for example, $Y = g(X, \theta) + \epsilon$ with g non-linear in θ, standard non-linear regression techniques are also available, as in Seber and Wild [56]. In both cases, it is customary to consider only the error that

arises from the estimation of the parameters. However, there exists many situations whereby we do not know which parametric family contains the target function. We may then wish to approximate the target function with a parametric family, and then find the best function from this parametric family that best fits. Thus, we have both an approximation error term (the bias) and an estimation error term (the variance) that contributes to the overall error.

These parametric families of functions are not restricted to a given parameter size. Rather, the dimension of the family is allowed to grow at a certain rate as a function of the sample size. Such families can be, for example, a family of single hidden layer feed-forward neural networks. It has been shown by Cybenko [16] and Hornik *et al* [29] that neural networks can be used to approximate continuous functions defined over bounded subsets of $\mathcal{R}^d$, to any arbitrary degree of accuracy by increasing the number of nodes. However, one also increases the number of parameters by increasing the number of nodes. Barron [1] showed how one can balance the two objectives of small approximation error and small estimation error. An approximation bound was obtained in Barron [3] and this was used together with [1] to obtain an overall mean squared estimation error bound in Barron [4].

In Barron [2, 3], single hidden layer neural net approximation bounds were derived for functions $f(x)$ defined over a bounded set $\mathcal{S}$ of $\mathcal{R}^d$, with Fourier representation $f(x) = \int_{\mathcal{R}^d} e^{i\omega \cdot t} \tilde{f}(\omega) d\omega$ and $|\omega \tilde{f}(\omega)|$ integrable. The bounds for a network f_T

that best approximates f are

$$\|f - f_T\|_2 \leq \frac{2C_{f,\mathcal{S}}}{T^{\frac{1}{2}}} \tag{1.3}$$

and

$$\|f - f_T\|_\infty \leq \gamma_d \frac{C_{f,\mathcal{S}}}{T^{\frac{1}{2}}} \tag{1.4}$$

where $C_{f,\mathcal{S}} = \int |\omega|_{\mathcal{S}} |\tilde{f}(\omega)| d\omega$, $|\omega|_{\mathcal{S}} = \sup_{x \in \mathcal{S}} |\omega \cdot x|$ and γ_d is some constant depending on the dimension d. An $\mathcal{L}_2$ estimation bound was also obtained in Barron [4]. This was

$$E\|f - \hat{f}_{T,N}\|_2^2 \leq K \left[\frac{C_{f,\mathcal{S}}^2}{T} + \frac{Td}{N} \log N \right], \tag{1.5}$$

where $\hat{f}_{T,N}$ is obtained by minimizing a sum of squared errors with suitable constraints on the parameter values. Here and elsewhere we use K to denote a constant. When $\hat{T}$ is selected by a penalized least squares criterion, the estimator $\hat{f} = \hat{f}_{\hat{T},N}$ achieves a risk bound of

$$E\|f - \hat{f}\|_2^2 \leq K C_{f,\mathcal{S}} \left(\frac{d \log N}{N} \right)^{\frac{1}{2}}$$

where K is a constant. The indices of resolvability in Barron [4] provide these bounds.

Here it is critical that the internal parameters (a_i, b_i) are adjusted to fit the target function. The bounds given above do not suffer from the curse of dimensionality in contrast to traditional methods of linear approximation and estimation. Indeed, suppose any T functions $g_1, \cdots, g_T$ are fixed (not adjusted to the target) as in the case of traditional polynomial or Fourier expansions, and that P_X is uniform on $\mathcal{S} = [-1, 1]^d$, then it is shown in Barron [4] that for at least one (and indeed for

most) f with $C_{f,\mathcal{S}} \leq V$ the $\mathcal{L}_2$ error of linear projection onto the span of the given functions g_i is at least $\frac{V}{dT^{1/d}}$. Correspondingly, the mean square error would be of order $\frac{1}{T^{2/d}} + \frac{T}{N}$, which at best is of order $\left(\frac{1}{N}\right)^{\frac{2}{2+d}}$. The number of terms T and the sample size N would need to be exponentially large in the dimension d to obtain accurate approximations and estimates. While such a rate $\left(\frac{1}{N}\right)^{\frac{2}{2+d}}$ (as in Ibragimov and Hasminskii [31], Nussbaum [43] and Hall [24]) is minimax optimal for estimation of functions with a bound on the gradient, we see that for $d > 2$ we can achieve a much better rate $\left(\frac{\log N}{N}\right)^{1/2}$ provided the gradient has an integrable transform and provided an estimation procedure is used that suitably adjusts the bases to the data (for example by fitting the internal parameters (a_i, b_i) in the nodes $\phi(a_i \cdot x - b_i)$ in the neural net model). Presumably adaptive selection and fitting of polynomial or trigonometric terms (with the frequencies serving as the internal parameters adjusted to the data) could achieve comparable performance to what is achieved here using the neural net, though we do not investigate that issue here.

Non-parametric curve estimates, which are nonadaptive, such as kernel methods and series expansions in which the bandwidths of the kernel or the first T terms in the series are preselected in accordance with a presumed smoothness class, have mean squared error that converges at the worst case rates (that is, the minimax rates) for functions in the standard smoothness classes. See for example, Härdle [25], Hall [24] and Stone [58]. The main problem is that they do not necessarily adapt to whatever additional regularity the target function may possess. Adjusting the choice of the

bandwidth or of the number of terms T by model selection criterion can provide some adaptivity, in which such an estimator achieves the minimax rates simultaneously in customary smoothness classes (for adaptive series methods see for example, Shibata [57], Li [36] and Polyak and Tsybakov [47], for kernel methods see for example, Müller and Stadtmüller [42], Schucany [54]). However, such mild adaptivities are not able to deal well with high-dimensionality. A greater degree of adaptivity is required, in which a subset of terms in series expansions are selected or in which parameters of nonlinear basis functions are adjusted in accordance with a penalized squared error criterion (see Yang and Barron [60], Barron *et al* [6]). In the spirit of such works on model selection and adaptation, we will derive in this thesis adaptive risk bounds that are more suitable for use with one and two hidden layer neural nets. We find more general types of regularity for target functions that allow the neural net estimates to perform at rates that do not exhibit the curse of dimensionality effects.

There are other methods of non-parametric estimation that attract current interest, such as projection pursuit, CART (classification and regression trees) and MARS (multivariate adaptive regression splines). A detailed discussion on the projection pursuit method is found in Huber [30] and the follow-up discussions. A relevant application of CART to function estimation is found in Breiman *et al* [10, Chapter 8] and the reader is referred to Friedman [19] for MARS. In projection pursuit, a large dimensional domain space is projected onto "interesting" low-dimensional spaces, and the function is fitted in that particular direction. The final fit is the sum of all the

fitted functions over these "interesting" directions. For example, a T term projection pursuit representation take the form

$$f_T(x) = \sum_{i=1}^{T} g_i(a_i \cdot x) \tag{1.6}$$

with the parameter a_i and the ridge functions g_i to be fitted from the data. Note the similarity between (1.1) and (1.6). (1.6) can be intepreted as a single neural net implementing different ridge functions g_i on its nodes. A more detailed discussion on projection pursuit regression is also found in Friedman and Stuetzle [20].

In CART, as applied to function estimation and the fitting of regression surfaces, splits are made in the domain space (assumed bounded) and the target function is estimated as a piecewise constant function in the various partitions bounded by these splits. In MARS, splines are fitted over these partitions instead. Like CART, the neural net (implementing the unit step function) produces piecewise constant function approximation. Usually CART selects cuts of regions oriented with the co-ordinate axes. In contrast the neural net selects jumps of global extent of arbitrary orientation and location. In both CART and MARS, the more partitions there are, the better the approximation. A similar situation occurs with neural nets; better approximation can be obtained by using more nodes and layers. However, this increases the number of parameters to be estimated and this does not necessarily decrease the overall estimation error. For CART and MARS, the fitted function to $f(x) = E[Y_i|X_i = x]$

is fitted iteratively as

$$\hat{f}_k(x) = \arg\min_{g \in \mathcal{G}_k} \Big(\frac{1}{N} \sum_{i=1}^{N} (Y_i - g(X_i))^2 + \lambda R(g) \Big), \tag{1.7}$$

where $\mathcal{G}_k$ is a class of splits of the partition used at the step $(k-1)$. In MARS, the splines meet one another where the partitions meet and $R(g)$ is a functional that increases with increasing roughness of g. It is usually the integrated squared Laplacian of the function g. For CART, the fitted function g is piecewise constant over these partitions. The more splits there are, the more jumps there are in g. Here $R(g)$ is a functional that increases with the number of splits in g. It can be intepreted as the cost of adding one more split to the tree. In the case of neural nets, we use the least squares estimator with a complexity penalty as in (3.3). This is obtained from (3.4). Likewise we will consider in chapter 4 iterative estimates where a new node is introduced at each step. Estimation of functions using neural nets is just one of the many nonparametric methods of function estimation.

There is also the connection between single layer neural nets and mixtures of logistic regression models for binary response variables. Suppose Y is a binary response variable, then such a model takes the form

$$P[Y = y|x, \beta] = \sum_{i=1}^{T} p_i \frac{\exp[y(\beta_i \cdot x - \beta_{0,i})]}{1 + \exp(\beta_i \cdot x - \beta_{0,i})} \tag{1.8}$$

When $Y = 1$, (1.8) is a single hidden layer neural net implementing the usual logistic sigmoid. Indeed, it is of the form (1.1).

1.3 New Results

Chapter 2 is about approximation bounds for two hidden layer feedforward neural nets. Some new results are found in this chapter. For example, one of our $\mathcal{L}_2$ approximation bounds in the case of the target function having variation V_f with respect to a class of ellipsoids is of order $1/T_1^{\frac{1}{2}} + 1/T_2^{\frac{1}{4}}$ as $T_1, T_2 \to \infty$. More precisely, the approximation error is shown to be bounded by a constant times $\frac{V_f}{T_1^{1/2}} + \frac{V_f d}{T_2^{1/4}}$, where T_1 is the number of nodes in the outer layer and T_2 is the number of nodes in the inner layer in the approximation f_{T_1,T_2}. A corresponding bound for the mean squared estimation error in chapter 3, when the parameter space is discretized in a suitable manner, yields $O(\|f - f_{T_1,T_2}\|_2^2) + O\left(\frac{m_{T_1,T_2}}{N} \log m_{T_1,T_2} N\right)$ where m_{T_1,T_2} is the dimension of the parameter space and N is the sample size. In these bounds, it can be seen that one need not have a large number of nodes (exponential in dimension) in order to achieve the desired accuracy.

In contrast, it does not appear to be possible to approximate well the indicator of a single ellipse (nor even a ball) by a single layer network. Thresholding certain single layer networks does provide an accurate approximation in this case. Such thresholding is a second layer of nonlinearity and we have used this technique to formulate the outer layer of our two layer approximations.

In chapter 3, we derive bounds for the mean square prediction error for two

hidden layer neural net estimators. We have data $(X_i, Y_i)_{i=1}^N$, which are independent with joint probability distribution $P_{X,Y}$. The target function is $f^*(x) = E[Y|X = x]$ and its range is assumed to be bounded. The estimator is selected over a class of suitable neural network models and it is the minimizer of the empirical estimation error plus a penalty term. The penalty term is added to to help the neural net estimator adapt the size of the network to the target function.

Chapter 3 is divided into two sections. In the first section, the parameter space of the estimator is discretized in the same manner as in Barron [4]. The estimated function takes the form of a two hidden layer neural network that implements ramp activation functions. These ramp functions are Lipshitz bounded. The parameter space for the estimator is discretized, with a fixed bound on the outer weights of the outer layer, and bounds on the inner weights of both layers that grow with the number of nodes in each repective layer. The penalty term in this case is the log cardinality of the discretized parameter space.

In the second section, we deal with function estimators that are in the class of neural networks implementing the step activation function with no restriction on the weights. Neural networks that implement the step activation functions do not satisfy the Lipshitz condition in the first section. In the single hidden layer case, our result is the extension of Lee *et al* [35] to include a penalty term, but we do not involve the bounded fan-in property that they assume. (The bounded fan-in property is the restriction for computational purposes that all but a small number of the input

weights are zero in each term.)

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. This includes functions f for which f/V is in the closure of the convex hull of signed (plus or minus) indicators of half-spaces. 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 is bounded above by

$$E\|f - \hat{f}_{\hat{T}}\|_2^2 \leq KV^2 \left(\frac{d \ln N}{N}\right)^{\frac{1}{2}}. \tag{1.9}$$

We 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. Our estimator in section 3.2 does not require advance knowledge of V_f. We show that the mean squared error between $\hat{f}_{\hat{T}}$ and f is bounded by

$$E\|f - \hat{f}_{\hat{T}}\|_2^2 \leq KV_f^2 \left(\frac{d \ln N}{N}\right)^{\frac{1}{2}} + \frac{K'V_f^4}{N}. \tag{1.10}$$

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

$$E\|f - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2 \leq Kd^{3/2}V_{f,\mathcal{E}}^2 \left(\frac{\ln N}{N}\right)^{\frac{1}{4}}. \tag{1.11}$$

A heuristic algorithm for fitting single hidden layer nets is presented in chapter 4. The target function (when normalized) is assumed to be in the closure of the convex hull of sigmoids. The sigmoids here are the odd-symmetric logistic sigmoid $\phi(z) =$

$(\exp(z) - \exp(-z))/(\exp(z) + \exp(-z))$ and $\psi(z) = 1/(1 + \exp(-z))$ which differs from the logistic sigmoid $\phi(z)$ by a simple rescaling of the output. The algorithm is based on Jones' [32] greedy approximation in Hilbert spaces. Our algorithm adopts the iterative procedure of Jones' [32] algorithm. A crucial step in his algorithm is the maximization of the cross-product of the residual from the previous fit with the candidate sigmoid for the current fit. The iterative procedure provides an apparant simplification in the computation (compared to overall least squares). However, the optimization that remains at each step is NP-complete. In our algorithm, we do not seek to maximize the cross-product of residuals and new sigmoid, but rather we maximize a concave lower bound to it. We are content with our fit as long as the resulting cross-product is sufficiently large. Some simulation results will be presented.

In chapter 5, we will mention possible ways of extending some of the results in this dissertation and potential difficulties that might arise.

Chapter 2

Approximation Bounds

2.1 Integral Representation Theorems

We begin this chapter by a short discussion on how integral representation theorems can be used to derive single hidden layer neural network approximations to given functions. Suppose the function $f(x)$ has the representation

$$f(x) = \int_S K(\alpha, x) v(d\alpha) \tag{2.1}$$

where v is a probability measure on α, then an approximation to f is

$$f_T(x) = \frac{1}{T} \sum_{i=1}^{T} K(\alpha_i, x) \tag{2.2}$$

with α_i sampled identically and independently from the distribution v. In particular, when $K(\alpha, x) = 1_{\{\alpha \cdot x > t\}}$ is the indicator of a half-space, a wedge $K(\alpha, x) = |\alpha \cdot x - b|$

or any ridge function $K(\alpha, x) = K(\alpha \cdot x)$ with a given K, this approximation is a single hidden layer neural net if K is the function implemented in its hidden nodes.

Lemma 2.1 *Let $P(\cdot)$ be a distribution on x and $||f_T - f||_2^2 = E_x|f_T(x) - f(x)|^2$. If $E_x E_\alpha K^2(\alpha, x) \leq C$ for some constant C, then there is an approximation $f_T(x) = \frac{1}{T}\sum_{i=1}^{T} K(\alpha_i, x)$ that satisfies*

$$\|f_T - f\|_2^2 \leq \frac{C}{T}.$$

Proof : If $E_x E_\alpha K^2(\alpha, x) \leq C$ for some constant C, then the cross-product terms in the following quadratic expansion vanish due to the independence of α_i and α_j, so that

$$\begin{aligned}
E_{\alpha_1,\cdots\alpha_T}\|f_T - f\|_2^2 &= E_{\alpha_1,\cdots\alpha_T}E_x\left|\frac{1}{T}\sum_{i=1}^{T}(K(\alpha_i, x) - E_{\alpha_i}K(\alpha_i, x))\right|^2 \\
&= \frac{1}{T^2}E_x E_{\alpha_1,\cdots\alpha_T}\left[\sum_{i=1}^{T}(K(\alpha_i, x) - E_{\alpha_i}K(\alpha_i, x))^2 + \right. \\
&\qquad \left. 2\sum_{i<j}^{T}(K(\alpha_i, x) - E_{\alpha_i}K(\alpha_i, x))(K(\alpha_j, x) - E_{\alpha_j}K(\alpha_j, x))\right] \\
&= \frac{1}{T^2}\sum_{i=1}^{T} E_x E_{\alpha_i}|K(\alpha_i, x) - f(x)|^2 \\
&\leq \frac{1}{T^2}\sum_{i=1}^{T} E_x E_{\alpha_i}|K(\alpha_i, x)|^2 \\
&\leq \frac{C}{T}.
\end{aligned}$$

Since the expected value of $\|f_T(\cdot, \alpha_1, \cdots, \alpha_T) - f(\cdot)\|_2^2$ has this bound, there exists some $\alpha_1, \cdots, \alpha_T$ such that $\|f_T(\cdot, \alpha_1, \cdots, \alpha_T) - f(\cdot)\|_2^2$ is not greater that the bound,

that is, there is an approximation f_T that satisfies

$$\|f_T - f\|_2^2 \leq \frac{C}{T}.$$

□

Here we see a probabilistic method used in a deterministic approximation problem. It is used to prove the existence of an accurate approximation. The existence of an integral representation (2.1) provides opportunity for Monte Carlo approximation with a dimension independent accuracy.

As an example, the representation given in Barron [2, Theorem 2] is

$$f(x) - f(0) =$$

$$v \int_{\mathcal{R}^d} \int_0^1 (1_{\{\alpha \cdot x < -t\}} \sin(-t|\omega|_S + \theta_\omega) - 1_{\{\alpha \cdot x > t\}} \sin(t|\omega|_S + \theta_\omega)) p(\omega, t) dt d\omega \quad (2.3)$$

for x in B, where $|\omega|_S = \sup_{x \in S} |\omega \cdot x|$, $\alpha = \omega/|\omega|_S$, and $p(\omega, t)$ is a probability density depending on the spectral representation of f. The marginal density of t is the uniform density over $[0, 1]$. More specifically,

$$p(\omega, t) = \frac{1}{v} |\omega|_S |\tilde{f}(\omega)|,$$

where the constant v is given by the spectral norm

$$v = C_{f,S} = \int_{\mathcal{R}^d} \int_0^1 |\omega|_S |\tilde{f}(\omega)| dt d\omega,$$

and $\tilde{f}(\omega) = \exp(i\theta_\omega)|\tilde{f}(\omega)|$ is the decomposition of the Fourier transform of f into the magnitude $|\tilde{f}(\omega)|$ and phase θ_ω. The integral representation (2.3) is valid for functions

f such that the spectral norm $\int_{\mathcal{R}^d} |\omega||\tilde{f}(\omega)|d\omega$ is finite, such as the Gaussian. The approximation is

$$f_T(x) - f(0) = \frac{v}{T}\sum_{i=1}^{T}\left(a_i 1_{\{\alpha_i \cdot x < -t_i\}} + b_i 1_{\{\alpha_i \cdot x > t_i\}}\right) \tag{2.4}$$

with α_i and t_i chosen from the density $p(\omega, t)$, and $a_i = \sin(-t|\omega_i|_S + \theta_i)$, $b_i = -\sin(t|\omega_i|_S + \theta_i)$. Note that (2.4) depends on x only through the step functions.

We illustrate the derivation of (2.3). First note that

$$f(x) - f(0) = \int (\exp(i\omega \cdot x) - 1)\tilde{f}(\omega)d\omega$$

and that

$$\begin{aligned} \exp(iz) - 1 &= i\int_0^z \exp(iu)du \\ &= \begin{cases} i\int_0^c 1_{\{z>u\}} \exp(iu)du, & 0 \le z \le c \\ -i\int_0^c 1_{\{z<-u\}} \exp(-iu)du, & -c \le z < 0. \end{cases} \end{aligned} \tag{2.5}$$

Because only one of the two expressions in (2.5) is non-zero depending on the sign of z, it follows that, for $|z| \le c$,

$$\exp(iz) - 1 = i\int_0^c \left(1_{\{z>u\}} \exp(iu) - 1_{\{z<-u\}} \exp(-iu)\right) du.$$

Substituting $z = \omega \cdot x$ and $c = \sup_S |\omega \cdot x| = |\omega|_S$ and integrating yields

$$\begin{aligned} &f(x) - f(0) = \\ &\quad i\int_{\mathcal{R}^d} \left(\int_0^1 \left(1_{\{z>u\}} \exp(iu) - 1_{\{z<-u\}} \exp(-iu)\right) du\right) \tilde{f}(\omega)d\omega. \end{aligned} \tag{2.6}$$

Finally, we take the real part of both sides of (2.6), do a change of variables with $u = |\omega|_S t$ for $0 \le t \le 1$ to obtain the integral representation in (2.3).

Barron's result provides an integral representation in terms of half-spaces, the result mentioned in Goodey and Weil [21] also provides integral representations of functions. Such representations for the support function

$$h_K(x) = \sup_{u \in K} u \cdot x, \qquad x \in \Sigma^{d-1}$$

of certain convex sets K that are centrally symmetric (where Σ^{d-1} is the surface of the unit ball) provide integral representations in terms of half-spaces. A zonotope Z_T is a convex body which can be represented as a set sum of line segments L_k, that is

$$Z_T = \{l_1 + \cdots l_T : l_k \in L_k\}.$$

Such zonotopes have support functions of the form $\sum_{i=1}^{T} |a_i \cdot x|$. More generally, zonoids K are the Hausdorff-metric limits of zonotopes with support functions of the form

$$h_K(x) = \int_{\Sigma^{d-1}} |x \cdot v| d\rho(v) \tag{2.7}$$

where ρ is a non-negative symmetric measure over the surface of the unit ball Σ^{d-1} in $\mathcal{R}^d$. This result is proven in Schneider [52]. Then a T-term approximation similar to a neural net is

$$h_T(x) = \frac{|\rho|}{T} \sum_{i=1}^{T} |x \cdot v_i| \tag{2.8}$$

with the nodes implementing the wedge function.

For a third setting in which integral representations of the form (2.1) arises, we consider classes of harmonic functions. A harmonic function can also be written as a linear combination of basis functions that are harmonic. Regular spherical harmonic

functions are homogeneous polynomials of some fixed degree in d-dimensions that satisfy Laplace's equation, that is, $\nabla^2 f(x) = 0$. Let $\{S_q(x)\}$ be a set of such regular spherical harmonics of degree q in x. Then $S_q(x)$ has the following representation, known as the Funk-Hecke formula,

$$\int_{\Sigma^{d-1}} g(\alpha \cdot x) S_q(x) \sigma_{d-1}(dx) = \lambda S_q(\alpha), \tag{2.9}$$

where $\lambda = \sigma_{d-1}(\Sigma^{d-1}) \int_{-1}^{1} g(t) P_q(t) (1-t^2)^{\frac{d-3}{2}} dt$, where σ_{d-1} is the surface area measure of the unit sphere in $\mathcal{R}^d$ and g is a continuous function, and P_q is the Legendre polynomial of degree q. This representation is of particular interest since $g(\alpha \cdot x)$ is a ridge function. It can be applied to neural net approximation of harmonic functions. Further discussion on harmonic functions may be found in Müller [41].

Although this work will address two hidden layer neural net approximation of ellipsoids and functions in the convex hull of the set of ellipsoids, such results depend on good single layer neural net approximation of the Gaussian function as we shall see in the next section. It is possible to express the Gaussian function (restricted over a bounded set $S \subset \mathcal{R}^d$) with an integral representation of the form (2.3). In the next section, we also give two other integral representations of the Gaussian, and we make use of one of these to obtain the upper bound to the neural net approximation of a ball. Integral representations of the form (2.7) and (2.9) open up the possibility of using single hidden layer neural net approximations that implement other types of ridge activation functions.

2.2 Approximating Balls and Ellipsoids with Neural Nets

2.2.1 The Classical Approach with Polytope Approximation

There already exists a rich literature on approximation of convex bodies with other sorts of convex bodies and polytopes. See, for example, Gruber [22], Fejes Tóth [18]. Like other convex bodies, a ball is an infinite intersection of tangent half-spaces. For a unit ball B in $\mathcal{R}^d$,

$$B = \bigcap_{a \in \Sigma^{d-1}} \{a \cdot x \leq 1\}, \tag{2.10}$$

where Σ^{d-1} is the unit sphere in $\mathcal{R}^d$. If we approximate it with the intersection of T (greater than $d+1$), of the half-spaces, then we are approximating the ball with a T-faced polytope $\mathcal{P}_T$.

There are results that bound the approximation error between convex bodies and their polytope approximators. Dudley [17] has shown that for each convex body B, there exists a constant c such that for every $T > d+1$ there is a polytope $\mathcal{P}_T$ achieving

$$\delta^H(B, \mathcal{P}_T) \leq \frac{c}{T^{\frac{2}{d-1}}}, \tag{2.11}$$

where δ^H is the Hausdorff metric. Results from Schneider and Wieacker [53], Gruber and Kenderov [23], have shown that for a convex body with sufficiently smooth

boundary such as the ball B, there exists a constant c such that for every polytope $\mathcal{P}_T$,

$$\delta(B, \mathcal{P}_T) \geq \frac{c}{T^{\frac{2}{d-1}}}, \tag{2.12}$$

where δ can be either the Hausdorff or the Lebesgue measure of the symmetric difference. Hence for an approximation error of ϵ, we would require a polytope with many faces of order $(\frac{1}{\epsilon})^{\frac{d-1}{2}}$, which is exponential in d. To avoid this curse of dimensionality, we will use T half-spaces in the approximation in a different manner.

To illustrate the idea, consider the set of points in at least k out of T given half-spaces. For instance, if we were given the $T = 9$ half-spaces determining the polygon approximation in figure (2.1), $k = 9$ yields the nonagon inscribed in the circle. In figure (2.2), we use $T = 9$ half-spaces, but we set the threshold at $k = 8$ to obtain the star-shaped approximation shown. In higher dimensions, our approximation will look somewhat like a jagged multi-faceted star-shaped object.

Here we can think of the T half-spaces as providing a test for membership in the set. Instead of requiring all T tests to be passed, we permit membership with at least k passed out of T. An extension of this idea is to weigh each test and determine membership by a weighted count exceeding a threshold.

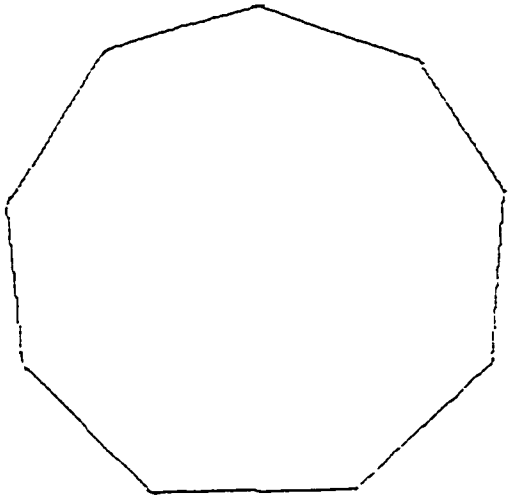

Figure 2.1

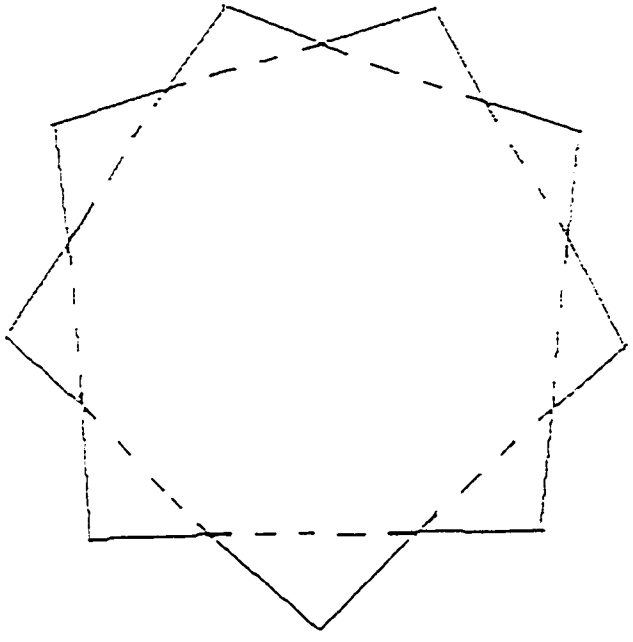

Figure 2.2

While polygon approximation may appear superior in the low-dimensional example given in the figure, in high dimensions, polytopes have extremely poor accuracy as shown in equation (2.12). In contrast we show that the use of a weighted count to determine membership in a set permits accuracy that avoids the curse of dimensionality. Indeed, with $2T = \frac{cd^2}{\epsilon^2}$ indicators of half-spaces, where c is a constant, we threshold a linear combination of them, in order to obtain accuracy ϵ. Note that the number of indicators of half-spaces needed is only quadratic in d and not exponential in d as in the classical method.

Our approximation to a ball takes the form

$$\mathcal{N}_{2T} = \{x \in \mathcal{R}^d : \sum_{i=1}^{2T} c_i 1_{\{a_i \cdot x \geq b_i\}} \geq k\}.$$

Let $\tilde{f}_{2T} = 1_{\mathcal{N}_{2T}}$ be the indicator (characteristic) function of this set. In neural network terminology, we are using a two layer perceptron approximation to the indicator of a ball, where the second layer thresholds the linear combination at the level k. We show that there is a constant c such that for every T and d, there is such an approximation $\mathcal{N}_{2T}$ such that the Hausdorff distance between a ball B_R of radius R and $\mathcal{N}_{2T}$ satisfies

$$\delta^H(B_R, \mathcal{N}_{2T}) \leq cR\sqrt{\frac{d(d+1)}{T}},$$

where c is a constant. A special role in the analysis is played by probabilistic methods and approximation of Gaussian functions.

2.2.2 Some background and the Gaussian function

A single hidden layer feedforward sigmoidal network is a family of real-valued functions $f_T(x)$ of the form

$$f_T(x) = \sum_{i=1}^{T} c_i \phi(a_i \cdot x + b_i) + k, x \in \mathcal{R}^d \tag{2.13}$$

parametrized by internal weight vectors a_i in $\mathcal{R}^d$, internal location parameter b_i in $\mathcal{R}$, external weights c_i and a constant term k (Cybenko [16], Haykin [27]). We choose to pull out the constant term k from the T-term neural network in (2.13) [compare with (1.1)] for convenience throughout this chapter since our integral representation of the Gaussian also has a constant term. By a sigmoidal function, we mean any nondecreasing functions on $\mathcal{R}$ with distinct finite limits at $+\infty$ and $-\infty$. Such a network has d inputs, T hidden nodes and a linear output unit. It implements ridge-functions $\phi(a_i \cdot x - b_i)$ on the nodes in the hidden layer. Here we will exclusively use the Heaviside function $\phi(z) = 1_{\{z \geq 0\}}$, in which case (2.13) is a linear combination of indicators of half spaces. Such a network is also called a perceptron network (Rosenblatt [49, 50]). Thresholding the output of a single hidden layer neural net at level k_1, we obtain $\tilde{f}_T(x) = \phi(f_T(x) - k_1)$ which equals

$$\tilde{f}_T(x) = \phi(\sum_{i=1}^{T} c_i \phi(a_i \cdot x + b_i) + k'). \tag{2.14}$$

For simplicity in the notation, we will often omit the parameters a_i, b_i, c_i and k in the arguments of f_T and $\tilde{f}_T$.

To approximate a ball we first consider approximation of the Gaussian function $f(x) = \exp(-\frac{|x|^2}{2})$ and then take level sets. A level set of a function f at level k, where k is real, is simply the set $\{x \in \mathcal{R}^d : f(x) \geq k\}$. Using the fact that the Gaussian is a positive definite function with Fourier transform $(2\pi)^{-\frac{d}{2}} \exp(-\frac{|\omega|^2}{2})$, so that f has a representation in the convex hull of sinusoids (sines and cosines), it is known that $f(x)$ can be expressed using the convex hull of indicators of half-spaces (see Barron [2, 3], Hornik *et al* [29], Yukich *et al* [61]). We take advantage of a similar representation here. We use $|\cdot|$ to denote the Euclidean $\mathcal{L}_2$ norm.

Let B_K be a ball of radius $K \geq 1$. Later we will arrange the construction of the neural net approximation $\mathcal{N}_{2T}$ of the unit ball B centered at the origin such that it is shown to be contained in B. We have the following lemma.

Lemma 2.2 *The Gaussian function on B_K satisfies*

$$f(x) = \int_{\mathcal{R}^d} \int_{-|a|K}^{|a|K} 1_{\{a\cdot x + b \geq 0\}} \sin(b) \frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} dbda + \exp(-\frac{K^2}{2}). \tag{2.15}$$

Proof : Starting with the right hand side of (2.13) and recalling that $|a \cdot x| \leq |a|K$ for all $x \in B_K$, we obtain

$$\begin{aligned} & \int_{\mathcal{R}^d} \int_{-|a|K}^{|a|K} 1_{\{a\cdot x + b \geq 0\}} \sin(b) \frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} dbda \\ = & -\text{Im} \int_{\mathcal{R}^d} \int_{-|a|K}^{|a|K} 1_{\{a\cdot x + b \geq 0\}} \exp(-ib) \frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} dbda \\ = & -\text{Im} \int_{\mathcal{R}^d} \left[\int_{a\cdot x - |a|K}^{a\cdot x + |a|K} 1_{\{s \geq 0\}} \exp(-is) ds \right] \frac{\exp(ia\cdot x) \exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} da \end{aligned} \tag{2.16}$$

$$= -\mathrm{Im} \int_{\mathcal{R}^d} \left[\int_0^{a\cdot x+|a|K} \exp(-is)ds \right] \frac{\exp(ia\cdot x)\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} da$$

$$= \mathrm{Im}\, i \int_{\mathcal{R}^d} [1 - \exp(-ia\cdot x)\exp(-i|a|K)] \frac{\exp(ia\cdot x)\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} da$$

$$= \exp(-\frac{|x|^2}{2}) - \int_{\mathcal{R}^d} \frac{\exp(-i|a|K)\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} da \tag{2.17}$$

$$= f(x) - \exp(-\frac{K^2}{2}). \tag{2.18}$$

In (2.16), we did a substitution $s = a\cdot x + b$.

□

Here $\exp(-\frac{K^2}{2})$ is the value of the Gaussian evaluated on the surface of the ball B_K. As we will see later, when approximating the unit ball B, we can arrange for the neural net level set $\mathcal{N}_{2T}$ to be entirely contained in B_K for $K \geq 1$, and hence take $K = 1$.

Decomposing the integral representation of f into positive and negative parts, we have

$$f(x) - \exp(-\frac{K^2}{2}) = f_1(x) - f_2(x) \tag{2.19}$$

$$= \int_{\mathcal{R}^d} \int_{-|a|K}^{|a|K} 1_{\{a\cdot x+b\geq 0\}} \sin^+(b) \frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} dbda$$

$$- \int_{\mathcal{R}^d} \int_{-|a|K}^{|a|K} 1_{\{a\cdot x+b\geq 0\}} \sin^-(b) \frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} dbda$$

$$= \nu_1 \int 1_{\{a\cdot x+b\geq 0\}} dV_1 - \nu_2 \int 1_{\{a\cdot x+b\geq 0\}} dV_2, \tag{2.20}$$

where V_1 is the probability measure for (a, b) on $\mathcal{R}^{d+1}$ with density

$1_{\{-|a|K<b<|a|K\}} \sin^+(b) \frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}\nu_1}$ with normalizing constant

$$\nu_1 = \int_{\mathcal{R}^d} \int_{-|a|K}^{|a|K} \sin^+(b) \frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} dbda$$

and similarly for V_2 and ν_2 (with $\sin^-(b)$ in place of $\sin^+(b)$). Here we use the convention $z^+ = z \vee 0$ and $z^- = (-z)^+$ for positive and negative parts. The total variation of the measure used to represent f is

$$\begin{aligned} \nu &= \nu_1 + \nu_2 \\ &= \int_{\mathcal{R}^d} \int_{-|a|K}^{|a|K} |\sin(b)| \frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} dbda \qquad (2.21) \\ &< \int_{\mathcal{R}^d} \frac{2|a|K \exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}} da \\ &\leq 2K\sqrt{d}. \qquad (2.22) \end{aligned}$$

An integral representation of the Gaussian as an expected value invites Monte Carlo approximation by a sample average. In particular, both $f_1(x)$ and $f_2(x)$ in (2.19) are expected values of indicators of half-spaces in $\mathcal{R}^d$. Thus a $2T$-term neural net approximation to $f(x)$ is then

$$f_{2T}(x) = \frac{\nu_1}{T} \sum_{i=1}^{T} \phi(a_i \cdot x + b_i) - \frac{\nu_2}{T} \sum_{i=n+1}^{2T} \phi(a_i \cdot x + b_i), \qquad (2.23)$$

where the parameters $(a_i, b_i)_{i=1}^T$ are drawn at random independently from the distribution V_1 and $(a_i, b_i)_{i=1}^{2T}$ from V_2. The sampling scheme is simple. For example, to obtain an approximation for $f_1(x)$, first draw a from a standard multivariate normal distribution over $\mathcal{R}^d$, then draw b from $[-|a|K, |a|K]$ with density proportional to

$\sin^+(b)$. We could have also used $\left(\frac{a}{|a|}, \frac{b}{|a|}\right)$ or (ka, kb), where k is positive, in place of (a, b) because of the scale invariant property $(\phi(z) = \phi(kz), k > 0)$ of the step activation function.

2.2.3 Bounding the $\mathcal{L}_\infty$-approximation for the Gaussian

We now bound the $\mathcal{L}_\infty$-approximation error between $f(x)$ and $f_{2T}(x)$. We will draw on symmetrization techniques and the concept of Orlicz norms in empirical process theory (see for example, Pollard [46]), and the theory of Vapnik-Červonenkis classes of sets (Vapnik and Červonenkis [59]). With the particular choice of $\Psi(x) = \frac{1}{5}\exp(x^2)$ used by Pollard [46], the Orlicz norm of a random variable Z is defined by

$$\|Z\|_\Psi = \inf\left\{C > 0 : E\exp\left(\frac{Z^2}{C^2}\right) \leq 5\right\}.$$

We examine the approximation error between $f_1(x)$ and $f_{1,T}(x)$, its T-term neural net approximation, first. From empirical process theory, the following lemma is obtained.

Let a parameterized class of sets $\mathcal{H} = \{H_\xi : \xi \in \Xi\}$ in $\mathcal{R}^d$ be given where Ξ is a measurable space. Let $\tilde{\mathcal{H}} = \{\tilde{H}_x : x \in \mathcal{R}^d\}$, where $\tilde{H}_x = \{\xi : x \in H_\xi\}$, be the dual class of sets in Ξ parametrized by x.

First we define some terms that will be used in the lemma. Let $\mathcal{G}$ be a class of functions mapping from $\mathcal{X}$ to $\mathcal{R}$ and let $x_1, \cdots, x_N \in \mathcal{X}$. We say that $x_1, \cdots, x_N$ are shattered by $\mathcal{G}$ if there exists $r \in \mathcal{R}^N$ such that for each $b = (b_1, \cdots, b_N) \in \{0,1\}^N$,

there is an $g \in \mathcal{G}$ such that for each i,

$$g(x_i) \begin{cases} \geq r_i & \text{if } b_i = 1 \\ < r_i & \text{if } b_i = 0. \end{cases}$$

The pseudo-dimension is defined as

$$\dim_P(\mathcal{G}) = \max\{N : \exists x_1, \cdots, x_N, \quad \mathcal{G} \text{ shatters } x_1, \cdots, x_N\} \tag{2.24}$$

if such a maximum exists, and ∞ otherwise. For the class of unit step functions $\phi(a \cdot x + b)$, the pseudo-dimension and the VC-dimension D coincide and is $d+1$. The ϵ-packing number $D_T(\epsilon, \mathcal{L}_p)$ for a subset of a metric space is defined as the largest number m for which there exist points $t_1, \cdots, t_m$ in the subset of the metric space with $d_p(t_i, t_j) > \epsilon$ for $i \neq j$, where d_p is the $\mathcal{L}_p$ metric.

Lemma 2.3 *If $\tilde{\mathcal{H}}$ has VC-dimension D and if h is a function in the convex hull of the indicators of sets in $\mathcal{H}$ which possesses an integral representation*

$$h(x) = \int 1_{H_\xi}(x) P(d\xi) \text{ for } x \in \mathcal{S},$$

then there is a choice of $\xi_1, \xi_2, \cdots, \xi_T$ such that the approximation $h_T(x) = \frac{1}{T} \sum_{i=1}^{T} 1_{H_{\xi_i}}(x)$ satisfies

$$\sup_{x \in \mathcal{S}} |h_T(x) - h(x)| \leq 34 \sqrt{\frac{D}{T}} \tag{2.25}$$

Remark : More generally if h has an integral representation $h(x) = \int g_x(\xi) P(d\xi)$ with $|g_x(\xi)| \leq 1$, in terms of a family of functions $G = \{g_x(\cdot), x \in \mathcal{R}^d\}$ with pseudo-

dimension D (as defined in Pollard [46]), then there exists $\xi_1, \cdots, \xi_T$ such that

$$\sup_{x \in \mathcal{S}} |h(x) - h_T(x)| \leq 34\sqrt{\frac{D}{T}}$$

where the approximation $h_T(x)$ equals $\frac{1}{T}\sum_{i=1}^{T} g_x(\xi)$. For classes of sets, the pseudo-dimension and the VC-dimension coincide.

Proof : Let $g_x(\xi) = 1_{\tilde{H}_x}(\xi) = 1_{H_\xi}(x)$ and let σ_i be independent random variable taking the values ± 1 with probablity $\frac{1}{2}$. Define $\underline{\xi} = (\xi_1, \xi_2, \cdots, \xi_T)$, where the ξ_i are independently and identically distributed with respect to $P(\cdot)$, and $\underline{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_T)$. By symmetrization, using Jensen's inequality as in Pollard [46, page 7], for all $C > 0$, we have

$$E\Psi\left(\frac{\sup_{x \in \mathcal{S}} \left|\sum_{i=1}^{T} g_x(\xi_i) - Th(x)\right|}{C}\right) \leq E_{\underline{\xi}} E_{\underline{\sigma}} \Psi\left(\frac{2\sup_{x \in \mathcal{S}} \left|\sum_{i=1}^{T} \sigma_i g_x(\xi_i)\right|}{C}\right). \quad (2.26)$$

Conditioning on $\underline{\xi}$, we seek a value of C for which $E_{\underline{\sigma}} \Psi\left(\frac{2\sup_{x \in \mathcal{S}} \left|\sum_{i=1}^{T} \sigma_i g_x(\xi_i)\right|}{C}\right)$ is not greater than 5. This involves bounding the Orlicz norm $\|2\sup_{x \in \mathcal{S}} \left|\sum_{i=1}^{T} \sigma_i g_x(\xi_i)\right| \|_\Psi$ with $\underline{\xi}$ fixed. Using a result in Pollard [46, pages 35 – 37],

$$\left\| 2\sup_{x \in \mathcal{S}} \left| \sum_{i=1}^{T} \sigma_i g_x(\xi_i) \right| \right\|_\Psi \leq 18\sqrt{T} \int_0^1 \sqrt{\log D_T(\epsilon, \mathcal{L}_2)}\, d\epsilon, \quad (2.27)$$

where $D_T(\epsilon, \mathcal{L}_2)$ is the $\mathcal{L}_2$ ϵ-packing number for $\tilde{\mathcal{H}}$, where the $\mathcal{L}_2$ norm on Ξ is taken with respect to the empirical probability measure on $\xi_1, \xi_2, \cdots, \xi_T$.

From Pollard [46, page 14],

$$D_T(\epsilon, \mathcal{L}_2) \leq \left(\frac{3}{\epsilon}\right)^D, \quad (2.28)$$

uniformly over all $\xi_1, \xi_2, \cdots, \xi_T$. We now work out an upper bound to $\int_0^1 \sqrt{\log D_T(\epsilon, \mathcal{L}_2)} d\epsilon$. From the Cauchy-Schwartz inequality,

$$\begin{aligned} \int_0^1 \sqrt{\log D_T(\epsilon, \mathcal{L}_2)} d\epsilon &\leq \sqrt{\int_0^1 \log D_T(\epsilon, \mathcal{L}_2) d\epsilon} \\ &\leq \sqrt{D \log 3 - D \int_0^1 \log \epsilon d\epsilon} \\ &= \sqrt{(1 + \log 3) D}. \end{aligned} \tag{2.29}$$

Substituting (2.29) into (2.26), we see that

$$\left\| 2 \sup_{x \in \mathcal{S}} \left| \sum_{i=1}^{T} \sigma_i g_x(\xi_i) \right| \right\|_{\Psi} \leq 18 \sqrt{(1 + \log 3) T D} \tag{2.30}$$

From the definition of the Orlicz norm, the choice of $C_0 = 18\sqrt{(1 + \log 3) T D}$ ensures that

$$E_\sigma \Psi \left(\frac{2 \sup_{x \in \mathcal{S}} \left| \sum_{i=1}^{T} \sigma_i g_x(\xi_i) \right|}{C_0} \right) \leq 1, \forall \underline{\xi}.$$

and hence,

$$E \Psi \left(\frac{\sup_{x \in \mathcal{S}} \left| \sum_{i=1}^{T} g_x(\xi_i) - T h(x) \right|}{C_0} \right) \leq 1. \tag{2.31}$$

Thus we conclude that there exists $\xi_1, \xi_2, \cdots, \xi_T$ such that

$$\Psi \left(\frac{\sup_{x \in \mathcal{S}} \left| \sum_{i=1}^{T} g_x(\xi_i) - T h(x) \right|}{C_0} \right) \leq 1,$$

whence

$$\begin{aligned} \sup_{x \in \mathcal{S}} \left| \frac{1}{T} \sum_{i=1}^{T} g_x(\xi_i) - h(x) \right| &\leq \frac{18 \sqrt{D(1 + \log 3) \log 5}}{\sqrt{T}} \\ &\leq 34 \sqrt{\frac{D}{T}}. \end{aligned} \tag{2.32}$$

□

In our case, $\xi = (a, b)$ and $g_x(\xi) = 1_{H_\xi}(x) = 1_{\{a \cdot x + b \geq 0\}}$. The dual class of sets in Ξ are $\tilde{H}_x = \{\xi : g_x(\xi) = 1\} = \{(a, b) : a \cdot x + b \geq 0\}$. Since $(a, b) \in \mathcal{R}^d \times \mathcal{R}$, which is a vector space of dimension $d + 1$, the class of sets $\tilde{\mathcal{H}} = \{\tilde{H}_x : x \in \mathcal{R}^d\}$ has VC-dimension $D = d + 1$ (Pollard [45, page 20], Haussler [26]). Thus we have the following corollary.

Corollary 2.1 *Let $\xi = (a, b)$ and let $g_x(\xi) = \phi(a \cdot x + b) = 1_{\{a \cdot x + b \geq 0\}}$. If $h(x) = \int \phi(a \cdot x + b) P(da, db)$ for $x \in S$ for some probability measure P on a, b, then there exists $\xi_1, \xi_2, \cdots, \xi_T$ such that*

$$\sup_{x \in S} \left| \frac{1}{T} \sum_{i=1}^{T} g_x(\xi_i) - h(x) \right| \leq 34 \sqrt{\frac{d+1}{T}} \tag{2.33}$$

Recall from section 2.2.2 that for the approximation of the Gaussian function, the $2T$-term neural network approximation f_{2T} can be split up into two parts, $f_{1,T}(x) = \frac{\nu_1}{T} \sum_{i=1}^{T} g_x(\xi_i)$ and $f_{2,T}(x) = \frac{\nu_2}{T} \sum_{i=T+1}^{2T} g_x(\xi_i)$, which approximate the positive and negative parts f_1 and f_2 respectively. Using Corollary 2.1 , we see that

$$\sup_{x \in B_K} \left| \frac{\nu_1}{T} \sum_{i=1}^{T} g_x(\xi_i) - f_1(x) \right| \leq 34 \nu_1 \sqrt{\frac{d+1}{T}} \tag{2.34}$$

and similarly,

$$\sup_{x \in B_K} \left| \frac{\nu_2}{T} \sum_{i=T+1}^{2T} g_x(\xi_i) - f_2(x) \right| \leq 34 \nu_2 \sqrt{\frac{d+1}{T}}. \tag{2.35}$$

Hence by the triangle inequality,

$$\begin{aligned}\sup_{x\in B_K}|f_{2T}(x)-f(x)| &\leq 34(\nu_1+\nu_2)\sqrt{\frac{d+1}{T}}\\ &= 34\nu\sqrt{\frac{d+1}{T}}\\ &\leq 68K\sqrt{\frac{d(d+1)}{T}}. \qquad (2.36)\end{aligned}$$

An upper bound on K will be determined later.

2.2.4 Bounding the Hausdorff distance of the approximation

The Hausdorff metric between two sets F and G is defined as

$$\delta^H(F,G)=\max\{\sup_{x\in F}\inf_{y\in G}|x-y|,\sup_{y\in G}\inf_{x\in F}|x-y|\}.$$

The norm $|\cdot|$ is the usual Euclidean norm in $\mathcal{R}^d$. We bound the Hausdorff distance between the ball and its approximating set $\delta^H(B,\mathcal{N}_{2T})$ in this section. The ball is assumed to be centered at the origin. However we apply the result later to other balls and ellipsoids that are not necessarily centered at the origin. Note that the unit ball B in $\mathcal{R}^d$ may be represented as

$$B=\{x:\exp(-\frac{|x|^2}{2})\geq\exp(-\frac{1}{2})\}.$$

We define $\mathcal{N}_{2T}$ as

$$\mathcal{N}_{2T}=\{x:f_{2T}(x)\geq\exp(-\frac{1}{2})+K\epsilon_T\}.$$

Let $f(x) = \exp(-\frac{|x|^2}{2})$ and $f_{2T}(x)$ be the approximation with T pairs of indicators. Here

$$\epsilon_T := 68\sqrt{\frac{d(d+1)}{T}},$$

for which we have the $\mathcal{L}_\infty$ bound between the Gaussian and its approximation bounded above by

$$\sup_{x \in B_K} |f_{2T}(x) - f(x)| \leq K\epsilon_T. \tag{2.37}$$

We are going to bound the Hausdorff distance between B and $\mathcal{N}_{2T}$, using this sup norm bound on the error between the functions f and f_{2T} which yield B and $\mathcal{N}_{2T}$ as level sets.

Theorem 2.1 *Let B_R be a ball of radius R in $\mathcal{R}^d$ centered at the origin, and let $\mathcal{N}_{2T}$ be the level set of the neural net approximation. For sufficiently large T, such that $\epsilon_T \leq \frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$,*

$$\delta^H(B_R, \mathcal{N}_{2T}) \leq 318R\sqrt{\frac{d(d+1)}{T}}.$$

Proof : The ball B coincides with the level set of f at the level $\exp(-\frac{1}{2})$. Let T be such that ϵ_T is less than $\frac{1}{2K}\exp(-\frac{1}{2})$. Choose r_0 such that $\exp(-\frac{r_0^2}{2}) = \exp(-\frac{1}{2}) + 2K\epsilon_T$. Let B_{r_0} be the ball of radius r_0 centered around the origin. If $x \in \mathcal{N}_{2T}$, then $\exp(-\frac{1}{2}) \leq f_{2T}(x) - K\epsilon_T \leq \exp(-\frac{|x|^2}{2})$ which implies that $x \in B$. Similarly if $x \in B_{r_0}$, then $\exp(-\frac{1}{2}) + K\epsilon_T \leq \exp(-\frac{|x|^2}{2}) - K\epsilon_T \leq f_{2T}(x)$, which implies that $x \in \mathcal{N}_{2T}$. Thus

$$B_{r_0} \subset \mathcal{N}_{2T} \subset B.$$

Both B and its approximating set $\mathcal{N}_{2T}$ are sandwiched between B_{r_0} and B. Consequently

$$\delta^H(B, \mathcal{N}_{2T}) \leq 1 - r_0.$$

The function $g(r) = \exp(-\frac{r^2}{2})$ has derivative $-rg(r)$ of magnitude less than the derivative at $r = 1$. Now

$$\begin{aligned} r_0 &= \sqrt{2\log\left(1/(e^{-\frac{1}{2}} + 2K\epsilon_T)\right)} \\ &= \sqrt{1 - 2\log(1 + 2K\epsilon_T e^{\frac{1}{2}})}, \end{aligned}$$

which is close to 1. By taking the Taylor expansion of r_0 in ϵ_T around 0, we see that

$$r_0 = 1 - 2K\epsilon_T e^{\frac{1}{2}} - O(\epsilon_T)^3$$

and that

$$\delta^H(B, \mathcal{N}_{2T}) \leq 2K\epsilon_T e^{\frac{1}{2}} - O(\epsilon_T)^3.$$

We now give a bound on the Hausdorff distance without the $O(\epsilon_T)^3$ term. If T is large enough that ϵ_T is less than $\frac{1}{2K}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$, then $r_0 g(r_0) \geq \frac{1}{\sqrt{2}}\exp(-\frac{1}{2})$, and hence using the mean-value theorem

$$\begin{aligned} \delta^H(B, \mathcal{N}_{2T}) &\leq 1 - r_0 \\ &\leq \frac{g(r_0) - g(1)}{r_0 g(r_0)} \\ &\leq 2\sqrt{2e}K\epsilon_T \\ &\leq 180\sqrt{2e}K\sqrt{\frac{d(d+1)}{T}}. \end{aligned} \tag{2.38}$$

Now we determine an upper bound to K. From section 2.2.3, B_K need only be large enough to cover both B and its approximation set $\mathcal{N}_{2T}$, thus we can take B_K to be B, whence $K = 1$. Again when $\epsilon_T \leq \frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$, we have

$$\delta^H(B, \mathcal{N}_{2T}) \leq 136\sqrt{2e}\sqrt{\frac{d(d+1)}{T}} \leq 318\sqrt{\frac{d(d+1)}{T}}. \tag{2.39}$$

For a ball B_R of radius R, the Hausdorff distance between it and its approximation set is simply bounded by $318R\sqrt{\frac{d(d+1)}{T}}$. This concludes the proof of theorem 2.1 .

□

2.2.5 An $\mathcal{L}_1$ Bound

Let B_R be a ball of radius R, $\mathcal{N}_{2T}$ the level set induced by the approximation as explained in section 2.2.4, μ is the Lebesgue measure, and δ is the Hausdorff distance between B_R and $\mathcal{N}_{2T}$ as obtained above.

Theorem 2.2 *The relative Lebesgue measure of the symmetric difference* $\frac{\mu(B_R \triangle \mathcal{N}_{2T})}{\mu(B_R)}$ *between* B_R *and its approximation set* $\mathcal{N}_{2T}$ *is bounded above by*

$$\frac{\mu(B_R \triangle \mathcal{N}_{2T})}{\mu(B_R)} \leq 318d\sqrt{\frac{d(d+1)}{T}}.$$

Proof : Since the symmetric difference $B_R \triangle \mathcal{N}_{2T}$ is included in the shell $B_R \backslash B_{R-\delta}$, we obtain

$$\int |1_{B_R} - 1_{\mathcal{N}_{2T}}| \frac{\mu(dx)}{\mu(B_R)} = \frac{\mu(B_R \triangle \mathcal{N}_{2T})}{\mu(B_R)}$$

$$
\begin{aligned}
&\leq \frac{\mu(B_R) - \mu(B_{R-\delta})}{\mu(B_R))} \\
&= 1 - (1 - \frac{\delta}{R})^d \\
&\leq d\frac{\delta}{R} \\
&\leq 318d\sqrt{\frac{d(d+1)}{T}}. \qquad (2.40)
\end{aligned}
$$

□

2.2.6 Ellipsoid approximation

Consider an ellipsoid $E = \{x : x'Mx \leq 1\}$ centered at the origin with $M = A'A$ strictly positive definite with a $d \times d$ positive definite square root A. Equivalently $E = \{x : \exp(-x'A'Ax/2) \geq \exp(-1/2)\}$ is the level set of a Gaussian surface. In a similar manner to the ball, it can also be accurately and parsimoniously approximated by a threshold of a single hidden layer neural net. Let the eigenvalues of A be $r_1 \leq r_2 \leq \cdots \leq r_d$ with the corresponding eigenvectors $\{\mathbf{r_1}, \mathbf{r_2}, \cdots \mathbf{r_d}\}$. If the approximating set for the unit ball takes the form $\{x : \sum_{i=1}^{2T} c_i 1_{\{a_i \cdot x \geq b_i\}} \geq k\}$, then the one for the ellipsoid E is

$$E_{2T} = \{x : \sum_{i=1}^{2T} c_i 1_{\{a_i \cdot Ax \geq b_i\}} \geq k\}$$

We are interested in bounding $\delta^H(E, E_{2T})$, the Hausdorff distance between the ellipsoid and its approximating set.

Theorem 2.3 *The Hausdorff distance between the ellipsoid E and its approximating set E_{2T} is bounded above by*

$$318 r_d \sqrt{\frac{d(d+1)}{T}}. \tag{2.41}$$

Proof : The matrix transformation A transforms the unit ball to an ellipsoid E by stretching the unit radius to length r_i in the $\mathbf{r}_i$ direction and the approximating set $\mathcal{N}_{2T}$ is similarly stretched in the same way to E_{2T}. For the ball B_{r_0} (as defined in the proof of Theorem 2.1), the matrix transformation A transforms it to an ellipsoid E' by stretching its radius to length $r_i r_0$ in the $\mathbf{r}_i$ direction. Thus the order of inclusivity is still preserved after the transformation and

$$E' \subset E_{2T} \subset E.$$

Note that the ellipsoids E and E' are similar, centered at the origin and aligned along the same axes. The only difference is in the scale.

The two extreme parts of the ellipsoids E and E' are along the directions $\mathbf{r}_1$ and $\mathbf{r}_d$. Thus $\delta^H(E, E_{2T})$ is bounded by the greatest distance between E and E', and this occurs along the direction of $\mathbf{r}_d$, and hence is bounded above by the Hausdorff distance between that of a ball of radius r_d (containing the ellipsoid) and a ball of radius $r_d r_0$, and that is in turn bounded above by

$$318 r_d \sqrt{\frac{d(d+1)}{T}}.$$

□

The error is the same as for approximation of a ball except that the radius of the ball

is replaced by the maximal eigenvalue (length of major axis).

Now consider an ellipsoid E with axial lengths $r_1 \leq \cdots \leq r_{d-1} \leq r_d = R$ and its approximating set E_{2T}. The ellipsoid $E^\delta = (1 - \frac{\delta}{R})E$ is a scaled down version of E and it has axial lengths $r_1(1-\frac{\delta}{R}) \leq \cdots \leq r_{d-1}(1-\frac{\delta}{R}) \leq r_d(1-\frac{\delta}{R}) = R - \delta$. Recall that the approximation set E_{2T} is obtained by scaling $\mathcal{N}_{2T}$ (the approximation set for the unit ball) by a factor of r_i along the i-th axis of the ellipsoid E. The Hausdorff distance between E and E_{2T} is δ which is bounded by $318R\sqrt{\frac{d(d+1)}{T}}$ from Theorem 2.3.

Corollary 2.2 *The measure of the symmetric difference $\mu(E \triangle E_{2T})$ between E and its approximation set E_{2T} is bounded above by*

$$\mu(E \triangle E_{2T}) \leq 318\mu(E)d\sqrt{\frac{d(d+1)}{T}}.$$

Proof : Since the difference $E \triangle E_{2T}$ is included in the shell $E \backslash E^\delta$, we obtain

$$\begin{aligned}
\int |1_E - 1_{E_{2T}}|\mu(dx) &= \mu(E) - \mu(E_{2T}) \\
&\leq \mu(E) - \mu(E^\delta) \\
&= \mu(E) - \left(1 - \frac{\delta}{R}\right)^d \mu(E) \\
&= \mu(E\left(1 - \left(1 - \frac{\delta}{R}\right)^d\right) \\
&\leq \mu(E)d\frac{\delta}{R}
\end{aligned}$$

$$\leq \quad 318\mu(E)d\sqrt{\frac{d(d+1)}{T}}, \tag{2.42}$$

□

2.2.7 Remarks

The integral representation to the Gaussian on B_K may also be written

$$\begin{aligned}\exp(-\frac{|x|^2}{2}) \quad = \quad & \int_{\mathcal{R}^d}\int_{-|a|K}^{|a|K} 1_{\{\mathrm{sgn}(b)a\cdot x+b\mathrm{sgn}(b)\geq 0\}}|\sin(b)|\frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}}dbda \\ & -\frac{1}{2}\int_{\mathcal{R}^d}\int_{-|a|K}^{|a|K}\sin^{-}(b)\frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{\frac{d}{2}}}dbda+\exp(-\frac{K^2}{2}).\end{aligned} \tag{2.43}$$

Sampling from the distribution V proportional to $|\sin(b)|\exp(-\frac{|a|^2}{2})$, the approximation to the ball takes the form

$$\mathcal{N}_T = \{x \in \mathcal{R}^d : \sum_{i=1}^{T} 1_{\{a_i\cdot x\geq b_i\}} \geq k\},$$

that is, x is in $\mathcal{N}_T$ if it is in at least k of the half-spaces. This approximation achieves

$$\delta^H(B,\mathcal{N}_T) \leq 318\sqrt{\frac{d(d+1)}{T}} \tag{2.44}$$

In particular, when $2T$ sigmoids are used in the approximation,

$$\delta^H(B,\mathcal{N}_{2T}) \leq 318\sqrt{\frac{d(d+1)}{2T}}$$

when the representation (2.43) is used, reducing the constant by a factor of $\frac{1}{\sqrt{2}}$ from the bound in Theorem 2.1.

It may be possible to extend our results to neural network approximation of other classes of closed convex sets with smooth boundaries, for example, to classes of sets of the form $\mathcal{D} = \{x \in \mathcal{R}^d : f(x) \cdot f(x) \leq 1\}$, where $f : \mathcal{R}^d \to \mathcal{R}^d$ has a strictly positive definite derivative. If this is achieved, the results pertaining to functions which have total finite variation with respect to a class of ellipsoids (in the following section) could be extended to those for a class of convex sets with some suitable smoothness properties.

2.3 Approximation Bounds for Two Layer Nets

2.3.1 Approximation with Heaviside Sigmoids

The second (outer) layer of a two layer net takes a linear combination of level sets H of functions represented by linear combinations on the first (inner) layer. The class of sets represented by level sets of combinations of first layer nodes include half-spaces and rectangles, and (as we have seen) approximations to ellipsoids. In this section we provide $\mathcal{L}_2$ approximation bounds for two layer networks for certain classes of functions. Our tools will be the $\mathcal{L}_\infty$ approximation results for level sets from the first layer combinations together with the idea of finite variation with respect to a class of sets.

A function f is said to have variation $V_{f,\mathcal{H}}$ with respect to a class of sets $\mathcal{H}$ if

$V_{f,\mathcal{H}}$ is the infimum of numbers V such that f/V is in the closure of the convex hull of signed indicators of sets in $\mathcal{H}$, where the closure is taken in $\mathcal{L}_2(P_X)$. A special case of finite variation is the case we call total variation with respect to a class of sets. Suppose $f(x)$ defined over a bounded region S in $\mathcal{R}^d$. We say that f has total variation V with respect to a class of sets $\mathcal{H} = \{H_\xi : \xi \in \Xi\}$ if there exist some signed measure v over the measurable space Ξ and

$$f(x) = \int_\Xi 1_{H_\xi}(x) v(d\xi) \text{ for } x \in \mathcal{S}, \tag{2.45}$$

and if v has finite total variation V. In the event that the representation (2.45) is not unique, we take the measure v that yields the smallest total variation V.

The function class $\mathcal{F}_{V,\mathcal{H}}$ of functions with variation $V_{f,\mathcal{H}}$ bounded by V arises naturally when thinking of the functions obtained by linear combinations on a layer of a network where the sum of absolute values of the coefficients of linear combination are bounded by V and the level sets from the preceding layer yield the sets in $\mathcal{H}$.

Lemma 2.4 *If f has variation V_f with respect to a class of sets $\mathcal{H}$ then for each T there exists $H_1, \cdots, H_T$ and $c_1, \cdots, c_T$ with $\sum_{i=1}^T |c_i| \leq V_f$ such that the approximation $f_T(x) = \sum_{i=1}^T c_i 1_{H_i}(x)$ achieves*

$$\|f - f_T\|_2 \leq \frac{V_f}{\sqrt{T}}. \tag{2.46}$$

A proof of this lemma and its use in approximation theory are in Barron [3], though the roots of the inequality in probability are classical. Pisier [44] attributed the result

in its classical form to B. Maurey.

Proof : The proof is based on the Monte Carlo sampling idea as in section 2.1. First fix T and suppose that f is not identically constant. (Equality occurs in (2.46) only if f is identically constant.) Since f is in the closure of the convex hull of $G = \{\pm V_f 1_H : H \in \mathcal{H}\}$, one takes a $\tilde{f}$ that is a (potentially very large) finite convex combination with $\|f - \tilde{f}\|_2 < \delta$. In particular we take $\delta = \frac{\epsilon}{\sqrt{T}}$ and ϵ small, say $\epsilon < V_f - \sqrt{V_f^2 - \frac{\|f\|^2}{4}}$, which is less than $\frac{\|f\|}{2}$.

By the triangle inequality,

$$\begin{aligned} \|f - f_T\|_2 &\leq \|f - \tilde{f}\|_2 + \|\tilde{f} - f_T\|_2 \\ &\leq \frac{\epsilon}{\sqrt{T}} + \|\tilde{f} - f_T\|_2. \end{aligned} \tag{2.47}$$

Suppose $\tilde{f} = \sum_i p_i g_i$ with g_i in G, and $p_i > 0$ with $\sum_i p_i = 1$. Apply the Monte Carlo sampling technique as in section 2.1. Draw indices $i_1, \cdots, i_T$ independently according to the distribution p_i in the representation of $\tilde{f}$ and let $f_T = \frac{1}{T}\sum_{j=1}^{T} g_{i_j}$. Then as in Lemma 2.1,

$$\begin{aligned} E_i\|\tilde{f} - f_T\|_2^2 &= \frac{E_i||g_i||^2 - ||\tilde{f}||^2}{T} \\ &\leq \frac{V_f^2 - ||\tilde{f}||^2}{T} \\ &\leq \frac{V_f^2 - \frac{\|f\|^2}{4}}{T}. \end{aligned} \tag{2.48}$$

and so there exists a choice of such an f_T with

$$\|\tilde{f} - f_T\|_2^2 < \frac{V_f^2 - \frac{\|f\|^2}{4}}{T}.$$

That is

$$\|\tilde{f} - f_T\|_2 < \frac{\sqrt{V_f^2 - \frac{\|f\|^2}{4}}}{\sqrt{T}}. \tag{2.49}$$

Substituting this bound back into (2.47) and letting ϵ go to zero completes the proof.

□

As a consequence of the lemma above, we have the following corollary involving approximation with a class of ellipsoids. Let ξ be the parameters that define the ellipsoids, and $1_{E_\xi}(x)$ the indicator of the ellipsoid.

Corollary 2.3 *If f has variation $V_f = V_{f,\mathcal{E}}$ with respect to the class $\mathcal{E}$ of ellipsoids then there is a choice of ellipsoids $E_1, \cdots, E_T$ and $s_1, \cdots, s_{T_1} \in \{-1, +1\}$, and $c_i = \frac{V_f s_i}{T_1}$ such that*

$$f_{T_1}(x) = \sum_{i=1}^{T_1} c_i 1_{E_i} \tag{2.50}$$

satisfies

$$\|f_{T_1} - f\|_2 \leq \frac{V_f}{\sqrt{T_1}} \tag{2.51}$$

The indicators of ellipsoids have two layer sigmoidal network approximations consisting of a single outer node and a single hidden inner layer. These approximations to 1_{E_i} may be substituted into the approximation in (2.50) to yield a two hidden layer approximation to f.

Let $\mathcal{E} = \{E_\xi : \xi \in \Xi\}$ be the set of ellipsoids with $\mu(E_\xi) \leq \mu(\mathcal{S})$ where μ is the Lebesgue measure. Let P_X be the uniform probability measure over $\mathcal{S}$, and let E_{2T_2}

be the neural net level set with $2T_2$ sigmoids that is used to approximate E. Using the bound in Corollary 2.2, for each $E \in \mathcal{E}$,

$$\begin{aligned}\int_S |1_E(x) - 1_{E_{2T_2}}(x)|^2 P_X(dx) &= \frac{\mu((E - E_{2T_2}) \cap S)}{\mu(S)} \\ &\leq \frac{\mu((E - E_{2T_2}) \cap S)}{\mu(S)} \\ &\leq \frac{\mu(E)}{\mu(S)} 318d\sqrt{\frac{d(d+1)}{2}} \\ &\leq 318d\sqrt{\frac{d(d+1)}{T_2}}. \end{aligned} \tag{2.52}$$

After replacing the indicators of the ellipsoids in (2.50) with their neural net approximations, we obtain

$$f_{T_1,2T_2} = \sum_{i=1}^{T_1} c_i \phi(\sum_{j=1}^{T_2} \omega_{ij} \phi(a_{ij} \cdot x - b_{ij}) - d_i). \tag{2.53}$$

The following theorem bounds the mean-squared approximation error. An ellipsoid in $\mathcal{E}$ is denoted by E.

Theorem 2.4 *If f has variation V_f with respect to the class of ellipsoids $\mathcal{E}$, with $\mu(E) \leq \mu(S)$ and P_X is the uniform probability measure over S, then there exist a choice of parameters $(a_{ij}, b_{ij}, c_i, d_i, \omega_{ij})$ such that a two hidden layer net with step activation function achieves approximation error bounded by*

$$\|f - f_{T_1,2T_2}\|_2 \leq \frac{V_f}{\sqrt{T_1}} + V_f \left(318d\sqrt{\frac{d(d+1)}{T_2}}\right)^{\frac{1}{2}}, \tag{2.54}$$

and

$$\|f - f_{T_1,2T_2}\|_1 \leq \frac{V_f}{\sqrt{T_1}} + V_f 318d\sqrt{\frac{d(d+1)}{T_2}},$$

where $||\cdot||_p$ *denotes the* $\mathcal{L}_p(P_X)$ *norm; provided that* T_2 *is large enough that* $68\sqrt{\frac{d(d+1)}{T_2}} \leq \frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$.

Proof : By the triangle inequality,

$$\|f - f_{T_1,2T_2}\|_2 \leq \|f - f_{T_1}\|_2 + \|f_{T_1} - f_{T_1,2T_2}\|_2. \tag{2.55}$$

Now

$$\|f - f_{T_1}\|_2 \leq \frac{V_f}{\sqrt{T_1}}$$

from Corollary 2.3. The other term on the right hand side of (2.55) is bounded as follows. Let $\tilde{E}_i$ be the neural net level set of the approximation to E_i from section 2.2.6. Then

$$\begin{aligned} \|f_{T_1} - f_{T_1,2T_2}\|_2 &= \left\| \sum_{i=1}^{T_1} c_i (1_{E_i} - 1_{\tilde{E}_i}) \right\|_2 \\ &\leq \frac{1}{T_1} \sum_{i=1}^{T_1} |c_i| \|1_{E_i} - 1_{\tilde{E}_i}\|_2 \\ &\leq V_f \left(318d\sqrt{\frac{d(d+1)}{T_2}} \right)^{\frac{1}{2}}, \end{aligned} \tag{2.56}$$

where (2.52) bounds the last inequality (2.56).

□

The proof of the $\mathcal{L}_1$ bound is similar (using $\|f - f_{T_1}\|_1 \leq \|f - f_{T_1}\|_2 \leq \frac{V_f}{\sqrt{T_1}}$) except that the square root is (2.56) is not used in bounding $\|1_{E_i} - 1_{\tilde{E}_i}\|_1$.

Example 2.1 Convex Combination of Balls.

Let $B(a,b)$ denote a ball centered at a with radius b. In $\mathcal{R}^3$, the function

$$\begin{aligned} f(x) &= \frac{4\pi}{3} - \sqrt{2}\pi(x_1^2+x_2^2+x_3^2)^{1/2} + \frac{\pi}{3\sqrt{2}}(x_1^2+x_2^2+x_3^2)^{3/2} \\ &= \int 1_{B(\theta,1)}(x)1_{B(0,1)}(\theta)d\theta \end{aligned} \tag{2.57}$$

is a convex combination of indicators of balls. Thus

$$f_{T_1}(x) = \frac{4\pi}{3T_1}\sum_{i=1}^{T_1} 1_{B(\theta_i,1)}(x) \tag{2.58}$$

is an approximation to $f(x)$ where the θ_i's are sample from the uniform distribution in a unit ball. We then approximate each ball $1_{B(\theta_i,1)}(x)$ with the form (2.14).

Example 2.2 A Radial Function.

Let $\mu \geq 2$,

$$f(x) = \begin{cases} \frac{1}{2}(|x| - \mu + 2), & \mu - 2 < |x| \leq \mu \\ \frac{1}{2}(\mu + 2 - |x|), & \mu < |x| \leq \mu + 2 \\ 0, & \text{otherwise} \end{cases} \tag{2.59}$$

then

$$f(x) = \int_{\mathcal{R}} 1_{(\theta-1,\theta+1)}(|x|)\frac{1}{2}1_{[-1,1]}(\theta-\mu)d\theta \tag{2.60}$$

and thus $f(x)$ can be approximated by

$$f_{T_1}(x) = \frac{1}{T_1}\sum_{i=1}^{T_1}\{1_{B(0,\theta_i+1)}(x) - 1_{B(0,\theta_i-1)}(x)\} \tag{2.61}$$

where $\theta_i \sim$ iid Uniform$(\mu-1, \mu+1)$.

2.3.2 Approximation with Ramp Sigmoids

A ramp sigmoid ϕ_ν with Lipshitz constant ν takes the form

$$\phi_\nu(z) = \begin{cases} 0 & \text{when } z < 0, \\ \nu z & \text{when } 0 \leq z \leq \frac{1}{\nu}, \\ 1 & \text{when } z > \frac{1}{\nu}. \end{cases} \tag{2.62}$$

In this subsection, we derive the analog of Theorem 2.4 using ramp sigmoids with Lipshitz constant ν_1 in the outer layer and ν_2 in the inner layer.

We first derive an analogous result to Theorem 2.1 for the unit ball. This will be extended to ellipsoids and finally to an analog of Theorem 2.4. Let $\mathcal{E} = \{E_\xi : \xi \in \Xi\}$ be a class of ellipsoids such that the Lebesgue volume satisfies $\mu(E) \leq \mu(S)$, where S is a given bounded domain. A two hidden layer neural net with ramp sigmoidal activation functions take the form

$$f_{T_1,T_2,\nu_1,\nu_2}(x) = \sum_{i=1}^{T_1} c_i \phi_{\nu_1}\Big(\sum_{j=1}^{T_2} \omega_{ij} \phi_{\nu_2}(a_{ij} \cdot x - b_{ij}) - d_i\Big). \tag{2.63}$$

The following theorem is the analog to Theorem 2.4, which bounds the approximation error of the above approximation, using ramp rather than step sigmoids.

Theorem 2.5 *If f has finite variation V_f with respect to the class of ellipsoids $\mathcal{E}$ where $\mu(E) \leq \mu(S)$, and P_X is the uniform probability measure over S, then there exist a choice of parameters $(a_{ij}, b_{ij}, c_i, d_i, \omega_{ij})$ such that the two hidden layer neural*

net f_{T_1,T_2,ν_1,ν_2} with ramp activation function achieves approximation error bounded by

$$\|f - f_{2T_1,T_2,\nu_1,\nu_2}\|_2 \leq \frac{2V_f}{\sqrt{T_1}} + 2V_f \left(328d\sqrt{\frac{d(d+1)}{T_2}}\right)^{\frac{1}{2}}, \tag{2.64}$$

provided that $\nu_1 \geq \max(4d\sqrt{2e}T_1, \sqrt{\frac{T_2}{d(d+1)}})$, and $\nu_2 \geq 2\sqrt{d}$, and T_2 large enough such that $70\sqrt{\frac{d(d+1)}{T_2}} \leq \frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$.

Proof : We modify the proofs from section 2.2 to show that the indicator of an ellipsoid can be approximated by the ramp ϕ_{ν_1} applied to a single layer approximation f_{T_2,ν_2} to a Gaussian using the ramp sigmoid ϕ_{ν_2} in the inner layer.

First we work with the case of the unit ball B. The Gaussian can be written as an integral of sinusoids (Fourier transform). Let $h(x) = \exp(-\frac{|x|^2}{2})$ and its T_2 term approximation with ramp sigmoids be $f_{T_2,\nu_2}(x) = \sum_{j=1}^{T_2} c_j \phi_{\nu_2}(a_j \cdot x + b_j) + d$. Consider $x \in B$. We want to bound

$$\sup_{x \in B} |h(x) - h_{T_2,\nu_2}(x)|.$$

From Barron [3], there is an integral representation of the Gaussian in terms of a family of cosines

$$h(x) - h(0) = \int \frac{C_h}{|a|}\left(\cos(|a|\frac{a \cdot x}{|a|}) - 1\right) p(a)da, \tag{2.65}$$

where $p(a) = \frac{|a|}{C_h}\frac{\exp(-\frac{|a|^2}{2})}{(2\pi)^{d/2}}$, where the normalizing constant for $p(a)$ is $C_h \leq \sqrt{d}$, the expectation of $|a|$ with respect to a standard multivariate normal on $\mathcal{R}^d$, Note that

$-1 \leq z = \frac{a \cdot x}{|a|} \leq 1$. Thus $h(x) - h(0)$ is a convex combination of functions in

$$\mathcal{G}_{\cos} = \left\{ \frac{\gamma}{|a|}(\cos(|a|z) - 1), |z| \leq 1 : |\gamma| \leq C_h, a \in \mathcal{R}^d \right\},$$

evaulated at linear combinations $z = \frac{a \cdot x}{|a|}$. Now consider the set of functions

$$\mathcal{G}_{\phi_{\nu_2}} = \left\{ \frac{\gamma}{|a|}\phi_{\nu_2}(z + b), |z| \leq 1 : |b| \leq 1, |\gamma| \leq 2C_h \right\}.$$

Note that functions in both $\mathcal{G}_{\cos}$ and $\mathcal{G}_{\phi_{\nu_2}}$ (when $\nu_2 \geq 2C_h$) have derivatives less than 1.

Now take any function $g_{|a|}$ from $\mathcal{G}_{\cos}$ and consider its increasing part and decreasing part separately, say

$$g_{|a|}(z) = g_{|a|,+}(z) - g_{|a|,-}(z).$$

The increasing part (and similarly, decreasing part) can be approximated by a linear combination of unit step functions,

$$g_{|a|,+,k}(z) = \sum_{i=1}^{k-1} [g(t_i) - g(t_{i-1})] 1_{\{z \geq t_i\}},$$

where $-1 = t_0 \leq t_1, \cdots, \leq t_{k-1} = 1$ form a partition. The position of the steps are chosen such that $g(t_i) - g(t_{i-1})$ partition the range space equally and that $g_{|a|,+}(t_i) = \frac{1}{2}[g(t_i) + g(t_{i-1})]$. That is, each jump is of equal height and the function $g_{|a|,+}(z)$ at the jump-point passes through exactly in the middle of the jump. Since the derivative of $g_{|a|,+}$ is bounded by C_h, it follows that the sum of absolute value of jump heights $\sum_{i=1}^{k-1} |g(t_i) - g(t_{i-1})|$ is bounded by C_h and adding up coefficients for the steps for

the decreasing part yields that the sum of absulute values of jump heights (for both parts combined) is no greater than $2C_h$. Now if we replace the above procedure with ϕ_{ν_2} instead of steps, and as long as ϕ_{ν_2} has a derivative no less than C_h, the error of such an approximation of $g_{|a|}$ with ϕ_{ν_2} is no greater than that of $g_{|a|}$ with steps. Thus $\mathcal{G}_{\cos} \subset \overline{\text{conv}}\mathcal{G}_{\phi_{\nu_2}}$ for $\nu_2 \geq 2C_h$. Here closure is achieved in $\mathcal{L}_\infty(P_X)$.

Let $g_{|a|}(z) = \frac{C_h}{|a|}(\cos(|a|z) - 1)$ be an element of $\mathcal{G}_{\cos}$. For each $g_{|a|}$ there exists an approximation

$$g_{|a|,\nu_2}(z) = \sum_{i=1}^{n_{|a|}} c_i \phi_{\nu_2}\left(z + b_{i,|a|}\right), \tag{2.66}$$

where $n_{|a|}$ may be very large, and $\sum_{i=1}^{n_{|a|}} |c_i| \leq 2C_h$, (for now it does not matter how many terms there are in $g_{|a|,\nu_2}$). We can choose the coefficients c_i in (2.66) such that the approximation $g_{|a|,\nu_2}$ achieves

$$\sup_{|z|\leq 1} |g_{|a|}(z) - g_{|a|,\nu_2}(z)| < \frac{2C_h}{\nu_2}\sqrt{\frac{d+1}{T_2}}.$$

Substituting (2.66) into (2.65), there is an approximation h_{ν_2} to the $h(x) - h(0)$ such that

$$\begin{aligned} h_{\nu_2}(x) &= E_a g_{|a|,\nu_2}\left(\frac{a \cdot x}{|a|}\right) \\ &= 2C_h E_a E_{i|a}\left[\text{sign}_{i,a}\phi_{\nu_2}\left(\frac{a \cdot x}{|a|} + b_i\right)\right], \end{aligned} \tag{2.67}$$

($\text{sign}_{i,a} \in \{-1, +1\}$) which is an infinite convex combination of elements of $\mathcal{G}_{\phi_{\nu_2}}$. Note that

$$h(x) - h(0) = E_a g_{|a|}\left(\frac{a \cdot x}{|a|}\right). \tag{2.68}$$

Thus using [3, Lemma 5],

$$\begin{aligned}\sup_{x\in B}|h(x)-h(0)-h_{\nu_2}(x)| &\leq E_a \sup_{x\in B}\left|g_{|a|}\left(\frac{a\cdot x}{|a|}\right)-g_{|a|,\nu_2}\left(\frac{a\cdot x}{|a|}\right)\right| \\ &\leq \frac{2C_h}{\nu_2}\sqrt{\frac{d+1}{T_2}}. \end{aligned} \tag{2.69}$$

We choose a T_2 term ramp sigmoidal neural net approximation to $h_{\nu_2}(x)$ by Monte Carlo sampling. From the remark after Lemma 2.2 (but now applied to functions bounded by 1 with pseudo-dimension $d+1$) and the techniques in deriving the sup bound between the Gaussian and its Heaviside sigmoid neural net approximation,

$$\sup_{x\in B}|h_{T_2,\nu_2}(x)-h_{\nu_2}(x)| \leq 2\sqrt{d}34\sqrt{\frac{d+1}{T_2}} \leq 68\sqrt{\frac{d(d+1)}{T_2}}. \tag{2.70}$$

Thus there exists an approximation $\frac{1}{T_2}\sum_{j=1}^{T_2} c_i\phi_{\nu_2}(a_i\cdot x+b_i)$ to the Gaussian such that

$$\begin{aligned}\sup_{x\in B}|h(x)-h(0)-h_{T_2,\nu_2}(x)| &\leq \sup_{x\in B}|h(x)-h(0)-h_{\nu_2}(x)|+\sup_{x\in B}|h_{\nu_2}(x)-h_{T_2,\nu_2}(x)| \\ &\leq \left(\frac{2}{\nu_2}+68\right)\sqrt{\frac{d(d+1)}{T_2}}. \end{aligned} \tag{2.71}$$

Let $f_{T_2,\nu_2}(x)=h_{T_2,\nu 2}(x)+h(0)$. The unit ball B is

$$B=\{x:\exp(-\frac{|x|^2}{2})\geq\exp(-\frac{1}{2})\}$$

as before. We define $\tilde{\mathcal{N}}_{T_2,\nu_1,\nu_2}$ as

$$\tilde{\mathcal{N}}_{T_2,\nu_1,\nu_2}:=\{x:f_{T_2,\nu_2}(x)\geq\exp(-\frac{1}{2})+\epsilon_{T_2}+\frac{1}{\nu_1}\}, \tag{2.72}$$

Set

$$\epsilon_{T_2} := \left(68 + \frac{2}{\nu_2}\right)\sqrt{\frac{d(d+1)}{T_2}},$$

and we see that

$$\tilde{\mathcal{N}}_{T_2,\nu_1,\nu_2} = \{x : \phi_{\nu_1}(f_{T_2,\nu_2}(x) - \exp(-\frac{1}{2}) - \epsilon_{T_2}) = 1\}. \tag{2.73}$$

We will set T_2 and ν_1 large enough that $\epsilon_{T_2} + \frac{1}{\nu_1}$ is less than $\frac{1}{2}\exp(-\frac{1}{2})$.

Choose r_0 such that $\exp(-\frac{r_0^2}{2}) = \exp(-\frac{1}{2}) + 2\epsilon_{T_2} + \frac{2}{\nu_1}$. Let B_{r_0} be the ball of radius r_0 centered around the origin. If $x \in \tilde{\mathcal{N}}_{T_2,\nu_1,\nu_2}$, then $\exp(-\frac{1}{2}) + \frac{1}{\nu_1} \leq f_{T_2,\nu_2}(x) - \epsilon_{T_2} \leq \exp(-\frac{|x|^2}{2})$ which implies that $x \in B$. Similarly if $x \in B_{r_0}$, then $\exp(-\frac{1}{2}) + \epsilon_{T_2} + \frac{1}{\nu_1} \leq \exp(-\frac{|x|^2}{2}) - \epsilon_{T_2} \leq f_{T_2,\nu_2}(x)$, which implies that $x \in \tilde{\mathcal{N}}_{T_2,\nu_1,\nu_2}$. Thus

$$B_{r_0} \subset \tilde{\mathcal{N}}_{T_2,\nu_1,\nu_2} \subset B$$

and consequently

$$\delta^H(B, \tilde{\mathcal{N}}_{T_2,\nu_1,\nu_2}) \leq 1 - r_0.$$

We also note that the set

$$\{x : f_{T_2,\nu_2}(x) \geq \exp(-\frac{1}{2}) + \epsilon_{T_2}\} \subset B.$$

Now

$$\begin{aligned} r_0 &= \sqrt{2\log\left(1/(e^{-\frac{1}{2}} + 2\epsilon_{T_2} + 2\frac{1}{\nu_1})\right)} \\ &= \sqrt{1 - 2\log(1 + 2(\epsilon_{T_2} + \frac{1}{\nu_1})e^{\frac{1}{2}})}, \end{aligned}$$

which is close to 1. Thus as in the derivation of (2.38) in section 2.2.4 if T_2 is large enough that $\epsilon_{T_2} + \frac{1}{\nu_1}$ is less than $\frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$, we have in this case

$$\begin{aligned} \delta^H(B, \tilde{\mathcal{N}}_{T_2,\nu_1,\nu_2}) &\leq 2\sqrt{2e}\left(\epsilon_{T_2} + \frac{1}{\nu_1}\right) \\ &\leq \left(318 + \frac{4\sqrt{2e}}{\nu_2}\right)\sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2e}}{\nu_1}. \end{aligned} \tag{2.74}$$

For an ellipsoid $E = \{x : x'A'Ax \leq 1\}$, with A strictly positive definite, and for which the largest eigenvalue of A is r_d, the ramp sigmoid approximation takes the form

$$\phi_{\nu_1}(f_{T_2,\nu_2}(Ax) - \exp(-\frac{1}{2}) - \epsilon_{T_2})$$

and

$$\tilde{E}_{T_2} = \{x : \phi_{\nu_1}(f_{T_2,\nu_2}(Ax) - \exp(-\frac{1}{2}) - \epsilon_{T_2}) = 1\}.$$

The Hausdorff distance between the ellipsoid and this approximating set is bounded above by

$$\delta^H(E, \tilde{E}_{T_2}) \leq \left(318 + \frac{4\sqrt{2e}}{\nu_2}\right) r_d\sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2e}r_d}{\nu_1}. \tag{2.75}$$

Suppose we approximate the ellipsoid E, which has major axis of length r_d, with a two layer neural net of the form

$$f_{1,T_2,\nu_1,\nu_2} = \phi_{\nu_1}\left(\frac{1}{T_2}\sum_{j=1}^{T_2}\omega_j\phi_{\nu_2}(a_j \cdot x - b_j) - d\right), \tag{2.76}$$

then using Corollary 2.2 (see also (2.52)) and the bound (2.75),

$$\int_S |1_E(x) - f_{1,T_2,\nu_1,\nu_2}(x)|^2 P_X(dx) \leq \int_S |1_E(x) - 1_{\tilde{E}_{T_2}}(x)|^2 P_X(dx) \tag{2.77}$$

$$\leq \left(318 + \frac{4\sqrt{2e}}{\nu_2}\right) d\sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2e}d}{\nu_1} \tag{2.78}$$

In (2.77), we have used the fact that $\tilde{E}_{T_2} = \{f_{1,T_2,\nu_1,\nu_2}(x) = 1\} \subset \{f_{1,T_2,\nu_1,\nu_2}(x) > 0\} \subset B$.

Now we examine what happens when $f_{T_1,\nu_1} = \sum_{i=1}^{T_1} c_i 1_{E_i}(x)$ replaces the indicators of ellipsoids in f_{T_1} with corresponding ramp functions of quadratic forms. We have via the triangle inequality

$$\|f - f_{T_1,\nu_1}\|_2 \leq \|f - f_{T_1}\|_2 + \|f_{T_1,\nu_1} - f_{T_1}\|_2. \tag{2.79}$$

We consider again the unit ball case, when the outer layer Heaviside sigmoid ϕ is replaced by ϕ_{ν_1}. An upper bound to

$$\left\|\phi\left(\exp\left(-\frac{|x|^2}{2}\right) - \exp\left(-\frac{1}{2}\right)\right) - \phi_{\nu_1}\left(\exp\left(-\frac{|x|^2}{2}\right) - \exp\left(-\frac{1}{2}\right)\right)\right\|_2$$

is

$$\left\|\phi\left(\exp\left(-\frac{|x|^2}{2}\right) - \exp\left(-\frac{1}{2}\right)\right) - \phi\left(\exp\left(-\frac{|x|^2}{2}\right) - \exp\left(-\frac{1}{2}\right) - \frac{1}{\nu_1}\right)\right\|_2.$$

Thus we seek first a bound on the Hausdorff distance between a unit ball and some smaller ball B_{r_1} of an appropriate radius r_1. The extension to ellipsoids follow from the techniques that we have use before in section 2.2.6. By solving for $r_1 = \sqrt{1 - 2\log(1 + \frac{e^{-1/2}}{\nu_1})}$, we see that $1 - r_1$ (that is, the Hausdorff distance) is bounded by $\frac{\sqrt{2e}}{\nu_1}$. For an ellipsoid with major axial length R, the Hausdorff distance

between such an ellipsoid and a smaller appropriate ellipsoid is $R\frac{\sqrt{2e}}{\nu_1}$. Using Corollary 2.2 (see also (2.52) for comparision), we see that

$$\left\| \phi\left(\exp\left(-\frac{x'A'Ax}{2}\right) - \exp\left(-\frac{1}{2}\right)\right) - \phi_{\nu_1}\left(\exp\left(-\frac{x'A'Ax}{2}\right) - \exp\left(-\frac{1}{2}\right)\right)\right\|_2$$
$$\leq d\frac{\sqrt{2e}}{\nu_1}. \tag{2.80}$$

Thus

$$\|f_{T_1} - f_{T_1,\nu_1}\|_2 \leq \sum_{i=1}^{T_1} |c_i| \|1_{E_i} - g_{\nu_1,E_i}\|_2 \tag{2.81}$$
$$\leq \frac{2V_f(d\sqrt{2e})^{1/2}}{\sqrt{\nu_1}}, \tag{2.82}$$

where in (2.81), g_{ν_1,E_i} is the sigmoid ϕ_{ν_1} applied to the Gaussian associated with ellipsoid E_i.

Finally, by adding up all the terms together,

$$\|f - f_{T_1,T_2,\nu_1,\nu_2}\|_2 \leq \frac{V_f}{\sqrt{T_1}} + \frac{2V_f(d\sqrt{2e})^{1/2}}{\sqrt{\nu_1}}$$
$$+2V_f\left(\left(318 + \frac{4\sqrt{2e}}{\nu_2}\right) d\sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2e}d}{\nu_2}\right)^{\frac{1}{2}}. \tag{2.83}$$

[Note that if we did let ν_1 and ν_2 go to infinity, we would obtain the bound for the step activation function case in Theorem 2.4.]

Now choose $\nu_1 \geq \max(4d\sqrt{2e}T_1, \sqrt{\frac{T_2}{d(d+1)}})$ and $\nu_2 \geq 2\sqrt{d}$. Then the bound from (2.83) yields

$$\|f - f_{T_1,T_2,\nu_1,\nu_2}\|_2 \leq \frac{2V_f}{\sqrt{T_1}} + 2V_f\left(328d\sqrt{\frac{d(d+1)}{T_2}}\right)^{\frac{1}{2}}. \tag{2.84}$$

Finally choosing T_2 to satisfy $70\sqrt{\frac{d(d+1)}{T_2}} \leq \frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$ ensures the requirement on $68\sqrt{\frac{d(d+1)}{T_2}} + \frac{2}{\nu_2}\sqrt{\frac{d(d+1)}{T_2}} + \frac{1}{\nu_1}$ given above inequality (2.74).

□

2.4 Other Approximation Results

A special case of two layer neural net approximation occurs when $f(x)$ is a composition of two functions which are both approximable by single layer neural nets, that is, $f(x) = f_1(f_2(x))$, where $f_1 : \mathcal{R}^{d_1} \to \mathcal{R}$ and $f_2 : B \subset \mathcal{R}^d \to I \subset \mathcal{R}^{d_1}$. We then obtain the following theorem which holds for any probability measure P_X and for $d_1 = 1$.

Theorem 2.6 *Let $f(x) = f_1(f_2(x))$, $f_1 : \mathcal{R} \to \mathcal{R}$ and $f_2 : B \subset \mathcal{R}^d \to I \subset \mathcal{R}$. Let ϕ_v be a sigmoid with Lipshitz bound v. Suppose*

1. *$f_1(z)$ has a single layer neural net approximation,*

$$f_{1,T_1,v}(z) = \sum_{i=1}^{T_1} c_i \phi_v(u_i \cdot z + d_i') + c_0 \tag{2.85}$$

and

$$\sup_z |f_1(z) - f_{1,T_1,v}(z)| \leq \frac{C_{f_1}}{\sqrt{T_1}}, \tag{2.86}$$

$\sum_{i=1}^{T_1} |c_i| \leq V$ and $|u| = \max_i |u_i|$;

2. *$f_2(x)$ has a single layer neural net approximation,*

$$f_{2,T_2}(x) = \sum_{j=1}^{T_2} k_j \phi(\omega_j \cdot x + b_j) + d \tag{2.87}$$

and

$$\|f_2 - f_{2,T_2}\|_2 \leq \frac{C_{f_2}}{\sqrt{T_2}} \tag{2.88}$$

then $f(x)$ has a two layer neural net approximation given by

$$\begin{aligned} f_{T_1,T_2,v}(x) &= \sum_{i=1}^{T_1} c_i \phi_v(u_i \cdot \sum_{j=1}^{T_2} k_j \phi(\omega_j \cdot x + b_j) + u_i d + d_i') + c_0 \\ &= \sum_{i=1}^{T_1} c_i \phi_v(\sum_{j=1}^{T_2} a_{ji} \phi(\omega_j \cdot x + b_j) + d_i) + c_0 \end{aligned} \tag{2.89}$$

and the approximation rate satisfies

$$\|f - f_{T_1,T_2,v}\|_2 \leq \frac{C_{f_1}}{\sqrt{T_1}} + V|u|v\frac{C_{f_2}}{\sqrt{T_2}} \tag{2.90}$$

Proof : Using the two layer neural net approximation in (2.89), one obtains

$$\begin{aligned} & \|f - f_{T_1,T_2,v}\|_2 \\ \leq\ & \|f - f_{1,T_1,v}(f_2)\|_2 + \|f_{1,T_1,\tau}(f_2) - f_{T_1,T_2,v}\|_2 \\ =\ & \|f_1(f_2) - f_{1,T_1,v}(f_2)\|_2 \\ & \quad + \left\| \sum_{i=1}^{T_1} c_i \phi_v(u_i \cdot f_2(x) + d_i') - \sum_{i=1}^{T_1} c_i \phi_v(u_i \cdot \sum_{j=1}^{T_2} k_j \phi(\omega_j \cdot x + b_j) + d + d_i') \right\|_2 \\ \leq\ & \frac{C_{f_1}}{\sqrt{T_1}} + \sum_{i=1}^{T_1} c_i u_i v \|f_2 - f_{2,T_2}\|_2 \\ \leq\ & \frac{C_{f_1}}{\sqrt{T_1}} + V|u|v\frac{C_{f_2}}{\sqrt{T_2}}. \end{aligned}$$

□

A fascinating result by Kolmogorov [33, in Russian] gives a decomposition of any continuous function of several variables into superpositions of functions of one variable and sums. See, for example, Lorentz [37, Chapter 11] for a discussion in English. The decomposition takes the form

$$f(x_1, \cdots, x_d) = \sum_{q=1}^{2d+1} g\left(\sum_{p=1}^{d} c_p \phi_q(x_p)\right). \tag{2.91}$$

Kolmogorov's representation actually uses a superposition of increasing functions with Lipshitz bounds for his inner layer, not unlike our neural network representation here. To handle arbitrary continuous functions, the choice of functions ϕ_q in the Kolmogorov representation, where ϕ_q does not depend on f, is typically not smooth. Kolmogorov has also shown that the functions ϕ_q used in the decomposition are less smooth compared to the target function. However, g_q depends on f. More recently, Kůrková [34] showed that if $f \in C[0,1]^d$, with a modulus of continuity ω_f, then for every $m \in \mathcal{N}$ such that $m \geq 2d+1$ and $\frac{d}{m-d} + v < \epsilon/||f||$ and $\omega_f(\frac{1}{m}) < \frac{v(m-d)}{2m-3d}$ for some positive real v, then f can be approximated with an accuracy ϵ by a two hidden layer neural net containing $dm(m+1)$ units in the inner layer and $m^2(m+1)^d$ units in the outer layer. Nevertheless, results like Theorem 2.7 can be used when our decomposition is sufficiently "nice" in the sense that it decomposes into functions with single layer neural net approximations. Target functions such as $\sin(a\exp(-|x|^2))$ satisfy these conditions.

Chapter 3

Estimation

This chapter consists of two sections. In the first section, the results of Barron [4] are extended to the two hidden layer case. The approximation results in chapter 2 are used in the derivation of the final estimation bound. In the second section, the results are extended to the case for hard-limiter sigmoids (unit-step sigmoids) as activation functions on the nodes. The results are also an extension of Lee *et al* [35] to include a penalty term.

In both sections, we have data $(X_i, Y_i)_{i=1}^N$, which is an independent random sample of size N from a joint probability distribution $P_{X,Y}$. The target function is $E[Y|X = x]$ and its range is assumed to be bounded. We are not interested in bounding the empirical estimation error $\frac{1}{N}\sum_{i=1}^N (Y_i - \hat{f}(X_i))^2$ *per se*, but rather the mean square error $E(f(X)-\hat{f}(X))^2$ or the mean square prediction error $E(Y-\hat{f}(X))^2$

averaging with respect to $(X_i, Y_i)_{i=1}^N$ and (X, Y) independent with distribution P_{XY}. In the absence of any further knowledge of the target function, our function estimator depends on the empirical estimation error based on the data. The estimator is selected over a class of suitable neural network models and it is the minimizer of the empirical estimation error plus a penalty term. The penalty term is added to to help the neural net adapt to the target function. Typically, the penalty increases as the number of nodes of the estimated function increases. That is, it is a measure of the model complexity. Working alone with the empirical estimation error without penalty, it is clear that the more nodes the function estimator has, the smaller the empirical estimation error. However it is not necessary the best predictor for the target function. With the penalty term, there is a trade off between minimizing the model complexity and the empirical estimation error. A well chosen penalty term will adapt the function estimator better to the target function.

In section 3.1, the target function is assumed to have finite variation with respect to a class of ellipsoids. The estimator to the target function takes the form of a two hidden layer neural network that implements ramp activation functions. These ramp functions are Lipshitz bounded. The approach is similar to what Barron [4] did for the single layer case. The parameter space for the estimator is discretized, with a fixed bound on the outer weights of the outer layer, and bounds on the inner weights of both layers that grow with the number of nodes in each respective layer. The penalty term in this case is the log cardinality of the discretized parameter space. A

disadvantage with this approach is that since the inner weights of the ramp sigmoids are bounded, the class of models from which the estimator is chosen does not include two layer neural nets with step activation functions.

In section 3.2, we deal with function estimators that are neural networks implementing the step activation function with bounded outer weights in the outer layer. 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. This difference is called the expected regret. We give a general theorem that gives the convergence rate of the expected regret to a multiple of the empirical regret as the sample size increases. A condition for this theorem is the existence of an exponential inequality (see Lemma 3.3) over each model class that utilizes the l_1-covering number of each class.

3.1 Two Hidden Layers with Ramp Sigmoids

3.1.1 The setting

In this section, we discuss how we will derive an upper bound for the mean integrated squared error between the estimated two layer network and the target function. We

will pursue a very similar approach that Barron [4] has done for the single layer case. The target function $f(x)$ is estimated from data $(X_i, Y_i)_{i=1}^N$, a random sample of size N from a joint probability distribution $P_{X,Y}$ with $f(x) = E[Y_i|X_i = x]$. The range of the target function is assumed to be in a given interval $I = [-B_o, B_o]$, and the estimated two layer network takes the form (1.2) and $\theta = (c_i, d_i, b_{ji}, \omega_{ji}, a_{ji})_{i=1}^{T_1} {}_{j=1}^{T_2} \in \Theta_{T_1,T_2} \subset \mathcal{R}^{2T_1+2T_1T_2+dT_1T_2}$. In this section, we use unit ramp sigmoids with unit Lipshitz bound. By adjusting the internal weights of these sigmoids, we will also be able to obtain ramp sigmoids with other Lipshitz bounds since $\phi_v(z) = \phi_1(vz)$ for positive v. The notation $f_{T_1,T_2}(x, \theta)$ is used as a convenient abbreviation for (1.2). We also replace $f_{T_1,T_2}(x, \theta)$ by $\bar{f}_{T_1,T_2}(x, \theta)$ where $\bar{f} = (f \vee -B_o) \wedge B_o$ in order to get a better fit, taking advantage of knowledge of an interval $[-B_o, B_o]$ containing the range of f.

3.1.2 Index of resolvability

Following Barron [4], the index of resolvability is defined to be

$$R_{T_1,T_2,N}(f) = \min_{\theta \in \Theta_{T_1,T_2}} \left(\|f - \bar{f}_{T_1,T_2}(\cdot, \theta)\|^2 + \lambda \frac{L_{T_1,T_2,N}(\theta)}{N} \right) \tag{3.1}$$

where λ exceeds a multiple of the square of the presumed range of f, that is, $\lambda \geq \frac{20B_o^2}{3}$ and $L_{T_1,T_2,N}(\theta)$ are positive numbers satifying $L_{T_1,T_2,N}(\theta) \geq \log 2$, and

$$\sum_{\theta \in \Theta_{T_1,T_2}} e^{-L_{T_1,T_2,N}(\theta)} \leq 1. \tag{3.2}$$

Note that (3.2) is the Kraft-MacMillan inequality for the existence of uniquely decodable codes. The information-theoretic interpretation for similar expressions in the single layer case has been discussed in Barron [4] and are also applicable here when the parameter space is discretized. The term $\exp(-L_{T_1,T_2,N}(\theta))$ can also be interpreted as a prior over the parameter space. The minimum complexity estimator of a two layer neural network of a given architecture (T_1, T_2) is then

$$\hat{f}_{T_1,T_2,N}(x) = \bar{f}_{T_1,T_2}(x, \hat{\theta}_{T_1,T_2,N}), \tag{3.3}$$

where

$$\hat{\theta}_{T_1,T_2,N} = \underset{\theta \in \Theta_{T_1,T_2}}{\arg\min} \Big(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{f}_{T_1,T_2}(X_i, \theta))^2 + \lambda \frac{L_{T_1,T_2,N}(\theta)}{N} \Big). \tag{3.4}$$

It is a least squares estimator with a complexity penalty.

3.1.3 Cardinality of the discretized parameter space

Some bounds are assumed on the parameters in the parameter space Θ_{T_1,T_2}. We let τ_1 and τ_2 be bounds on the internal weights of the sigmoids in the outer and the inner layers respectively. These will be allowed to grow large at a specified rate with respect to the number of nodes in the respective layers. Let

$$\begin{aligned} \Theta_{T_1,T_2,\tau_1,\tau_2,C} = \; & \{\theta \in \Theta_{T_1,T_2} : \textstyle\sum_{i=1}^{T_1} |c_i| \leq C, \sum_{j=1}^{T_2} |a_{ji}| \leq \tau_1, \\ & |\omega_{ji}|_1 \leq \tau_2, |b_{ji}| \leq \tau_2 \text{and } |d_i| \leq \tau_1\}. \end{aligned} \tag{3.5}$$

$\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$ is a set of parameter points that ϵ-covers $\Theta_{T_1,T_2,\tau_1,\tau_2,C}$, that is, for any parameter point θ in $\Theta_{T_1,T_2,\tau_1,\tau_2,C}$, there is a θ^* in $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$ such that

$$\begin{aligned} |\omega_{ji} - \omega^*_{ji}|_1 &\leq \epsilon \\ |b_{ji} - b^*_{ji}| &\leq \epsilon \\ |d_i - d^*_i| &\leq \epsilon \\ \sum_{j=1}^{T_2} |a_{ji} - a^*_{ji}| &\leq \tau_1 \epsilon \end{aligned} \tag{3.6}$$

and

$$\sum_{i=1}^{T_1} |c_i - c^*_i| \leq C\epsilon.$$

In (3.6), ϵ may be used instead of $\tau_1\epsilon$ for the bound on the a_{ji}s. We then have the following lemma.

Lemma 3.1 *If* (3.6) *holds, then for each* θ *in* $\Theta_{T_1,T_2,\tau_1,\tau_2,C}$, *there is a* θ^* *in* $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$ *such that uniformly for* $x \in B$

$$|f_{T_1,T_2}(x,\theta) - f_{T_1,T_2}(x,\theta^*)| \leq 4C\tau_1\epsilon \tag{3.7}$$

where $\sum_{i=1}^{T_1} |c_i| \leq C$ *and* $f_{T_1,T_2}(x,\theta)$ *is a family of sigmoids of the form* (2.63).

Proof : Consider θ in $\Theta_{T_1,T_2,\tau_1,\tau_2,C}$ and θ^* in $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$. We use the fact the ramp sigmoid ϕ is Lipshitz, with Lipshitz constant 1, and it is bounded by 1. For any $x \in B$,

$$|f_{T_1,T_2}(x,\theta) - f_{T_1,T_2}(x,\theta^*)|$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{T_1} |c_i| \Big| \sum_{j=2}^{T_2} a_{ji}\phi(\omega_{ji}\cdot x + b_{ji}) - d_i - \sum_{j=2}^{T_2} a^*_{ji}\phi(\omega^*_{ji}\cdot x + b^*_{ji}) + d^*_i \Big| \\
&\qquad + \sum_{i=1}^{T_1} |c_i - c^*_i| \\
&\leq \sum_{i=1}^{T_1} |c_i| \Big[\sum_{j=2}^{T_2} |a_{ji}||\phi(y) - \phi(y^*)| + \sum_{j=2}^{T_2} |a_{ji} - a^*_{ji}||\phi(y^*)| + |d_i - d^*_i| \Big] \\
&\qquad + \sum_{i=1}^{T_1} |c_i - c^*_i| \qquad (3.8) \\
&\leq \sum_{i=1}^{T_1} |c_i| \Big[\sum_{j=2}^{T_2} |a_{ji}||y - y^*| + \sum_{j=2}^{T_2} |a_{ji} - a^*_{ji}| + |d_i - d^*_i| \Big] + \sum_{i=1}^{T_1} |c_i - c^*_i| \\
&\leq \sum_{i=1}^{T_1} |c_i| \sum_{j=2}^{T_2} |a_{ji}||y - y^*| + \sum_{i=1}^{T_1} |c_i| \sum_{j=2}^{T_2} |a_{ji} - a^*_{ji}| + \sum_{i=1}^{T_1} |c_i||d_i - d^*_i| \\
&\qquad + \sum_{i=1}^{T_1} |c_i - c^*_i| \qquad (3.9)
\end{aligned}
$$

where $y = \omega_{ji}\cdot x + b_{ji}$ and $y^* = \omega^*_{ji}\cdot x + b^*_{ji}$. From (3.5) and (3.6), $|y - y^*|$ is bounded by some multiple of ϵ. τ_1 can be assumed to be greater than 1. Hence it follows that

$$|f_{T_1,T_2}(x,\theta) - f_{T_1,T_2}(x,\theta^*)| \leq 4C\tau_1\epsilon \qquad (3.10)$$

where $\sum_{i=1}^{T_1} |c_i| \leq C$.

□

We also have a corollary to Lemma 3.1.

Corollary 3.1 *For functions f that have neural network approximations of the form*

$$f_{2T_1,2T_2} = \sum_{i=1}^{2T_1} c_i\phi\Big(\sum_{j=1}^{2T_2} \omega_{ij}\phi(a_{ij}\cdot x - b_{ij}) - d_i\Big)$$

with $\sum_{i=1}^{T_1} |c_i| \leq C$, *and with the ramp sigmoid* ϕ, *then there exists a parameter* θ^* *restricted to* $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$ *for which the approximation error is*

$$\|f - f_{T_1,T_2}(\cdot,\theta^*)\|_2 \leq \|f - f_{T_1,T_2}\|_2 + 4C\tau_1\epsilon, \tag{3.11}$$

where $f_{T_1,T_2}(\cdot,\theta)$ *is the best approximation to* f, *with* θ *chosen from* $\Theta_{T_1,T_2,\tau_1,\tau_2,C}$. *Consequently*

$$\|f - f_{T_1,T_2}(\cdot,\theta^*)\|_2^2 \leq 2\|f - f_{T_1,T_2}\|_2^2 + 32(C\tau_1\epsilon)^2 \tag{3.12}$$

Next, we examine the cardinality of the finite set $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$. We can actually take $\tau = \max(\tau_1,\tau_2)$ and consider the cardinality of $\Theta_{T_1,T_2,\epsilon,\tau,C}$ instead. The following lemma bounds the log-cardinality of $\Theta_{T_1,T_2,\epsilon,\tau,C}$ and hence that of $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$. As in the single layer case in Barron [4], a scaling property used in the count makes this log-cardinality independent of C.

Lemma 3.2 *For each* $\epsilon > 0$ *and* $C \geq 1$, *there is a set* $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$ *that satisfies* (4.6) *and has log-cardinality bounded by*

$$\log|\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}| \leq (2T_1 + 2T_1T_2 + dT_1T_2)\log\Big(\frac{2e(\tau+1)}{\epsilon}\Big). \tag{3.13}$$

Proof : We use similar counting techniques used in Barron [4, Lemma 2]. For the w_{ji}s, the cardinality is upper bounded by $\left(\frac{2e(\tau_2+\epsilon)}{\epsilon}\right)^{dT_1T_2}$. For the b_{ji}s, it is $\left(\frac{2e(\tau_2+\epsilon)}{\epsilon}\right)^{T_1T_2}$. The upper bound for the cardinality of the d_is is $\left(\frac{2(\tau_1+\epsilon)}{\epsilon}\right)^{T_1}$. For

the c_is, it is $\left(\frac{2(1+\epsilon)}{\epsilon}\right)^{T_1}$ and finally for the a_{ji}s, it is $\left(\frac{2(1+\epsilon)}{\epsilon}\right)^{T_1T_2}$. Thus the cardinality of $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$ is

$$|\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}| \leq \left(\frac{2e(\tau_2+\epsilon)}{\epsilon}\right)^{dT_1T_2}\left(\frac{2e(\tau_2+\epsilon)}{\epsilon}\right)^{T_1T_2}\left(\frac{2(\tau_1+\epsilon)}{\epsilon}\right)^{T_1} \left(\frac{2(1+\epsilon)}{\epsilon}\right)^{T_1}\left(\frac{2(1+\epsilon)}{\epsilon}\right)^{T_1T_2} \tag{3.14}$$

Choosing $\tau = \max(\tau_1, \tau_2)$ to be greater than 1, and ϵ less than 1 yields (3.13).

□

Estimation by two layer networks give us the flexibility in choosing the number of nodes in the inner and outer layers and also the way they are connected to one another. In the case of the unit ball example, (2.63) has only one outer node. In the case of the function composition example, the same inner layer nodes are fed forward to the outer layer. In the unit ball case, the bound in (3.7) is $k\tau_1\epsilon$ and in the function composition example it is $kC\tau_1\epsilon$ for some constant k. Lemma 3.2 generalizes to

$$\log|\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}| \leq m_{T_1,T_2}\log\left(\frac{2e(\tau+1)}{\epsilon}\right) \tag{3.15}$$

where m_{T_1,T_2} is the dimension of the parameter space.

3.1.4 The Risk Bound

We set $\lambda = 8B_o^2$, where $[-B_o, B_o]$ is the assumed bound on the support of $|Y|$. Let $\gamma = 7$. These choices of γ and λ will be used when we apply a complexity

regularization theorem in Barron [1] in the proof of the theorem below. The theorem in [1] bounds the mean square prediction error in terms of the resolvability, specifically

$$E\|f - \hat{f}_{T_1,T_2}(\cdot,\theta)\|_2^2 \leq \gamma R_{T_1,T_2,N}(f) + \frac{2\gamma\lambda}{N}.$$

Throughout this whole chapter, expressions of the form $O(g(\cdot))$ refer to quantities bounded by a constant times $g(\cdot)$, where $g(\cdot)$ is an expression involving several variables and the constant is independent of those variables. Dependence of the bounds on B_o may be hidden in these "constants" but can be made explicit from examination of the proofs. In particular, we require the constant to not depend on T_1, T_2, N, d or f. We are now equipped to prove the following theorem.

Theorem 3.1 *The minimum complexity estimator $\hat{f}_{T_1,T_2}$ of a two layer neural net of a given architecture (T_1, T_2) with parameters restricted to $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$ with $\epsilon = \frac{1}{8C\tau_1}(\frac{\lambda}{N}m_{T_1,T_2})^{1/2}$ has risk bound*

$$E\|f - \hat{f}_{T_1,T_2}(\cdot,\theta)\|_2^2 \leq O\left(\|f - f_{T_1,T_2}\|_2^2\right) + O\Big(\frac{m_{T_1,T_2}}{N}\log(\frac{\tau^4 N}{m_{T_1,T_2}})\Big), \tag{3.16}$$

where N is the sample size, m_{T_1,T_2} is the dimension of the parameter space Θ, and $f_{T_1,T_2} = f_{T_1,T_2,\tau_1,\tau_2}$ is the best approximation to f in $\Theta_{T_1,T_2,\tau_1,\tau_2,C}$ and $\tau = \max(\tau_1,\tau_2)$ is the bound on the sum of internal weights to each ramp sigmoid. In particular with $C = V$, under the additional conditions of Theorem 2.5 for the target function, if $\tau_1 \geq \max(4d\sqrt{2e}T_1, \sqrt{\frac{T_2}{d(d+1)}})$ and $\tau_2 \geq \max(2\sqrt{d})$, if P_X is the uniform distribution on S, and if $V_{f,\mathcal{E}} \leq V$ where $\mathcal{E}$ is the class of ellipsoids with Lebesgue measure $\mu(E) \leq$

$\mu(\mathcal{S})$, *then*

$$E\|f - \hat{f}_{T_1,T_2}(\cdot,\theta)\|_2^2 \le O\Big(\frac{V^2}{T_1}\Big) + O\Big(\frac{V^2 d^2}{\sqrt{T_2}}\Big) + O\Big(B_o^2 \frac{dT_1T_2}{N}\log(dT_1T_2N)\Big). \quad (3.17)$$

Proof : From the above corollary and Lemma 3.2, there exists θ^* in $\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}$,

$$\begin{aligned} R_{T_1,T_2,N}(f) &\le \|f - f_{T_1,T_2}(\cdot,\theta^*)\|_2^2 + \frac{\lambda}{N}\log|\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}| \\ &\le 2\|f - f_{T_1,T_2}\|_2^2 + 32(C\tau_1\epsilon)^2 + \frac{\lambda}{N}m_{T_1,T_2}\log\Big(\frac{2e(\tau+1)}{\epsilon}\Big). \end{aligned} \quad (3.18)$$

The penalty is $\lambda \log|\Theta_{T_1,T_2,\epsilon,\tau_1,\tau_2,C}|$. The choice of ϵ that optimizes this bound is

$$\epsilon = \frac{1}{8C\tau_1}\Big(\frac{\lambda}{N}m_{T_1,T_2}\Big)^{1/2} \quad (3.19)$$

which yields

$$\begin{aligned} R_{T_1,T_2,N}(f) &\le 2\|f - f_{T_1,T_2}\|_2^2 + \frac{\lambda m_{T_1,T_2}}{2N} \\ &\qquad + \frac{\lambda m_{T_1,T_2}}{2N}\log\Big((16Ce(\tau+1))^2\tau_1^2\frac{N}{\lambda m_{T_1,T_2}}\Big) \\ &\le 2\|f - f_{T_1,T_2}\|_2^2 + \frac{\lambda m_{T_1,T_2}}{2N} \\ &\qquad + \frac{\lambda m_{T_1,T_2}}{2N}\log\Big((16Ce(\tau+1))^2\tau^2\frac{N}{\lambda m_{T_1,T_2}}\Big) \end{aligned} \quad (3.20)$$

where $\tau = \max(\tau_1,\tau_2)$. Thus the bound is of order

$$R_{T_1,T_2,N}(f) \le 2\|f - f_{T_1,T_2,\tau_1,\tau_2}\|_2^2 + \frac{K\lambda m_{T_1,T_2}}{N}\log\frac{\tau^4 N}{m_{T_1,T_2}} \quad (3.21)$$

where K some positive constant. Under the conditions assumed for Theorem 2.5, $V_{f,\mathcal{E}} \leq V$, $\tau_1 = K_1 \max(dT_1, \frac{\sqrt{T_2}}{d})$, $\tau_2 = 2\sqrt{d}$ and $\lambda = 8B_o^2$; we obtain

$$R_{T_1,T_2,N}(f) \leq k_1 \frac{V^2}{T_1} + k_2 \frac{V^2 d^2}{\sqrt{T_2}} + k_3 B_o^2 \frac{dT_1T_2}{N} \log(dT_1T_2N). \tag{3.22}$$

Then using the complexity regularization theorem in Barron [1], we obtain

$$E\|f - \hat{f}_{T_1,T_2}(\cdot,\theta)\|_2^2 \leq O(R_{T_1,T_2,N}(f)) \leq O\Big(\frac{V^2d^2}{T_1}\Big) + O\Big(\frac{V^2d^2}{\sqrt{T_2}}\Big) + O\Big(B_o^2 \frac{dT_1T_2}{N} \log(dT_1T_2N)\Big).$$

The estimation bound for the other cases can be worked out in a similar manner. In terms of the approximation rate, the dimension of the parameter space and the sample size, it is

$$E\|f - \hat{f}_{T_1,T_2}(\cdot,\theta)\|_2^2 \leq O(\|f - f_{T_1,T_2}\|_2^2) + O\Big(\frac{m_{T_1,T_2}}{N} \log(m_{T_1,T_2}N)\Big).$$

□

3.1.5 Selecting the Size of the Network

The bounds above can be extended to the case when the size of the two hidden-layer network architecture is not preselected. Let $L(T_1, T_2)$ be numbers satisfying $\sum_{T_1} \sum_{T_2} e^{-L(T_1,T_2)} \leq 1$.The index of resolvability is then

$$R_N(f) = \min_{(T_1,T_2)} \Big(R_{T_1,T_2,N}(f) + \lambda \frac{L(T_1,T_2)}{N}\Big) \tag{3.23}$$

and the minimum complexity estimator with both θ and (T_1, T_2) estimated is $\hat{f}_{\hat{T}_1,\hat{T}_2,N}(x)$ where

$$(\hat{T}_1, \hat{T}_2) = \underset{(T_1,T_2)}{\arg\min} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{f}_{T_1,T_2,N}(X_i))^2 + \lambda \frac{L_{T_1,T_2,N}(\hat{\theta})}{N} + \lambda \frac{L(T_1, T_2)}{N} \right). \quad (3.24)$$

Again we use the same values of γ and λ when applying the complexity regularization theorem from Barron [1]. Thus we have the following corollary when the number of nodes are estimated from the data.

Corollary 3.2 *Under the conditions of theorem* 3.1 *and with the choice of* $L(T_1, T_2) = 2\log(T_1 T_2) + 2\beta$, *where* $\beta = \log \frac{\pi^2}{6}$, *the minimum complexity estimator with* (T_1, T_2) *estimated satisfies*

$$E\|f - \hat{f}_{\hat{T}_1,\hat{T}_2,N}\|_2^2 \leq O\left(B_o^2 V^{3/2} d^{5/4} \left(\frac{\log N}{N} \right)^{\frac{1}{4}} \right). \quad (3.25)$$

Proof : Take the penalty to be $L(T_1, T_2) = 2\log(T_1 T_2) + 2\beta \leq 2T_1 T_2 + 2\beta$. From (3.22),

$$\begin{aligned} R_N(f) &\leq \min_{(T_1,T_2)} \left\{ k_1 \frac{V^2}{T_1} + k_2 \frac{V^2 d^2}{\sqrt{T_2}} + k_3 B_o^2 \frac{d T_1 T_2}{N} \log(d T_1 T_2 N) + \lambda \frac{2T_1 T_2 + 2\beta}{N} \right\} \\ &\leq \min_{(T_1,T_2)} \left\{ k_1 \frac{V^2}{T_1} + k_2 \frac{V^2 d^2}{\sqrt{T_2}} + k_3' B_o^2 \frac{d T_1 T_2}{N} \log(d T_1 T_2 N) \right\}. \end{aligned}$$

Plugging in suitable values of T_1 and T_2 in terms of N yields

$$E\|f - \hat{f}_{\hat{T}_1,\hat{T}_2,N}\|_2^2 = O(R_N(f)) \leq O\left(B_o^2 V^{3/2} d^{5/4} \left(\frac{\log N}{N} \right)^{\frac{1}{4}} \right),$$

which goes to zero as $N \to \infty$. The values of T_1 and T_2 are of order $V^{1/2}\left(\frac{N}{d\log(dVN)}\right)^{\frac{1}{4}}$ and $Vd^{5/4}\left(\frac{N}{d\log(dVN)}\right)^{\frac{1}{2}}$ respectively.

□

In the same manner, one may use a penalty term to select from the data a suitable C for the sum of the absolute value of the output weights. Then the result would be as above but we would not need prior knowledge of the value of $V = V_f$.

3.2 Estimation with Heaviside Sigmoids

3.2.1 Preliminaries

In this section, we extend the results of Barron [4] to the case of unit step sigmoids and that of Lee *el al* [35] to include a penalty term. 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 $\mathcal{L}_2(P_X)$. The support of each X_i is in some $\mathcal{X} \subset \mathcal{R}^d$. For a given sample from $X_1, \cdots, X_N$, we write $\underline{x} \in \mathcal{X}^N$.

Before specializing to neural nets, we give a general theorem bounding the risk of penalized least squares estimators under entropy conditions on the component models. 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

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{f}_{M,N}(X_i))^2 + \text{pen}_N(M) \tag{3.26}$$

where the form of the penalty will be specified later.

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 $\mathcal{L}_2(P_X)$). 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)

$$\begin{aligned} r(f) = r(f, f^*) &:= E(Y - f(X))^2 - \inf_{f\in\mathcal{F}} E(Y - f(X))^2 \\ &= E(Y - f(X))^2 - E(Y - f^*_{\mathcal{F}}(X))^2 \end{aligned} \tag{3.27}$$

and the empirical loss function

$$\hat{r}(f) := \frac{1}{N}\sum_{i=1}^{N}(Y_i - f(X_i))^2 - \frac{1}{N}\sum_{i=1}^{N}(Y_i - f^*_{\mathcal{F}}(X_i))^2 \tag{3.28}$$

Note that the mean square prediction error satisfies $E(Y - f(X))^2 = ||f^* - f||_2^2 + E(\text{Var}(Y|X))$ for every f. Thus the relative regret $r(f, f^*)$ measures the regret in $\mathcal{L}_2$ approximation of f^* by f compared to the best approximation in $\mathcal{F}$,

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

In particular if f^* is in $\mathcal{F}$ then

$$r(f, f^*) = \|f^* - f\|_2^2.$$

We select $\hat{f}_{\hat{M}}$ from $\bigcup_M \mathcal{F}_M$, and bound the expected value of the relative regret of the estimator $E[r(\hat{f}_{\hat{M}})]$. The choice of $\hat{f}_{\hat{M}}$ minimizes the penalized empirical mean squared error

$$\hat{r}(f) + \text{pen}_N(M).$$

Correspondingly it is natural to examine the performance in terms of a penalized approximation error

$$r(f) + \text{pen}_N(M).$$

Thus we define an index of resolvability

$$R_{N,M}(f^*) := \min_{f \in \mathcal{F}_M} \{r(f, f^*) + \text{pen}_N(M)\}. \tag{3.29}$$

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 lemma is needed for our theorem.

Lemma 3.3 *Suppose the distribution of* (X_i, Y_i) *is such that* $|Y_i| \leq B_o$, *that* $|f(x)| \leq$

B_M, for all $f \in \mathcal{F}_M$, and that $|f^*_{\mathcal{F}}(x)| \leq B_1$, where $f^*_{\mathcal{F}}$ is the $\mathcal{L}_2(P_X)$ projection of f^* onto a convex class of functions $\mathcal{F}$ that includes $\mathcal{F}_M$, and is used in the definition of the regrets r and $\hat{r}$, then for each $v, \delta > 0$,

$$\begin{aligned} & P\{\exists f_M \in \mathcal{F}_M \quad \text{with} \quad r(f_M) \geq 2\hat{r}(f_M) + v + \delta\} \\ \leq \; & 6 \sup_{\underline{x} \in \mathcal{X}^{2N}} \mathcal{N}(\frac{\delta}{8}, \mathcal{F}_{M|\underline{x}}) \exp\left(-\frac{3vN}{10496\bar{B}_M^2}\right) \\ = \; & 6\mathcal{N}_{2N}(\frac{\delta}{8}, M) \exp\left(-\frac{3vN}{10496\bar{B}_M^2}\right), \end{aligned} \tag{3.30}$$

where $\bar{B}_M = \max(B_M, B_o, B_1, 1)$.

Remark : This is actually adapted from a result in Lee *et al* [35, Theorem 3] by rescaling some of the variables. We will not reproduce their proof here. We note that $|r(f_M)| \leq 8\bar{B}_M^2$ and $|\hat{r}(f_M)| \leq 8\bar{B}_M^2$. Let $K_1 = 8\bar{B}_M^2$ and $K_2 = 16\bar{B}_M^2$. From Lee *et al* [35, Theorem 6] applied to $\frac{1}{K_1}r(f_M)$, $\frac{1}{K_1}\hat{r}(f_M)$, $\frac{v}{K_1}$ and $\frac{\delta}{K_1}$ and setting $\alpha = \frac{1}{2}$ in [35, Theorem 6], we obtain

$$\begin{aligned} & P\{\exists f_M \in \mathcal{F}_M \quad \text{with} \quad r(f_M) \geq 2\hat{r}(f_M) + v + \delta\} \\ \leq \; & 2 \sup_{\underline{x} \in \mathcal{X}^{2N}} \mathcal{N}(\frac{\delta}{8}, \mathcal{F}_{M|\underline{x}}) \exp\left(-\frac{3vN}{10496\bar{B}_M^2}\right) \\ & +4 \sup_{\underline{x} \in \mathcal{X}^{2N}} \mathcal{N}(\frac{\delta}{8}, \mathcal{F}_{M|\underline{x}}) \exp\left(-\frac{vN}{64\bar{B}_M^2}\right) \\ \leq \; & 6\mathcal{N}_{2N}(\frac{\delta}{8}, M) \exp\left(-\frac{3vN}{10496\bar{B}_M^2}\right). \end{aligned}$$

In our application of this result, the choices of v and δ will depend on our entropy bounds for the models and resulting penalty terms.

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 resolvability. First we cover the case that there is a fixed upper bound B to the values of B_o and B_M for all $M \in \mathcal{M}$. Next we cover the case that B_M unbounded for $M \in \mathcal{M}$. (In that case clipping the values to a fixed range would violate the convexity requirement for $\bigcup_M \mathcal{F}_M$.)

Theorem 3.2

A. 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

$$\sum_M 6\mathcal{N}_{2N}(\frac{\delta_{M,N}}{8}, M) \exp\left(-\frac{3(\text{pen}_N(M) - \delta_{M,N}/2)N}{5248B^2}\right) \leq 1, \tag{3.31}$$

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

$$E[r(\hat{f}_{\hat{M}})] \leq 2R_N(f^*) + \frac{c_1 B^2}{N}, \tag{3.32}$$

where $c_1 = 20992$.

B. 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*

$$\sum_{M \in \mathcal{M}} 6\mathcal{N}_{2N}(\frac{\delta_{M,N}}{8}, M) \exp\left(-\frac{3\left(\text{pen}_N(M) - \frac{\delta_{M,N}}{2} - \frac{1312\bar{B}_M^4}{N}\right) N}{5248\bar{B}_M^2} \right) \leq 1, \quad (3.33)$$

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*

$$E[r(\hat{f}_{\hat{M}})] \leq 7R_N(f^*). \quad (3.34)$$

If each term in the summand (3.31) 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

$$\text{pen}_N(M) = \frac{5248B^2}{3N} \ln\left[\frac{6\bar{\mathcal{N}}_{2N}(\frac{\delta_{M,N}}{8}, M)}{g(M)}\right] + \frac{\delta_{M,N}}{2}. \quad (3.35)$$

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 $\mathcal{F}_M$.

Proof of part A : We first bound the difference between the theoretical loss and a multiple of the empirical loss. Let $v_M = t + 2\text{pen}_N(M) - \delta_{M,N}$. From Lemma 3.3 and (3.35),

$$\begin{aligned}
& P\left\{\exists M, \exists f_M \in \mathcal{F}_M, r(f_M) \geq 2\hat{r}(f_M) + 2\text{pen}_{M,N} + t\right\} \\
= \; & P\left\{\exists M, \exists f_M \in \mathcal{F}_M, r(f_M) \geq 2\hat{r}(f_M) + (2\text{pen}_N(M) - \delta_{M,N} + t) + \delta_{M,N}\right\} \\
\leq \; & \sum_M P\left\{\exists f_M \in \mathcal{F}_M, r(f_M) \geq 2\hat{r}(f_M) + v_M + \delta_{M,N}\right\}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_M 6\bar{\mathcal{N}}_{2N}(\frac{\delta_{M,N}}{8}, M) \exp\left(-\frac{3(\mathrm{pen}_{M,N} - \frac{\delta_{M,N}}{2} + \frac{t}{2})N}{5248B^2}\right) \\
&\leq \sum_M 6\bar{\mathcal{N}}_{2N}(\frac{\delta_{M,N}}{8}, M) \exp\left(-\frac{3\left(\mathrm{pen}_M(N) - \frac{\delta_{M,N}}{2}\right)}{5248B^4}\right) \exp\left(-\frac{3tN}{10496B^2}\right) \\
&\leq \exp\left(-\frac{3tN}{10496B^2}\right). \qquad (3.36)
\end{aligned}
$$

In other words,

$$
r(f_M) \leq 2\hat{r}(f_M) + 2\mathrm{pen}_N(M) + t, \text{for all } f_M \in \mathcal{F}_M, \text{ for all } M,
$$

except for data in a set of probability not greater than $\exp(-\frac{3tN}{10496B^2})$.

Taking $\hat{M}$ and $\hat{f} = \hat{f}_{\hat{M}}$ to be the choice that minimizes $\hat{r}(f_M) + \mathrm{pen}_N(M)$, the following bounds hold on the loss $r(\hat{f}, f^*)$,

$$
\begin{aligned}
r(\hat{f}) &< 2\hat{r}(\hat{f}) + 2\mathrm{pen}_N(\hat{M}) + t \\
&\leq 2\hat{r}(f^*_{M^*}) + 2\mathrm{pen}_N(M^*) + t,
\end{aligned}
$$

except for data in a set of probability not greater than $\exp(-\frac{3tN}{10496B^2})$. Here $f^*_{M^*}$ minimizes the resolvability $r(f_M) + \mathrm{pen}_N(M)$. Thus

$$
\begin{aligned}
&P\left\{r(\hat{f}) > 2\hat{r}(f^*_{M^*}) + 2\mathrm{pen}_N(M^*) + t\right\} \\
\leq\ &P\left\{r(\hat{f}) \geq 2\hat{r}(\hat{f}) + 2\mathrm{pen}_N(M) + t\right\} \\
\leq\ &\exp\left(-\frac{3tN}{10496B^2}\right). \qquad (3.37)
\end{aligned}
$$

Now

$$
P\left\{\hat{r}(f^*_{M^*}) \geq r(f^*_{M^*}) + t\right\} \leq \exp\left(-\frac{3tN}{48B^2}\right) \qquad (3.38)
$$

(see remark after the proof). Since $\exp\left(-\frac{3tN}{48B^2}\right) \leq \exp\left(-\frac{3tN}{10496B^2}\right)$, we obtain

$$P\left\{\hat{r}(f^*_{M^*}) \geq r(f^*_{M^*}) + t\right\} \leq \exp\left(-\frac{3tN}{10496B^2}\right). \tag{3.39}$$

By summing up (3.37) and (3.39),

$$P\left\{r(\hat{f}) > 2r(f^*_{M^*}) + 2\mathrm{pen}_{M^*,N} + 3t\right\} \leq 2\exp\left(-\frac{3tN}{10496B^2}\right). \tag{3.40}$$

Choose $f^*_{M^*}$ to attain $R_N(f^*)$, the minimum value of the resolvability. Integrating (3.40) out with respect to t, we obtain

$$\begin{aligned} & E[r(\hat{f})] - 2R_N(f^*) \\ \leq & \int_0^\infty P\{r(\hat{f}) - 2R_N(f^*) \geq 3t\}3dt \\ \leq & 6\int_0^\infty \exp\left(-\frac{3tN}{10496B^2}\right)dt \\ = & \frac{20992B^2}{N}. \end{aligned} \tag{3.41}$$

Thus

$$E[r(\hat{f})] \leq 2R_N(f^*) + \frac{c_1B^2}{N},$$

when $c_1 = 20992$.

□

For part B, we may take the penalty to be

$$\mathrm{pen}_N(M) = \frac{5248\bar{B}_M^2}{3N}\ln\left[\frac{6\bar{\mathcal{N}}_{2N}(\frac{\delta_{M,N}}{8}, M)}{g(M)}\right] + \frac{\delta_{M,N}}{2} + \frac{1312\bar{B}_M^4}{N}, \tag{3.42}$$

where $g(M)$ satisfies $\sum_m g(M) \leq 1$ as before. What is different here is the presence of the $\frac{\bar{B}_M^4}{N}$ term in the penalty that we include to handle the case of unknown B_M, $M \in \mathcal{M}$.

Proof of part B : The proof is similar in essence to part A. We first bound the difference between the theoretical loss and a multiple of the empirical loss. Let $v_M = \frac{t\bar{B}_M^2}{N} + 2\text{pen}_{M,N} - \delta_{M,N} - \frac{2624\bar{B}_M^4}{N}$. Note that $\frac{t^2}{10496N} \geq \frac{t\bar{B}_M^2}{N} - \frac{2624\bar{B}_M^4}{N}$. Thus from Lemma 3.3 and (3.42),

$$
\begin{aligned}
& P\left\{\exists M, \exists f_M \in \mathcal{F}_M, r(f_M) \geq 2\hat{r}(f_M) + 2\text{pen}_N(M) + \frac{t^2}{10496N}\right\} \\
\leq\ & P\left\{\exists M, \exists f_M \in \mathcal{F}_M, r(f_M) \geq 2\hat{r}(f_M) + 2\text{pen}_{M,N} + \frac{t\bar{B}_M^2}{N} - \frac{2624\bar{B}_M^4}{N}\right\} \\
=\ & P\left\{\exists M, \exists f_M \in \mathcal{F}_M, r(f_M) \geq 2\hat{r}(f_M) + \left(2\text{pen}_N(M) - \delta_{M,N} - \frac{2624\bar{B}_M^4}{N} + \frac{t\bar{B}_M^2}{N}\right)\right. \\
& \qquad \left. + \delta_{M,N}\right\} \\
\leq\ & \sum_M P\left\{\exists f_M \in \mathcal{F}_M, r(f_M) \geq 2\hat{r}(f_M) + v_M + \delta_{M,N}\right\} \\
\leq\ & \sum_M 6\bar{\mathcal{N}}_{2N}\left(\frac{\delta_{M,N}}{8}, M\right) \exp\left(-\frac{3\left(2\text{pen}_N(M) - \delta_{M,N} - \frac{2624\bar{B}_M^4}{N} + \frac{t\bar{B}_M^2}{N}\right)N}{10496\bar{B}_M^2}\right) \\
\leq\ & \sum_M 6\bar{\mathcal{N}}_{2N}\left(\frac{\delta_{M,N}}{8}, M\right) \exp\left(-\frac{3\left(\text{pen}_N(M) - \frac{\delta_{M,N}}{2} - \frac{1312B_M^4}{N}\right)N}{5248\bar{B}_M^2}\right) \exp\left(-\frac{3t}{10496}\right) \\
\leq\ & \exp\left(-\frac{3t}{10496}\right).
\end{aligned}
$$

Note that we modified the assumption on the penalty so that the factor $\exp(-\frac{3t}{10496})$

would not depend on B_M and hence could be factored out of the sum. Thus,

$$P\left\{\exists M, \exists f_M \in \mathcal{F}_M, r(f_M) \geq 2\hat{r}(f_M) + 2\text{pen}_N(M) + \frac{t^2}{10496N}\right\} \leq \exp(-\frac{3t}{10496}). \tag{3.43}$$

Taking $\hat{M}$ and $\hat{f} = \hat{f}_{\hat{M}}$ to be the choice that minimizes $\hat{r}(f_M) + \text{pen}_N(M)$, it follows from (3.43) that

$$P\left\{r(\hat{f}) > 2\hat{r}(f^*_{M\cdot}) + 2\text{pen}_N(M^*) + \frac{t^2}{10496N}\right\} \leq \exp\left(-\frac{3t}{10496}\right). \tag{3.44}$$

Now

$$P\left\{\hat{r}(f^*_{M\cdot}) \geq r(f^*_{M\cdot}) + \frac{t\bar{B}^2_{M\cdot}}{N}\right\} \leq \exp\left(-\frac{3t}{48}\right) < \exp\left(-\frac{3t}{10496}\right) \tag{3.45}$$

(see remark after the proof). By summing up (3.44) and (3.45),

$$P\left\{r(\hat{f}) > 2r(f^*_{M\cdot}) + 2\text{pen}_N(M^*) + \frac{2t\bar{B}^2_{M\cdot}}{N} + \frac{t^2}{10496}\right\} \leq 2\exp\left(-\frac{3t}{10496}\right), \tag{3.46}$$

and hence,

$$P\left\{r(\hat{f}) > 2r(f^*_{M\cdot}) + 2\text{pen}_N(M^*) + \frac{5248\bar{B}^4_{M\cdot}}{N} + \frac{3t^2}{10496N}\right\} \leq 2\exp\left(-\frac{3t}{10496}\right). \tag{3.47}$$

Integrating (3.47) out with respect to t, we obtain

$$\begin{aligned} & E[r(\hat{f})] - 2R_N(f^*) - \frac{5248\bar{B}^4_{M\cdot}}{N} \\ \leq & \int_0^\infty P\{r(\hat{f}) - 2R_N(f^*) - \frac{5248\bar{B}^4_{M\cdot}}{N} \geq \frac{3t}{10496N}\}\frac{3t}{5248N}dt \end{aligned}$$

$$\leq \frac{3}{2624N}\int_0^\infty t\exp\left(-\frac{3t}{10496}\right)dt$$
$$= \frac{41984}{3N}. \tag{3.48}$$

Thus

$$E[r(\hat{f})] \leq 2R_N(f^*) + \frac{\bar{5}248B_{M\cdot}^4}{N} + \frac{41984}{3N}$$
$$\leq 7R_N(f^*).$$

□

Remark : The lemma below is used in the proof of Lemma 3.3 and in the derivation of (3.38) and (3.45). Define $d_v(r,s) := \frac{r-s}{v+r}$.

Lemma 3.4 (Lee *et al* [35, Lemma 8]). *Let $V_1, \cdots, V_d$ be independent identically distributed random variables with $|V_i| < Q_1$, $EV_i \geq 0$ and $EV_i^2 < Q_2 EV_i$, $Q_2 \geq 1$, for $i = 1, \cdots, d$. Then for $0 < \alpha \leq 1$,*

$$P\left\{d_v\left(E\left[\frac{1}{N}\sum_{i=1}^N V_i\right], \frac{1}{N}\sum_{i=1}^N V_i\right) \geq \alpha\right\} \leq \exp\left(-\frac{3\alpha^2 vN}{2(Q_1+Q_2)}\right). \tag{3.49}$$

Lemma 3.4 is derived from a result in Haussler [26]. In our application of Lemma 3.4, $V_i = E(Y - f_{M\cdot}^*(X))^2 - (Y_i - f_{M\cdot}^*(X))^2$, and $|Y| \leq B_o$, where $B_o \geq 1$, and $|f_{M\cdot}^*| \leq B_{M\cdot}$. Let $\bar{B}_{M\cdot} = \max(B_o, B_{M\cdot})$. We can set Q_1 to be $8B_{M\cdot}^2$ and Q_2 can be set to $16\bar{B}_{M\cdot}^2$, and $\alpha = 1$, thus yielding

$$P\left\{\hat{r}(f_{M\cdot}^*) \geq 2r(f_{M\cdot}^*) + t\right\} \leq \exp\left(-\frac{3tN}{48\bar{B}_{M\cdot}^2}\right),$$

for all $f > 0$ and hence

$$P\left\{\hat{r}(f^*_{M\cdot}) \geq 2r(f^*_{M\cdot}) + \frac{t\bar{B}^2_{M\cdot}}{N}\right\} \leq \exp\left(-\frac{3t}{48}\right).$$

3.2.2 Main Result for Single Layer Networks

In this section, we apply the result from Theorem 3.2 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 (1.1).

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 $\mathcal{F}_B$ = closure $(\bigcup_T \mathcal{F}_{B,T})$. Then the indices for application of Theorem 3.2 are integers $\mathcal{F}_B = M = \{1, 2, ...\}$.
$M = \{1, 2, ...\}$.

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 $\mathcal{M} = \{1, 2, ...\}^2$. Now $\bigcup_{B,T} \mathcal{F}_{B,T}$ is convex and includes $\bigcup_T \mathcal{F}_T$. In this case, by Hornik *et al* [29], its closure $\mathcal{F} = \text{closure}\left(\bigcup_{B,T} \mathcal{F}_{B,T}\right) = \text{closure}\left(\bigcup_T \mathcal{F}_T\right)$ contains all $\mathcal{L}_2(P_X)$ functions. 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 as we see below, 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 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.

A further refinement in the estimator is obtained by taking the model $\mathcal{F}_T^o$ to be the collection of all $f_T \in \mathcal{F}_T$ for which $|f_T(x)| \leq B_o$ for $x \in \mathcal{S}$, where $\mathcal{S}$ is the presumed bounded support of P_X and B_o is a known bound on the support of the

response variable Y. Once again $\bigcup_T \mathcal{F}_T^o$ is convex and $\mathcal{F} = \text{closure} \bigcup_T \mathcal{F}_T^o$ contains all continuous functions on S that are bounded by B_o (by application of the result of Cybenko [16]). The advantage of this refinement is that we get better control over the l_1-entropy of $\mathcal{F}_T^o$ (without the appearance of the potentially large B in the penalty term).

We state some results from Lee *el al* [35] that we will use to prove our main result.

Lemma 3.5 *Let $\mathcal{F}_{T|\underline{x}}$ be a class of single layer neural networks with T hidden units and range restricted to $[-B, B]$. Then the l_1 ϵ-covering number for $\mathcal{F}_{T|\underline{x}}$ and any sequence $\underline{x} \in \mathcal{X}^N$ is*

$$\mathcal{N}(\epsilon, \mathcal{F}_{T|\underline{x}}) \leq 2 \left(\frac{eNd}{d+1}\right)^{T(d+1)} \left(\frac{2eB}{\epsilon} \ln \frac{2eB}{\epsilon}\right)^{T}. \tag{3.50}$$

Thus a bound on $\mathcal{N}_{2N}(\epsilon, \mathcal{F}_{T|\underline{x}})$ is $2 \left(\frac{2eNd}{d+1}\right)^{T(d+1)} \left(\frac{2eB}{\epsilon} \ln \frac{2eB}{\epsilon}\right)^{T}$. In a version of this result in Lee *et al* [35, Lemma 3], the sum of absolute values of output weights is bounded by B and the bound on the covering number was given in terms of B. We do not prove this Lemma here. However we will prove a similar version of this result for the two hidden layer case (see Lemma 3.8).

We also make use of the following result from Lee *el al* [35]. This result extends Jones' [32] iterative approximation algorithm for a target function that does not

necessarily belong to the convex hull of classes of functions in some Hilbert space.

Lemma 3.6 (Lee *et al* [35, Theorem 2]). *Let H be a Hilbert space with norm $||\cdot||$. Let G be a subset of H with $||g|| \leq b$ for each $g \in G$. Let $\mathrm{conv}(G)$ be the convex hull of G. For any $f \in H$, let $d_f = \inf_{g' \in \mathrm{conv}(G)} ||g' - f||$. Suppose that f_1 is chosen to satisfy*

$$||f_1 - f||^2 \leq \inf_{g \in G} ||g - f||^2 + \epsilon_1$$

and iteratively, f_k is chosen to satisfy

$$||f_k - f||^2 \leq \inf_{g \in G} ||(1-\alpha)f_{k-1} + \alpha g - f||^2 + \epsilon_k$$

where $\alpha = 2/(k+1)$, $c \geq b^2$ and $\epsilon_k \leq \frac{4(c-b^2)}{(k+1)^2}$. Then for every $k \geq 1$,

$$||f - f_k||^2 - d_f^2 \leq \frac{4c}{k}. \tag{3.51}$$

Typically f_k is chosen in the form $(1-\alpha)f_{k-1} + \alpha g$ with g chosen to achieve the minimum of $||(1-\alpha)f_{k-1} + \alpha g - f||^2$. Then we may take $c = B^2$. Note that at each step, f_k is in G. In this chapter we do not make use of the algorithm *per se*, but we do use the bound (3.41). If f_k^* is the best approximation to f using a convex combination of k points from $\mathcal{G}$, then

$$||f - f_k^*||^2 - d_f^2 \leq ||f - f_k||^2 - d_f^2 \leq \frac{4c}{k}.$$

When $\mathcal{H} = \mathcal{L}_2(P_X)$ and when f has projection $f_{\mathcal{F}}$ onto $\mathcal{F} = \overline{\mathrm{conv}}(\mathcal{G})$ achieving $d_f^* = ||f - f_{\mathcal{F}}||^2$, one has the inequality $||f - f_{\mathcal{F}}||^2 \leq ||f - f_k||^2 - d_f^2$ and hence

$$||f_k - f_{\mathcal{F}}||^2 \leq \frac{4c}{k}.$$

Now we return our attention to neural net estimation. Recall that a single hidden layer sigmoidal neural network in $\mathcal{F}_{B,T}$ takes the form

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

Denoting the parameter space by $\Theta_{T,B} \subset \mathcal{R}^{T(d+1)}$, where $\Theta_{T,B} = \{\theta = (a_i, b_i, c_i)_{i=1}^T : \sum_{i=1}^T |c_i| \leq B\}$, the penalized least squares estimator with T and θ estimated and B fixed is

$$\hat{f}_{\hat{T},B}(x) = f_{\hat{T}}(x,\hat{\theta}), \tag{3.52}$$

where

$$(\hat{\theta},\hat{T}) = \underset{\theta_T \in \Theta_{T,B}, T \in \{1,2,...\}}{\arg\min} \left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{f}_T(X_i,\theta_T))^2 + \text{pen}_{B,N}(T)\right). \tag{3.53}$$

We will see that a valid choice for the penalty, when the constrain on the sum of absolute value of output weights B is greater than the bound B_o on $|Y_i|$, is

$$\begin{aligned}\text{pen}_{B,N}(T) &= \frac{5248B^2}{3N}\ln\left(2T^2\pi^2\left(\frac{2eNd}{d+1}\right)^{T(d+1)}\left(\frac{3Ne}{656BT}\ln\left(\frac{3Ne}{656BT}\right)\right)^T\right) \\ &\quad + \frac{5248B^2T}{3N}\end{aligned} \tag{3.54}$$

uniformly for $\theta_T \in \Theta_{T,B}$. This corresponds in (3.35) (with $M = T$) to the choices of $g(M) = \frac{6}{\pi^2 T^2}$ and the bound on $\mathcal{N}\left(\frac{\delta_{M,N}}{8}, M\right)$ from Lemma 3.5 with $\delta_{M,N} = \frac{10496B^2T}{3N}$ to optimize the resulting bounds. In place of (3.54), one could use any penalty that is at least as large, for example,

$$\text{pen}_{B,N}(T) = \frac{KB^2 m_T}{N}\ln N, \tag{3.55}$$

where K is some constant, $N \geq 2$, and where $m_T = T(d+1)$ is the dimension of the parameter space.

Theorem 3.3 *Let the data be $(X_i, Y_i)_{i=1}^N$, independently distributed with joint probability distribution $P_{X,Y}$ and $f^*(x) = E(Y_i|X_i = x)$, $|Y| \leq B_o$. Then an upper bound to the expected regret of the estimator $\hat{f}_{\hat{T}}$ compared to the best $g \in \mathcal{F}_B$ with $B \geq B_o$ is*

$$\begin{aligned} & E\|f^* - \hat{f}_{\hat{T}}\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 \\ \leq \; & 2 \min_T \left\{ \|f^* - f_{T,B}\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 + \frac{K m_T B^2}{N} \ln N \right\}, \end{aligned} \tag{3.56}$$

where $f_{T,B}$ is the best approximation to f^* in $\mathcal{F}_{B,T}$. Using the bound from Lemma 3.6, *then*

$$\begin{aligned} E\|f^* - \hat{f}_{\hat{T}}\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 \; &\leq \; 2 \min_T \left\{ \frac{4B^2}{T} + \frac{K m_T B^2}{N} \ln N \right\} \\ &\leq \; O\left(B^2 \left(\frac{d \ln N}{N} \right)^{\frac{1}{2}} \right). \end{aligned} \tag{3.57}$$

By choosing $\hat{f} = \hat{f}_{\hat{T},\hat{B}}$ with $\hat{T}$ and $\hat{B}$ among the intergers T, B with $B \geq B_o$ so to minimize

$$\frac{1}{N} \sum_{i=1}^{N} \left(Y_i - f_{\hat{T},B}(X_i) \right)^2 + \mathrm{pen}_N(T, B), \tag{3.58}$$

with

$$\mathrm{pen}_N(T, B) = \mathrm{pen}_B, N(T) + (2 \ln B + \ln \frac{\pi^2}{6}) \frac{5248 B^2}{N} + \frac{1312 B^4}{N},$$

the mean square error $E\|f^* - f\|_2^2$ converges to zero for every f^* bounded by B_o at

the rate

$$E\|f^* - \hat{f}\|_2^2 \leq 7 \min_B \left\{ \|f^* - f_B^*\|^2 + KB^2 \left(\frac{d \ln N}{N} \right)^{\frac{1}{2}} + K' \frac{B^4}{N} \right\}, \tag{3.59}$$

where for each B, f_B^* is the projection of f^* onto $\mathcal{F}_B$.

Proof : With the choice of $g(T) = \frac{6}{\pi^2 T^2}$, $\delta = \frac{10496B^2T}{3N}$ and $\epsilon = \frac{1312B^2T}{3N}$, and using Lemma 3.5, the penalty is

$$\begin{aligned} \text{pen}_{B,N}(T) &= \frac{5248B^2}{3N} \ln \left(2T^2\pi^2 \left(\frac{eNd}{d+1} \right)^{T(d+1)} \left(\frac{3Ne}{656BT} \ln \left(\frac{3Ne}{656BT} \right) \right)^T \right) \\ &\quad + \frac{5248B^2}{3N}. \end{aligned}$$

It follows from Theorem 3.2 and the upper bound (3.55) to the penalty that

$$\begin{aligned} E[r(\hat{f}_{\hat{T}})] &= E\|f^* - \hat{f}_{\hat{T}}\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 \\ &\leq 2R_N(f^*) + c_1 \left(\frac{B^2}{N} \right) \\ &\leq 2 \min_T \left\{ \|f^* - f_T\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 + \frac{Km_T B^2}{N} \ln N \right\} + c_1 \left(\frac{B^2}{N} \right), \end{aligned}$$

Now $\|f^* - f_{T,B}\|_2^2$ above (and in (3.56)) is the best approximation error between f^* and a T-term neural net approximation. This is bounded above by the square error if the T-term approximation were to be chosen iteratively, thus from Lemma 3.6,

$$\|f^* - f_T\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 \leq \frac{4B^2}{T}, \tag{3.60}$$

and hence

$$E\|f^* - \hat{f}_{\hat{T},B}\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 \leq 2 \min_T \left\{ \frac{4B^2}{T} + \frac{Km_T B^4}{N} \ln N \right\},$$

where K is a constant. Optimizing over T yields the bound in (3.57), which is $K'B^2\left(\frac{d\ln N}{N}\right)^{\frac{1}{2}}$, which decreases to zero as $N \to \infty$.

Finally choosing $\hat{f} = \hat{f}_{\hat{T},\hat{B}}$ with $\hat{B} \geq B_o$ selected to minimize

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - f_{\hat{T},B}(X_i)\right)^2 + \text{pen}_N(\hat{T}, B),$$

we get that $E\|f^* - \hat{f}\|^2$ converges to zero for every $f^* \in \mathcal{L}_2(P_X)$ at rate controlled by the index of resolvability

$$\begin{aligned} E\|f^* - \hat{f}\|^2 &\leq 7 \min_{T,B}\left\{\inf_{g\in\mathcal{F}_{B,T}}\|f^* - g\|^2 + \frac{KdTB^2\ln N}{N} + \frac{K'B^4}{N}\right\} \\ &\leq 7 \min_{B}\left\{\|f^* - f_B^*\|^2 + \min_{T}\left\{\frac{4B^2}{T} + \frac{KdTB^2}{N}\ln N\right\} + \frac{K'B^4}{N}\right\} \\ &\leq 7 \min_{B}\left\{\|f^* - f_B^*\|^2 + KB^2\left(\frac{d\ln N}{N}\right)^{\frac{1}{2}} + \frac{K'B^4}{N}\right\}, \end{aligned}$$

where for each B, f_B^* is the projection of f^* onto $\mathcal{F}_B$.

□

If f^* has finite variation V_{f^*} with respect to half-spaces then we achieve rate $O\left(V_{f^*}^2\left(\frac{d\ln N}{N}\right)^{\frac{1}{2}} + \frac{V_{f^*}^4}{N}\right)$ by automatic selection of B and T without prior knowledge of V_{f^*}. If V_{f^*} is infinite, we still have consistency, but at a slower rate.

If we used $\mathcal{F}_T^o$ consisting of T term networks with range bounded by the fixed B_o (in place of controlling the sum of absolute output weights through B), then we would achieve similar consistency for all $f \in$ closure $(\bigcup_T \mathcal{F}_T^o)$ (which includes all bounded

continuous f on a bounded support $\mathcal{S}$), with a somewhat better resolvability

$$E\|f^* - \hat{f}\|_2^2 \leq 2 \min_T \left\{ \inf_{g \in \mathcal{F}_T^o} \|f^* - g\|_2^2 + \frac{K' dT B_o^2 \ln N}{N} \right\}. \tag{3.61}$$

3.2.3 Main Result for Two Layer Networks

In this section, the results from the previous section is extended to two hidden layer feedforward neural nets with step activation functions. The definitions from the previous sections extend to the present setting. 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 (1.2).

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

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

We may restrict $\sum_{j=1}^{T_2} |\omega_{ij}| \leq 1$ 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.

We state some results that we are using in this section. However, we first define

the concept of pseudo-dimension. Let $\mathcal{G}$ be a class of functions mapping from $\mathcal{X}$ to $\mathcal{R}$ and let $x_1, \cdots, x_N \in \mathcal{X}$. We say that $x_1, \cdots, x_N$ are shattered by $\mathcal{G}$ if there exists $r \in \mathcal{R}^N$ such that for each $b = (b_1, \cdots, b_N) \in \{0,1\}^N$, there is an $g \in \mathcal{G}$ such that for each i,

$$g(x_i) \begin{cases} \geq r_i & \text{if } b_i = 1 \\ < r_i & \text{if } b_i = 0. \end{cases}$$

The pseudo-dimension is defined as

$$\dim_P(\mathcal{G}) = \max\{N : \exists x_1, \cdots, x_N, \quad \mathcal{G} \text{ shatters } x_1, \cdots, x_N\}$$

if such a maximum exists, and ∞ otherwise. For the class of unit step functions $\phi(a \cdot x + b)$, the pseudo-dimension and the VC-dimension coincide and is $d+1$.

Lemma 3.7 (Lee *et al* [35, Lemma 1], Haussler [26]) *Let F be a class of functions from a set Z into $[-B, B]$ and suppose the pseudo-dimension of F is D for some $1 \leq D \leq \infty$. Then for all $0 < \epsilon \leq B$ and any finite sequence $\underline{z}$ of points in $\mathcal{Z}$, the l_1 ϵ-covering number $\mathcal{N}(\epsilon, F_{|\underline{z}})$ is bounded by*

$$\mathcal{N}(\epsilon, F_{|\underline{z}}) < 2\left(\frac{2eB}{\epsilon}\ln\frac{2eB}{\epsilon}\right)^D.$$

This will be used to prove the following lemma.

Lemma 3.8 *Let $\mathcal{F}_{T_1,T_2|\underline{x}}$ be a class of two layer neural networks with T_1 outer hidden units and T_2 inner hidden units, with range of the neural net output restricted to*

$[-B, B]$. *Then the* l_1 ϵ*-covering number for* $\mathcal{F}_{T_1,T_2|\underline{x}}$ *and any sequence* $\underline{x} \in \mathcal{X}^N$ *is*

$$\mathcal{N}_N(\epsilon, \mathcal{F}_{T_1,T_2|\underline{x}}) \leq 2\left(\frac{eNd}{W}\right)^{T_1 W}\left(\frac{2eB}{\epsilon}\ln\frac{2eB}{\epsilon}\right)^{T_1}, \tag{3.62}$$

where $W = T_2 d + 2T_2 + 1$.

Note that W is the total number of parameters in the inner layer per node, when that node is in the outer layer.

Proof : Let $G = \{x \to \phi\left(\sum_{j=1}^{T_2} \omega_{ij}\phi(a_{ij} \cdot x - b_{ij}) - d_i\right)\}$. From Cover [14, 15], and Baum and Haussler [8], the function class G has VC-dimension bounded by $W = T_2 d + 2T_2 + 1$. The VC-dimension bound of a multilayer neural net (of step activation functions) is the same as that of one with all the nodes stringed out together in a single hidden layer. Fix a sequence $\underline{x} \in \mathcal{X}^N$, $\mathcal{X} \subset \mathcal{R}^d$. From Baum and Haussler [8, Theorem 1], the cardinality of G restricted to $\underline{x}$ is bounded by $|G_{|\underline{x}}| \leq \left(\frac{eNd}{W}\right)^W$.

There are at most $\left(\frac{eNd}{W}\right)^{WT_1}$ ways of picking $(g_1, \cdots, g_{T_1})$ which will give functions in $G_{|\underline{x}}$. Let $f = \sum_{i=1}^{T_1} c_i g_i$ be an arbitrary function in $\mathcal{F}_{T_1,T_2}$; with range restricted to $[-B, B]$. Momentarily fix one such set of function $(g_1, \cdots, g_{T_1})$. Evaluation of these functions at $x_1, \cdots, x_N$ in $\mathcal{R}^d$ yields N points in $\mathcal{R}^{T_1}$, where $\underline{z}_1 = (g_1(x_1), \cdots, g_T(x_1)), \cdots, \underline{z}_N = (g_1(x_N), \cdots, g_T(x_N))$. For linear functions with T_1 inputs (the inputs are now a points in $\mathcal{R}^{T_1}$), the size of an l_1 ϵ-cover of $\{\underline{z} \to \underline{c} \cdot \underline{z} : |\underline{c} \cdot \underline{z}| \leq B, \underline{c} \in \mathcal{R}_1^T\}_{|\underline{z}_1,\cdots,\underline{z}_N}$ is no more than $2\left(\frac{2eB}{\epsilon}\ln\frac{2eB}{\epsilon}\right)^{T_1}$, from Lemma 3.7. This is because the class of all linear functions $\{\underline{z} \to \underline{c} \cdot \underline{z} : \underline{c} \in \mathcal{R}^{T_\infty}\}$ with domain $\mathcal{R}^{T_1}$ has pseudo-dimension T_1 by Pollard [46] and so restricting the domain (to the set $\bigcup_{\underline{x}} G_{|\underline{x}}^{T_1}$

of points realizable as outputs $\underline{z} = (g_1, \cdots, g_T)$ from the first layer) and restricting the $\underline{c}$ (so that the range of $\underline{c} \cdot \underline{z}$ is bounded by B) will have not larger pseudo-dimension, thus Lemma 3.7 applies. Thus

$$\mathcal{N}_N(\epsilon, \mathcal{F}_{T_1,T_2|x}) \leq 2\left(\frac{eNd}{W}\right)^{T_1 W}\left(\frac{2eB}{\epsilon}\ln\frac{2eB}{\epsilon}\right)^{T_1}$$

□

Much to our initial surprise this covering bound does not necessarily require constraint on the sum of the absolute values of the output weights. If the values of $\underline{z} = (g_1(\underline{x}), \cdots, g_{T_1}(\underline{x}))$ ranged over all points in $\{0,1\}^{T_1}$ then requiring $|\underline{c} \cdot \underline{z}| \leq B$ would be the same as $|\underline{c}|_1 \leq B$. However, not all points in $\{0,1\}^{T_1}$ are necessarily represented in the range of $(g_1(\underline{x}), \cdots, g_{T_1}(\underline{x}))$.

Recall that a two hidden layer neural network takes the form

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

Denoting the parameter space by $\Theta_{T_1,T_2,B} \subset \mathcal{R}^{2T_1+2T_1T_2+dT_1T_2}$, where $\Theta_{T_1,T_2,B} = \{\theta_{T_2,T_2} = (c_i, d_i, b_{ji}, \omega_{ji}, a_{ji})_{i=1\,j=1}^{T_1\;T_2} : \sum_{i=1}^{T_1}|c_i| \leq B, \sum_{j=1}^{T_2}|ji| \leq 1\}$, the minimum complexity estimator with (T_1, T_2) and θ estimated is

$$\hat{f}_{\hat{T}_1,\hat{T}_2}(x) = f_{\hat{T}_1,\hat{T}_2}(x, \hat{\theta}), \tag{3.63}$$

where

$$(\hat{T}_1, \hat{T}_2, \hat{\theta}) = \underset{T_1,T_2,\theta\in\Theta_{T_1,T_2,B}}{\arg\min}\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{f}_{T_1,T_2}(X_i, \theta))^2 + \text{pen}_{B,N}(T_1, T_2)\right). \tag{3.64}$$

We will see that with a fixed $B \geq B_o$. We may set

$$\begin{aligned} & \mathrm{pen}_{B,N}(T_1, T_2) \\ = \; & \frac{5248B^2}{3N} \ln \left[\frac{\pi^4 T_1^2 T_2^2}{3} \left(\frac{2eNd}{T_2 d + 2T_2 + 1} \right)^{T_1(T_2 d + 2T_2 + 1)} \left(\frac{3Ne}{656BT_1} \ln \left(\frac{3Ne}{656BT_1} \right) \right)^{T_1} \right] \\ & + \frac{5248B^2 T_1}{3N} \end{aligned} \tag{3.65}$$

uniformly for $\theta_{T_1,T_2} \in \Theta_{T_1,T_2,B}$. This corresponds in (3.35) to the choices of $g(M) = \frac{36}{\pi^4 T_1^2 T_2^2}$ and the bound on $\bar{\mathcal{N}}_{2N}\left(\frac{\delta}{8}, M\right)$ from Lemma 3.8 with $\delta_{M,N} = \frac{10496B^2T_1}{3N}$ to optimize the resulting bounds In place of (3.65), one could use any penalty that is at least as large, for example,

$$\mathrm{pen}_{B,N}(T_1, T_2) = \frac{KB^2 m_{T_1,T_2}}{N} \ln N, \tag{3.66}$$

where K is some constant, $N \geq 2$, and where $m_{T_1,T_2} = 2T_1 + 2T_1T_2 + dT_1T_2$ is the dimension of the parameter space. The following theorem is the two hidden layer analogue of Theorem 3.3.

Theorem 3.4 *Let the data be $(X_i, Y_i)_{i=1}^N$, identically and randomly distributed with joint probability distribution $P_{X,Y}$ and $f^*(x) = E(Y_i|X_i = x)$, and $|Y| \leq B_o$. Let $B \geq B_o$. Then an upper bound to the expected regret compared to the best $g \in \mathcal{F}_B =$ closure$\bigcup_{T_1,T_2} \mathcal{F}_{B,T_1,T_2}$ is*

$$\begin{aligned} & E\|f^* - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 \\ \leq \; & 2 \min_{T_1,T_2} \left\{ \|f^* - f_{T_1,T_2}\|_2^2 - \inf_{g \in \mathcal{F}} \|f^* - g\|_2^2 + \frac{K m_{T_1,T_2} B^2}{N} \ln N \right\}, \end{aligned} \tag{3.67}$$

where f_{T_1,T_2} is the best approximation to f^* in $\mathcal{F}_{B,T_1,T_2}$.

If $\frac{1}{B}f^*$ is in $\mathcal{H}$, the closure of the convex hull of signed indicators of ellipsoids E with Lebesque measure $\mu(E) \leq \mu(S)$, and P_X is the uniform probability measure, then

$$E\|f^* - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2 \leq 2 \min_{T_1,T_2} \left\{ \frac{K_1 B^2}{T_1} + \frac{K_2^2 B^2 d^2}{\sqrt{T_2}} + \frac{K B^2 m_{T_1,T_2}}{N} \ln N \right\}. \tag{3.68}$$

which yields

$$E\|f^* - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2 \leq O\left(d^{3/2} B^2 \left(\frac{\ln N}{N} \right)^{\frac{1}{4}} \right). \tag{3.69}$$

If $\frac{1}{B}f^*$ is not in $\mathcal{H}$. then the risk compared to the best $g \in \mathcal{F}_B = closure \bigcup_{T_1,T_2} \mathcal{F}_{B,T_1,T_2}$ satisfies

$$\begin{aligned} E\|f^* - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2 \;\leq\; & \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2 \\ & + 2 \min_{T_1,T_2} \left\{ \frac{4B^2}{T_1} + \frac{K_2^2 B^2 d^2}{\sqrt{T_2}} + 2 \frac{K_2 B d}{T_2^{\frac{1}{4}}} \sqrt{\frac{4B^2}{T_1} + \inf_{g \in \mathcal{F}} \|f^* - g\|_2^2} \right. \\ & \left. + \frac{K m_{T_1,T_2} B^4}{N} \ln N \right\}. \end{aligned} \tag{3.70}$$

Proof : With the choice of $g(T_1,T_2) = \frac{36}{\pi^4 T_1^2 T_2^2}$, $\delta = \frac{10496 B^2 T_1}{3N}$ and $\epsilon = \frac{1312 B^2 T_1}{3N}$, and using Lemma 3.8, the penalty is

$$\begin{aligned} & \text{pen}_{B,N}(T_1,T_2) \\ = \; & \frac{5248 B^2}{3N} \ln \left[\frac{\pi^4 T_1^2 T_2^2}{3} \left(\frac{eNd}{T_2 d + 2T_2 + 1} \right)^{T_1(T_2 d + 2T_2 + 1)} \left(\frac{3Ne}{656 B T_1} \ln \left(\frac{3Ne}{656 B T_1} \right) \right)^{T_1} \right] \\ & + \frac{5248 B^2 T_1}{3N}. \end{aligned}$$

It follows from Theorem 3.2 and the upper bound (3.56) to the penalty that

$$E[r(\hat{f}_{\hat{T}_1,\hat{T}_2})] \;=\; E\|f^* - \hat{f}_{\hat{T}_1,\hat{T}_2}\|_2^2 - \inf_{g \in \mathcal{F}_B} \|f^* - g\|_2^2$$

$$\leq 3R_N(f^*) + O\left(\frac{B^4}{N}\right)$$
$$\leq 2\min_{T_1,T_2}\left\{\|f^* - f_{T_1,T_2}\|_2^2 - \inf_{g\in\mathcal{F}}\|f^* - g\|_2^2 + \frac{Km_{T_1,T_2}B^2}{N}\ln N\right\}.$$

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 Theorem 2.4 and using the bound there on the approximation error, we obtain

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

Optimizing over T_1 and T_2 yields the bound in (3.69), which is $K'd^{3/2}B^2\left(\frac{\ln N}{N}\right)^{\frac{1}{4}}$. 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.

Let $d^* = \inf_{g\in\mathcal{F}_B}\|f^* - g\|_2^2$. Suppose $\frac{1}{B}f^*$ is not in the closure of convex hull of ellipsoids with bounded surface area, then (scaling down to $B = 1$ first)

$$\|f^* - f_{T_1,T_2}\|_2^2 - d^*$$
$$\leq (\|f^* - f_{T_1}\|_2 + \|f_{T_1} - f_{T_1,T_2}\|_2)^2 - d^*$$
$$= \|f^* - f_{T_1}\|_2^2 - d^* + \|f_{T_1} - f_{T_1,T_2}\|_2^2 + 2\|f_{T_1} - f_{T_1,T_2}\|_2\|f^* - f_{T_1}\|_2 \quad (3.71)$$
$$\leq \frac{4}{T_1^2} + \frac{K_2^2d^2}{\sqrt{T_2}} + 2\frac{K_2d}{T_2^{\frac{1}{4}}}\sqrt{\frac{4}{T_1^2} + d^*}. \quad (3.72)$$

In going from (3.71) to (3.72), we use the bound (3.53) in Lemma 3.6 to bound $\|f^* - f_{T_1}\|_2^2 - d^* \leq \frac{4}{T_1}$. The bound for $\|f_{T_1} - f_{T_1,T_2}\|_2 \leq \frac{K_2^{1/2}d}{T_2^{1/4}}$ is obtained from

Theorem 2.4 when sums of indicators of ellipsoids are approximated with two hidden layer nets. Substituting (3.72) back into (3.67) and rescaling back to $B \geq 1$, we obtain the bound

$$\begin{aligned} E[r(\hat{f}_{\hat{T}_1,\hat{T}_2})] &\leq 2 \min_{T_1,T_2} \left\{ \|f^* - f_{T_1,T_2}\|_2^2 - \inf_{g \in \mathcal{F}} \|f^* - g\|_2^2 + \frac{K m_{T_1,T_2} B^4}{N} \ln N \right\} \\ &\leq 2 \min_{T_1,T_2} \left\{ \frac{4B^2}{T_1^2} + \frac{K_2 B^2 d^2}{\sqrt{T_2}} + 2 \frac{K_2^{1/2} B d}{T_2^{\frac{1}{4}}} \sqrt{\frac{4B^2}{T_1} + d^*} \right. \\ &\qquad \left. + \frac{K m_{T_1,T_2} B^2}{N} \ln N \right\}. \end{aligned}$$

□

If we proceed to modify the penalty to account for selection of B by penalized least squares then we would obtain an estimator $\hat{f} = \hat{f}_{\hat{T}_1,\hat{T}_2,\hat{B}}$ with risk $E\|f^* - \hat{f}\|_2^2$ bounded by the minimum over $B \geq B_o$ of $\|f^* - f_B^*\|_2^2$ plus a constant times the right side of expression (3.67) (modified to include an order $\frac{B^4}{N}$ term). In particular we would have the minimum over $B \geq B_o$ of the right side of (3.70) as a bound on the risk that tends to zero for all target functions bounded by B_o. The resulting minimization over T_1, T_2, B express the trade-off between the approximation error and the size parameters of the two layer networks relative to the sample size. As the case of convex combinations of indicators of ellipsoids illustrates, two layer networks provide accurate estimators in cases where accurate one layer representations are not necessarily available.

Chapter 4

A Greedy Algorithm

4.1 Preliminaries

The material in this chapter was presented at the 1995 World Congress of Neural Networks in Washington, DC. This is based on the joint work by Barron and Cheang [7]. An algorithm is presented for implementing the single hidden layer approximation. It takes advantage of the assumption that the target function (when normalized) is in the closure of the set of convex combinations of sigmoids.

A single hidden layer feedforward sigmoidal network $f_T(x)$ of the form

$$f_T(x,\theta) = \sum_{i=1}^{T} c_i \phi(a_i \cdot x - b_i), x \in \mathcal{R}^d, \tag{4.1}$$

parametrized by $\theta = (a_i, b_i, c_i)_{i=1}^T$ with internal weight vectors a_i in $\mathcal{R}^d$, internal location parameter b_i in $\mathcal{R}$, external weights c_i is considered. We will use the odd-

symmetric logistic sigmoid (hyperbolic tangent) $\phi(z) = (\exp(z) - \exp(-z))/(\exp(z) + \exp(-z))$. With this choice we may and do restrict attention to positive weights c_i without loss of generality because $-\phi(z) = \phi(-z)$, so that the negative sign may be absorbed into the choice of the internal weights.

A discussion of the $\mathcal{L}_2$ bounds for function approximation by single hidden layer feedforward neural networks is found in chapters 1 and 2. Those bounds could be theoretically obtainable using non-linear least squares. Although the non-linear least squares estimate is provably accurate, its computation is problematic as there are usually many local minima in the error surface. A number of algorithms based on backpropagation exist to perform this minimization. But no provably computationally feasible algorithms have been demonstrated to have the level of performance guaranteed in (1.3).

4.2 An accurate greedy algorithm

Let f be the target function. Suppose f/c is in the closure of the convex hull $\overline{\text{conv}}G$ of some subset G of a Hilbert space, that is, $f \in c\overline{\text{conv}}G$. Here G may be the set of sigmoids $G = \{\phi(a \cdot x - b) : a \in \mathcal{R}^d, b \in \mathcal{R}\}$. The algorithm updates

$$f_k = \alpha_k f_{k-1} + \beta_k c g_k \tag{4.2}$$

where $g_k \in G$, and $||g_k|| \leq 1$. Thus at each step we introduce one more sigmoid. The resulting f_k is a single hidden layer network of the form (4.1). The heart of the matter is the choice of the internal weights a_k and b_k of the sigmoid determining g_k, For each choice of g_k, optimal external weights α_k and β_k are readily determined by least squares projection. To see how the internal weights may be chosen, we examine the improvement in the approximation error that result from including the new term.

$$\begin{aligned} \|f - f_k\|^2 &= \|f - (\alpha_k f_{k-1} + \beta_k c g_k)\|^2 \\ &\leq \|f - (1 - \frac{1}{m}) f_{k-1} - \frac{1}{m} c g_k\|^2 \qquad (4.3) \\ &= \|(1 - \frac{1}{m})(f - f_{k-1}) + \frac{1}{m}(c g_k - f)\|^2 \\ &= \{(1 - \frac{1}{m})^2 \|f - f_{k-1}\|^2 - 2\frac{1}{m}(1 - \frac{1}{m})\langle f - f_{k-1}, c g_k - f\rangle \\ &\qquad + (\frac{1}{m})^2 \|c g_k - f\|^2\} \qquad (4.4) \end{aligned}$$

Now $f/c \in \overline{\text{conv}} G$ implies that there exists $g_k \in G$ such that $\langle f - f_{k-1}, c g_k - f\rangle \geq 0$. With such g in (4.4) we have

$$\begin{aligned} \langle f - f_{k-1}, g\rangle &\geq \frac{1}{c}\langle f - f_{k-1}, f\rangle \\ &= \frac{1}{c}\langle f - f_{k-1}, f - f_{k-1}\rangle \\ &= \frac{1}{c}\|f - f_{k-1}\|^2 \qquad (4.5) \end{aligned}$$

and we obtain

$$\begin{aligned} \|f - f_k\|^2 &\leq (1 - \frac{1}{m})^2 \|f - f_{k-1}\|^2 + (\frac{1}{m})^2 \|c g_k - f\|^2 \\ &\leq (1 - \frac{1}{m})^2 \|f - f_{k-1}\|^2 + 2(\frac{1}{m})^2 c^2 \qquad (4.6) \end{aligned}$$

Then the bound

$$\|f - f_k\|^2 \leq \frac{2c^2}{k} \tag{4.7}$$

readily follows by induction. This is the standard Jones [32] proof except for the explicit use of $\alpha = \frac{1}{m}$ in deriving the bound. Thus if at each step we find a sigmoid g satisfying $\langle f - f_{k-1}, g \rangle \geq \frac{1}{c}\|f - f_{k-1}\|^2$, then we get a good approximation with $\|f - f_{k-1}\| \leq \frac{2c^2}{k}$. An extension of this iterative bound in Lee *et al* [35] shows that f need not be in $\overline{\text{conv}}G$. Iterations yield

$$\|f_k - f_*\|_2^2 \leq \|f - f_k\|^2 - \|f - f_*\|^2 \leq \frac{2c^2}{k} \tag{4.8}$$

where f_* is the projection of f onto $\overline{\text{conv}}G$.

We now assume that the function f is sampled from the data points $(X_i, f(X_i))_{i=1}^N$. The theory above may also be used to give a bound on the average squared training error

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{f}_{k,N}(X_i))^2 \leq \frac{2c^2}{k} \tag{4.9}$$

in the case that $Y_i = f(X_i)$ with $f/c \in \overline{\text{conv}}G$. The point here is that it is sufficient to obtain at each step a sufficiently large value for the inner product

$$\frac{1}{N}\sum_{i=1}^{N}(f(X_i) - f_{k-1}(X_i))\phi(a_k \cdot X_i - b_k) \tag{4.10}$$

Indeed if at the kth step, say, we get

$$\sum_{i=1}^{N}(f(X_i) - f_{k-1}(X_i))\phi(a_k \cdot X_i - b_k) \geq \frac{1}{c}\sum_{i=1}^{N}(f(X_i) - f_{k-1}(X_i))^2 \tag{4.11}$$

then (4.9) follows as desired. This follows from (4.4) and (4.5). If $f/c \in \overline{\text{conv}}G$, we know that such a, b exist. The problem is to find them.

One tactic is to maximize the inner product (4.10) directly. Here we have a considerable reduction in the search space (from $T(d+2)$ parameters in the full T term network) down to $d+1$ parameters internal to the current node. Nevertheless, this inner product may still have multiple local optima that inhibit the ability to search for a maximizer with a suitable level of performance. One could try to simply use a local gradient algorithm (one unit backpropagation) from some random initialization of a and b. The condition (4.11) can be checked to acertain whether we are successful at this step. Unfortunately, due to the multiplicity of local optima, there is no guarantee that it will achieve a desirably large value of the inner product (4.10). Attempting restarts at new random initializations should improve on the search. However, there are no theoretical bounds on the number of such restarts required.

4.3 Theoretical Basis for a Heuristic Algorithm

Recall our assumption that f/c is in the closed convex hull $\overline{\text{conv}}G$, where G is as before. Without loss of generality, we assume that $c = 1$. Here we use the logistic sigmoid $\psi(z) = 1/(1+\exp(-z))$ which differs from the hyperbolic tangent sigmoid $\phi(z)$ by a simple rescaling of the output. As before, we assume that we have data points $(X_i, f(X_i))_{i=1}^N$. By a suitable output rescaling, there exists $f^*(x) = \frac{1}{2} + \frac{\rho}{2}(f(x) -$

1) such that $f^* \in \overline{\text{conv}} G_\psi$, where $0 < \rho < 1$. Instead of maximizing

$$\sum_{i=1}^{N} (f^*(X_i) - f^*_{k-1}(X_i))(\psi(a_k \cdot X_i - b_k) - f^*(X_i)) \tag{4.12}$$

we will maximise a lower bound on it. Let $r_i = f^*(X_i) - f^*_{k-1}(X_i)$, $\bar{\psi}(x) = 1 - \psi(x)$, $\bar{f}^*(x) = 1 - f^*(x)$, $r^+ = \max(0, r)$ and $r^- = (-r)^+$. Then

$$\begin{aligned}
& \sum_{i=1}^{N} r_i(\psi(a_k \cdot X_i - b_k) - f^*(X_i)) \\
= & \sum_{i=1}^{N} r_i^+(\psi(a_k \cdot X_i - b_k) - f^*(X_i)) + \sum_{i=1}^{N} r_i^-(\bar{\psi}(a_k \cdot X_i - b_k) - \bar{f}^*(X_i)) \\
= & \sum_{i=1}^{N} r_i^+ f^*(X_i)(\frac{\psi(a_k \cdot X_i - b_k)}{f^*(X_i)} - 1) + \sum_{i=1}^{N} r_i^- \bar{f}^*(X_i)(\frac{\bar{\psi}(a_k \cdot X_i - b_k)}{\bar{f}^*(X_i)} - 1) \\
\geq & \sum_{i=1}^{N} r_i^+ f^*(X_i) \log \frac{\psi(a_k \cdot X_i - b_k)}{f^*(X_i)} + \sum_{i=1}^{N} r_i^- \bar{f}^*(X_i) \log \frac{\bar{\psi}(a_k \cdot X_i - b_k)}{\bar{f}^*(X_i)} \qquad (4.13)
\end{aligned}$$

This expression (4.13) is strictly concave in a and b, so it is readily maximized. The concavity of (4.13) in a_k and b_k follows from the concavity of $\log \psi(z)$ and $\log(1-\psi(z))$ and the positivity of the coefficients $r_i^+ f^*(X_i)$ and $r_i^- \bar{f}^*(X_i)$.

The heuristic is based on the fact that as long as we obtain a_k and b_k such that

$$\sum_{i=1}^{N} r_i^+ f^*(X_i) \log \frac{\psi(a_k \cdot X_i - b_k)}{f^*(X_i)} + \sum_{i=1}^{N} r_i^- \bar{f}^*(X_i) \log \frac{\bar{\psi}(a_k \cdot X_i - b_k)}{\bar{f}^*(X_i)} \geq 0 \tag{4.14}$$

holds, we do not reject them and do a restart, even though these may not be the maximizers for the k-step. Even if (4.14) turns out to be negative, we can check to see if (4.12) is positive. The maximizing values of a_k and b_k for (4.14) will not necessarily maximize (4.12). Nevertheless, each iterative fitting still results in the reduction of the overall mean squared error to the fit.

4.4 Schematic Representation of the Algorithm

Here we present a schematic representation of the algorithm. Recall that $\phi(z) = (\exp(z) - \exp(-z))/(\exp(z) + \exp(-z))$ and $\psi(z) = 1/(1 + \exp(-z))$. We have data $(X_i, Y_i)_{i=1}^N$, assumed to be sampled without statistical error. If some of the Y_i are negative, we assume that $Y_i = f(X_i)$ in $c\overline{\text{conv}}G$. We then scale the output so that it is transformed to $f^*(X_i)$ in $c\overline{\text{conv}}G_\psi$ and apply the algorithm to $(X_i, f^*(X_i))_{i=1}^N$. If Y_i are all positive, then we assume that $Y_i = f^*(X_i)$ in $c\overline{\text{conv}}G_\psi$. The algorithm below assumes that we are fitting the logistic sigmoids ψ iteratively to $(X_i, f^*(X_i))_{i=1}^N$.

From section 4.2, the following is what we do in principle.

0. $f_0^*(x_i) := 1$

1. Choose a_k and b_k to maximize

$$\frac{1}{N}\sum_{i=1}^{N}(f^*(X_i) - f_{k-1}^*(X_i))\psi(a_k \cdot X_i - b_k) \tag{4.15}$$

2. Verify that

$$\sum_{i=1}^{N}(f^*(X_i) - f_{k-1}^*(X_i))\psi(a_k \cdot X_i - b_k) \geq \frac{1}{c}\sum_{i=1}^{N}(f^*(X_i) - f_{k-1}^*(X_i))^2 \tag{4.16}$$

If not, redo Step 1.

3. Update outer weights in $f_k^* = \alpha_k f_{k-1}^* + \beta_k c\psi_k$ using least squares projection.

4. $k = k + 1$. Go back to Step 1.

However, as explained in section 4.3, the following is what is actually done.

1. Maximize the lower bound of (4.15), which is

$$\sum_{i=1}^{N} r_i^+ f^*(X_i) \log \frac{\psi(a_k \cdot X_i - b_k)}{f^*(X_i)} + \sum_{i=1}^{N} r_i^- \bar{f}^*(X_i) \log \frac{\bar{\psi}(a_k \cdot X_i - b_k)}{\bar{f}^*(X_i)} \qquad (4.17)$$

2. Is the maximum good enough ? That is, is it positive ?

   No: maximize (4.15) with random starting values (or initially with the maximizer of (4.17) as starting values) until a good enough value is obtained.

   Yes: use the maximizing values of a_k and b_k of (4.17) in step 1.

3. Update outer weights in $f_k^* = \alpha_k f_{k-1}^* + \beta_k c\psi_k$ using least squares projection.

4. $k = k + 1$, go back to Step 1.

4.5 Examples

Recall again that $\psi(z) = 1/(1 + \exp(-z))$.

Example 1. The target function is

$$f^*(x) = \frac{1}{3}\psi(20(-x_1 + x_2 + x_3 - x_4 + x_5 + x_6 - 0.5))$$

$$+\frac{1}{3}\psi(20(x_1+x_2+x_3+x_4-x_5-x_6-0.5))$$
$$+\frac{1}{3}\psi(20(x_1-x_2-x_3+x_4+x_5+x_6-0.5))$$

The sample consists of 500 points $(X_i, f^*(X_i))$ drawn independently and randomly from the uniform distribution over $[-1,1]^6$. Table 4.1 shows how the training error decreases and the number of restarts needed. The criterion at the k-step is the positivity of

$$\sum_{i=1}^{N}(f^*(X_i)-f^*_{k-1}(X_i))\phi(a_k\cdot X_i-b_k)\geq\sum_{i=1}^{N}(f^*(X_i)-f^*_{k-1}(X_i))^2,$$

that is (4.16) with $c=1$. For the first restart, the maximizer of (4.17) is used and substituted in (4.15) as a starting value for maximizing (4.15) directly. In case we are trapped in a local maxima, it does not matter as long as the criterion is satisfied. If the criterion is not satisfied, we repeat direct maximization of (4.15) using random restart values until we get end up with a convergence that satisfies the criterion.

Table 4.1

k	Error SS	Restarts	Criterion
0	20.69	–	–
1	13.87	0	0.29
2	11.53	1	2.20
3	10.16	2	0.53
4	8.61	1	2.39
5	6.33	1	7.59
6	4.22	1	8.13
7	4.22	1	0.95
8	3.25	1	5.58

The zeroth step is when the intercept term (mean) is fitted. Figure 4.1 shows how the error sum of squares (training error) decreases with the number of sigmoids fitted.

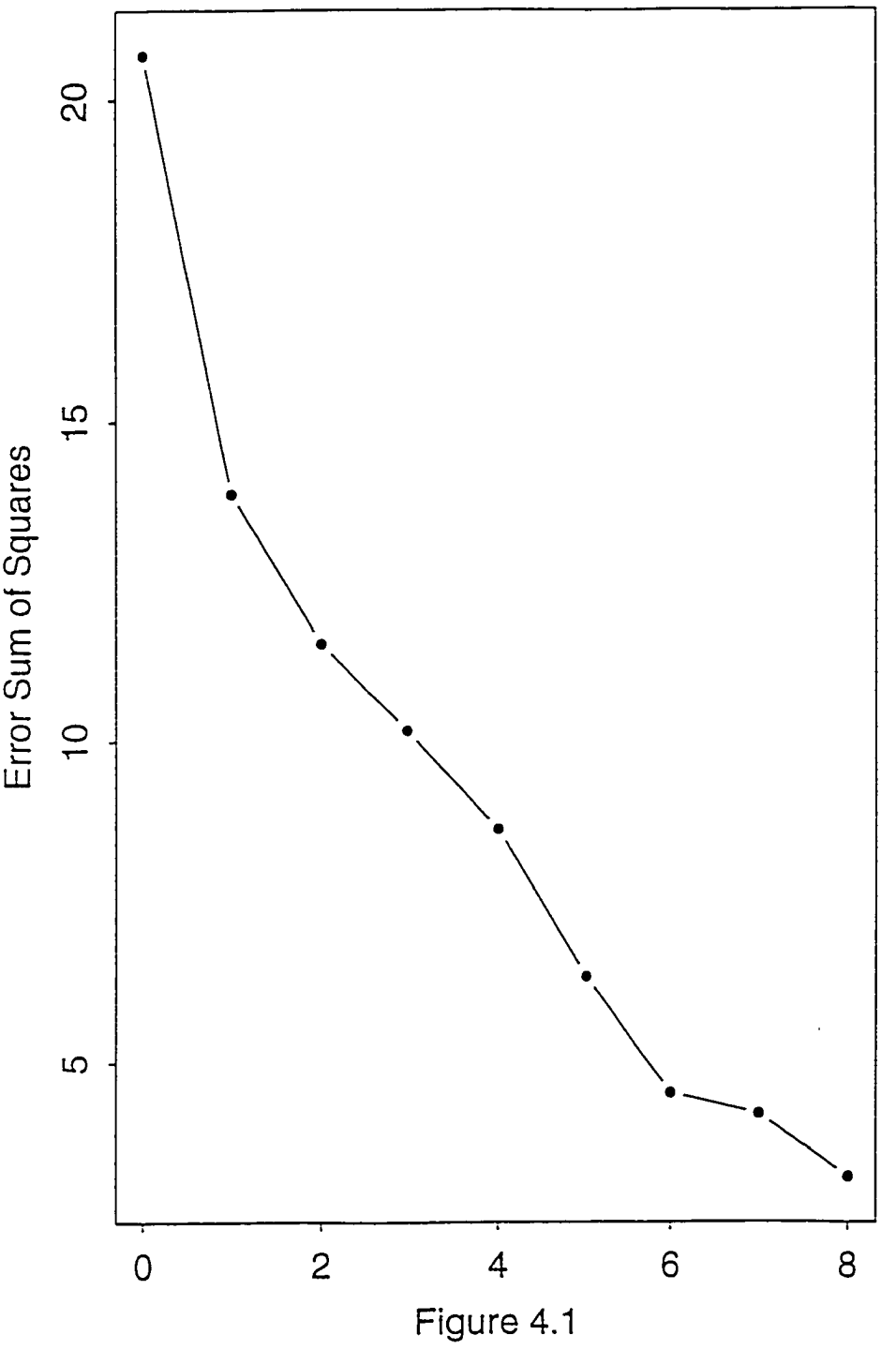

Figure 4.1

Example 4.2. The function is

$$\begin{aligned} f^*(x) &= \frac{1}{4}\psi(100(x_1 + x_2 - x_3 - x_4 + x_5 - x_6 + x_7 - 0.5)) \\ &= +\frac{1}{4}\psi(100(x_1 - x_2 + x_3 - x_4 + x_5 + x_6 - x_7 - 0.5)) \\ &= +\frac{1}{4}\psi(100(x_1 + x_2 + x_3 - x_4 - x_5 + x_6 - x_7 - 0.5)) \\ &= +\frac{1}{4}\psi(100(x_1 - x_2 + x_3 + x_4 + x_5 - x_6 + x_7 - 0.5)) \end{aligned}$$

The sample consists of 700 points $(X_i, f^*(X_i))$ drawn independently and randomly from the uniform distribution over $[-1, 1]^7$. Table 4.2 shows how the training error decreases and the number of restarts needed, with the restarts done in the same manner as in example 4.1.

Table 4.2

k	Error SS	Restarts	Criterion
0	40.76	–	–
1	17.27	0	9.52
2	15.05	4	0.89
3	13.04	1	4.40
4	11.21	2	1.48
5	9.87	1	3.86
6	8.93	2	2.03
7	7.93	2	3.69

The simulation in example 4.2 is done in two ways as indicated in Figure 4.2. One way is via the algorithm outlined above as for example 4.1 The training error at each step is shown by the solid line. In the second method, we take the maximizer of (4.17) regardless of whether the criterion is satisfied. The training error from this method is indicated by the broken line. In practice, the maximizer of (4.17) is vastly different from that of (4.15). Nevertheless, the second method, which has a training

error almost twice as large but reasonable for the amount of data, has the advantage that no restarts are required.

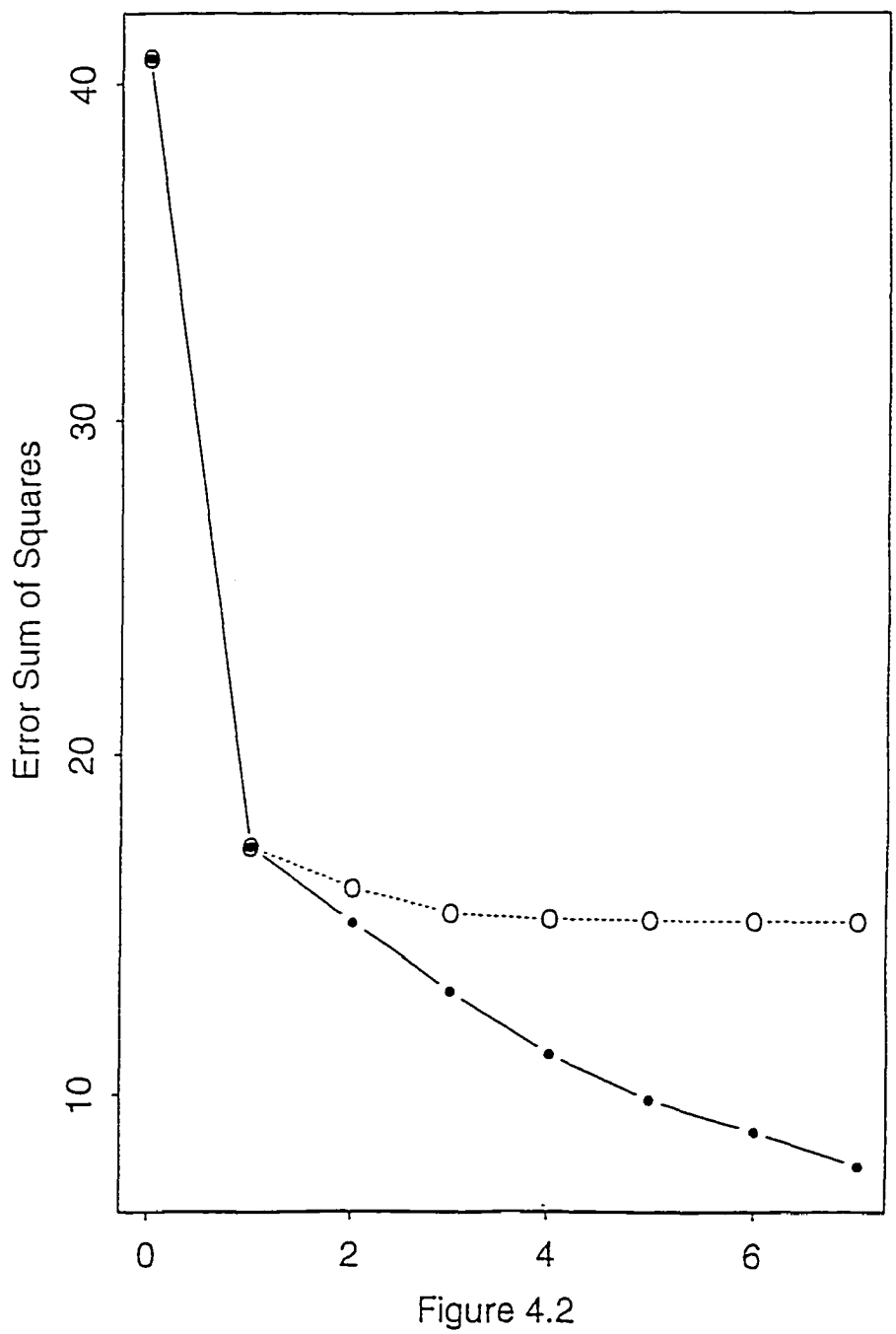

Figure 4.2

In both cases, the first term fitted is actually the intercept term. This is just the empirical mean of f^* given the data points $(X_1, X_2, \cdots, X_N)$. Example 4.1 is a linear combination of 3 sigmoids and example 4.2, of 4 sigmoids. The algorithm does not yield the exact true parameter values of the sigmoids. However, if the function is a small linear combination of sigmoids as it is in our two cases, each subsequent fitting reduces the error sum of squares quite dramatically for the first few fits. The sigmoids in the later fits tend to be close to the sigmoids fitted in the beginning. The first sigmoid fitted accomodates the combined effects of the sigmoids that sum up to the actual function. Subsequent fitted sigmoids either align themselves close to the true sigmoids in the given functions or seek to annul the effects of errors from earlier fits.

Chapter 5

Conclusion and Further Research Problems

In this chapter, we conclude by looking at possible ways of extending our results in the previous chapters.

5.1 Approximation Bounds

Theoretically, there is no reason why it is not possible to derive approximation bounds for approximation of functions with three or more hidden layer neural nets. By considering the target function to be in a class of function compositions of functions approximable by single layer neural nets, we can build up functions that can be approximated accurately with a k hidden layer neural network. However, we question

the utility of this exercise, since Kolmogorov's [33] result (see section 2.4) suggests that at most two hidden layers would suffice for approximating a continuous function. It would be better instead to concentrate on better methods of approximation with two hidden layer networks instead.

It would be desirable to extend the approximation results of Theorems 2.1 and 2.3 to that of a smooth closed convex set (with some conditions on the smoothness of the set). That is, we would desire to give a neural net approximation to the indicator of more general convex sets than the balls and ellipsoids. One way to do this is perhaps to consider sets of the form $\mathcal{D} = \{x \in \mathcal{R}^d : f(x) \cdot f(x) \leq 1\}$, where $f : \mathcal{R}^d \to \mathcal{R}^d$ has a strictly positive definite derivative, or even $\mathcal{D} = \{x \in \mathcal{R}^d : f(x) \leq 1\}$, where $f : \mathcal{R}^d \to \mathcal{R}$ and f is a convex function. It is expected that a better understanding of differential geometry would be needed for these problems.

5.2 Lower Bounds

It would be desirable to have some examples of functions that can be approximated well with two hidden layer neural nets but not with only single hidden layer nets. For example, consider the indicator of a unit cube in $\mathcal{R}^d$ which is enclosed in some bounded bigger space. The cube has $2d$ faces. If we choose the sigmoids (indicators of half-spaces) for the inner layer such that each sigmoid is aligned to each face and is $+1$ on the inside, 0 on the outside, then the sum of inner layer sigmoids $\sum_{j=1}^{2d} \phi(a_j \cdot x - b_j)$

is $2d$ on the inside of the cube and less than $2d$ on the outside. By thresholding the inner layer at $2d$, we obtain an exact representation to the cube

$$1_{\text{Cube}}(x) = \phi(\sum_{j=1}^{2d} \phi(a_j \cdot x - b_j) - 2d). \tag{5.1}$$

Thus a cube can be represented as a two hidden layer neural net with no approximation error. However it is not known whether a linear combination of sigmoids (single layer neural net) is able to approximate a cube with small error. We anticipate that many nodes, say of order $(\frac{1}{\epsilon})^d$, would be needed to obtain accuracy ϵ.

The second example is a tensor product of cosines

$$f(x) = \prod_{i=1}^{d} \cos(\pi x_i), \tag{5.2}$$

where $x \in [-1, 1]^d$. The choice of the constant multiplier is such that $||f|| = 1$, where $||\cdot||$ is the L_2 norm with respect to the uniform probablity measure over $[-1, 1]^d$. An attempt was made to obtain a lower bound for the approximation error between f and its single hidden layer neural net approximation. The idea is that since many terms (exponential in d) are required for the Fourier expansion of f, then perhaps just as many sigmoidal terms would be needed for the single layer neural net approximation in order to obtain an accurate approximation.

We note by induction and by using basic trigonometric identities that (5.2) can be expressed as

$$f(x) = \prod_{i=1}^{d} \sqrt{2}\cos(\pi x_i)$$

$$= \sum_{k\in\{1\}\times\{-1,1\}^d} \cos(\pi k\cdot x)$$
$$= 2^{-\frac{d-1}{2}} \sum_{k\in\{1\}\times\{-1,1\}^{d-1}} \sqrt{2}\cos(\pi k\cdot x). \quad (5.3)$$

The right hand side of (5.3) is the Fourier expansion of (5.2) with ridge trigonometric functions. Since all terms in the summand (5.3) have equal weights and are orthonormal, a best T term approximation f_T would be any of the T terms in the sum in (5.3), when $T < 2^{d-1}$. that is

$$f_T(x) = 2^{-\frac{d-1}{2}} \sum_{\mathcal{K}} \sqrt{2}\cos(\pi k\cdot x), \quad (5.4)$$

where $\mathcal{K}$ is any subset of $\{1\}\times\{-1,1\}^{d-1}$ of size T. The squared approximation error is then

$$||f - f_T||^2 = (1 - \frac{T}{2^{d-1}}). \quad (5.5)$$

T needs to be greater than $2^{d-1}(1-\delta)$ for a small squared approximation error less than δ. This implies that many ridge cosines (of order exponential in d) are needed to approximate well a tensor product of cosines.

We examine the approximation of $f(x) = \prod_{i=1}^d \sqrt{2}\cos(\pi x_i)$ by a single hidden layer net. Let a K-term approximation be

$$f_K(x) = \sum_{i=1}^{K} c_i\phi(a_i\cdot x - b_i), \quad (5.6)$$

where $|c_i| \leq C$. Without loss of generality, the parameter values of a_i and b_i can be restricted to $(a_i, b_i) \in \mathcal{Z}^{d+1}$, since $\phi(z) = \phi(\kappa z)$ for any $\kappa > 0$. This is to avoid problems with fractional frequencies when we approximate the sigmoids in (5.6) with ridge trigonometric functions.

We now take the Fourier expansion of each hard-limiter sigmoid in (5.6). Taking the first KT terms with the largest absolute-value co-efficients, we obtain

$$f^*_{KT}(x) = \sum_{i=1}^{KT} c'_i \cos(a'_i \cdot x - b'_i), \tag{5.7}$$

where $(a'_i, b'_i) \in \pi \mathcal{Z}^{d+1}$. Now let $f_{KT}(x)$ be the best KT term approximation using terms only in $\{1\} \times \{-1, 1\}^{d-1}$ as in (5.4). Then

$$\begin{aligned} ||f - f^*_{KT}||^2 &\geq ||f - f_{KT}||^2 \\ &= (1 - \tfrac{KT}{2^{d-1}}) \end{aligned} \tag{5.8}$$

from (5.5), since $f_{KT}(x)$ is the best KT term approximation for $f(x)$ using cosines in (5.3) and since $f \in \text{Span}\{\cos(\pi k \cdot x) | k \in \{-1, 1\}^d\}$, projecting the the terms in f^*_{KT} that are not in the subspace spanned by cosines in (5.3) onto the null vector improves the fit.

There is a need to examine the approximation rate for approximating the sigmoids in (5.7) with trigonometric functions. At present we are not able to determine if the parameters a'_i and a_i are aligned with each axis since f is a tensor product of cosines aligned along each axis. If this were the case, this would greatly simplify the calculation of the upper bound to the approximation error $||f_K - f^*_{KT}||$. Then we would be able to conclude by noting that a lower bound to the squared approximation rate of $f_K(x)$ to $f(x)$ is

$$||f - f_K||^2 \geq \frac{1}{2}||f - f^*_{KT}||^2 - ||f_K - f^*_{KT}||^2 \tag{5.9}$$

$$\geq \frac{1}{2}(1 - \frac{KT}{2^{d-1}}) - ||f_K - f^*_{KT}||^2 \tag{5.10}$$

By optimizing over T in (5.17), we would like to obtain an approximation rate in (5.17) of the form $C\left(1-\frac{K}{2^{\gamma(d-1)}}\right)$, where K has to be exponential in d in order for the rate to be independant of d. This would then be an indication that large numbers of nodes are needed for the single layer neural net approximation of f.

5.3 Heuristic Algorithm

A greedy algorithm (even if heuristic) for iteratively fitting the nodes of two hidden layer neural networks would be highly interesting from a computational point of view. For a start, we could consider target functions (when normalized) which are in the closed convex hull of indicators of ellipsoids. From chapter 2, we see how two hidden layer approximation can be split up first into approximation by sums of indicators of ellipsoids and then by a further approximation of the indicators of ellipsoids with thresholds of single hidden layer nets. The indicator of an ellipsoid $1_E(x)$ can be written as

$$1_E(x) = 1_{\{x\in R^d:\sum_{k<l}\eta_{kl}x_kx_l-k'\leq 1\}}.$$

In fitting the outer layer ellipsoids, it might be possibly to apply the algorithm to such functions and use $\psi(1-\sum_{k<l}\eta_{kl}x_kx_l+k')$ in place of $\psi(a\cdot x-b)$ as in chapter 4. We take the target function as f^*. Here ψ is the logistic sigmoid as in chapter 4. The parameter values obtained from the algorithm can then be substituted into $1_E(x)$. Since $\psi(\kappa z)$ converges to the heaviside function as $\kappa\to\infty$, we could possibly

enhance the heuristic algorithm by fitting $\psi(\kappa(1-\sum_{k<l}\eta_{kl}x_kx_l+k'))$ for some large κ. What is needed now is an iterative algorithm to fit the inner layer. Parameter values thus obtained can be used to initialize subsequent searches, for example, by gradient descent (back propagation algorithm) or perturbation methods, for finer adjustments to fit the data.

5.4 Conclusion

We summarize the results in this work. Bounds for two layer neural net approximation are obtained for functions that have variation with respect to indicators of balls and ellipsoids. These indicate that such functions can be approximated well with two layer neural nets. The approximation bounds are used in the calculation of the overall estimation error. Two techniques are used in obtaining the estimation bounds. In one case, a minimum complexity estimator is used. In the other case, a general theorem bounding the risk using penalized least squares estimator is derived. The risk is bounded under entropy conditions on the component models of the class of candidate models from which the estimator is chosen. This theorem is then applied to neural net estimation. Finally, a heuristic algorithm for fitting single hidden layer nets iteratively to a class of target functions is also given. Functions in this class (when normalized) lie in the closed convex hull of sigmoids. Some simulation results are presented.

Bibliography

[1] A. R. Barron, *Complexity regularization with applications to artificial neural networks*, Nonparametric functional estimation (ed. G. Rousas), Kluwer Academic Publishers, Dordrecht, the Netherlands and Boston, MA. (1991) 561 – 576

[2] A. R. Barron, *Neural net approximation*, Proc of the 7th Yale workshop on adaptive and learning systems. (1992).

[3] A. R. Barron. *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Info. Thy. 39 (1993) 930 – 944

[4] A. R. Barron, *Approximation and estimation bounds for artificial neural networks*, Machine Learning 14 (1994) 113 – 143

[5] A. R. Barron and R. L. Barron, *Statistical learning networks : a unifying view*, Computing science and statistics : a 20th symposium on the interface (ed. E. J. Wegman, D. T. Gantz and J. J. Miller), American Statistical Association; Fairfax, VA. (1988) 192 – 203

[6] A. R. Barron, L. Birgé and P. Massart, *Risk bounds for model selection via penalization*, to appear in Prob. Thy. Rel. Fields

[7] A. R. Barron and G. H. L. Cheang, *Greedy Algorithms for Single Layer Network Training*. Invited talk, WCNN '95, Washington, D.C.

[8] E. R. Baum and D. Haussler, *What size net gives valid generalization ?* Neural Comp. **1** (1989) 151 – 160

[9] C. M. Bishop, *Neural networks for pattern recognition*, Clarendon Press, Oxford. 1995

[10] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and regression trees*. Wadsworth. Belmont, CA . 1984

[11] W. L. Buntine and A. S. Weigend, *Bayesian backpropagation*, Complex Systems **5** (1991) 603 – 643

[12] G. H. L. Cheang, *Neural network approximation and estimation of functions*, Proc. of the 1994 IEEE–IMS Workshop on Info. Theory and Stat, 1994

[13] B. Cheng and D. M. Titterington *Neural networks : a review from a historical perspective*, Statistical Science 9 (1994) 2 – 30

[14] T. M. Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*, IEEE Trans. Elec. Comp. **14** (1965) 326 – 334

[15] T. M. Cover, *Capacity problems for linear machines*, Pattern recognition (ed. L. Kanal), Thompson Book Co. Washington (1968) 283 – 289

[16] G. Cybenko, *Approximations by superpositions of a sigmoidal function*, Mathematics of Control, Signals, and Systems **2** (1989) 303 – 314

[17] R. Dudley, *Metric entropy of some classes of sets with differentiable boundaries*, J. Approximation Theory **10** (1974) 227–236, Correction, J. Approximation Theory **26** (1979) 192–193

[18] L. Fejes Tóth. *Lagerungen in der Ebene, auf der Kugel und im Raum*, Springer Verlag, Berlin 1953, 1972

[19] J. H. Friedman, *Multivariate adaptive regression splines*, Ann. Stats. **19** (1991) 1 – 68

[20] J. H. Friedman and W. Stuetzle. *Projection pursuit regression*, Jour. Amer. Stats. Assoc. **20** (1981) 817 – 823

[21] P. Goodey and W. Weil, *Zonoids and generalisations*, Handbook of convex geometry, (eds. P. M. Gruber, J. M. Wills), Elsevier, Amsterdam. (1993) 1297 – 1326

[22] P. M. Gruber,*Approximation of convex bodies*, Convexity and its appplications, (eds. P. M. Gruber, J. M. Wills), Birkhäuser, Basel (1983) 131–162

[23] P. M. Gruber and P. Kenderov,*Approximation of convex bodies by polytopes*, Rend. Circolo Mat. Palermo **31** (1982) 195–225

[24] P. Hall, *On convergence rates in nonparametric problems*, International Statistical Review (1989) **57** 45 – 58

[25] W. Härdle, *Applied nonparametric regression*, Cambridge University Press, Cambridge, U.K. 1990

[26] D. Haussler, *Decision theoretic generalizations of the PAC model for neural net and other learning applications*, Inform. Comp. **100(1)** (1992) 78–150

[27] S. S. Haykin, *Neural networks : a comprehensive foundation*, Macmillan, New York 1994

[28] J. J. Hopfield. *Neural networks and physical systems with emergent collective computational abilities*, Proc. Nat. Acad. Sci. **79** (1982) 2554 – 2558

[29] K. Hornik, M. Stinchcombe and H. White, *Multi-layer feed forward networks are universal approximators*, Neural Networks **2** (1988) 359 – 366

[30] P. Huber. *Projection pursuit*, Ann. Stats. **13** (1985) 435 – 474

[31] I. A. Ibragimov and R. Z. Hasminskii, *On nonparametric estimation of regression*, Dokl. Akad. Nauk. SSSR. **252** (1980) 780 – 784

[32] L. K. Jones, *A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training*, Ann. Stats. **20** (1990) 608 – 613

[33] A. N. Kolmogorov. *On the representation of continuous functions of several variables by superpositions of continuous functions of one variable and addition*, Dokl. Akad. Nauk. SSSR. **114** (1957) 679 – 681

[34] V. Kůrková, *Kolmogorov's theorem and multilayer neural networks*, Neural Networks **5** (1992) 501–506

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

[36] K. C. Li, *Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation : discrete index set*, Ann. Stat. **15** (1987) 958 – 975

[37] G. G. Lorentz. *Approximation of Functions*, Holt, Rinehart and Winston.

[38] Y. Makovoz, *Random approximants and neural networks*, J. Approx. Thy. **85** (1996) 98 – 109

[39] D. F. McCaffrey and A. R. Gallant, *Convergence rates for single hidden layer feedforward networks*. Rand Corporation working paper, Santa Monica, CA and Institute of Statistics Mimeograph Series 2207, North Carolina State University. (1991)

[40] W. S. McCullough and W. Pitts, *A logical calculus of ideas immanent in nervous activity*, Bull. Maths. Biophysics **5** (1943) 115 – 133

[41] C. Müller, *Spherical harmonics,* Lecture notes in mathematics. Springer–Verlag. New York **17** (1966)

[42] H.-G. Müller and U. Stadtmüller, *Variable bandwidth kernel estimators of regression curves.* Ann. Stat. **15** (1987) 182 – 201

[43] M. Nussbaum, *On nonparametric estimation of a regression function that is smooth in a domain on* $\mathcal{R}^k$, Theory of Probability and its Applications **31** (1986) 118 – 125

[44] G. Pisier, *Remarques sur un résultat non publié de B. Maurey,* Séminaire d'analyse fonctionelle 1980 - 1981, École Polytechnique, Centre de Mathématiques. Palaiseau.

[45] D. Pollard, *Convergence of stochastic processes*, Springer Verlag, Berlin 1984

[46] D. Pollard, *Empirical Processes : theory and applications,* NSF-CBMS Regional Conference Series in Prob. and Stats. **2** (1990)

[47] B. T. Polyak and A. B. Tsybakov, *Asymptotic optimality of the* C_p *test for the orthogonal series estimation of regression*, Thy. Prob. App. **35** (1990) 293 – 306

[48] B. D. Ripley, *Neural networks and related methods for classification*, Jour. Roy. Stat. Soc. B **56** (1994) 409 – 433

[49] F. Rosenblatt. *The perceptron : a probabilistic model for information storage and organization in the brain*, Psych. Rev. **65** (1958) 386–408

[50] F. Rosenblatt. *Principles of neurodynamics.* Spartan books, Washington, D.C. 1962

[51] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning internal representation by error propagation*, Parallel distributed processing : explorations in the microstructure of cognition (ed. D. E. Rumelhart, J. L. McClelland and the PDP research group), MIT Press. 1986

[52] R. Schneider, *Über eine Integralgleichung in der Theorie der konvexen Körper*, Math. Nachr. 44 (1970) 55 – 75

[53] R. Schneider and J.A. Wieacker, *Approximation of convex bodies by polytopes*, Bull. London Math. Soc. **13** (1981) 149–156

[54] W. R. Schucany. *Adaptive bandwidth choice for kernel regression*, Jour. Am. Stat. Assoc. **90** (1995) 535 – 540

[55] S. R. Searle, *Linear Models*, Wiley. New York. 1971

[56] G. A. F. Seber and C. M. Wild, *Nonlinear Regression*, Wiley, New York. 1989

[57] R. Shibata, *An optimal selection of regression variables*, Biometrika **68** (1981) 45 – 54

[58] C. J. Stone, *The use of polynomial splines and their tensor products in multivariate function estimation*, Ann. Stat. **22** (1994) 118 – 184

[59] V. N. Vapnik and A. Ya. Červonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory Prob. Appl. **16** No. 2 (1971) 264–280

[60] Y. Yang and A. R. Barron, *An asymptotic property of model selection criteria*, IEEE Trans. Info. Thy. 44 (1998) 95 – 116

[61] J. E. Yukich, M. B. Stinchcombe and H. White, *Sup-norm approximation bounds for networks through probabilistic methods*, IEEE Tran. Info. Theory **41** (1995) 1021–1027

IMAGE EVALUATION TEST TARGET (QA-3)

150mm

6"

APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989