Error Bounds for Functional Approximation and Estimation Using Mixtures of Experts

Assaf Zeevi, Ron Meir, and Vitaly Maiorov

This research was partially supported by a grant from the Israel Science Foundation.
Support from the Ollendorff center of the Department of EE at the Technion is also acknowledged.
Assaf Zeevi is with the Information Systems Lab, Stanford University, CA 94305-4055
Ron Meir is with Department of Electrical Engineering, Technion, Haifa 32000, Israel
Vitaly Maiorov is with the Department of Mathematics, Technion, Haifa 32000, Israel

Abstract

We examine some mathematical aspects of learning unknown mappings with the Mixture of Experts Model (MEM). Specifically, we observe that the MEM is at least as powerful as a class of neural networks, in a sense that will be made precise. Upper bounds on the approximation error are established for a wide class of target functions. The general theorem states that $\|f - f_n\|_p \leq c/n^{r/d}$ for $f \in W_p^r(L)$ (a Sobolev class over $[-1,1]^d$), and f_n belongs to an n-dimensional manifold of normalized ridge functions. The same bound holds for the MEM as a special case of the above. The stochastic error, in the context of learning from i.i.d. examples, is also examined. An asymptotic analysis establishes the limiting behavior of this error, in terms of certain pseudo-information matrices. These results substantiate the intuition behind the MEM, and motivate applications.

Keywords

Mixture of Experts, Estimation Error, Approximation Error.

I. Introduction

For several years now, neural network models have enjoyed wide popularity, being applied to problems of regression, classification and time series analysis. The theoretical aspects of these models have been studied by several researchers [3], [5], [13], [22], [24], to name but a few. These results substantiated the, already widespread, use of these models in many application.

Although neural networks are universal function approximators [5],[13],[24], and statistical aspects related to learning are well understood by now [3],[6],[22] the practitioner is still faced with quite a few problems. Perhaps one of the main concerns is understanding the structure and the parameterization of the model. Ultimately, one would like to deduce conclusive statements on the data structure, by inspection and analysis of the actual performance and application results (i.e., residual error on the training set, and prediction results).

Recently, a novel non-linear model, termed the Mixture of Experts Model (MEM) was introduced by Jacobs *et al.* [10]. The idea underlying this model is to combine several local estimators, or experts, each 'specializing' in some region of the input space. The framework of this model originates in the field of Statistics. More specifically it is an adaptation of standard mixture models, a field of study which is applied to problems of density estimation, pattern classification and clustering [19].

The MEM architecture is composed of n expert networks, each of which solves a function approximation problem over a local region of the input space. A stochastic model, that relates input vectors $\mathbf{x} \in \mathbb{R}^d$ to output vectors $\mathbf{y} \in \mathbb{R}^s$, is associated with each expert. We denote the conditional probability model of each expert as follows $p(\mathbf{y}\,|\mathbf{x};\boldsymbol{\theta}_j)$ with $j = 1, 2, \ldots, n$, where the $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}$ are parameter vectors associated with each expert. Typically, these densities are chosen from the exponential family. Thus, the overall stochastic model assumes the form of a mixture density

$$p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{n} g_j(\mathbf{x};\boldsymbol{\theta}_g)p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_j). \tag{1}$$

Each expert network produces as output a vector $\boldsymbol{\mu}_j$ where

$$\boldsymbol{\mu}_j = \psi(\mathbf{x};\boldsymbol{\theta}_j) \qquad j = 1, 2, \ldots, n$$

that is $\psi : \mathbb{R}^d \times \mathbf{\Theta} \longrightarrow \mathbb{R}^s$. The function ψ may be a simple linear transformation, or a more general non-linear mapping. In most formulations of this architecture, the function ψ was taken to be linear in the parameters, a structure which is better suited to the learning algorithm. An additional requirement is that $\boldsymbol{\mu}_j$ be the conditional expectation taken w.r.t the underlying jth component density in the mixture, i.e., $\boldsymbol{\mu}_j = \mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_j]$. Although more restrictive, this imposition allows a more natural interpretation of the output, viewed as a mixture of regressors.

The MEM also utilizes an auxiliary network, termed the *gating network*, whose objective is to partition the input space into regions, corresponding to the various experts. This task is assumed by assigning a probability vector $[\alpha_1, \alpha_2, \ldots, \alpha_n]^T$ to each point in the input space. The implementation is by a multiple output extension of the logistic regression model, or multinomial logit device, defined as follows

$$\alpha_j = g(\mathbf{x}, \boldsymbol{\theta}_g) \stackrel{\triangle}{=} \frac{\exp\{s_j\}}{\sum_{i=1}^{n} \exp\{s_i\}} \qquad j = 1, 2, \ldots, n \tag{2}$$

where $s_j : \mathbb{R}^d \times \mathbf{\Theta}^g \longrightarrow \mathbb{R}$, and is typically taken to be a linear mapping $s_j = \boldsymbol{\theta}_{g_j}{}^T \mathbf{x} + \theta_{g_{j,0}}$. Note that by definition of $g(\cdot)$, we have $\sum_{j=1}^{n} \alpha_j = 1$ for all $\mathbf{x}$.

There are several advantages associated with the probabilistic formulation of the model, one of the most important being the availability of an efficient learning algorithm. Jordan and Jacobs [11] demonstrated the applicability of the Expectation - Maximization (EM) algorithm to the learning phase. This optimization technique is extremely well suited to mixture model estimation problems, by breaking down the global optimization into several re-estimation equations. These equations are in many cases insightful, driven by the intrinsic properties of the mixture model. In many cases, this yields substantially simpler and more straightforward estimation than gradient methods, and more robust behavior than second order algorithms [16]. The method is also much less intensive computationally than gradient descent, making the MEM an attractive candidate, and contender to neural network models, where gradient descent has been the popular optimization technique.

In the sequel we will be mainly concerned with the model class $\mathcal{H}_n$ (defined formally in (7))

$$f_n(\mathbf{x}; \boldsymbol{\theta}) = \mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}; \mathcal{H}_n] = \sum_{j=1}^{n} g_j(\mathbf{x}; \boldsymbol{\theta}_g) \psi(\mathbf{x}; \boldsymbol{\theta}_j) \tag{3}$$

that is, the parametric mapping induced by taking the conditional expectation w.r.t. the conditional density in (1). As the choice of $\psi(\cdot)$ is arbitrary, we will restrict attention to the simple case where $\psi(\mathbf{x}; \boldsymbol{\theta}_j) = \boldsymbol{\theta}_j^T \mathbf{x} + \theta_{j,0}$, following the original formulation of Jacobs *et al.* [10]. Thus, we propose to model the 'true' regression function by a parametric function from the class $\mathcal{H}_n$, and therefore one may view $f_n \in \mathcal{H}_n$ as a functional parametric estimator.

The main concern of this paper is to study some theoretical properties of the MEM model. In particular, we will focus on two question: One, given a target function f in a prescribed function class $\mathcal{F}$, can we approximate it to arbitrary accuracy using function $f_n \in \mathcal{H}_n$. In fact, we will be interested in a somewhat sharper answer, that is, how large should n be (alternatively, how many experts should one choose) so as to have a prescribed accuracy level in the approximation. A second question relates to the statistical properties

of the estimation error, in learning the function f from a given sample set. The following setting embodies both these questions. We concentrate on the problem of least squares estimation (as opposed to the estimation procedures of the stochastic model (1) via maximum-likelihood and the EM algorithm). The focus is on learning some unknown function, belonging to a certain class, by means of parametric models taken from the model class $\mathcal{H}_n$. The *target* function in this setting is the 'true' underlying regression function associated with some observable, noisy, input-output process.

A typical result we obtain, is that the MEM is capable of approximating any target function in a certain Sobolev class. Bounds on the approximation error are established, demonstrating that

$$\|f - f_n\|_p \leq \frac{c}{n^{r/d}} \qquad 1 \leq p \leq \infty$$

where f is the target function. In general $f : \mathbb{R}^d \longrightarrow \mathbb{R}^s$, and we concentrate on the case $s = 1$ for simplicity. The function f_n is given in (3), c is an absolute constant, d is the dimension of the input space and r is the number of continuous derivatives in L_p we assume f to possess. In this formulation, f is the 'true' regression function, associated with the observable stochastic process, and f_n is the model used to approximate it. This statement follows from a general result (Theorem 1), concerning the degree of approximation characteristics of a class of linear combinations of normalized ridge functions. A recent paper by Mhaskar [14], points out that this bound is of optimal order under further conditions. We do not make any such claim in the setting we analyze herein, though we will briefly digress to discuss this point following the presentation of the main results.

The asymptotic estimation error is determined as well, and an upper bound is thus established by combining the two error terms. We note in passing that the estimation is generally assumed to be in a *misspecified* framework [21], that is we conceive that the model we have f_n, differs from the 'true' regression function f, associated with the data generating mechanism. Moreover, we do not assume $f \in \mathcal{H}_n$ for any n. Finally, we note that the estimation error is analyzed under asymptotic assumptions, and therefor we must be careful in interpreting these results, in particular when only small sample sets are available.

Based on the derived upper bounds, a model selection criterion is introduced, inspired by the method of structural risk minimization [20]. This method has recently been pursued in the context of neural networks by Murata and Amari [15], based on Amari *et al.*'s work on learning curves [3], and has been termed by these authors as NIC - Network Information Criterion. The question of how to determine the number of experts, best suited to solve a given problem (available in the form of a sample set), can be similarly addressed in a systematic manner.

The remainder of the paper is organized as follows. Section II is devoted to some preliminaries and definitions which are essential to the statement of the main results. In Section III we present the main theorems, concerning the degree of approximation results and the estimation error. In Section IV, we introduce a model selection criterion based on the results of these theorems. Finally, we discuss the results and some open problems. Technical proofs are relegated to the Appendix for continuity of ideas.

II. Definitions, Notation and Problem Statement

Let $(\mathbf{X}, Y)$ be random variables, defined over an underlying probability space $(E, \mathcal{E}, P)$, such that $(\mathbf{X}, Y) : E \to I^d \times \mathbb{R}$, with $I^d \equiv [-1, 1]^d$. Let P be chosen so that $\mathbb{E}Y^2 < \infty$. The following induced probability measures on I^d and $\mathbb{R}$ may then be defined: $\nu(A) = P(\mathbf{X} \in A)$ for all $A \in \mathcal{B}(I^d)$ (the Borel σ-field on I^d), and $\eta_x(B) = P(Y \in B | \mathbf{X} = \mathbf{x})$ for all $\mathbf{x} \in I^d$ and $B \in \mathcal{B}(\mathbb{R})$. We will be concerned with the following problem. Given a random sample set $\mathcal{D}_N = \{\mathbf{X}_t, Y_t\}_{t=1}^N$, consisting of N i.i.d. copies of $(\mathbf{X}, Y)$, our objective is to come up with the 'best' possible estimate of the regression function $f(\mathbf{x}) \equiv \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Here f is a deterministic unknown mapping $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, in some prescribed class of functions $\mathcal{F}$. In view of the above definitions, one may view the sample set as being generated by the relation

$$Y_t = f(\mathbf{X}_t) + \varepsilon_t, \qquad t = 1, 2, \ldots, N, \tag{4}$$

where $\{\varepsilon_t\}$, is a zero mean, finite variance, noise process. Since the focus will be on the regression function f, one may view the noise process as the residual randomness $\varepsilon_t = Y_t - f(\mathbf{X}_t)$.

We attempt to reconstruct f, over I^d, using estimators from the MEM parametric family. A complexity index n is assigned to the MEM estimators, referring to the number of experts in the architecture. Throughout this paper we use upper case letters to denote random variables and correspondingly lower case letters to denote realizations. Boldface type will be used to denote vector valued quantities.

We note in passing that the restriction on $\mathbf{x}$ can be to any compact domain $K \subset \mathbb{R}^d$. The selection of $K = I^d$ has been chosen in order to simplify the mathematical analysis, and make it more transparent. However, the fact that the support of $\mathbf{x}$ is compact is crucial to the proof techniques. As for the commonly used i.i.d. assumption, we note that it may be replaced with much weaker assumptions, provided the uniform strong law of large numbers and certain formulations of the central limit theorem still hold. For instance, in the case of correlated data as in time series we will typically assume jointly stationary ergodic vectors with appropriate mixing conditions (see [26] for details in the context of the MEM class and time series prediction).

Define the L_p norm over I^d as follows

$$\|f - g\|_p \stackrel{\triangle}{=} \left(\int_{I^d} (f - g)^p d\lambda \right)^{1/p} \qquad 1 \leq p < \infty$$

and the L_∞ norm as

$$\|f - g\|_\infty \stackrel{\triangle}{=} \operatorname*{ess\,sup}_{\mathbf{x} \in I^d} |f(\mathbf{x})| \quad ,$$

where λ denotes the Lebesgue measure on $\mathbb{R}^d$. Let $L_p(I^d, \lambda)$ denote the vector space of measurable functions f which have $\|f\|_p < \infty$, where we identify $f = g$ if the functions are equal a.e.-λ.

Define the following *risk function*, w.r.t. the squared loss,

$$L_{\nu,\eta}(\mathcal{F}, f_n) \stackrel{\triangle}{=} \int_{\mathbb{R}} \int_{I^d} [y(\mathbf{x}) - f_n(\mathbf{x}; \boldsymbol{\theta})]^2 \eta(dy|\mathbf{x}) \nu(d\mathbf{x}) \quad , \tag{5}$$

where $\mathcal{F}$ is the class of target functions, and the dependence of y on $\mathbf{x}$ has been made explicit. Define the *empirical risk* function

$$l(\mathcal{D}_N, f_n) \stackrel{\triangle}{=} \frac{1}{N}\sum_{t=1}^{N}[Y_t - f_n(\mathbf{X}_t;\boldsymbol{\theta})]^2 \tag{6}$$

which follows from taking the integration in (5) w.r.t. the empirical distribution $\mu_N(\mathbf{x}, y) \equiv 1/N \sum_{t=1}^{N} \mathbf{1}(\mathbf{x}-\mathbf{x}_t, y-y_t)$.

The purpose of learning is to find a function f_n^* that minimizes (5), w.r.t. a class of estimators defining f_n. In this work we concentrate on the following class

$$\mathcal{H}_n = \left\{ f_n \mid f_n(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{n} g_j(\mathbf{x};\boldsymbol{\theta}_g)[\boldsymbol{\theta}_j^T\mathbf{x} + \theta_{j,0}],\ [\boldsymbol{\theta}_j^T, \theta_{j,0}]^T \in \boldsymbol{\Theta}_n,\ \boldsymbol{\theta}_g \in \boldsymbol{\Theta}_n^g \right\} \tag{7}$$

where $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_g^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_n^T, \theta_{1,0}, \ldots, \theta_{n,0}]^T$, and $\boldsymbol{\Theta}_n, \boldsymbol{\Theta}_n^g$ are compact subsets of $\mathbb{R}^{n(d+1)}$ defined as follows

$$\begin{aligned} \boldsymbol{\Theta}_n &= \left\{ (\boldsymbol{\theta}_j, \theta_j)_{j=1}^n \in \mathbb{R}^d \times \mathbb{R} \mid \|\boldsymbol{\theta}_j\|_\infty \le c_\theta,\ |\theta_j| \le 2Le^{3dn(1+o(1))},\ j = 1, 2, \ldots, n \right\} \\ \boldsymbol{\Theta}_n^g &= \left\{ (\boldsymbol{\theta}_{g_j}, \theta_{g_{j,0}})_{j=1}^n \mid \|\boldsymbol{\theta}_{g_j}\|_\infty \le 1,\ |\theta_{g_{j,0}}| \le 1,\ j = 1, 2, \ldots, n \right\} \quad . \end{aligned} \tag{8}$$

Here $c_\theta \in \mathbb{R}^+$ is arbitrary, and L is defined in Theorem 3 and Assumption 1. Note that the restrictions on $\boldsymbol{\theta}_g$ are a consequence of the conditions stated in Assumption 2, applied to the function $\sigma(t) = e^t$. We use the term $o(1)$, appearing in the exponent, to abbreviate terms whose growth is dominated by the term e^n. The explicit expressions for these terms appear in the Appendix. We note that the somewhat unusual bound on the size of the parameters θ_j arises because of certain technical conditions required to achieve the correct degree of approximation. This issue is expanded on in Remark 3 in Section III.

This class definition follows from the formulation of the MEM as in (3), where the gating network is implemented as a 'softmax' function as in [10]. That is

$$g_j(\mathbf{x};\boldsymbol{\theta}_g) \stackrel{\triangle}{=} \frac{\exp\{\boldsymbol{\theta}_{g_j}^T\mathbf{x} + \theta_{g_{j,0}}\}}{\sum_{i=1}^{n}\exp\{\boldsymbol{\theta}_{g_i}^T\mathbf{x} + \theta_{g_{i,0}}\}} \quad .$$

The vector of parameters $\boldsymbol{\theta}_g$ is composed of n sub-vectors $\boldsymbol{\theta}_{g_j}$; $j = 1, 2, \ldots, n$ and n constants $\theta_{g_{j,0}}$. The choice of 'softmax' functions is due to the inherent positivity and normalization, two properties imposed on the output of the gating network in [10] and [11].

Obviously, there is no hope in attempting to approximate any target function using this class of approximants, unless we restrict the target class by imposing some regularity conditions. The following assumption is useful in characterizing the target class.

Assumption 1: The target function f belongs to the Sobolev class, $f \in W_p^r(L)$

$$W_p^r(L) \stackrel{\triangle}{=} \left\{ h(\mathbf{x}) \mid \|h\|_{W_p^r} = \sum_{|\alpha| \le r} ||h^{(\alpha)}(\mathbf{x})||_p \le L \right\}$$

where

$$h^{(\alpha)} \equiv \frac{\partial^{\|\alpha\|_1} h}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}$$

and $\alpha \in \mathbf{Z}_+^d$ is a multi integer $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$.

Now, the objective of seeking f_n^* - the minimizer of the risk function (5) - is not feasible, since we are only able to define the empirical risk, based on the sample set $\mathcal{D}_N$. We thus minimize $l(\mathcal{D}_N, f_n)$, the empirical risk, and obtain a least squares estimator $\hat{f}_{n,N}$. Note that by the following decomposition we can define the risk function w.r.t. the regression function f.

$$\begin{aligned} L_{\nu,\eta}(\mathcal{F}, f_n) &= \mathbb{E}_{\nu,\eta}[Y - f_n(\mathbf{X};\boldsymbol{\theta})]^2 \\ &= \mathbb{E}_\eta\{Y - \mathbb{E}[Y|\mathbf{X}]\}^2 + \mathbb{E}_\nu\{\mathbb{E}[Y|\mathbf{X}] - f_n(\mathbf{X};\boldsymbol{\theta})\}^2 \\ &= \sigma^2 + \mathbb{E}_\nu\{f(\mathbf{X}) - f_n(\mathbf{X};\boldsymbol{\theta})\}^2 \\ &= \sigma^2 + L_\nu(\mathcal{F}, f_n). \end{aligned} \tag{9}$$

Here σ^2 is the variance of the zero mean additive noise $\{\varepsilon_t\}$. Obviously, minimizing (5) is equivalent to minimizing (9). The vector of parameters associated with the minimizer of (9), f_n^*, will be denoted $\boldsymbol{\theta}_n^*$

$$\boldsymbol{\theta}_n^* \stackrel{\triangle}{=} \arg\min_{\boldsymbol{\theta}\in\Omega_n} L_\nu(\mathcal{F}, f_n),$$

and the vector of least square estimates, derived from the minimization of the empirical risk function

$$\hat{\boldsymbol{\theta}}_{n,N} \stackrel{\triangle}{=} \arg\min_{\boldsymbol{\theta}\in\Omega_n} l(\mathcal{D}_N, f_n),$$

where $\Omega_n \equiv \boldsymbol{\Theta}_n \cup \boldsymbol{\Theta}_n^g$, with $\boldsymbol{\Theta}_n$ and $\boldsymbol{\Theta}_n^g$ are defined in (8). Plugging $\hat{\boldsymbol{\theta}}_{n,N}$ into f_n we obtain $\hat{f}_{n,N}$, the estimator of f, based on the sample set $\mathcal{D}_N$.

III. Main Results

Having defined the estimator $\hat{f}_{n,N}$, our objective is to assess its performance by examining the mean integrated squared error between f and $\hat{f}_{n,N}$. Denote the total error as $\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N})$, where

$$\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N}) \stackrel{\triangle}{=} \|f - \hat{f}_{n,N}\|^2_{L_2(I^d,\nu)} = \int_{I^d} [f(\mathbf{x}) - f(\mathbf{x};\hat{\boldsymbol{\theta}}_{n,N})]^2 \nu(d\mathbf{x}) \tag{10}$$

and define $\mathcal{L}(\boldsymbol{\theta}_n^*)$, the total error evaluated at the point $\boldsymbol{\theta} = \boldsymbol{\theta}_n^*$ as

$$\mathcal{L}(\boldsymbol{\theta}_n^*) \stackrel{\triangle}{=} \|f - f_n^*\|^2_{L_2(I^d,\nu)} = \int_{I^d} [f(\mathbf{x}) - f_n(\mathbf{x};\boldsymbol{\theta}_n^*)]^2 \nu(d\mathbf{x}). \tag{11}$$

We start the derivation of the main results by considering the following decomposition of the total error $\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N})$, by means of a second order stochastic Taylor series expansion around $\boldsymbol{\theta}_n^*$. Since $f_n(\mathbf{x};\boldsymbol{\theta})$ is clearly three times continuously differentiable w.r.t. $\boldsymbol{\theta}$, the expansion exists.

$$\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N}) = \underbrace{\mathcal{L}(\boldsymbol{\theta}_n^*)}_{\text{(i)}} + \underbrace{\nabla^T \mathcal{L}(\boldsymbol{\theta}_n^*)(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)}_{(*)} + \underbrace{\frac{1}{2}(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}_n^*)(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*) + r_{n,N}}_{\text{(ii)}}, \tag{12}$$

where the remainder term $r_{n,N}$ is given by

$$r_{n,N} = \sum_{i,j,k=1}^{2n(d+1)} \frac{1}{3!} \frac{\partial^3 \mathcal{L}(\tilde{\boldsymbol{\theta}}_n)}{\partial\theta_i \partial\theta_j \partial\theta_k} (\hat{\theta}_i - \tilde{\theta}_i)(\hat{\theta}_j - \tilde{\theta}_j)(\hat{\theta}_k - \tilde{\theta}_k) \quad . \tag{13}$$

Here, and in the sequel, we denote $\nabla\mathcal{L}(\boldsymbol{\theta}_n^*) = \nabla_\theta \mathcal{L}(\boldsymbol{\theta})|_{\theta=\theta_n^*}$, to avoid cluttering the equations. Under further assumptions imposed in the sequel, and the results of Lemma 1 (which we discuss in Section III-A) this remainder term will be shown to be uniformly bounded and its expected value $o(\kappa_n/N)$. In the above expressions all gradients are taken w.r.t. $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}}$ is a point on the line segment connecting $\hat{\boldsymbol{\theta}}_{n,N}$ and $\boldsymbol{\theta}_n^*$. Note that we have used the generic symbol θ_i to denote the ith component of the parameter vector, and did not distinguish between the different origins of the components as we have done previously (gating network parameters, different experts, etc.). Also in what follows we will denote $S_{n,N} \stackrel{\triangle}{=} \frac{1}{2}(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}_n^*)(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)$ for brevity.

- The first term on the right hand side (r.h.s.) of (12), labeled (i), is the *approximation* term, measuring the deviation from zero of the minimal risk. Here we induce an error due to the limits of the approximation class $\mathcal{H}_n$.
- The second term on the r.h.s., labeled (*), is zero by definition.
- The third term on the r.h.s., labeled (ii), is the *estimation* error induced by a parameter estimate which is based on a sample of size N. This error term is also referred to as the *stochastic error*. Note, that this term includes the remainder term $r_{n,N}$.

Our next task is to estimate the magnitude of these error terms, and establish some bounds which will lead to a bound on the total error $\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N})$.

A. Statistical Properties of the Estimation Error

Since the sample set $\mathcal{D}_N$ has been drawn at random, and both the parameter estimator $\hat{\boldsymbol{\theta}}_{n,N} = \hat{\boldsymbol{\theta}}(\mathcal{D}_N)$ as well as $\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N})$ are measurable functions, we will be interested in their statistical properties in what follows.

The parameter estimator $\hat{\boldsymbol{\theta}}_{n,N}$ is subject to a distribution $Q_{n,N}(\cdot)$. It is shown by White [21],[22], as part of a general theory of misspecified models, that $Q_{n,N} \Rightarrow Q_n$ as N tends to infinity, where $\Rightarrow$ denotes weak convergence. Moreover this limit distribution is a Gaussian distribution centered around $\boldsymbol{\theta}_n^*$, the minimizer of the expected risk function. We present the following lemma, adapted from White [23], without proof.

Lemma 1: Let $\hat{\boldsymbol{\theta}}_{n,N}$ be a sequence of least squares estimators (i.e., minimizers of $l(\mathcal{D}_N, f_n)$), and assume that $L_\nu(\mathcal{F}, f_n)$ has a unique minimum at $\boldsymbol{\theta}_n^*$ in Ω_n, a compact subset of $\mathbb{R}^{2n(d+1)}$, then $\hat{\boldsymbol{\theta}}_{n,N} \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_n^*$. Assume further that the matrices A_n^* and B_n^* (defined below) are nonsingular and that $\boldsymbol{\theta}_n^*$ is interior to Ω_n. Then the r.v. $\sqrt{N}(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)$ is

asymptotically normal, i.e.

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*) \xrightarrow{d} \mathbf{Z}_n \sim N(0, C(\boldsymbol{\theta}_n^*)). \tag{14}$$

where $C_n^* \stackrel{\triangle}{=} C(\boldsymbol{\theta}_n^*) = (A_n^*)^{-1} B_n^* (A_n^*)^{-1}$,

$$A_n^* \stackrel{\triangle}{=} A(\boldsymbol{\theta}_n^*) = \int \nabla^2 [f(\mathbf{x}) - f_n(\mathbf{x}; \boldsymbol{\theta}_n^*)]^2 \nu(d\mathbf{x}),$$

and

$$B_n^* \stackrel{\triangle}{=} B(\boldsymbol{\theta}_n^*) = 4 \int [(f(\mathbf{x}) - f_n(\mathbf{x}; \boldsymbol{\theta}_n^*))^2 + \sigma^2] \nabla f_n(\mathbf{x}; \boldsymbol{\theta}_n^*) \nabla^T f_n(\mathbf{x}; \boldsymbol{\theta}_n^*) \nu(d\mathbf{x}) \tag{15}$$

Here, all gradients are taken w.r.t. the parameter vector $\boldsymbol{\theta}$, and $\sigma^2 = \mathbb{E}\varepsilon_t^2$. Also, $\hat{C}_{n,N} = \hat{A}_{n,N}^{-1} \hat{B}_{n,N} \hat{A}_{n,N}^{-1}$ is a strongly consistent estimator of $C(\boldsymbol{\theta}_n^*)$ where

$$\hat{A}_{n,N} \stackrel{\triangle}{=} \frac{1}{N} \sum_{t=1}^{N} \nabla^2 [Y_t - f_n(\mathbf{X}_t; \hat{\boldsymbol{\theta}}_{n,N})]^2$$

and

$$\hat{B}_{n,N} \stackrel{\triangle}{=} \frac{1}{N} \sum_{t=1}^{N} \nabla [Y_t - f_n(\mathbf{X}_t; \hat{\boldsymbol{\theta}}_{n,N})]^2 \nabla^T [Y_t - f_n(\mathbf{X}_t; \hat{\boldsymbol{\theta}}_{n,N})]^2.$$

Remark 1: For the lemma to hold as stated, one must verify the following three conditions (see also [23]). We denote $r_\theta(\mathbf{x}, y) = [y - f_n(\mathbf{x}; \boldsymbol{\theta}_n)]^2$ for brevity.
(1) $r_\theta(\mathbf{x}, y) \leq m(\mathbf{x}, y)$ for all $\boldsymbol{\theta}_n \in \Omega_n$ and $\mathbf{x} \in I^d$, $y \in \mathbb{R}$ where $\int m(\mathbf{x}, y) \nu(d\mathbf{x}) \eta(dy|\mathbf{x}) < \infty$.
(2) $\partial r_\theta(\mathbf{x}, y)/\partial \theta_i$ are measurable functions of $(\mathbf{x}, y)$, and continuously differentiable functions of θ for each $(\mathbf{x}, y)$, and $i = 1, 2, \ldots, 2n(d+1)$.
(3) $|\frac{\partial r_\theta(\mathbf{x},y)}{\partial \theta_i} \frac{\partial r_\theta(\mathbf{x},y)}{\partial \theta_j}|$ and $|\frac{\partial^2 r_\theta(\mathbf{x},y)}{\partial \theta_i \partial \theta_j}|$ are dominated by functions integrable w.r.t. $\nu(d\mathbf{x})\eta(dy|\mathbf{x})$.
Since Ω_n is compact, the first condition holds trivially, and we may set $m(\mathbf{x}, y) = 2y^2 + 2 \sup_{\theta \in \Omega_n} f_n^2(\mathbf{x}; \boldsymbol{\theta})$, where the supremum is finite since f_n is continuous in θ and Ω_n is compact. That $m(\mathbf{x}, y)$ is integrable is obvious. Since $f_n(\mathbf{x}; \cdot)$ is twice continuously differentiable (by inspection), and $\mathbf{x} \in I^d$, the second and third condition hold by the same argumentation, and thus the results of the lemma follow, given the specified assumptions concerning the matrices A_n^*, B_n^* and the uniqueness of the minimizer $\boldsymbol{\theta}_n^*$.

Lemma 1 establishes the strong consistency, and asymptotic distribution of $\hat{\boldsymbol{\theta}}_{n,N}$, and explicitly defines its statistical properties (i.e., the mean vector and asymptotic covariance matrix). Moreover, Lemma 1 defines consistent estimators of the information matrices. The statement of the lemma is reassuring in the face of a misspecified estimation framework. In most scenarios the estimator $\hat{\boldsymbol{\theta}}_{n,N}$ will not 'lead us' to the true parameter (characterizing the target function), as no such parameterization exists in general. On the other hand, we are assured that the estimator will consistently reach the optimal parameter in the class of functional estimators $(\mathcal{H}_n)$, as the sample set becomes large.

Remark 2: White's results actually hold under more general conditions than the ones specified above. In particular, consistency and asymptotic normality of the estimator are established also for non i.i.d. data. In White's monograph [21], the misspecified framework is given a rigorous mathematical and conceptual treatment. The main results

are seen to hold also for mixing processes, certain ergodic stationary processes and other typical divergences from the classical i.i.d. assumption. Thus, an extension to the case of correlated signals or time series would be straightforward (see also [26]).

Unlike the well specified case, in which $S_{n,N}$ (in the estimation term, (ii) in (12)), would asymptotically follow a Chi-squared distribution, in the misspecified case (i.e., where the model cannot fully approximate the target) this term is asymptotically given by a quadratic form in normal random variables. The distribution of quadratic forms has been studied, and a summary of their properties can be found in [12]. To elucidate the analysis of the estimation term, we shall make use only of basic results concerning first and second order moments. The following lemma establishes the statistical properties of the stochastic error term given in (12). We use the notation $x_N = o(a_N)$ if $(x_N/a_N) \to 0$ and $x_N = O(a_N)$ if $\exists C < \infty$ such that $\overline{\lim}(x_N/a_N) \leq C$. With some abuse of notation, we will write $\mathbb{E}|\mathbf{X}| < \infty$ to mean $\mathbb{E}|X_i| < \infty$ for all $i = 1, 2, \ldots, p$ where $\mathbf{X}$ is a r.v. mapping from the underlying sample space to $\mathbb{R}^p$.

Lemma 2: Let the conditions of Lemma 1 hold. Assume further that for any fixed n, $\exists \delta > 0$ s.t. $\sup_N \mathbb{E}|\mathbf{Z}_{n,N}|^{4+\delta} < \infty$ (i.e., the inequality is assumed to hold for each coordinate of $\mathbf{Z}_{n,N}$), with $\mathbf{Z}_{n,N} \equiv \sqrt{N}(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)$. Then,

$$\mathbb{E}\left[S_{n,N} + r_{n,N}\right] \leq O\left(\frac{1}{2N}\mathrm{Tr}\{B(\boldsymbol{\theta}_n^*)A^{-1}(\boldsymbol{\theta}_n^*)\}\right) + o\left(\frac{\kappa_n}{N}\right) \quad . \tag{16}$$

If in addition for every fixed n, $\sup_N \mathbb{E}|\mathbf{Z}_{n,N}|^{8+\delta} < \infty$ then,

$$\mathrm{Var}\left[S_{n,N} + r_{n,N}\right] \leq O\left(\frac{1}{2N^2}\mathrm{Tr}\{B(\boldsymbol{\theta}_n^*)A^{-1}(\boldsymbol{\theta}_n^*)B(\boldsymbol{\theta}_n^*)A^{-1}(\boldsymbol{\theta}_n^*)\}\right) + o\left(\frac{\kappa_n'}{N^2}\right). \tag{17}$$

where κ_n, κ_n' are constants independent of N, and the matrices $A(\boldsymbol{\theta}_n^*)$ and $B(\boldsymbol{\theta}_n^*)$ are defined in Lemma 1. Here $S_{n,N} = \frac{1}{2}(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}_n^*)(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)$, and $r_{n,N}$ is the remainder term as in (13).

Proof. See Appendix.

An obvious result of Lemma 1 is the following

Corollary 1: Let the assumptions of Lemma 2, needed for (16), hold. Then, for any fixed n the estimation error converges to zero almost surely

$$\left\{\frac{1}{2}(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*)^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}_n^*)(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*) + r_{n,N}\right\} \xrightarrow{\text{a.s.}} 0 \ .$$

Proof. First, note that since $\hat{\boldsymbol{\theta}}_{n,N} \to \boldsymbol{\theta}_n^*$ almost surely, and since $S_{n,N}$ is a measurable function of $\hat{\boldsymbol{\theta}}_{n,N}$, then $S_{n,N} \to 0$ almost surely. By application of the Markov inequality, for all $\epsilon > 0$ we have

$$\mathbb{P}\{|r_{n,N}| > \epsilon\} \leq \frac{\mathbb{E}|r_{n,N}|}{\epsilon}$$

In the proof of Lemma 2, we establish that $\mathbb{E}|r_{n,N}| \leq \kappa_n N^{-3/2}$ and therefor by the Borel-Cantelli Lemma $r_{n,N} \to 0$ almost surely, and the result follows. □

These results, concerning the statistical properties of the estimation error term, will be the basis of the bounds, established in Section III-C. We will utilize the first and second moment calculation to formulate bounds on the mean integrated squared error, and bounds in probability on the integrated squared error.

B. Degree of Approximation Results

The main task now is to bound the magnitude of the approximation term (part (i) of the r.h.s. of (12)). We first state a general theorem, concerning the approximation of functions in the Sobolev class by a manifold of normalized ridge functions.

Definition 1: A function $h : \mathbb{R}^d \to \mathbb{R}$ is called a *ridge* function if it may be expressed as

$$h(\mathbf{x}) = \sigma(\mathbf{a}^T\mathbf{x} + b) \quad ,$$

with $\mathbf{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $\sigma : \mathbb{R} \to \mathbb{R}$.
Thus, a ridge function takes constant values on hyper-planes in $\mathbb{R}^d$. In what follows we will use the term ridge function to mean the function $\sigma(t)$, suppressing the explicit argument, where $t \in \mathbb{R}$. The following definition will be used in the statement of the theorem.

Definition 2: A superposition of normalized ridge functions (σ) is a manifold of the form

$$\mathcal{Q}_n \stackrel{\triangle}{=} \left\{ q_n(\mathbf{x}) \mid q_n(\mathbf{x}) = \frac{\sum_{k=1}^n c_k \sigma(\mathbf{a}_k^T\mathbf{x} + b_k)}{\sum_{k=1}^n \sigma(\mathbf{a}_k^T\mathbf{x} + b_k)},\ c_k, b_k \in \mathbb{R}, \mathbf{a}_k \in \mathbb{R}^d \right\} . \tag{18}$$

Note that $\mathcal{Q}_n$ differs from $\mathcal{H}_n$ in (7) by making the normalization explicit, and taking the linear functions to be constants. Also, in the case of $\mathcal{H}_n$ the ridge functions are taken to be $\sigma(t) = e^t$, and the parameters take values in an explicit compact subset of $\mathbb{R}^d$ and $\mathbb{R}$ respectively. The statement of the approximation theorem, establishing upper bounds on the approximation error between functions in the Sobolev class and classes of superpositions of normalized ridge functions, requires the following assumption concerning the admissibility of ridge functions.

Assumption 2: The ridge function $\sigma(t)$ satisfies the following conditions:
1. For any bounded subset $K \subset \mathbb{R}$ there exists a positive constant c such that $\sigma(t) \geq c > 0 \ \ \forall t \in K$.
2. $\exists b \in \mathbb{R}$ such that $\forall k \in \mathbf{Z}_+$, the k'th order derivative $\sigma^{(k)}(b) \neq 0$. Moreover, there is a $\delta > 0$, and a finite interval $[b-\delta, b+\delta]$ where σ is infinitely many times differentiable.
3. For any bounded subset $K \subset \mathbb{R}$, $|\sigma^{(k)}(t)| < \infty$ for all $t \in K$ and $k \in \mathbf{N}$. Moreover, $\bar{C}_\sigma \stackrel{\triangle}{=} \max_{1\leq i\leq k} \sup_{t\in[b-\delta,b+\delta]} |\sigma^{(i)}(t)| \leq \bar{c}_b e^k$, and $\underline{C}_\sigma = \min_{1\leq i\leq k} \sup_{t\in[b-\delta,b+\delta]} |\sigma^{(i)}(t)| \geq \underline{c}_b e^{-k}$.
Now we state the main approximation result concerning the above functional classes.

Theorem 1: Let Assumption 1 and Assumption 2 hold, then

$$\sup_{f\in W_p^r} \inf_{q_n\in\mathcal{Q}_n} \|f(\mathbf{x}) - q_n(\mathbf{x})\|_{L_p(I^d,\lambda)} \leq \frac{c}{n^{r/d}} \qquad 1 \leq p \leq \infty \tag{19}$$

where c is an absolute constant. Moreover, the coefficients c_k, $\mathbf{a}_k$ and b_k defining the class $\mathcal{Q}_n$ in (18) may be chosen, without loss of generality, so that $\|\mathbf{a}_k\|_\infty \leq \delta/2d$, $|b_k - b| \leq \delta/2$, and $\max_{1\leq k\leq n} |c_k| \leq 2Le^{3dn(1+o(1))}$, where b and δ are as in Assumption 2.
Proof: See Appendix.

Note that this result does not yet establish a bound on the approximation term in (12). We postpone this derivation to Section III-C, where we establish upper bounds on the total error induced by the estimator $\hat{f}_{n,N}$. However, this result immediately extends to the class of approximants defined by the MEM architecture, with degenerate experts (i.e., each expert is a constant, and not a linear function of the input), since the conditions of

Assumption 2 are easily verified for $\sigma(t) = e^t$, i.e. Theorem 1 is applicable to the 'softmax network'. In fact, the statement in Theorem 1 applies to a large class of approximants, since any ridge function obeying the conditions of Assumption 2 will allow this result to hold true.

At this point we digress to make several remarks concerning these results.

Remark 3: The result established in Theorem 1 is in the spirit of the results obtained by Mhaskar [14], where the same bound is seen to hold for a standard neural network, under similar assumptions. Since the last condition of Assumption 2 is not satisfied for a sigmoidal function, Theorem 1 does not yield as an immediate corollary, that normalized neural networks are characterized by the same degree of approximation results as the standard neural networks. Note however, that we have imposed the last condition in Assumption 2 in order to obtain bounds on the magnitudes of the parameters appearing in the definition of the functional class (18), as in Theorem 1. These bounds are essential for the analysis of the estimation error, but are otherwise superfluous for the analysis of the approximation error. Consequently, from the point of view of the approximation results per se, we have that neural networks, using normalized sigmoidal units, are characterized by the same degree of approximation results, as are neural networks that employ sigmoidal units. Moreover, while the bounds are finite and explicit, we have made no attempt to optimize them, as this seems rather difficult within the particular approximation scheme we are using here. We believe that these bounds can be substantially improved using alternative techniques from the theory of function approximation.

Remark 4: Recently upper bounds of the order of $c/n^{1/2}$ have been established by Barron [5], w.r.t. feedforward neural networks. This bound was seen to hold for a class of target functions that are effectively band-limited (i.e., absolute value first order moments of the bandwidth are finite, and upper bounded by a global constant). This result has been established in the L_2 norm [5], and extended to the sup-norm (L_∞) by Yukich *et al.* [24]. Both proofs employ a random coding argument. The interesting property of these bounds is their independence of the dimensionality, compared to classical results obtained by Mhaskar [14], and the results proved herein. The simple explanation for this seeming dissonance lies in the definition of the target class. The now classical result of Barron [5] is driven by the restriction of target functions to a fairly limited class, while Mhaskar's analysis [14] assumes functions are in a Sobolev class. One should note, however, that the constant factor in Barron's bounds could be exponential in the dimensionality of the problem, thus requiring an exponentially large number of terms in the approximant. It should also be mentioned that as shown in [5] in the case where $r = d/2 + 2$ partial derivatives of $f(\mathbf{x})$ are known to exist then the class under study is a sub-set of the class studied by Barron, for which approximation rates of order $c/n^{1/2}$ can be achieved. It is interesting that in this case, there is no requirement for the square integrability of the derivatives.

Remark 5: The following lower bound is a consequence of [8] (see also [14] for further details and discussion)

$$\sup_{f \in W_p^r} \inf_{\tilde{q}_n \in \tilde{\mathcal{Q}}_n} \|f(\mathbf{x}) - \tilde{q}_n(\mathbf{x})\|_p \geq \frac{c}{n^{r/d}}$$

where $\tilde{\mathcal{Q}}_n$ is the standard sigmoidal neural network, with n nonlinear sigmoidal units in the so-called hidden layer (i.e., linear combinations of n terms of sigmoidal functions).

This lower bound is valid if the parameterization of the neural network is such that the linear parameters are continuous functionals of the unknown mapping f. In a sense, this limits the effects of small fluctuations (around the true target function) on the choice of the parameterization of $\tilde{q}_n(\cdot)$ - the neural network approximator. Recently, Mhaskar [14] established optimal degree of approximation results for sigmoidal neural networks, by establishing an upper bound which is of the same order, using a parameterization determined by continuous linear functionals on f. In our setting, we have not focused on the issue of optimality, and the parameterization studied both for the approximation bound , as well as for the bound in Theorem 3, are not restricted to be continuous linear functionals of the mapping f. The upper bound does *suggest* however that it may be optimal in order, but this is left at best as a conjecture.

Remark 6: Obviously, the result of Theorem 1 holds for the case of $\psi(\cdot;\boldsymbol{\theta}_j)$ that is linear in the parameters, or any other non-linear function that can be reduced to a constant. In fact, in the statement of Theorem 1, we have eliminated some degrees of freedom in the original construction of the MEM class $\mathcal{H}_n$ by taking ψ to be constants. The case of linear experts is particularly important since local linear regression can be interpreted more directly than global non-linear models. Consider as an example the case of non-linear models for time series (or more general temporal signals for that matter). A local linear approximation, in the form of an Autoregressive (AR) model, allows insight and analysis of localized time scale phenomena. In [26] we demonstrate that these results carry over to the framework of prediction in time series, thus local linearization is in some sense sufficient, if an adequate partition function is implemented. This statement can be made rigorous with the aid of Theorem 1, and its implications as to the choice of gating networks.
Note that in the well specified case, the approximation error is zero, and all that remains is the stochastic error, which can be straightforwardly analyzed with the aid of classical large sample properties of the LS estimator. In the next section we derive an expression for the total error bound, based on the bounds and statistical properties that have been developed and studied in the previous sections.

C. Total Error Bounds

In some of the results presented in this section we will need the following technical condition

Assumption 3: Assume that $\nu \ll \lambda$ where λ is the Lebesgue measure in $\mathbb{R}^d$. Furthermore, let the associated density function be uniformly bounded over I^d.
We are now ready to derive the complete error bounds, combining the estimation and approximation bounds obtained thus far.

Theorem 2: Suppose assumption 2, 3, and the conditions needed for (16) hold. Assume further that $f \in W_2^r(L)$, then for N sufficiently large we have

$$\mathbb{E}\|f - \hat{f}_{n,N}\|^2_{L_2(I^d,\nu)} \leq \frac{c}{n^{2r/d}} + O\left(\frac{1}{2N}\mathrm{Tr}\{B(\boldsymbol{\theta}_n^*)A^{-1}(\boldsymbol{\theta}_n^*)\}\right) + o\left(\frac{\kappa_n}{N}\right). \tag{20}$$

where c is an absolute constant (see Appendix), and κ_n is a constant appearing in Lemma 2, independent of N. Here n is the complexity index (i.e., the number of additive terms in the approximating manifold). The parameter r is the number of continuous derivatives in L_2 that f is assumed to possess and d is the dimensionality of the input. The matrices A_n^* and B_n^* are defined in Lemma 1 and N is the sample size.

Proof. By the second order Taylor series expansion of the mean squared error, we have

$$\|f - \hat{f}_{n,N}\|^2_{L_2(I^d,\nu)} = \|f - f^*_n\|^2_{L_2(I^d,\nu)} + S_{n,N} + r_{n,N} \tag{21}$$

where $S_{n,N}$ is the stochastic error term, and $r_{n,N}$ is the remainder. The first term on the r.h.s is simply $\mathcal{L}(\boldsymbol{\theta}^*_n)$, the approximation error term. The bound on this term is established with the aid of Assumption 3 and Theorem 1 as follows:

$$\begin{aligned}
\|f - f^*_n\|^2_{L_2(I^d,\nu)} &= \int_{I^d} |f - f^*_n|^2 d\nu \\
&\leq K \int_{I^d} |f - f^*_n|^2 d\lambda \\
&= K\|f - f^*_n\|^2_{L_2(I^d,\lambda)} \\
&\leq \frac{c'}{n^{2r/d}} \quad .
\end{aligned}$$

where the first inequality follows from Assumption 3 with K the uniform upper bound, and the second inequality follows from Theorem 1. Note, that since $\mathcal{Q}_n \subseteq \mathcal{H}_n$, we have $\inf_{f_n \in \mathcal{H}_n} \|f - f_n\| \leq \inf_{f_n \in \mathcal{Q}_n} \|f - f_n\|$. Taking the expectation and applying the results of Lemma 2 we have

$$\mathbb{E}\|f - f^*_n\|^2_{L_2(I^d,\nu)} \leq \frac{c'}{n^{2r/d}} + O\left(\frac{1}{2N}\mathrm{Tr}\{B(\boldsymbol{\theta}^*_n)A^{-1}(\boldsymbol{\theta}^*_n)\}\right) + o\left(\frac{\kappa_n}{N}\right)$$

which concludes the proof. □.

The following corollary asserts that if we restrict $f \in W^r_\infty$, then Assumption 3 may be dropped.

Corollary 2: Suppose assumption 2, and the conditions of (16) of Lemma 2 hold. Assume further that $f \in W^r_\infty(L)$ then, for N sufficiently large we have

$$\mathbb{E}\|f - \hat{f}_{n,N}\|^2_{L_2(I^d,\nu)} \leq \frac{c}{n^{2r/d}} + O\left(\frac{1}{2N}\mathrm{Tr}\{B(\boldsymbol{\theta}^*_n)A^{-1}(\boldsymbol{\theta}^*_n)\}\right) + o\left(\frac{\kappa_n}{N}\right) \tag{22}$$

where all the parameters are as in Theorem 2.

Proof. Immediately follows from the fact that

$$\int_{I^d} |f - f^*_n|^2 d\nu \leq \|f - f^*_n\|_\infty$$

and we apply the result of Theorem 1 with $p = \infty$. The final bound then follows straightforwardly. □

In [6] Barron obtains an upper bound on the estimation error, w.r.t. the class of neural network functional estimators, which is $O(nd \log N/N)$. This bound, unlike the bound obtained herein, is not asymptotic in N, rather it holds for finite values of N. Moreover, this bound is explicit in expressing the relation between the dimensionality complexity and sample size. In the setting we pursue herein, these relations are only implicit in the form of the derived upper bound.

For the overall bound in (22) to actually decrease to zero, we must specify $n(N)$. Since the increase rate of $n(N)$ is restricted by stringent requirements, i.e., the limiting behavior

of $\hat{\boldsymbol{\theta}}_{n,N}$, the solution is not obvious to us at the moment. The question of consistent estimation, can be addressed by use of sieves (the reader is referred to Geman and Hwang's paper [9] for a general overview, White's work in the context of neural networks [22], and the work of Barron [4], [6]). The general results of [9] and [22] suggest that consistency in the case of nonlinear regression (on i.i.d. data), can be established by taking the sequence $n(N) = O(N^{1-\epsilon})$ for any $\epsilon > 0$. The growth of the parameter space is also limited by bounding the sum of absolute valued linear coefficients to be $O(\log N)$. In the process of revising this paper, we have established a result along these lines, proving the above in general form (see [25]).

An alternative formulation of Theorem 2 is established, as the total error is bounded in probability.

Corollary 3: Let the conditions of Theorem 2 needed for (17) hold. Then, for N sufficiently large, and $\alpha \in (0,1)$ we have

$$\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N}) \leq \frac{c}{n^{2r/d}} + O\left(\frac{1}{N}\left[\frac{\mathrm{Tr}\{B(\boldsymbol{\theta}_n^*)A^{-1}(\boldsymbol{\theta}_n^*)\}}{2} + \sqrt{\frac{\mathrm{Tr}\{B(\boldsymbol{\theta}_n^*)A^{-1}(\boldsymbol{\theta}_n^*)B(\boldsymbol{\theta}_n^*)A^{-1}(\boldsymbol{\theta}_n^*)\}}{2\alpha}}\right]\right) + o\left(\frac{\sqrt{\kappa_n''/\alpha}}{N}\right) \tag{23}$$

with probability exceeding $1-\alpha$, where all the parameters are as in Theorem 2.

Proof. The result follows trivially from the Chebychev inequality,

$$\mathbb{P}\left\{ \mid \mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N}) - \mathbb{E}[\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N})] \mid \leq \sqrt{\frac{\mathrm{Var}[\mathcal{L}(\hat{\boldsymbol{\theta}}_{n,N})]}{\alpha}} \right\} \geq 1-\alpha, \qquad \forall \alpha \in (0,1) \quad .$$

Plugging in the bounds on the mean and variance derived in Lemma 2 completes the proof. □.

IV. Model Selection by Complexity Regularization

The problem of model selection, in the context of the MEM, can be stated as follows. We are given two parametric models, one denoted by $f_{n_1}(\mathbf{x};\boldsymbol{\theta}_1)$ and the other $f_{n_2}(x;\boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_1 \in \Omega_{n_1} \subset \mathbb{R}^{2n_1(d+1)}$ and $\boldsymbol{\theta}_2 \in \Omega_{n_2} \subset \mathbb{R}^{2n_2(d+1)}$, $(n_1 < n_2)$. We assume that one is a sub-model of the other,

$$\{f_{n_1}(\mathbf{x};\boldsymbol{\theta}_1);\boldsymbol{\theta}_1 \in \Omega_{n_1}\} \subset \{f_{n_2}(\mathbf{x};\boldsymbol{\theta}_2);\boldsymbol{\theta}_2 \in \Omega_{n_2}\}.$$

This implies, that by restricting some components of $\boldsymbol{\theta}_2$ to fixed values, or within fixed relations, we obtain the first sub-model. In the case of the MEM, by clamping expert parameters to zero we eliminate, for all practical purposes, some of the experts and obtain a restricted sub-model. Alternatively, one can obtain the same effect by choosing the parameters of the gating network so that $g_j(\mathbf{x},\cdot) > 0$ for some values of j and zero for others, in which case we again have a reduced complexity model. This can be done, for example, by choosing the parameters $\theta_{g_j,0} \ll -\|\boldsymbol{\theta}_{g_j}\|_\infty$ in the representation (2) with $s_j = \boldsymbol{\theta}_{g_j}^T\mathbf{x} + \theta_{g_j,0}$.

When the parameters of the two competing models are estimated, based on a common training set $\mathcal{D}_N$, the problem is to decide which model is superior. We are interested in properties beyond the measure of fit to the data set $\mathcal{D}_N$, and therefore define as our objective to choose the model which will eventually *generalize* better to future data points. We shall concentrate on a model selection criterion, based on the method of complexity regularization, in the spirit of Akaike's AIC [2], Rissanen's MDL [17]. Instead of minimizing the empirical risk function (i.e., the average sum of squares), we add a regularization term, and attempt to minimize the sum of the two terms. The complexity is understood in the sense of the number of free parameters, characterizing the model. We note in passing that a similar methodology has also been suggested by Vapnik, who termed it *structural risk minimization* [20]. There, an attempt is made to minimize some bound on the sample size needed for consistent learning (i.e., establishing conditions so that the uniform law of large numbers holds).

We follow Murata *et al.* [15], who suggested the following regularization scheme. Let $\mathcal{M}_i = \{f_{n_i}(\mathbf{x};\boldsymbol{\theta}_i);\boldsymbol{\theta}_i \in \Omega_{n_i}\}$ denote a hierarchical series of models $\mathcal{M}_1 \subset \mathcal{M}_2 \cdots \subset \mathcal{M}_m \subset \cdots$. Let $\hat{\boldsymbol{\theta}}_{n_i,N}$ denote the parameter vector of the model $\mathcal{M}_i$ obtained by minimizing the following complexity regularized risk function

$$\mathcal{R}(\mathcal{D}_N, f_{n_i}) \stackrel{\triangle}{=} \frac{1}{N}\sum_{t=1}^{N}[Y_t - f_{n_i}(\mathbf{X}_t;\boldsymbol{\theta}_i)]^2 + \frac{1}{2N}\mathrm{Tr}\left\{\hat{B}_N \hat{A}_N^{-1}\right\} \tag{24}$$

where $\hat{B}_N$ and $\hat{A}_N$ are the misspecified model information matrices, defined in Lemma 1. Note that as the size of $\mathcal{D}_N$ becomes large, minimizing $\mathcal{R}(f_{n_i}, \mathcal{D}_N)$ will be equivalent to minimizing the bound on the expected total error, given in Theorem 2, as all quantities in (24) converge almost surely to their expectations. Therefore, minimizing $\mathcal{R}(f_{n_i}, \mathcal{D}_N)$ is consistent with minimizing the upper bounds on the expected total error as the sample size increases. The questions concerning statistical properties of this complexity regularized estimator, are still under investigation.

Remark 7: Note that the penalty term in the definition of $\mathcal{R}(\mathcal{D}_N, f_{n_i})$ (24), is itself of asymptotic nature, since it is the estimator of the expected stochastic error, based on the asymptotic normality of the esti mator $\hat{\boldsymbol{\theta}}_{n,N}$. This may contradict the application of this penalty term in small sample sets, and mars the generality of the argument. We note that the same reasoning applies both to Akaike's AIC and Rissanen's MDL, two popular methods of model selection by complexity regularization. A possible solution to this may be sought in the more general framework of uniform convergence of means to their expectations, introduced and studied by Vapnik [20]. This pioneering work, allows bounds to be established, much like the bound in Theorem 3, sidestepping asymptotic. These bounds can then serve as regularization terms, and enable model selection criteria which are more robust in the face of finite sample size. In a similar vein, the framework of complexity regularization introduced by Barron and Cover [4] and Barron [6] is another framework, which similarly to the MDL has its roots in Information theoretic considerations, in which finite sample effects may be handled very elegantly (as opposed to the AIC and MDL)

V. Discussion

We have studied some of the properties of a novel non-linear model, the so called Mixture of Experts Model (MEM), in the context of multivariate regression. Extensions are

straightforward to other modeling frameworks such as time series and nonlinear signal processing. The model is characterized by a simple architecture , and offers the practitioner intuition and insight, two features which are absent in most non-linear models (such as neural networks). The main task of this work was to illuminate some of the theoretical foundation, underlying the MEM.

In the derivation of the approximation bound, we observe that the MEM may be regarded as 'equivalent' to a class of neural networks with normalized ridge function units (where 'equivalence' is taken in the sense that both classes are characterized by the same degree of approximation). We complement the approximation results by examining the stochastic (estimation) error. The asymptotic bound on the estimation error term is established using the point estimation results in a misspecified framework. Thus, the bound is characterized by quantities related to the asymptotic variance of the least squares estimator, via certain psudo-information matrices.

Several fundamental questions are still unresolved. For one, it is not clear to us whether the approximation bounds that have been derived are in fact optimal, i.e., does there exist a lower bound of the same order of magnitude. A related issue concerns the restrictions we have imposed on the MEM function class in deriving the degree of approximation results. Namely, we have forced the linear experts to be constants. Is there a loss of generality, and can it be quantified? We also expect that the coarse bounds on the parameters can be made tighter with the use of other approximation techniques.

The results that have been obtained in the analysis of the estimation term are quite restrictive, both in the conditions needed for them to hold, as well as in the interpretation they may have in face of a finite sample size. These issues indeed mar the generality of the arguments, and we believe that it should be possible to rephrase most of this work in terms of the uniform convergence framework (c.f., Vapnik [20]), thus obtaining finite sample results.

Acknowledgment The authors thank Hrushikesh Mhaskar for his helpful suggestions, careful reading, and constructive comments on the derivation of the approximation bound. Helpful comments from Kurt Hornik concerning the consistency of the method are also gratefully acknowledged. Finally, we thank the anonymous reviewers for their detailed and constructive comments and suggestions, which have greatly improved the content and style of the manuscript.

Appendix

I. Proof of Lemma 2

Let $H_n \equiv \nabla^2 \mathcal{L}(\boldsymbol{\theta}_n^*)$. We first establish (16). Since, $\mathbf{Z}_{n,N} \xrightarrow{d} \mathbf{Z}_n$ we have $\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N} \xrightarrow{d} \mathbf{Z}_n^T H_n \mathbf{Z}_n$, by the fact that $\psi(\mathbf{z}) = \mathbf{z}^T H_n \mathbf{z}$ is continuous in $\mathbf{z}$. The random variable $\mathbf{Z}_n^T H_n \mathbf{Z}_n$ is a quadratic form in Gaussian random variables, since $\mathbf{Z}_n \sim N(0, C_n^*)$ by Lemma 1. Thus, we have the representation

$$\mathbf{Z}_n^T H_n \mathbf{Z}_n = \sum_{i=1}^{2n(d+1)} \lambda_i R_i^2$$

with R_i i.i.d. Gaussian random variables with zero mean and unit variance (see [12, p. 150 - 153]). The λ_i's are the eigenvalues of the matrix $H_n C_n^*$. Using the above, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_n^T H_n \mathbf{Z}_n] &= \mathbb{E}\left[\sum_{i=1}^{2n(d+1)} \lambda_i R_i^2\right] \\
&= \sum_{i=1}^{2n(d+1)} \lambda_i \\
&= \mathrm{Tr}\{H_n C_n^*\} \\
&= \mathrm{Tr}\{B_n^*(A_n^*)^{-1}\}
\end{aligned}$$

where the last step follows from observing that $H_n = A_n^*$, by definition, and $C_n^* = (A_n^*)^{-1} B_n^* (A_n^*)^{-1}$. Now, write $\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N} = \sum_{i,j} H_{ij} Z_i Z_j$ where we have simplified the notation by omitting the dependence on n and N, and the scalars Z_i are the components of the vector $\mathbf{Z}_{n,N}$. Since, $\mathbb{E}[Z_i Z_j] \leq \sqrt{\mathbb{E} Z_i^2}\sqrt{\mathbb{E} Z_j^2}$, and by assumption $\sup_N \mathbb{E}|\mathbf{Z}_{n,N}|^3 < \infty$, it follows that $\mathbb{E}|\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N}|^{1+\delta} < \infty$ and thus $\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N}$ is u.i. (uniformly integrable). Since $\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N} \xrightarrow{d} \mathbf{Z}_n^T H_n \mathbf{Z}_n$, and $\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N}$ is u.i., it follows (c.f. [7, Proposition 25.12]) that $\mathbb{E}[\mathbf{Z}_n^T H_n \mathbf{Z}_n] < \infty$ and $\mathbb{E}[\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N}] \to \mathbb{E}[\mathbf{Z}_n^T H_n \mathbf{Z}_n]$ in $\mathbb{R}$. Thus, we have $\mathbb{E}[S_{n,N}] = O\left((2N)^{-1}\mathrm{Tr}\{B_n^*(A_n^*)^{-1}\}\right)$ (in fact $\mathbb{E}[S_{n,N}] \sim (2N)^{-1}\mathrm{Tr}\{B_n^*(A_n^*)^{-1}\}$, with $a_N \sim b_N$ if $\lim a_N/b_N = 1$). Now,

$$\begin{aligned}
\mathbb{E}|r_{n,N}| &\leq \mathbb{E}\left[\sum_{i,j,k=1}^{2n(d+1)} \frac{1}{3!}\left|\frac{\partial^3 \mathcal{L}(\tilde{\boldsymbol{\theta}})}{\partial\theta_i \partial\theta_j \partial\theta_k}\right| |\hat{\theta}_i - \tilde{\theta}_i||\hat{\theta}_j - \tilde{\theta}_j||\hat{\theta}_k - \tilde{\theta}_k|\right] \\
&\leq \frac{\kappa_n}{N^{3/2}}
\end{aligned}$$

where the second step follows from noting that (a) the function $\mathcal{L}(\boldsymbol{\theta})$ is three times continuously differentiable, over the compact domain Ω_n, therefore the third order derivatives are uniformly bounded; (b) the components of the random variable $|\mathbf{Z}_{n,N}|^4$ are u.i., thus applying the Cauchy-Schwartz inequality the result follows. Therefore,

$$\mathbb{E}[S_{n,N} + r_{n,N}] \leq O\left(\frac{1}{2N}\mathrm{Tr}\{B_n^*(A_n^*)^{-1}\}\right) + o\left(\frac{\kappa_n}{N}\right) \quad .$$

The proof of (17) follows along the same lines. First, we have

$$(\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N})^2 \xrightarrow{d} (\mathbf{Z}_n^T H_n \mathbf{Z}_n)^2 \quad .$$

Now, since $|\mathbf{Z}_{n,N}|^4$ is u.i. (by assumption), we have $\mathbb{E}[\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N}]^2 \rightarrow \mathbb{E}[\mathbf{Z}_n^T H_n \mathbf{Z}_n]^2$ in $\mathbb{R}$, and consequently $\mathrm{Var}[\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N}] \rightarrow \mathrm{Var}[\mathbf{Z}_n^T H_n \mathbf{Z}_n]$, where

$$\begin{aligned} \mathrm{Var}[\mathbf{Z}_{n,N}^T H_n \mathbf{Z}_{n,N}] &= \mathrm{Var}\left[\sum_{i=1}^{2n(d+1)} \lambda_i R_i^2\right] \\ &= 2\sum_{i=1}^{2n(d+1)} \lambda_i^2 \\ &= 2\mathrm{Tr}\{B_n^*(A_n^*)^{-1}B_n^*(A_n^*)^{-1}\} \quad . \end{aligned} \tag{25}$$

Here we have used the fact that $\mathrm{Var}[R_i^2] = 2$. To bound the remaining components of the variance of the estimation error we note that $\mathbb{E}|r_{n,N}|^2 \leq c_n/N^3$ (as $|\mathbf{Z}_{n,N}|^8$ is u.i. by assumption, and applying the Cauchy-Schwartz inequality), and

$$\begin{aligned} \mathrm{Cov}(S_{n,N}, r_{n,N}) &\leq \sqrt{\mathrm{Var}[S_{n,N}]}\sqrt{\mathrm{Var}[r_{n,N}]} \\ &\leq O\left(\frac{\kappa_n'}{N^{5/2}}\right) \quad . \end{aligned} \tag{26}$$

Thus, combining the above statements we have

$$\mathrm{Var}[S_{n,N} + r_{n,N}] \leq O\left(\frac{1}{2N^2}\mathrm{Tr}\{B_n^*(A_n^*)^{-1}B_n^*(A_n^*)^{-1}\}\right) + o\left(\frac{\kappa_n'}{N^2}\right)$$

which concludes the proof. $\square$.

II. Proof of Approximation Bounds

A. Preliminaries

We repeat some of the definitions and notation introduced in the main section of the paper. We assume the target function f belongs to the Sobolev class

$$W_p^r(L) \stackrel{\triangle}{=} \left\{ f(\mathbf{x}) \mid ||f||_{W_p^r} = \sum_{|\alpha| \leq r} ||f^{(\alpha)}(\mathbf{x})||_p \leq L \right\} \tag{27}$$

$\mathbf{x} \in I^d = [-1, 1]^d$. Define the manifold $\mathcal{Q}_n$

$$\mathcal{Q}_n \stackrel{\triangle}{=} \left\{ q(\mathbf{x}) \mid q(\mathbf{x}) = \frac{\sum_{k=1}^n c_k \sigma(\mathbf{a}_k^T \mathbf{x} + b_k)}{\sum_{k=1}^n \sigma(\mathbf{a}_k^T \mathbf{x} + b_k)},\ c_k, b_k \in \mathbb{R}, \mathbf{a}_k \in \mathbb{R}^d \right\}. \tag{28}$$

The ridge functions $\sigma(\cdot)$ are chosen to satisfy Assumption 2. The distance between the class W_p^r and the manifold $\mathcal{Q}_n$ is defined as

$$\mathrm{dist}\{W_p^r, \mathcal{Q}_n\} \stackrel{\triangle}{=} \sup_{f \in W_p^r} \inf_{q \in \mathcal{Q}_n} ||f - q||_p \tag{29}$$

where the $L_p(I^d, \lambda)$ norm is defined as $\|f - q\|_p \equiv [\int_{I^d} |f - q|^p d\lambda]^{1/p}$.

B. Proof of the Main Theorem

We now present the main result of the appendix.

Theorem 3: For every integer p, $1 \leq p \leq \infty$, there holds

$$\mathrm{dist}\{W_p^r, \mathcal{Q}_n\} \leq \frac{c}{n^{r/d}}, \tag{30}$$

where $c = c(r, d, p)$. Moreover, the parameters $\{\mathbf{a}_k, b_k, c_k\}$ in (28) can be bounded for each $1 \leq k \leq n$ as follows: $\|\mathbf{a}_k\|_\infty \leq \delta/2d$, $|b_k - b| \leq \delta/2$ and $|c_k| \leq 2Ln^4 \exp\left(3n + dn^{1/d}\right)$, with b and δ as in Assumption 2.

For clarity, we outline the proof by stating two lemmas without proof, and a proposition concerning the properties of an auxiliary function, to be defined. Combining the results by use of the triangle inequality concludes the proof of the theorem. The proof of these lemmas and the proposition is given preceding the outline. The following definitions are necessary for the statement of the first lemma. We denote by $T_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^d \cos(k_i \arccos x_i)$, $k_i = 0, 1, \ldots,$ the d dimensional Chebychev polynomials restricted to I^d, and define

$$\varphi(\mathbf{x}) = \varphi_\rho(\mathbf{x}) \stackrel{\triangle}{=} (2\rho)^{-d} \int_{[-\rho,\rho]^d} \sigma(\mathbf{w}^T\mathbf{x} + b) d\mathbf{w} \qquad (0 \leq \rho \leq 1), \tag{31}$$

where b is defined in Assumption 2. We also introduce the manifold

$$\mathcal{T}_n \stackrel{\triangle}{=} \left\{ t(\mathbf{x}) \mid t(\mathbf{x}) = \sum_{0 \leq \mathbf{k} \leq \mathbf{m}} d_{\mathbf{k}} \frac{T_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})},\ d_{\mathbf{k}} \in \mathbb{R},\ \mathbf{k} \in \mathbf{Z}_+^d \right\}. \tag{32}$$

where $\mathbf{M} = (m, m, \ldots, m)^T$, $m = \lceil n^{1/d} \rceil$, and $0 \leq \mathbf{k} \leq \mathbf{M}$ means $0 \leq k_i \leq m \quad \forall i = 1, 2, \ldots, d$. We note that the Chebychev polynomials may be expressed as

$$T_{\mathbf{k}}(\mathbf{x}) = \sum_{0 \leq \mathbf{p} \leq \mathbf{k}} \tau_{\mathbf{k},\mathbf{p}} \mathbf{x}^{\mathbf{p}}. \tag{33}$$

In the sequel we will need to bound the coefficients $\tau_{\mathbf{k},\mathbf{p}}$. For this purpose we introduce the following simple result.

Proposition 1: For any multi-integer $0 \leq \mathbf{k} \leq \mathbf{M}$ the coefficients $\tau_{\mathbf{k},\mathbf{p}}$ in (33) are bounded as follows:

$$C_\tau \stackrel{\triangle}{=} \max_{0 \leq \mathbf{k} \leq \mathbf{M}} \max_{0 \leq \mathbf{p} \leq \mathbf{k}} |\tau_{\mathbf{k},\mathbf{p}}| \leq \left(\frac{m}{2}\right)^d (2e)^{m^d} \tag{34}$$

Proof Consider the one-dimensional Chebychev polynomial given by [27]

$$T_k(x) = \cos(k \cos^{-1} x) = \frac{k}{2} \sum_{j=0}^{\lfloor k/2 \rfloor} (-1)^j \frac{(k-j-1)!}{j!(k-2j)!} (2x)^{k-2j} \stackrel{\triangle}{=} \sum_{j=0}^{\lfloor k/2 \rfloor} t_{k,j} (2x)^{k-2j}.$$

Using the inequality $\binom{k}{j} \leq (ek/j)^j$ and simple algebra we obtain $|t_{k,j}| \leq \frac{k}{2}(2e)^k$. The result follows by taking the tensor product needed to define the d-dimensional Chebychev functions, namely $T_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^d T_{k_i}(x_i)$. □

We first present a proposition concerning the properties of $\varphi(\mathbf{x})$ defined in (31):

Proposition 2: For the function $\varphi(\mathbf{x})$, the following holds:
1. $\varphi(\mathbf{x}) \geq c_1 > 0 \ \ \forall \mathbf{x} \in [-1,1]^d$ and $\forall \rho > 0$.
2. $\forall f \in W_p^r$ we have the following inequality

$$||f\varphi||_{W_p^r} = \sum_{|\alpha| \leq r} ||D^\alpha [f\varphi]||_p \leq K_1$$

where $K_1 = K_1(r,d)$, i.e. $f\varphi \in W_p^r$.
We now state the first lemma:

Lemma 3: For every positive integers $p, n > 0$ and $r \geq 0$

$$\text{dist}\{W_p^r, \mathcal{T}_n\} \leq \frac{c}{n^{r/d}} \quad . \tag{35}$$

Moreover, the coefficients $d_{\mathbf{k}}$ in the definition of the class $\mathcal{T}_n$ in (32) may, without loss of generality, be assumed to be bounded as follows: $|d_{\mathbf{k}}| \leq 2Ln$.
The following lemma states that the functions $T_{\mathbf{k}}(\mathbf{x})/\varphi(\mathbf{x})$ can be approximated to arbitrary accuracy by a linear combination of n normalized ridge functions $\sigma(\cdot)$. That is, we establish that the distance between the manifold $\mathcal{Q}_n$ and the functions constituting the manifold $\mathcal{T}_n$ is arbitrarily small.

Lemma 4: For every $\mathbf{k} \in [0,m]^d$ and $h \in (1, (3md)^{-1})$ there exist a vector $\mathbf{p} = [p_1, p_2, \ldots, p_d]^T \in [0,m]^d$ and a set of bounded coefficients $\{a_{\mathbf{j},\mathbf{k}}\}$ such that

$$\left\| \frac{T_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} - \frac{\sum_{0 \leq \mathbf{j} \leq \mathbf{p}} a_{\mathbf{j},\mathbf{k}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x} + b]}{(\hat{p})^{-1} \sum_{0 \leq \mathbf{j} \leq \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x} + b]} \right\|_\infty \leq c'_{d,\sigma} h, \tag{36}$$

where $\mathbf{j} \in \mathbf{Z}_+^d$. The vector inequality $\mathbf{j} \leq \mathbf{p}$ is defined coordinate-wise, and we have defined $\hat{p} \equiv (p_1+1)(p_2+1)\cdots(p_d+1)$ and $\hat{p} \leq (m+1)^d$ is implicit. The exact bound on $|a_{\mathbf{j},\mathbf{k}}|$ is given in Lemma 5 below.
We now present the proof of the main theorem.

Proof of Theorem 3 From the second property of φ, stated in Proposition 2, and Lemma 3 we have $\forall f \in W_p^r$, $\exists d_{\mathbf{k}} = d_{\mathbf{k}}(f, \varphi)$ such that

$$\left\| f(x)\varphi(x) - \sum_{0 \leq \mathbf{k} \leq \mathbf{M}} d_{\mathbf{k}} T_{\mathbf{k}}(x) \right\|_p \leq cn^{-r/d}. \tag{37}$$

Now, from (37) and the result of Lemma 4 we have the following chain of inequalities:

$$\begin{aligned}
\Delta &\equiv \left\| f(\mathbf{x}) - \sum_{0 \leq \mathbf{k} \leq \mathbf{M}} d_{\mathbf{k}} \frac{\sum_{0 \leq \mathbf{j} \leq p} a_{\mathbf{j},\mathbf{k}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x} + b]}{\sum_{0 \leq \mathbf{j} \leq \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x} + b]} \right\|_p \\
&\overset{(a)}{\leq} \left\| f(\mathbf{x}) - \sum_{0 \leq \mathbf{k} \leq \mathbf{M}} d_{\mathbf{k}} \frac{T_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} \right\|_p \\
&+ \left\| \sum_{0 \leq \mathbf{k} \leq \mathbf{M}} d_{\mathbf{k}} \frac{T_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} - \sum_{0 \leq \mathbf{k} \leq \mathbf{M}} d_{\mathbf{k}} \frac{\sum_{0 \leq \mathbf{j} \leq \mathbf{p}} a_{\mathbf{j},\mathbf{k}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x} + b]}{\sum_{0 \leq \mathbf{j} \leq \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x} + b]} \right\|_p
\end{aligned}$$

$$\begin{aligned} &\overset{(b)}{\leq} \frac{cn^{-r/d}}{\|\varphi(\mathbf{x})\|_p} + \sum_{0\leq\mathbf{k}\leq\mathbf{M}} |d_{\mathbf{k}}| c'_{d,\sigma} h \\ &\overset{(c)}{\leq} c' n^{-r/d}. \end{aligned} \tag{38}$$

Step (a) follows from the triangle inequality. Step (b) Follows from the bounds obtained in Lemma 4 and (37), and step (c) is established on setting $h = n^{-r/d}/(c'_{d,\sigma} \sum_{0\leq\mathbf{k}\leq\mathbf{M}} |d_{\mathbf{k}}|)$ and $c' \equiv c/||\varphi(\mathbf{x})||_p$ by using $\|\varphi\|_p \geq c_1$ together with the bounds on $d_{\mathbf{k}}$ established in Lemma 3. Since h is arbitrary, we may select it to be so small that the second term is at most of the order of magnitude of the first. Finally, the upper bound on the coefficients c_k appearing in the definition of the class $\mathcal{Q}_n$ is obtained by noting that they are upper bounded by $\sum_{0\leq\mathbf{k}\leq\mathbf{M}} |d_{\mathbf{k}} a_{\mathbf{j},\mathbf{k}}|$ and using the bounds already derived for $|d_{\mathbf{k}}|$, and for $|a_{\mathbf{j},\mathbf{k}}|$ in Lemma 3 in conjunction with the above choice of h. The upper bound on the parameters of the ridge function $\sigma(t)$, $\mathbf{a}_k$ and b_k, follows from Assumption 2. Thus, for example, we may set $b_k = b$ and since $\mathbf{x} \in I^d$, set $\|\mathbf{a}_k\|_\infty \leq \delta/2d$, ensuring $|\mathbf{a}_k^T \mathbf{x}| \leq \delta$ as required. □

C. Proof of Lemmas

We shall now give the proof of Proposition 2, Lemma 3 and Lemma 4. In the process we introduce two auxiliary lemmas (5, 6), which are proved as well.

Proof of Proposition 2 The first property follows trivially by the assumption 2 on the lower boundedness of $\sigma(t)$. The second property is proved as follows

$$\begin{aligned} \|f\varphi\|_{W_p^r} &= \sum_{|\alpha|\leq r} \left(\int_{I^d} |D^\alpha (f\varphi)(\mathbf{x})|^p d\mathbf{x} \right)^{1/p} \\ &\overset{(a)}{=} \sum_{|\alpha|\leq r} \left(\int_{I^d} | \sum_{\alpha'+\alpha''=\alpha} A_{\alpha',\alpha''} [D^{\alpha'} f D^{\alpha''} \varphi](\mathbf{x})|^p d\mathbf{x} \right)^{1/p} \\ &\overset{(b)}{\leq} \sum_{|\alpha|\leq r} \sum_{\alpha'+\alpha''=\alpha} |A_{\alpha',\alpha''}| \left(\int_{I^d} |[D^{\alpha'} f D^{\alpha''} \varphi](\mathbf{x})|^p d\mathbf{x} \right)^{1/p} \\ &\overset{(c)}{\leq} \sum_{|\alpha|\leq r} \sum_{\alpha'+\alpha''=\alpha} |A_{\alpha',\alpha''}| \; ||D^{\alpha'} f(\mathbf{x})||_p ||D^{\alpha''} \varphi(\mathbf{x})||_\infty \\ &\overset{(d)}{\leq} c_{r,\sigma} \sum_{|\alpha|\leq r} ||D^\alpha f||_p \\ &\overset{(e)}{=} c_{r,\sigma} \|f\|_{W_p^r} \leq C \end{aligned} \tag{39}$$

where (a) follows from the chain rule of differentiation and the coefficients $A_{\alpha',\alpha''}$ depend only on α' and α''. Steps (b) and (c) follow from Minkowski's and Hölder's inequalities (with $p = 1$ and $q = \infty$), respectively. Step (d) follows from the boundedness of the derivatives of $\sigma(\mathbf{w}^T\mathbf{x} + b)$, that is

$$\|D^{\alpha''}\varphi\|_\infty = \left\| (2\rho)^{-d} \int_{[-\rho,\rho]^d} D_x^{\alpha''} \sigma(\mathbf{w}^T\mathbf{x} + b) d\mathbf{w} \right\|_\infty \leq c_\sigma$$

and rewriting the summation over the derivatives of f. Finally, step (e) is established by the assumption of $f \in W_p^r(L)$ so that $||f||_{W_p^r} \leq L$. By the proof of the two properties, Proposition 2 is proved. $\square$

Proof of Lemma 3 The proof of the lemma is straightforward, based on the results of Proposition 2

$$\begin{aligned}
\text{dist}\{W_p^r, \mathcal{T}_n\} &= \sup_f \inf_{d_{\mathbf{k}}} \left\| f(\mathbf{x}) - \sum_{0 \leq \mathbf{k} \leq \mathbf{M}} d_{\mathbf{k}} \frac{T_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} \right\|_p \\
&= \sup_f \inf_{d_{\mathbf{k}}} \left\| \frac{f(\mathbf{x})\varphi(\mathbf{x}) - \sum_{0 \leq \mathbf{k} \leq \mathbf{M}} d_{\mathbf{k}} T_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} \right\|_p \\
&\overset{(a)}{\leq} (c_1)^{-1} \sup_f \inf_{d_{\mathbf{k}}} \left\| f(\mathbf{x})\varphi(\mathbf{x}) - \sum_{0 \leq \mathbf{k} \leq \mathbf{M}} d_{\mathbf{k}} T_{\mathbf{k}}(\mathbf{x}) \right\|_p \\
&\overset{(b)}{\leq} c_1' n^{-r/d}. \qquad (40)
\end{aligned}$$

Step (a) is established by plugging in the lower bound $||\varphi||_p \geq c_1$ (see property 1 of φ in Proposition 2). Step (b) is a consequence of a well known fact in approximation theory, stating that any function in W_p^r can be closely approximated by a linear combination of Chebychev polynomials. The degree of approximation is related to the number of terms in the combination (n), the dimensionality (d) and the class W_p^r, as in (40). This result may be applied since in Proposition 2 we have established that $f(\mathbf{x})\varphi(\mathbf{x}) \in W_p^r$. The boundedness of the coefficients $d_{\mathbf{k}}$ can be directly demonstrated by making use of the results of Mhaskar in [14]. In particular, note that the coefficients $d_{\mathbf{k}}$ are identical to the parameters $V_{\mathbf{k}}(f)$ defined through (3.10) in [14]. From the construction in [14], one may show using straightforward algebra, that $|d_{\mathbf{K}}| = O(n)$. We omit the details of this derivation. Note also that this factor is 'washed out' by the exponential growth of the coefficients $a_{\mathbf{j},\mathbf{k}}$ which dominates the final bound on the linear parameters c_k.

Proof of Lemma 4 The main idea behind the proof of Lemma 4 is to show that the two expressions in the denominator and numerator in (36) can be made arbitrarily close. ¿From Lemma 5 below we know that for any $h \in (0, (3md)^{-1})$ and $\mathbf{k} \in [0, m]^d$ there exist a $\mathbf{p} \in [0, m]^d$ and bounded coefficients $a_{\mathbf{j},\mathbf{k}}$ such that

$$\left\| T_{\mathbf{k}}(\mathbf{x}) - \sum_{0 \leq \mathbf{j} \leq \mathbf{p}} a_{\mathbf{j},\mathbf{k}} \sigma[h(2\mathbf{j} - \mathbf{p})^T \mathbf{x} + b] \right\|_\infty \leq Kh. \qquad (41)$$

That is, the numerator expressions in (36) can be made arbitrarily close. Let us define $\rho \equiv \hat{p}^{1/d} h$, and proceed to evaluate the normed difference of the denominator expressions.

$$\begin{aligned}
\Delta_1 &\equiv \left\| \varphi_\rho(\mathbf{x}) - (\hat{p})^{-1} \sum_{0 \leq \mathbf{j} \leq \mathbf{p}} \sigma[h(2\mathbf{j} - \mathbf{p})^T \mathbf{x} + b] \right\|_\infty \\
&\overset{(a)}{=} \left\| (2\rho)^{-d} \int_{[-\rho,\rho]^d} \sigma(\mathbf{w}^T \mathbf{x} + b) d\mathbf{w} - (\hat{p})^{-1} \sum_{0 \leq \mathbf{j} \leq \mathbf{p}} \sigma[h(2\mathbf{j} - \mathbf{p})^T \mathbf{x} + b] \right\|_\infty
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} \left\| \hat{p}^{-1}(2h)^{-d} \sum_{0\le \mathbf{j}\le \mathbf{p}} \int_{[0,2h]^d} \sigma([(\mathbf{w}+h(2\mathbf{j}-\mathbf{p})+b]^T\mathbf{x}) d\mathbf{w} - (\hat{p})^{-1} \sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b] \right\|_\infty \\
&\overset{(c)}{=} \hat{p}^{-1}(2h)^{-d} \left\| \sum_{0\le \mathbf{j}\le \mathbf{p}} \int_{[0,2h]^d} \{\sigma([(\mathbf{w}+h(2\mathbf{j}-\mathbf{p})+b]^T\mathbf{x}) - \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]\} d\mathbf{w} \right\|_\infty \\
&\overset{(d)}{\le} c_{d,\sigma} h
\end{aligned} \tag{42}
$$

In step (a) we simply plug in the definition of φ_ρ so it appears explicitly in the expression. Step (b) consists of partitioning the integration region $[-\hat{p}^{1/d}h, \hat{p}^{1/d}h]^d$ into cells of size $[0,2h]^d$. The number of these cells is equal to the cardinality of $\mathbf{p}$ (i.e., the number of terms in the summation). In step (c) we represent the second term as an integral over $\mathbf{w}$ in the region $[0,2h]^d$, and utilize the linearity of the integration operator. Step (d) follows from the mean value theorem, applied to the integrand (i.e., the difference of sigmoid functions). Formally we have

$$\|\sigma([(\mathbf{w}+h(2\mathbf{j}-\mathbf{p})+b]^T\mathbf{x}) - \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]\| \le c\|\nabla\sigma\|_\infty \|\mathbf{w}\|_1 \le c_{d,\sigma}h$$

where the second inequality follows from the definition of $\mathbf{w}\in[0,2h]^d$, thus $\|\mathbf{w}\|_1 \le 2dh$. A corollary of (42) is

$$
\begin{aligned}
(\hat{p})^{-1} \sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b] &\ge c_1 - \Delta_1 \\
&\ge c_1 - c_{d,\sigma}h \ge c_1/2
\end{aligned} \tag{43}
$$

where the third inequality follows from taking $|h| \le c_1/(2c_{d,\sigma})$. As a result of (42) and (43) we have

$$\left\| \frac{1}{\varphi_\rho(\mathbf{x})} - \frac{1}{(\hat{p})^{-1}\sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]} \right\|_\infty \le \frac{c_{d,\sigma}}{(c_1/2)\|\varphi_\rho(\mathbf{x})\|} \le c^* h \tag{44}$$

where $\rho = \hat{p}^{1/d}h$, and the bound on φ_ρ follows from Proposition 2. The following series of inequalities establishes of Lemma 4.

$$
\begin{aligned}
\Delta_2 &\equiv \left\| \frac{T_\mathbf{k}(\mathbf{x})}{\varphi(\mathbf{x})_\rho} - \frac{\sum_{0\le \mathbf{j}\le \mathbf{p}} a_{\mathbf{j},\mathbf{k}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]}{(\hat{p})^{-1}\sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]} \right\|_\infty \\
&\overset{(a)}{\le} \left\| \frac{T_\mathbf{k}(\mathbf{x})}{\varphi_\rho(\mathbf{x})} - \frac{T_\mathbf{k}(\mathbf{x})}{(\hat{p})^{-1}\sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{r}-\mathbf{p})^T\mathbf{x}+b]} \right\|_\infty + \\
&\quad + \left\| \frac{T_\mathbf{k}(\mathbf{x})}{(\hat{p})^{-1}\sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]} - \frac{\sum_{0\le \mathbf{j}\le p} a_{\mathbf{j},\mathbf{k}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]}{(\hat{p})^{-d}\sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]} \right\|_\infty \\
&\overset{(b)}{\le} |T_\mathbf{k}(\mathbf{x})| \left\| \frac{1}{\varphi_\rho(\mathbf{x})} - \frac{1}{(\hat{p})^{-1}\sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]} \right\|_\infty + \\
&\quad + \left\| \frac{1}{(\hat{p})^{-1}\sum_{0\le \mathbf{j}\le \mathbf{p}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b]} \right\|_\infty \left\| T_\mathbf{k}(\mathbf{x}) - \sum_{0\le \mathbf{j}\le \mathbf{p}} a_{\mathbf{j},\mathbf{k}} \sigma[h(2\mathbf{j}-\mathbf{p})^T\mathbf{x}+b] \right\|_\infty \\
&\overset{(c)}{\le} c^* h + \frac{2}{c_1} K h \le c'_{d,\sigma} h
\end{aligned} \tag{45}
$$

Step (a) is established by simply adding and subtracting the term $T_{\mathbf{k}}/\Sigma_2$, and applying the triangle inequality, where $T_{\mathbf{k}}$ denotes by the numerator of the first expression and Σ_2 the denominator of the second expression in the definition of Δ_2. Step (b) is derived by factoring out the common terms in each normed expression, and step (c) follows from the bounds established in Lemma 5, (42) and (43). We also use the fact that $|T_{\mathbf{k}}(\mathbf{x})| = \prod_{i=1}^{d} |\cos(k \arccos x_i)| \leq 1$. Thus, we have proved Lemma 4 $\square$

We present now the lemma establishing the claim in (41).

Lemma 5: For any multi-integer $\mathbf{k} \in [0, m]^d$ and $h \in (0, (3md)^{-1})$, there exists a vector $\mathbf{p} \in [0, m]^d$ and a bounded set of coefficients $a_{\mathbf{j},\mathbf{k}}$ such that

$$\left\| T_{\mathbf{k}}(\mathbf{x}) - \sum_{0 \leq \mathbf{j} \leq \mathbf{p}} a_{\mathbf{j},\mathbf{k}} \sigma[h(2\mathbf{j} - \mathbf{p})^T \mathbf{x} + b] \right\|_{\infty} \leq Kh, \tag{46}$$

where $K \leq dn^{2+1/d} e^{4n}$. Moreover, the parameters $a_{\mathbf{j},\mathbf{k}}$ can be upper bounded as follows: $|a_{\mathbf{j},\mathbf{k}}| \leq n^2 e^{3n + dn^{1/d}} \left(\frac{1}{h}\right)^{dn^{1/d}}$.

Proof The proof relies on the approach pursued in Lemma 3.2 of [14]. However, we offer a slightly modified proof which will, for completeness, be presented in full. Moreover, we believe that (3.20) occurring in the proof in [14] is erroneous, although this affects only the constants and not the essential points. In any event, a major point in the proof, not stressed in [14], is the boundedness of the coefficients $a_{\mathbf{j},\mathbf{k}}$ which is required for the estimation bound.

We follow [14] and define for each multi-integer $\mathbf{p} = (p_1, \ldots, p_d)$, $p_i \geq 0$,

$$\sigma^{(\mathbf{p})}(\mathbf{w}^T \mathbf{x} + b) \stackrel{\triangle}{=} \frac{\partial^{(|\mathbf{p}|)}}{\partial w_1^{p_1} \cdots \partial w_d^{p_d}} \left[\sigma(\mathbf{w}^T \mathbf{x} + b)\right] = \mathbf{x}^{\mathbf{p}} \sigma^{(|\mathbf{p}|)}(\mathbf{w} \cdot \mathbf{x} + b), \tag{47}$$

where $|\mathbf{p}| = p_1 + p_2 + \cdots + p_d$ and $\mathbf{x}^{\mathbf{p}} = \prod_{i=1}^{d} x_i^{p_i}$. Furthermore, let

$$\sigma_{\mathbf{p},x}(b) = \sigma^{(\mathbf{p})}(\mathbf{w}^T \mathbf{x} + b)|_{\mathbf{w}=0} = \mathbf{x}^{\mathbf{p}} \sigma^{(|\mathbf{p}|)}(b), \tag{48}$$

and thus

$$\mathbf{x}^{\mathbf{p}} = \sigma_{\mathbf{p},x}(b) \left(\sigma^{(|\mathbf{p}|)}(b)\right)^{-1}. \tag{49}$$

For any fixed b, consider a finite difference of order $\mathbf{p}$ [18]:

$$\Delta_{h,x}^{\mathbf{p}} \sigma(b) = \sum_{0 \leq \mathbf{l} \leq \mathbf{p}} (-1)^{|\mathbf{l}|} \binom{\mathbf{p}}{\mathbf{l}} \sigma[h(2\mathbf{l} - \mathbf{p})^T \mathbf{x} + b]. \tag{50}$$

Note that $\Delta_{h,x}^{\mathbf{p}} \sigma(b)$ represents a ridge function 'neural network' with $\prod_{i=1}^{d}(p_i + 1)$ hidden units. In Lemma 6 below we show that

$$\left| \sigma_{\mathbf{p},x}(b) - (2h)^{-|\mathbf{p}|} \Delta_{h,x}^{\mathbf{p}} \sigma(b) \right| \leq \|\sigma^{(|\mathbf{p}|+1)}\|_{\infty,b} |\mathbf{p}| h, \tag{51}$$

where $\|\cdot\|_{\infty,b}$ is the supremum norm restricted to the interval of size 2δ centered at b (see Assumption 2). Now, the Chebychev polynomial $T_{\mathbf{k}}(\mathbf{x})$ can be expanded as in (33) ,

where the coefficients $\tau_{\mathbf{k},\mathbf{p}}$ are constants dependent only on $\mathbf{k}$ and $\mathbf{p}$. From (49) we then conclude that

$$T_{\mathbf{k}}(\mathbf{x}) = \sum_{0\leq\mathbf{p}\leq\mathbf{k}} \tau_{\mathbf{k},\mathbf{p}}\sigma_{\mathbf{p},x}(b)\left(\sigma^{(|\mathbf{p}|)}(b)\right)^{-1}. \tag{52}$$

Adding and subtracting $h^{-|\mathbf{p}|}\Delta^{\mathbf{p}}_{h,x}\sigma(b)$ on the r.h.s. of (52) and using the triangle inequality we obtain

$$\begin{aligned}
&\left|T_{\mathbf{k}}(\mathbf{x}) - \sum_{0\leq\mathbf{p}\leq\mathbf{k}} \tau_{\mathbf{k},\mathbf{p}}\left(\sigma^{(|\mathbf{p}|)}(b)\right)^{-1} h^{-|\mathbf{p}|}\Delta^{\mathbf{p}}_{h,x}\sigma(b)\right| \\
\leq &\sum_{0\leq\mathbf{p}\leq\mathbf{k}} \tau_{\mathbf{k},\mathbf{p}}\left(\sigma^{(|\mathbf{p}|)}(b)\right)^{-1}\left|\sigma_{\mathbf{p},x}(b) - h^{-|\mathbf{p}|}\Delta^{\mathbf{p}}_{h,x}\sigma(b)\right| \\
\leq &\sum_{0\leq\mathbf{p}\leq\mathbf{k}} \tau_{\mathbf{k},\mathbf{p}}\left(\sigma^{(|\mathbf{p}|)}(b)\right)^{-1}\|\sigma^{(|\mathbf{p}|+1)}\|_{\infty,b}|\mathbf{p}|h \\
\leq &(m+1)^d\frac{C_\tau\bar{C}_\sigma}{\underline{C}_\sigma}mdh,
\end{aligned} \tag{53}$$

where we have used Lemma 6 in the final step, imposing the constraint $h \leq (3md)^{-1}$. The bound on K may be obtained from (53) by using the bounds on C_τ, $\underline{C}_\sigma$ and $\bar{C}_\sigma$.

Now, from (50) and (53) we conclude that for $h \leq (3md)^{-1}$ we have

$$\left|T_{\mathbf{k}}(\mathbf{x}) - \sum_{0\leq\mathbf{p}\leq\mathbf{k}}\sum_{0\leq\mathbf{l}\leq\mathbf{p}} c_{\mathbf{p},\mathbf{l}}\sigma[h(2\mathbf{l}-\mathbf{p})^T\mathbf{x}+b]\right| < Kh, \tag{54}$$

where $c_{\mathbf{p},\mathbf{l}} = \tau_{\mathbf{k},\mathbf{p}}\left(\sigma^{(|\mathbf{p}|)}(b)\right)^{-1} h^{-|\mathbf{p}|}(-1)^{|\mathbf{l}|}\binom{\mathbf{p}}{\mathbf{l}}$. At this point we observe that the double sum in (54) may be reduced to a single sum using the identity

$$\sum_{0\leq\mathbf{p}\leq\mathbf{k}}\sum_{0\leq\mathbf{l}\leq\mathbf{p}} c_{\mathbf{p},\mathbf{l}}\sigma[h(2\mathbf{l}-\mathbf{p})^T\mathbf{x}+b] = \sum_{-\mathbf{k}\leq\mathbf{j}\leq\mathbf{k}} b_{\mathbf{j}}\sigma(h\mathbf{j}^T\mathbf{x}+b),$$

where $b_{\mathbf{j}} = \sum_{\{\mathbf{l}:2\mathbf{l}-\mathbf{p}=\mathbf{j}\}} c_{\mathbf{p},\mathbf{l}}$. In order to complete the proof we need to show that the coefficients $b_{\mathbf{j}}$ are bounded. This is easily established on noting that $|b_{\mathbf{j}}| \leq m^d \max_{0\leq\mathbf{p},\mathbf{l}\leq\mathbf{M}}|c_{\mathbf{p},\mathbf{l}}|$ and noting that

$$\begin{aligned}
|c_{\mathbf{p},\mathbf{l}}| &= \left|\tau_{\mathbf{k},\mathbf{p}}\left(\sigma^{(|\mathbf{p}|)}(b)\right)^{-1} h^{-|\mathbf{p}|}\binom{\mathbf{p}}{\mathbf{l}}\right| \\
&\leq \frac{C_\tau}{\underline{C}_\sigma}\left(\frac{1}{h}\right)^{dm} e^{dm},
\end{aligned} \tag{55}$$

which establishes the desired result upon using the bounds on C_τ and $\underline{C}_\sigma$. □

Finally, we present the proof of (51).

Lemma 6: Let $\sigma_{\mathbf{p},x}(b)$ and $\Delta^{\mathbf{p}}_{h,x}\sigma(b)$ be defined as in (48) and (50), respectively. Then for $h \leq 1/3md$ there holds

$$\left|\sigma_{\mathbf{p},x}(b) - (2h)^{-|\mathbf{p}|}\Delta^{\mathbf{p}}_{h,x}\sigma(b)\right| \leq \|\sigma^{(|\mathbf{p}|+1)}\|_\infty|\mathbf{p}|h.$$

Proof Using standard results from the theory of approximation [18] allows us to replace the difference operator $\Delta^{\mathbf{p}}_{h,x}$ by an integral representation

$$\Delta^{\mathbf{p}}_{h,x}\sigma(b) = \mathbf{x}^{\mathbf{p}} I_h^{|\mathbf{p}|}\sigma(\mathbf{x}), \tag{56}$$

where

$$I_h^{|\mathbf{p}|}\sigma(\mathbf{y}) \stackrel{\triangle}{=} \int_{-h}^{h}\cdots\int_{-h}^{h} \sigma^{(|\mathbf{p}|)}\left[\left((\tau_1+\cdots+\tau_{p_1})y_1+\cdots+(\tau_{|\mathbf{p}|-p_d+1}+\cdots\tau_{|\mathbf{p}|})y_d\right)+b\right] d\boldsymbol{\tau}, \tag{57}$$

where $\boldsymbol{\tau} = (\tau_1,\ldots,\tau_{|\mathbf{p}|})$. Note that by Assumption 2 the derivative in the integrand exists for $h < 1/3md$. We then have

$$\begin{aligned} A_x &\stackrel{\triangle}{=} \sigma^{(\mathbf{p})}(b) - (2h)^{-|\mathbf{p}|}\Delta^{\mathbf{p}}_{h,x}\sigma(b) \\ &= \mathbf{x}^{\mathbf{p}}\left(\sigma^{(|\mathbf{p}|)}(b) - (2h)^{-|\mathbf{p}|} I_h^{\mathbf{p}}\sigma^{(|\mathbf{p}|)}(\mathbf{x})\right), \end{aligned} \tag{58}$$

where we have used (47) and (56). Using (57) and the mean value theorem we conclude that there is a $\xi \in [0, |\mathbf{t}\cdot\mathbf{x}|]$, where $\mathbf{t} = (p_1 h,\ldots,p_d h)$, such that

$$I_h^{|\mathbf{p}|}\sigma(\mathbf{x}) = (2h)^{|\mathbf{p}|}\sigma^{(|\mathbf{p}|)}(b+\xi). \tag{59}$$

Since $\|\mathbf{x}\|_\infty \le 1$ we then have

$$|A_x| \le |\sigma^{(|\mathbf{p}|)}(b) - \sigma^{(|\mathbf{p}|)}(b+\xi)| \le \|\sigma^{(|\mathbf{p}|+1)}\|_\infty |\xi| \le \|\sigma^{(|\mathbf{p}|+1)}\|_\infty |\mathbf{p}| h \tag{60}$$

which concludes the proof. □

References

[1] Adams, R.A. *Sobolev Spaces*, Academic Press, New York, 1975.

[2] Akakike, H. "A New Look at the Statistical Model Identification", *IEEE Trans. AC*, vol. 19:6, 716-723, 1974.

[3] Amari, S.I. and Murata, N. "Statistical Theory of Learning Curves under Entropic Loss Criterion", *Neural Computation*, vol. 5: 140-153, 1993.

[4] Barron, A.R. and Cover, T.M. "Minimum Complexity Density Estimation," *IEEE Trans. Inf. Theory*, vol. IT-37:4, 1034-1054, 1991.

[5] Barron, A.R. "Universal Approximation Bound for Superpositions of A Sigmoidal Function," *IEEE Trans. Inf. Theory*, vol. IT-39, pp. 930-945, 1993.

[6] Barron, A.R. "Approximation and Estimation Bounds for Artificial Neural Networks", *Machine Learning*, vol. 4, pp. 115-133, 1994.

[7] Billingsley, P. *Probability and Measure*, Wiley, 1979.

[8] DeVore, R., Howard, R. and Micchelli, C.A. "Optimal Non-linear Approximation", *Manuscripta Mathematica*, vol. 63, 469-478, 1989.

[9] Geman S. and Hwang, C.R. "Nonparameteric Maximum Likelihood Estimation by the Method of Sieves", *Annals of Stats.*, vol. 10:2, 401-414, 1982.

[10] Jacobs, R.A., Jordan, M.I., Nowlan, S.J. and Hinton, G.E. "Adaptive Mixtures of Local Experts",*Neural Computation*, vol. 3:79-87, 1991.

[11] Jordan, M.I. and Jacobs, R.A. "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation*, vol. 6:181-214, 1994.

[12] Johnson, N.L. and Kotz, S. *Distributions in Statistics. Continuous Univariate Distributions - 2*. Wiley, New-York, 1972.

[13] Leshno, M., Lin, V., Pinkus, A. and Schocken, S. "Multilayer Feedforward Networks with a Non-polynomial Activation Function Can Approximate any Function", *Neural Networks*, vol. 6, 861-867, 1993.

[14] Mhaskar, H. "Neural Networks for Optimal Approximation of Smooth and Analytic Functions", *Neural Computation*, vol.8:1, pp.164-177, 1996.

[15] Murata, N., Yoshizawa, S. and Amari, S.I. "Network Information Criterion - Determining the Number of Hidden Units for an Artificial Neural Network", *IEEE Trans. on Neural Networks*, vol. 5:6, pp. 865-872, 1994.
[16] Redner, R.A. and Walker, H.F. "Mixture Densities, Maximum Likelihood and the EM Algorithm", *SIAM Review*, vol. 26, 195-239, 1984.
[17] Rissanen, J. "Universal Coding, Information Prediction and Estimation," *IEEE Trans. Inf. Theory*, vol. IT-30:4, pp. 629-636, 1984.
[18] Timan, A.F. *Theory of Approximation of Functions of a Real Variable*, Macmillan, New York, 1963.
[19] Titterington, D.M., Smith, A.F.M., and Makov, U.E. *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York, 1985.
[20] Vapnik, V. *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York, 1982.
[21] White, H. *Estimation, Inference and Specification Analysis*, Cambridge university press, 1994.
[22] White, H. "Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings", *Neural Networks*, vol. 3, 535-549, 1991.
[23] White, H. "Consequences and Detection of Misspecified Nonlinear Regression Models", *J. Amer. Statis. Assoc.*, vol. 76, 419-433, 1981.
[24] Yukich, J.E., Stinchcombe, M.B. and White, H. " Sup-Norm Approximation Bounds for Networks Through Probabilistic Methods", *IEEE Trans. on Info. Theory*, vol. IT-41:4, 1995.
[25] Zeevi, A.J. "A Consistent Learning Property of Mixtures of Experts", submitted for publication, 1997.
[26] Zeevi, A.J., Meir, R. and Adler, R. "Time Series Prediction Using Mixtures of Experts", in *Advances in Neural Information Processing Systems 9*, Ed. M. Jordan, MIT Press, 1997.
[27] D. Zwillinger, *SRC Standard Mathematical Tables and Formulae*, 30th Ed., CRC Press, 1996.