Approximation with neural networks activated by ramp sigmoids

Gerald H.L. Cheang

School of Mathematics and Statistics, University of South Australia, Adelaide, Australia

Received 21 October 2008; received in revised form 22 February 2010; accepted 17 March 2010
Available online 23 March 2010

Abstract

Accurate and parsimonious approximations for indicator functions of \(d\)-dimensional balls and related functions are given using level sets associated with the thresholding of a linear combination of ramp sigmoid activation functions. In neural network terminology, we are using a single-hidden-layer perceptron network implementing the ramp sigmoid activation function to approximate the indicator of a ball. In order to have a relative accuracy \(\epsilon\), we use \(T = c(d^2/\epsilon^2)\) ramp sigmoids, a result comparable to that of Cheang and Barron (2000) [4], where unit step activation functions are used instead. The result is then applied to functions that have variation \(V_f\) with respect to a class of ellipsoids. Two-hidden-layer feedforward neural nets with ramp sigmoid activation functions are used to approximate such functions. The approximation error is shown to be bounded by a constant times \(V_f/T_1^{1/2} + V_f d/T_2^{1/4}\), where \(T_1\) is the number of nodes in the outer layer and \(T_2\) is the number of nodes in the inner layer of the approximation \(f_{T_1,T_2}\).

1. Introduction

In this paper, we consider the approximation of certain classes of functions \(f : \mathcal{S} \to \mathbb{R}\), where \(\mathcal{S} \subseteq \mathbb{R}^d\) is some bounded space with finite Lebesgue measure \(\mu(\mathcal{S})\). The functions \(f\)
have bounded variation $V_f$ with respect to a class of ellipsoids $\mathcal{E}$ that are contained in $\mathcal{S}$. We show that such functions can be approximated accurately and parsimoniously with classes of two-hidden-layer neural networks activated by ramp sigmoids.

The approximating functions take the form

$$f_{T_1,T_2,v_1,v_2}(x) = \sum_{i=1}^{T_1} c_i \phi_{v_1} \left( \sum_{j=1}^{T_2} \omega_{ij} \phi_{v_2}(a_{ij} \cdot x - b_{ij}) - d_i \right), \quad x \in \mathbb{R}^d, \tag{1}$$

where

$$\phi_v(z) = \begin{cases} 0 & \text{when } z < 0, \\ vz & \text{when } 0 \leq z \leq \frac{1}{v}, \\ 1 & \text{when } z > \frac{1}{v}. \end{cases} \tag{2}$$

The function $\phi_v : \mathbb{R} \to \mathbb{R}$, with $v > 0$ is called a ramp sigmoid and it is a piecewise-linear approximation to the Heaviside function. As $v \to \infty$, the ramp sigmoid $\phi_v$ converges pointwise to the Heaviside function. In the approximating functions (1), there are $T_1$ nodes in the outer layer and $T_2$ nodes in the inner layer. The positive Lipschitz constants are $v_1$ and $v_2$ for the ramp sigmoids in the outer and the inner layers respectively. The parameters are the external weight $\{c_1, \ldots, c_{T_1}\}$ in $\mathbb{R}$, of the outer hidden layer; external weights $\omega_{ij}$ in $\mathbb{R}$, of the inner hidden layer, which are elements of a $T_1 \times T_2$ matrix; internal weights $\{a_{ij} \in \mathbb{R}^d : 1 \leq i \leq T_1, 1 \leq j \leq T_2\}$ in the inner hidden layer; and real-valued location parameters $\{b_{ij} \in \mathbb{R} : 1 \leq i \leq T_1, 1 \leq j \leq T_2\}$ and $\{d_1, \ldots, d_{T_1}\}$. In the terminology of Cybenko [5] and Haykin [10], the approximating functions (1) are called two-hidden-layer feedforward neural networks. Such a network is also called a perceptron network (Rosenblatt [17,18]). We show that for functions with variation $V_f$ with respect to a class of ellipsoids, the approximation error $\|f - f_{T_1,T_2,v_1,v_2}\|_{L_2(\mathcal{S})}$ is bounded by constant times $V_f/T_1^{\frac{1}{2}} + V_f/d/T_2^{\frac{1}{2}}$. In this case, the constant depends on the Lipschitz constants $v_1$ and $v_2$. In the limiting case when $v_1$ and $v_2$ go to infinity, the ramp sigmoids become the Heaviside function, and the approximation error bound coincides with the bound in [4].

A key element of the analysis is the approximation of the indicator function $\mathbb{1}_B$ of the unit ball $B$ centered at the origin with an approximation function of the form

$$\phi_{v_1} \left( \sum_{j=1}^{T} \omega_{ij} \phi_{v_2}(a_{ij} \cdot x - b_{ij}) + k \right). \tag{3}$$

The approximation function (3) can be considered as a smoothed version of the level set

$$\mathcal{N}_T = \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{T} c_i \mathbb{1}_{\{a_{ij} \cdot x \geq b_{ij}\}} \geq k \right\}, \tag{4}$$

that was used to approximate the unit ball $B$ in [4]. A level set of a function $f$ at level $k$ is simply the set $\{x \in \mathbb{R}^d : f(x) \geq k\}$. The approximating set (4) is constructed by taking the set of points $x \in \mathbb{R}^d$ such that these points are in some number of linear combinations of half-spaces. Its shape is generally not convex. In higher dimensional spaces, it takes the form of a multi-faceted star-shaped object. Artstein-Avidan et al. [1] call such sets zigzag sets.

A detailed survey of existing literature on approximating a unit ball (and other similar convex bodies) with the traditional polytope approximation (see [6,19,7,8]) is found in [4]. In [4], it is
shown that a threshold of a linear combination of \( c(d^2/\epsilon^2) \) indicators of half-spaces are needed in order to obtain a relative accuracy of \( \epsilon \) for such an approximation. Artstein-Avidan et al. [1] improve on the result of [4] to show that only \( C(d \log(1/\epsilon)/\epsilon^2) \) indicators of half-spaces are needed. However, their result applies only with probability \( 1 - e^{-cd} \).

This paper is organized as follows. In Section 2, we provide an \( L_\infty \) bound for a single-hidden-layer neural net for the Gaussian function \( \exp\left(-\frac{x^2}{2}\right) \) since the unit ball can be expressed as some suitable level set of the Gaussian function. In Section 3, we bound the Hausdorff distance of the level set of the approximation function derived in Section 2 and that of the unit ball. The Hausdorff distance is then used to derive an \( L_1 \) bound between the indicator function of the ball and its approximation set. This is extended to the case of an ellipsoid in Section 5. The main result of the paper is presented in Section 6 and other approximation results are noted in Section 7.

### 2. Approximation of the Gaussian function

A unit ball in \( \mathbb{R}^d \), centered at the origin, may be represented as

\[
B = \left\{ x \in \mathbb{R}^d : \exp\left(-\frac{|x|^2}{2}\right) \geq \exp\left(-\frac{1}{2}\right) \right\},
\]

where \( |\cdot| \) is the Euclidean norm. A key role in our analysis is the use of probabilistic methods and the approximation of the Gaussian function \( h(x) = \exp\left(-\frac{|x|^2}{2}\right) \), and thus the representation of the unit ball as a level set of a Gaussian function in (5).

Using the fact that the Gaussian is a positive definite function with Fourier transform \( (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{|\omega|^2}{2}\right) \), and so \( h \) has a representation in the convex hull of sinusoids, it is known that \( h(x) \) can be expressed using the convex hull of indicators of half-spaces (see [2,3,11,21]). Thus we show that we can find a good single-hidden-layer net approximation (with \( T \) ramp sigmoids \( \phi_{v_2} \) as activation functions) of the form

\[
\sum_{j=1}^{T} \omega_j \phi_{v_2}(a_j \cdot x - b_j) + k
\]

for the Gaussian \( h(x) \) first. Then by thresholding the output of (6) at a certain level \( k' \) we show that the indicator function of the unit ball \( 1_B \) can be approximated well by (3), as shown in Lemma 2.1.

**Lemma 2.1.** Let \( h(x) = \exp\left(-\frac{|x|^2}{2}\right) \) for \( x \) in \( B \), a unit ball in \( \mathbb{R}^d \); then there exists a \( T_2 \)-term ramp sigmoidal neural net approximation \( f_{T_2,v_2}(x) = \sum_{j=1}^{T_2} c_j \phi_{v_2}(a_j \cdot x + b_j) + k \) activated by ramp sigmoids with Lipschitz constant \( v_2 \) such that the approximation error satisfies

\[
\sup_{x \in B} |h(x) - f_{T_2,v_2}(x)| \leq \left( \frac{2}{v_2} + 68 \right) \sqrt{\frac{d(d+1)}{T_2}}.
\]

The proof of Lemma 2.1 uses an integral representation of the Gaussian \( h(x) \) of the form

\[
h(x) = \int_{\mathbb{R}^d} \cos(a \cdot x) \frac{\exp\left(-\frac{|a|^2}{2}\right)}{(2\pi)^{d/2}} da, \quad x \in \mathbb{R}^d,
\]
Lemma 2.1

From [3], it was shown that \( h(x) - h(0) \) is a convex combination of functions in

\[
G_{\cos} = \left\{ \frac{\nu}{|a|} \cos(|a| z - 1), |z| \leq 1 : |\nu| \leq C_h, a \in \mathbb{R}^d \right\}.
\]

It was further shown that \( G_{\cos} \) is in the convex hull of some suitable set of step functions, i.e. scaled and shifted Heaviside functions. The first part of the proof of Lemma 2.1 follows that of [3, Theorem 2]. However, in this paper, we are interested in approximating the Gaussian as a linear combination of ramp sigmoids. Thus adjustments to the results of [3] are necessary.

**Proof.** From (8), there is an integral representation of the Gaussian in terms of a family of cosines

\[
h(x) - h(0) = \int_{\mathbb{R}^d} \frac{C_h}{|a|} \left( \cos \left( |a| \frac{a \cdot x}{|a|} - 1 \right) \right) p(a) da, \quad x \in \mathbb{R}^d,
\]

where

\[
p(a) = \frac{|a| \exp \left( -\frac{|a|^2}{2} \right)}{C_h (2\pi)^{d/2}},
\]

where the normalizing constant for \( p(a) \) is

\[
C_h = \mathbb{E}|a| = \int_{\mathbb{R}^d} |s| \exp \left( -\frac{|s|^2}{2} \right) ds = \sqrt{2} \frac{\Gamma \left( \frac{d+1}{2} \right)}{\Gamma \left( \frac{d}{2} \right)},
\]

the expectation of \( |a| \) with respect to a standard multivariate normal with zero mean vector and identity covariance matrix on \( \mathbb{R}^d \), which is bounded above by \( \sqrt{d} \) via the Cauchy–Schwarz inequality. Note that \(-1 \leq z = \frac{a \cdot x}{|a|} \leq 1 \) for \( x \) in the unit ball \( B \). Thus \( h(x) - h(0) \) is a convex combination of functions in \( G_{\cos} \) evaluated at \( z = \frac{a \cdot x}{|a|} \).

Now consider the set of functions

\[
G_{\phi_{v_2}} = \left\{ \frac{\nu}{|a|} \phi_{v_2} (z + b), |z| \leq 1 : |b| \leq 1, |\nu| \leq \frac{2C_h}{v_2} \right\}.
\]

Note that functions in both \( G_{\cos} \) and \( G_{\phi_{v_2}} \) (when \( v_2 \geq 2C_h \)) have derivatives less than 1. Take any function \( g_{|a|} \) from \( G_{\cos} \) and consider its increasing part and decreasing part separately, say

\[
g_{|a|}(z) = g_{|a|},+(z) - g_{|a|},-(z).
\]

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

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

where \(-1 = t_0 \leq t_1 \leq \cdots \leq t_{k-1} = 1 \) form a partition. The positions of the steps are chosen such that \( g(t_i) - g(t_{i-1}) \) partition the range space equally and that \( g_{|a|},+(t_i) = \frac{1}{2} \left[ g(t_i) + g(t_{i-1}) \right] \). That is, each jump is of equal height and the function \( g_{|a|},+(z) \) at the jump point passes through exactly in the middle of the jump. Since the derivative of \( g_{|a|},+ \) is bounded by \( C_h \), it follows that the sum of the absolute values of the jump heights \( \sum_{i=1}^{k-1} |g(t_i) - g(t_{i-1})| \) is bounded by \( C_h \), and adding up coefficients for the steps for the decreasing part yields that the sum of absolute values of jump heights (for both parts combined) is no greater than \( 2C_h \).
Now if we repeat the above procedure with \( \phi_{v_2} \) instead of steps, and as long as the increasing part of \( \phi_{v_2} \) has a derivative no less than \( 2C_h \), the error of such an approximation of \( g_{[a]} \) with \( \phi_{v_2} \) is no greater than that of \( g_{[a]} \) with steps. Thus \( \mathcal{G}_{\cos} \subseteq \text{conv}\,\mathcal{G}_{\phi_{v_2}} \) for \( v_2 \geq 2C_h \). Here closure is achieved in \( \mathcal{L}_\infty \).

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

\[
g_{[a], v_2}(z) = \sum_{i=1}^{n_{[a]}} c_i \phi_{v_2}(z + b_{i, [a]}),
\]

(11)

where \( n_{[a]} \) may be very large, and \( \sum_{i=1}^{n_{[a]}} |c_i| \leq 2C_h \) (for now it does not matter how many terms there are in \( g_{[a], v_2} \)). We can choose the coefficients \( c_i \) in (11) such that the approximation \( g_{[a], v_2} \) achieves

\[
\sup_{|z| \leq 1} |g_{[a]}(z) - g_{[a], v_2}(z)| \leq \frac{2C_h}{v_2} \sqrt{\frac{d + 1}{T_2}}.
\]

(12)

The choice of an arbitrarily small bound in (12) that contains the factor \( \sqrt{\frac{d + 1}{T_2}} \) becomes apparent later in the proof, so we have a common factor of \( \sqrt{\frac{d + 1}{T_2}} \) in the sum of two terms in (17).

Substituting (11) for the corresponding \( g_{[a]}(z) \) in (9), there is an approximation \( h_{v_2} \) to the \( h(x) - h(0) \) such that

\[
h_{v_2}(x) = \mathbb{E}_a g_{[a], v_2} \left( \frac{a \cdot x}{|a|} \right)
\]

\[
= 2C_h \mathbb{E}_a \mathbb{E}_{i, [a]} \left[ \text{sign}_{i, a} \phi_{v_2} \left( \frac{a \cdot x}{|a|} + b_i \right) \right].
\]

(13)

(where \( \text{sign}_{i, a} \in \{-1, +1\} \)) which is an infinite convex combination of elements of \( \mathcal{G}_{\phi_{v_2}} \). Note that

\[
h(x) - h(0) = \mathbb{E}_a g_{[a]} \left( \frac{a \cdot x}{|a|} \right).
\]

(14)

Using [3, Lemma 5], the bound

\[
\sup_{x \in B} |h(x) - h(0) - h_{v_2}(x)| \leq \mathbb{E}_a \sup_{x \in B} \left| g_{[a]} \left( \frac{a \cdot x}{|a|} \right) - g_{[a], v_2} \left( \frac{a \cdot x}{|a|} \right) \right|
\]

\[
\leq \frac{2C_h}{v_2} \sqrt{\frac{d + 1}{T_2}}
\]

(15)

follows.

We choose a \( T_2 \) term ramp sigmoidal neural net approximation to \( h_{v_2}(x) \) in (13) by Monte Carlo sampling using the distribution \( \frac{\mathbb{I}}{2} \) for the \( \pm 1 \) values taken by the variable \( \text{sign}_{i, a} \) in (13). From [4, Lemma 4] (applied to classes of functions bounded by 1 with pseudo-dimension \( d + 1 \)), there exists an approximation \( h_{T_2, v_2}(x) = \frac{1}{T_2} \sum_{j=1}^{T_2} c_i \phi_{v_2}(a_i \cdot x + b_i) \)

\[^{1}\text{A modified version of the lemma in [4] is found in the Appendix of this paper and we explain how it can also be applied here since the statement of the original lemma in [4] is given for indicator functions of sets.}

\[\]
Lemma 2.1 is equivalent to the level set of the neural net approximation that approximates the indicator function $1_{\tilde{N}}$.

Let $f_{T_2,v_2}(x) = h_{T_2,v_2}(x) + h(0)$. Thus there exists an approximation $f_{T_2,v_2}(x)$ to the Gaussian such that

$$\sup_{x \in B} |h(x) - f_{T_2,v_2}(x)| \leq \sup_{x \in B} |h(x) - h(0) - h_{v_2}(x)| + \sup_{x \in B} |h_{v_2}(x) - h_{T_2,v_2}(x)| \leq \left(\frac{2}{v_2} + 68\right) \sqrt{\frac{d(d + 1)}{T_2}}. \quad (17)$$

3. Bounding the Hausdorff distance of the approximation

The Hausdorff distance between two sets $F$ and $G$ is defined as

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

The norm $|\cdot|$ is the usual Euclidean norm in $\mathbb{R}^d$. We bound the Hausdorff distance between the ball and a suitable approximating set in this section. We define $\tilde{N}_{T_2,v_1,v_2}$ as a level set

$$\tilde{N}_{T_2,v_1,v_2} := \left\{ x : f_{T_2,v_2}(x) \geq \exp \left( -\frac{1}{2} \right) + \epsilon_{T_2} + \frac{1}{v_1} \right\}, \quad (18)$$

where $f_{T_2,v_2}(x)$ is chosen from the Monte Carlo sampling scheme as described in the proof of Lemma 2.1. From the definition of the ramp sigmoid (2), we see that the set (18) is equivalent to

$$\tilde{N}_{T_2,v_1,v_2} = \left\{ x : \phi_{v_1} \left( f_{T_2,v_2}(x) - \exp \left( -\frac{1}{2} \right) - \epsilon_{T_2} \right) = 1 \right\}. \quad (19)$$

From this point onwards, set

$$\epsilon_{T_2} := \left( 68 + \frac{2}{v_2} \right) \sqrt{\frac{d(d + 1)}{T_2}},$$

the upper bound to the $L_\infty$ error between the Gaussian and its approximation in Lemma 2.1. We will bound the Hausdorff distance $\delta^H(B, \tilde{N}_{T_2,v_1,v_2})$ between the ball $B$ and the level set $\tilde{N}_{T_2,v_1,v_2}$.

Lemma 3.1. Let $B_R$ be a ball of radius $R$ in $\mathbb{R}^d$ centered at the origin, and let $\tilde{N}_{T_2,v_1,v_2}$ be the level set of the neural net approximation that approximates the indicator function $1_{B_R}$. For sufficiently large $T_2$ and $v_1$ such that $\epsilon_{T_2} + \frac{1}{v_1} \leq \frac{1}{2} \exp \left( -\frac{1}{2} \right)$,

$$\delta^H(B_R, \tilde{N}_{T_2,v_1,v_2}) \leq \left( 318 + \frac{4\sqrt{2}e}{v_2} \right) R \sqrt{\frac{d(d + 1)}{T_2}} + R \frac{2\sqrt{2}e}{v_1}. \quad (20)$$
We set $T_2$ and $v_1$ large enough that $\epsilon_T_2 + \frac{1}{v_1}$ is less than $\frac{1}{2} \exp \left( -\frac{1}{2} \right)$. Choose $r_0$ such that $\exp \left( -\frac{r_0^2}{2} \right) = \exp \left( -\frac{1}{2} \right) + 2\epsilon_T_2 + \frac{2}{v_1}$. Let $B_{r_0}$ be the ball of radius $r_0$ centered around the origin. If $x \in \tilde{N}_{T_2,v_1,v_2}$, then $\exp \left( -\frac{1}{2} \right) + \frac{1}{v_1} \leq f_{T_2,v_2} (x) - \epsilon_T_2 \leq \exp \left( -\frac{1}{2} \right)$ which implies that $x \in B$. Similarly if $x \in B_{r_0}$, then $\exp \left( -\frac{1}{2} \right) + \epsilon_T_2 + \frac{1}{v_1} \leq \exp \left( -\frac{1}{2} \right) - \epsilon_T_2 \leq f_{T_2,v_2} (x)$, which implies that $x \in \tilde{N}_{T_2,v_1,v_2}$. Thus $B_{r_0} \subset \tilde{N}_{T_2,v_1,v_2} \subset B$ and consequently

$$\delta^H (B, \tilde{N}_{T_2,v_1,v_2}) \leq 1 - r_0.$$ 

We also note that the set

$$\left\{ x : f_{T_2,v_2} (x) \geq \exp \left( -\frac{1}{2} \right) + \epsilon_T_2 \right\} \subset B.$$

Now

$$r_0 = \sqrt{2 \log \left( \frac{1}{\left( e^{-\frac{1}{2}} + 2\epsilon_T_2 + \frac{2}{v_1} \right)} \right)}$$

$$= \sqrt{1 - 2 \log \left( 1 + 2 \left( \epsilon_T_2 + \frac{1}{v_1} \right) e^{\frac{1}{2}} \right)},$$

which is close to 1. Thus if $T_2$ is large enough that $\epsilon_T_2 + \frac{1}{v_1}$ is less than $\frac{1}{2} \left( \exp \left( -\frac{1}{4} \right) - \exp \left( -\frac{1}{2} \right) \right)$ (the choice of $T_2$ such that $70 \sqrt{\frac{d(d+1)}{T_2}} \leq \frac{1}{2} \left( \exp \left( -\frac{1}{4} \right) - \exp \left( -\frac{1}{2} \right) \right)$ suffices), we have in this case

$$\delta^H (B, \tilde{N}_{T_2,v_1,v_2}) \leq 2\sqrt{2}e \left( \epsilon_T_2 + \frac{1}{v_1} \right)$$

$$\leq \left( 318 + \frac{4\sqrt{2}e}{v_2} \right) \sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2}e}{v_1}. \quad (21)$$

Suppose we have a ball $B_R$ of radius $R$ instead, and we approximate the indicator function of $B_R$ with a suitable level set $\tilde{N}_{T_2,v_1,v_2}$ of a similar form to (19); then using an analysis similar to that for a unit ball, the Hausdorff distance between $B_R$ and its approximation set can be shown to be bounded by

$$\left( 318 + \frac{4\sqrt{2}e}{v_2} \right) R \sqrt{\frac{d(d+1)}{T_2}} + R \frac{2\sqrt{2}e}{v_1}. \quad \Box$$

4. An $\mathcal{L}_1$ bound

Let $B_R$ be a ball of radius $R$, $\tilde{N}_{T_2,v_1,v_2}$ the level set induced by the approximation as explained in Section 3, $\mu$ the Lebesgue measure, and $\delta$ the Hausdorff distance between $B_R$ and its approximation as obtained in (20). Here we bound the relative Lebesgue measure of
the symmetric difference $\frac{\mu(B_R \triangle \tilde{N}_{T_2,v_1,v_2})}{\mu(B_R)}$. The symmetric difference $B_R \triangle \tilde{N}_{T_2,v_1,v_2}$ is the set $(B_R \cap \tilde{N}_{T_2,v_1,v_2}^C) \cup (B_R^C \cap \tilde{N}_{T_2,v_1,v_2})$. The following lemma provides a $\mathcal{L}_1$ bound between the indicators of the ball $\mathbb{1}_{B_R}$ and the indicator of some approximation set $\mathbb{1}_{\tilde{B}_R}$, where $\tilde{B}_R \subset B_R$.

**Lemma 4.1.** Let $B_R$ be a ball or radius $R$ in $\mathbb{R}^d$. Suppose there is an approximating set $\tilde{B}_R$ to $B_R$ such that $\tilde{B}_R \subset B_R$, and that the Hausdorff distance between $B_R$ and $\tilde{B}_R$ is $\delta$; then the relative Lebesgue measure of the symmetric difference $\frac{\mu(B_R \triangle \tilde{B}_R)}{\mu(B_R)}$ is bounded above by $d \frac{\delta}{R}$.

**Proof.** Since the symmetric difference $B_R \triangle \tilde{B}_R$ is included in the shell $B_R \setminus B_{R-\delta}$, one has

$$
\int \left| \mathbb{1}_{B_R} - \mathbb{1}_{\tilde{B}_R} \right| \frac{\mu(dx)}{\mu(B_R)} = \frac{\mu(B_R \triangle \tilde{B}_R)}{\mu(B_R)} \\
\leq \frac{\mu(B_R) - \mu(B_{R-\delta})}{\mu(B_R)} \\
\leq 1 - \left( 1 - \frac{\delta}{R} \right)^d \\
\leq d \frac{\delta}{R}. \quad \Box
$$

(22)

In our application of Lemma 4.1, note that the approximating set is $\tilde{N}_{T_2,v_1,v_2}$ from (18), or equivalently (19). An upper bound (20) to the Hausdorff distance between $B_R$ and its approximating set is given in Section 3. Thus the following corollary.

**Corollary 4.1.** The relative Lebesgue measure of the symmetric difference $\frac{\mu(B_R \triangle \tilde{N}_{T_2,v_1,v_2})}{\mu(B_R)}$ between $B_R$ and its approximation set $\tilde{N}_{T_2,v_1,v_2}$ is bounded above by

$$
\frac{\mu(B_R \triangle \tilde{N}_{T_2,v_1,v_2})}{\mu(B_R)} \leq \left( 318 + \frac{4\sqrt{2}e}{v_2} \right) \sqrt{\frac{d(d+1)}{T_2}} + \frac{2d\sqrt{2}e}{v_1}.
$$

5. Ellipsoid approximation

Consider an ellipsoid $E = \{ x \in \mathbb{R}^d : x^T M x \leq 1 \}$ centered at the origin with $M = A^T A$ strictly positive definite with the $d \times d$ positive definite square root $A$. Equivalently

$$
E = \{ x \in \mathbb{R}^d : \exp(-x^T A^T A x/2) \geq \exp(-1/2) \}
$$

(23)
is the level set of a Gaussian surface. Like for the ball, it can be accurately and parsimoniously approximated by thresholding a single-hidden-layer neural net.

Suppose the eigenvalues of $A$ are $r_1 \leq r_2 \leq \cdots \leq r_d$ with the corresponding eigenvectors $\{r_1, r_2, \ldots, r_d\}$; then as in [4], the ellipsoid $E$ described by (23) can be constructed from the unit ball by stretching its unit radii to $r_1, r_2, \ldots, r_d$ along the directions $\{\pm r_1, \pm r_2, \ldots, \pm r_d\}$ respectively. The axes of the ellipsoid $E$ are oriented along $\{\pm r_1, \pm r_2, \ldots, \pm r_d\}$.
If the approximating set for the unit ball takes the form \( \{ x : f_{T_2, v_2}(x) \geq \exp\left(-\frac{1}{2}\right) + \epsilon_{T_2} + \frac{1}{v_1} \} \) as in (18), then the one for the ellipsoid \( E \) is
\[
\tilde{E}_{T_2, v_1, v_2} = \left\{ x : f_{T_2, v_2}(Ax) \geq \exp\left(-\frac{1}{2}\right) + \epsilon_{T_2} + \frac{1}{v_1} \right\}.
\] (24)

Equivalently to (24), the indicator function \( \mathbb{1}_E \) of the ellipsoid is approximated by
\[
\phi_{v_1}\left(f_{T_2, v_2}(Ax) - \exp\left(-\frac{1}{2}\right) - \epsilon_{T_2}\right)
\]
and
\[
\tilde{E}_{T_2, v_1, v_2} = \left\{ x : \phi_{v_1}\left(f_{T_2, v_2}(Ax) - \exp\left(-\frac{1}{2}\right) - \epsilon_{T_2}\right) = 1 \right\}.
\]

We are interested in bounding \( \delta^H(E, \tilde{E}_{T_2, v_1, v_2}) \), the Hausdorff distance between the ellipsoid and its approximating set, as well as the measure of its symmetric difference \( \mu(E \triangle \tilde{E}_{T_2, v_1, v_2}) \).

**Lemma 5.1.** Suppose there is an approximating set \( \tilde{E} \) to \( E \) such that \( \tilde{E} \subset E \), and that the Hausdorff distance between \( E \) and \( \tilde{E} \) is \( \delta \); then the relative Lebesgue measure of the symmetric difference \( \frac{\mu(E \triangle \tilde{E})}{\mu(E)} \) is bounded above by \( d \frac{\delta}{rd} \). If \( T_2 \) is chosen to satisfy \( 70\sqrt{\frac{d(d+1)}{T_2}} \leq \frac{1}{2} \left( \exp\left(-\frac{1}{4}\right) - \exp\left(-\frac{1}{2}\right) \right) \), so that the requirement \( 68\sqrt{\frac{d(d+1)}{T_2}} + \frac{2}{v_2} \sqrt{\frac{d(d+1)}{T_2}} + \frac{1}{v_1} \leq \frac{1}{2} \left( \exp\left(-\frac{1}{4}\right) - \exp\left(-\frac{1}{2}\right) \right) \) holds, then the Hausdorff distance between the ellipsoid \( E \) and its approximating set \( \tilde{E}_{T_2, v_1, v_2} \) is bounded above by
\[
\delta^H(E, \tilde{E}_{T_2, v_1, v_2}) \leq \left( 318 + \frac{4\sqrt{2e}}{v_2} \right) rd \sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2er_d}}{v_1}.
\] (25)

In particular, the measure of the symmetric difference \( \mu(E \triangle \tilde{E}_{T_2, v_1, v_2}) \) between \( E \) and its approximation set \( \tilde{E}_{T_2, v_1, v_2} \) is bounded above by
\[
\mu(E \triangle \tilde{E}_{T_2, v_1, v_2}) \leq \left( 318 + \frac{4\sqrt{2e}}{v_2} \right) \mu(E) d \sqrt{\frac{d(d+1)}{T_2}} + \frac{2\mu(E) \sqrt{2ed}}{v_1}.
\] (26)

**Proof.** The proof applies the same argument for stretching a unit ball \( B \) to the ellipsoid \( E \) along the directions of the eigenvectors \( \{ \pm r_1, \pm r_2, \ldots, \pm r_d \} \), which give the orientation of the axes of the ellipsoid as in [4]. From Lemma 2.1, it follows that
\[
\sup_{x \in B} |h(Ax) - f_{T_2, v_2}(Ax)| \leq \left( \frac{2}{v_2} + 68 \right) \sqrt{\frac{d(d+1)}{T_2}} = \epsilon_{T_2}.
\]

Thus the difference between the Gaussian \( \exp(-x^T A^T Ax/2) \) associated with the ellipsoid and its approximation \( f_{T_2, v_2}(Ax) \) is bounded above by \( \left( \frac{2}{v_2} + 68 \right) \sqrt{\frac{d(d+1)}{T_2}} \) in the \( L_\infty \) norm. The results of Lemma 5.1 follow in the same manner as the respective bounds on the unit ball and its approximation are derived in Section 3, Lemma 4.1 and Corollary 4.1 of this paper. □
6. Approximation bounds for two-layer nets

A function \( f \) is said to have variation \( V_{f,\mathcal{H}} \) with respect to a class of sets \( \mathcal{H} \) if \( V_{f,\mathcal{H}} \) is the infimum of numbers \( V \) such that \( f/V \) is in the closure of the convex hull of signed indicators of sets in \( \mathcal{H} \), where the closure is taken in \( L_2(P_X) \). A special case of finite variation is the case that we call total variation with respect to a class of sets. Suppose that \( f(x) \) is defined over a bounded region \( S \) in \( \mathbb{R}^d \). We say that \( f \) has total variation \( V \) with respect to a class of sets \( \mathcal{H} = \{ H_\xi : \xi \in \Xi \} \) if there exist some signed measure \( v \) over the measurable space \( \Xi \) and

\[
f(x) = \int_{\Xi} \mathbb{1}_{H_\xi}(x)v(d\xi) \quad \text{for } x \in S, \tag{27}
\]

and if \( v \) has finite total variation \( V \). The sets \( H_\xi \) are parametrized by \( \xi \) in \( \Xi \). In our context, the \( H_\xi \) are half-spaces in \( \mathbb{R}^d \) where the \( \xi \) consist of the location and orientation parameters. In the event that the representation (27) is not unique, we take the measure \( v \) that yields the smallest total variation \( V \).

The function class \( \mathcal{F}_{V,\mathcal{H}} \) of functions with variation \( V_{f,\mathcal{H}} \) bounded by \( V \) arises naturally when thinking of the functions obtained by linear combinations on a layer of a network where the sums of absolute values of the coefficients of a linear combination are bounded by \( V \) and the level sets from the preceding layer yield the sets in \( \mathcal{H} \). In the main result of this paper, we consider functions \( f \), defined over some bounded domain \( S \), that have variation \( V_{f,\mathcal{H}} \) with respect to the class of ellipsoids \( \mathcal{E} \). The class \( \mathcal{E} = \{ E_\xi \in \mathbb{R}^d : \xi \in \Xi \} \) consists of ellipsoids \( E_\xi \) with \( \mu(E_\xi) \leq \mu(S) \) where \( \mu \) is the Lebesgue measure over \( S \). In [4], such functions \( f \) are approximated by two-hidden-layer neural nets activated by the Heaviside functions of the form

\[
f_{T_1,T_2}(x) = \sum_{i=1}^{T_1} c_i \phi \left( \sum_{j=1}^{T_2} \omega_{ij} \phi(a_{ij} \cdot x - b_{ij}) - d_i \right) \tag{28}
\]

In the form of the approximation (28), the second (outer) layer of the two-layer net takes a linear combination of level sets of functions represented by linear combinations on the first (inner) layer. The class of sets represented by level sets of combinations of first-layer nodes in this case are the approximations to ellipsoids. Here we take the approximation further to the case where ramp sigmoids \( \phi_{v_1} \) and \( \phi_{v_2} \) replace the Heaviside functions \( \phi \) in the approximation (28) so that the approximation to \( f \) takes the form (1).

In our analysis, we will take advantage of both \( L_\infty \) approximation bounds (used to yield approximations to the indicators of ellipsoids in the inner layer) and \( L_2 \) approximation bounds for convex hulls of indicators of ellipsoids (essentially achieved by the outer layer of the network). First we state a simple \( L_2 \) approximation bound found in [4].

**Lemma 6.1.** If \( f \) has variation \( V_f = V_{f,\mathcal{E}} \) with respect to the class \( \mathcal{E} \) of ellipsoids then there is a choice of ellipsoids \( E_1, \ldots, E_T \) and \( s_1, \ldots, s_T \in \{-1, +1\} \), and \( c_i = \frac{V_f s_i}{T_1} \) such that

\[
f_{T_1}(x) = \sum_{i=1}^{T_1} c_i \mathbb{1}_{E_i} \tag{29}
\]

satisfies

\[
\| f_{T_1} - f \|_2 \leq \frac{V_f}{\sqrt{T_1}}. \tag{30}
\]
Lemma 6.1 is actually a corollary of the a lemma found in [4, Lemma 2] which is based on an earlier form found in [12].

The indicators of ellipsoids have two-layer sigmoidal network approximations consisting of a single outer node and a single hidden inner layer. These approximations to $\mathbb{1}_E$ may be substituted into the approximation in (29) to yield a two-hidden-layer approximation to $f$. The following theorem bounds the approximation error using ramp sigmoids.

**Theorem 6.1.** If $f$ has finite variation $V_f$ with respect to the class of ellipsoids $\mathcal{E}$ where $\mu(E) \leq \mu(S)$, and $P_X$ is the uniform probability measure over $S$, then there exist a choice of parameters $(a_{ij}, b_{ij}, c_i, d_i, w_{ij})$ such that the two-hidden-layer neural net $f_{T_1, T_2, v_1, v_2}$ with ramp activation function achieves an approximation error bounded by

$$
\| f - f_{T_1, T_2, v_1, v_2} \|_2 \leq \frac{2V_f}{\sqrt{T_1}} + 2V_f \left( 328d \sqrt{\frac{d(d+1)}{T_2}} \right)^{\frac{1}{2}},
$$

provided that $v_1 \geq \max \left( 4d \sqrt{2eT_1}, \sqrt{\frac{T_2}{d(d+1)}} \right)$, and $v_2 \geq 2\sqrt{d}$, and $T_2$ large enough that $70 \sqrt{\frac{d(d+1)}{T_2}} \leq \frac{1}{2} \left( \exp \left( -\frac{1}{4} \right) - \exp \left( -\frac{1}{2} \right) \right)$.

**Proof.** First we approximate $f$ with $f_{T_1}$ in (29). From Lemma 6.1, the bound to the approximation error is

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

Then we examine what happens when $f_{T_1, v_1}$ replaces the indicators of ellipsoids in $f_{T_1}$ in (29) with corresponding ramp functions of quadratic forms. We have via the triangle inequality

$$
\| f - f_{T_1, v_1} \|_2 \leq \| f - f_{T_1} \|_2 + \| f_{T_1, v_1} - f_{T_1} \|_2.
$$

We consider again the unit ball case, when the outer layer Heaviside sigmoid $\phi$ is replaced by $\phi_{v_1}$. An upper bound to

$$
\left\| \phi \left( \exp \left( -\frac{|x|^2}{2} \right) - \exp \left( -\frac{1}{2} \right) \right) - \phi_{v_1} \left( \exp \left( -\frac{|x|^2}{2} \right) - \exp \left( -\frac{1}{2} \right) \right) \right\|_2
$$
is

$$
\left\| \phi \left( \exp \left( -\frac{|x|^2}{2} \right) - \exp \left( -\frac{1}{2} \right) \right) - \phi \left( \exp \left( -\frac{|x|^2}{2} \right) - \exp \left( -\frac{1}{2} \right) - \frac{1}{v_1} \right) \right\|_2.
$$

In (33), the term $\phi \left( \exp \left( -\frac{|x|^2}{2} \right) - \exp \left( -\frac{1}{2} \right) - \frac{1}{v_1} \right)$ is the indicator function of the set

$$
\left\{ x \in \mathbb{R}^d : \exp \left( -\frac{|x|^2}{2} \right) \geq \exp \left( -\frac{1}{2} \right) + \frac{1}{v_1} \right\},
$$

which is a ball $B_{r_1}$ of radius $r_1 = \sqrt{1 - 2 \log \left( 1 + \frac{e^{-1/2}}{v_1} \right)}$, where $B_{r_1} \subset B$. Thus we seek first a bound on the Hausdorff distance between the unit ball and $B_{r_1}$. Applying the same manner of analysis as in Section 3, we see that $1 - r_1$ (that is, the Hausdorff distance) is bounded by $\frac{\sqrt{2e}}{v_1}$. 

Remark. Lemma 6.1 is actually a corollary of the a lemma found in [4, Lemma 2] which is based on an earlier form found in [12].
Lemma 5.1

Again, and similar techniques (35) to the squared $\sum \parallel A \parallel$ bound to the term $\sum g_i E_i - g_{v_1, E_i} \parallel 2$

Thus

$$\parallel f_{T_1, v_1} - f_{T_1, v_1, v_2} \parallel 2 \leq \sum_{i=1}^{T_1} |c_i| \parallel \mathbb{1}_{E_i} - g_{v_1, E_i} \parallel 2$$

$$\leq \frac{2V_f \left( d \sqrt{2e} \right)^{1/2}}{\sqrt{v_1}},$$

where in (35), the term $g_{v_1, E_i}$ is the sigmoid $\phi_{v_1}$ applied to the Gaussian associated with ellipsoid $E_i$.

Now recall that the goal is to use a two-hidden-layer ramp sigmoidal neural net

$$f_{T_1, T_2, v_1, v_2}(x) = \sum_{i=1}^{T_1} c_i \phi_{v_1} \left( \sum_{j=1}^{T_2} \omega_{ij} \phi_{v_2}(a_{ij} \cdot x - b_{ij}) - d_i \right)$$

to approximate $f(x)$. Thus by the triangle inequality,

$$\parallel f - f_{T_1, T_2, v_1, v_2} \parallel 2 \leq \parallel f - f_{T_1, v_1} \parallel 2 + \parallel f_{T_1, v_1} - f_{T_1, T_2, v_1, v_2} \parallel 2.$$  

A bound to the term $\parallel f_{T_1, v_1} - f_{T_1, T_2, v_1, v_2} \parallel 2$ is

$$\parallel f_{T_1, v_1} - f_{T_1, T_2, v_1, v_2} \parallel 2 \leq \sum_{i=1}^{T_1} |c_i| \parallel \mathbb{1}_{v_1, E_i} - g_{v_1, E_i} \parallel 2 \left( \sum_{j=1}^{T_2} \omega_{ij} \phi_{v_2}(a_{ij} \cdot x - b_{ij}) - d_i \right) \parallel 2,$$

where $\sum_{j=1}^{T_2} \omega_{ij} \phi_{v_2}(a_{ij} \cdot x - b_{ij}) - d_i$ approximates the Gaussian associated with ellipsoid $E_i$ in the second hidden layer. For each ellipsoid $E_i$, let the approximation to $g_{v_1, E_i}$ be $f_{i, T_2, v_1, v_2}(x) = \phi_{v_1} \left( \sum_{j=1}^{T_2} \omega_{ij} \phi_{v_2}(a_{ij} \cdot x - b_{ij}) - d_i \right)$.

Now the approximation set $\tilde{E}_{i, T_2} = \{ f_{i, T_2, v_1, v_2}(x) = 1 \} \subset \{ f_{i, T_2, v_1, v_2}(x) > 0 \} \subset E_i$. Hence the squared $L_2$ approximation error between $g_{v_1, E_i}$ and $f_{i, T_2, v_1, v_2}$ is bounded above by

$$\int_S \parallel \mathbb{1}_{E_i}(x) - f_{i, T_2, v_1, v_2}(x) \parallel 2 P_X(dx) \leq \int_S \parallel \mathbb{1}_{E_i}(x) - \mathbb{1}_{\tilde{E}_{i, T_2}}(x) \parallel 2 P_X(dx)$$

$$\leq 318 + \frac{4\sqrt{2e}}{v_2}$$

$$d \sqrt{\left( \frac{d(d+1)}{T_2} \right) + \frac{2\sqrt{2e}}{v_1}}.$$  

For an ellipsoid with major axial length $R$, application of Lemma 5.1 and similar techniques of stretching the unit ball to the ellipsoid in [4] yields the upper bound $R \frac{\sqrt{2e}}{v_1}$ to the Hausdorff distance between such an ellipsoid and a smaller appropriate ellipsoid. From Lemma 5.1 again, we see that

$$\parallel \phi \left( \exp \left( \frac{-x^T A^T A x}{2} \right) \right) - \phi_{v_1} \left( \exp \left( \frac{-x^T A^T A x}{2} \right) \right) \parallel 2 \leq d \frac{\sqrt{2e}}{v_1}.$$
Thus
\[
\| f_{T_1, v_1} - f_{T_1, T_2, v_1, v_2} \|_2 \leq \sum_{i=1}^{T_1} |c_i| \left( (318 + \frac{4\sqrt{2e}}{v_2}) d \sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2ed}}{v_1} \right)^{1/2}
\]
\[
\leq 2V_f \left( (318 + \frac{4\sqrt{2e}}{v_2}) d \sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2ed}}{v_2} \right)^{1/2}.
\]

Finally, by adding up all the terms together,
\[
\| f - f_{T_1, T_2, v_1, v_2} \|_2 \leq \frac{V_f}{\sqrt{T_1}} + \frac{2V_f(d\sqrt{2e})^{1/2}}{\sqrt{v_1}}
\]
\[
+ 2V_f \left( (318 + \frac{4\sqrt{2e}}{v_2}) d \sqrt{\frac{d(d+1)}{T_2}} + \frac{2\sqrt{2ed}}{v_2} \right)^{1/2}.
\] (40)

Now choose \(v_1 \geq \max \left( 4d\sqrt{2e}T_1, \sqrt{\frac{T_2}{d(d+1)}} \right)\) and \(v_2 \geq 2\sqrt{d}\). Then the bound from (40) yields
\[
\| f - f_{T_1, T_2, v_1, v_2} \|_2 \leq \frac{2V_f}{\sqrt{T_1}} + 2V_f \left( 328d \sqrt{\frac{d(d+1)}{T_2}} \right)^{1/2}. \quad \Box \] (41)

Note that if we let \(v_1\) and \(v_2\) go to infinity in (40), we obtain the bound for the step activation function case in [4]. Although sharper bounds on the Hausdorff distance between the unit ball and that of its zigzag approximation set are reported in [1], the zigzag approximation set there is constructed from a suitable intersection of half-spaces. This does not extend to the case where ramp sigmoids are used in place of half-spaces. Furthermore, we need to exploit the representation of the Gaussian as a convex combination of sinusoids (e.g. from [3]) to carry the results through from [4].

The reader can refer to two examples of continuous functions that have bounded variation with respect to classes of balls (with different radii and/or centers) in [4]. With the approximation of these functions given in [4] by two-hidden-layer neural nets activated by Heaviside functions there, the approximations are not continuous functions. However, if the Heaviside functions are further approximated by ramp sigmoids, the approximations will also be continuous functions.

7. Other approximation results

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

**Theorem 7.1.** Let \(f(x) = f_1(f_2(x)), f_1 : \mathbb{R} \rightarrow \mathbb{R}\) and \(f_2 : B \subset \mathbb{R}^d \rightarrow I \subset \mathbb{R}\). Let \(\phi_v\) be a sigmoid with Lipschitz bound \(v\). Suppose
1. \( f_1(z) \) has a single-layer neural net approximation,

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

and

\[
    \sup_z |f_1(z) - f_{1, T_1, v}(z)| \leq \frac{C_{f_1}}{\sqrt{T_1}},
\]

\[
    \sum_{i=1}^{T_1} |c_i| \leq V \quad \text{and} \quad |u| = \max_i |u_i|;
\]

2. \( f_2(x) \) has a single-layer neural net approximation,

\[
    f_{2, T_2}(x) = \sum_{j=1}^{T_2} k_j \phi(w_j \cdot x + b_j) + d
\]

and

\[
    \|f_2 - f_{2, T_2}\|_2 \leq \frac{C_{f_2}}{\sqrt{T_2}};
\]

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

\[
    f_{T_1, T_2, v}(x) = \sum_{i=1}^{T_1} c_i \phi_v \left( u_i \cdot \sum_{j=1}^{T_2} k_j \phi(w_j \cdot x + b_j) + u_i d + d_i' \right) + c_0
\]

\[
    = \sum_{i=1}^{T_1} c_i \phi_v \left( \sum_{j=1}^{T_2} a_{ji} \phi(w_j \cdot x + b_j) + d_i \right) + c_0
\]

and the approximation rate satisfies

\[
    \|f - f_{T_1, T_2, v}\|_2 \leq \frac{C_{f_1}}{\sqrt{T_1}} + V |u| v \frac{C_{f_2}}{\sqrt{T_2}}.
\]

**Proof.** Using the two-layer neural net approximation in (46), one obtains

\[
    \|f - f_{T_1, T_2, v}\|_2 \leq \|f - f_{1, T_1, v}(f_2)\|_2 + \|f_{1, T_1, v}(f_2) - f_{1, T_1, v}\|_2
\]

\[
    = \|f_1(f_2) - f_{1, T_1, v}(f_2)\|_2 + \left\| \sum_{i=1}^{T_1} c_i \phi_v \left( u_i \cdot \sum_{j=1}^{T_2} k_j \phi(w_j \cdot x + b_j) + d + d_i' \right) \right\|_2
\]

\[
    \leq \frac{C_{f_1}}{\sqrt{T_1}} + \sum_{i=1}^{T_1} c_i |u_i| f_2 - f_{2, T_2}\|_2
\]

\[
    \leq \frac{C_{f_1}}{\sqrt{T_1}} + V |u| v \frac{C_{f_2}}{\sqrt{T_2}}. \quad \square
\]
A fascinating result of Kolmogorov [13] gives a decomposition of any continuous function of several variables into superpositions of functions of one variable and sums. See, for example, Lorentz [14] for a discussion in English. The decomposition takes the form

\[ f(x_1, \ldots, x_d) = \sum_{q=1}^{2d+1} g \left( \sum_{p=1}^{d} c_p \phi_q(x_p) \right). \]

Kolmogorov’s representation actually uses a superposition of increasing functions with Lipschitz bounds for his inner layer, not unlike our neural network representation here. For handling arbitrary continuous functions, the functions \( \phi_q \) chosen in the Kolmogorov representation, where \( \phi_q \) does not depend on \( f \), are typically not smooth. Kolmogorov has also shown that the functions \( \phi_q \) used in the decomposition are less smooth compared to the target function. However, \( g_q \) depends on \( f \).

Acknowledgments

The author thanks the editor and the two anonymous referees for their helpful comments and corrections.

Appendix

The following lemma from [4] is used in the proof of Lemma 2.1. It draws upon the theory of Vapnik–Červonenkis classes of sets [20], and the concepts of VC-dimension and pseudo-dimension (see [15,16,9]).

**Lemma A.1.** Let a parametrized class of sets \( \mathcal{H} = \{ H_\xi : \xi \in \Xi \} \) in \( \mathbb{R}^d \) be given where \( \Xi \) is a measurable space. Let \( \tilde{\mathcal{H}} = \{ \tilde{H}_x : x \in \mathbb{R}^d \} \) with \( \tilde{H}_x = \{ \xi : x \in H_\xi \} \) be the dual class of sets in \( \Xi \) parametrized by \( x \). If \( \tilde{\mathcal{H}} \) has pseudo-dimension \( D \) and if \( h \) is a function in the convex hull of the indicators of sets in \( \mathcal{H} \) which possesses an integral representation

\[ h(x) = \int \mathbb{1}_{H_\xi}(x) P(d\xi) \quad \text{for } x \in B, \]

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

\[ \sup_{x \in B} |h_T(x) - h(x)| \leq 34 \sqrt{\frac{D}{T}}. \] (49)

For the class of unit step functions \( \phi(a \cdot x + b) \) (as well as the classes of half-spaces \( \mathbb{1}_{\{ \phi_v(a \cdot x + b) > 0 \}} \) generated by the ramp sigmoids \( \phi_v(a \cdot x + b) \)), the pseudo-dimension and the VC-dimension \( D \) coincide and take the value \( d + 1 \) [15,9].

References