Nonclosedness of the Set of Neural Networks in Sobolev Space

Scott Mahan*†  Emily J. King‡  Alex Cloninger*§

Abstract

We examine the closedness of the set of realized neural networks of a fixed architecture in Sobolev space. For an exactly $m$-times differentiable activation function $\rho$, we construct a sequence of neural networks $(\Phi_n)_{n \in \mathbb{N}}$ whose realizations converge in order-$(m-1)$ Sobolev norm to a function that cannot be realized exactly by a neural network. Thus, the set of realized neural networks is not closed in the order-$(m-1)$ Sobolev space $W^{m-1,p}$. We further show that this set is not closed in $W^{m,p}$ under slightly stronger conditions on the $m$-th derivative of $\rho$. For a real analytic activation function, we show that the set of realized neural networks is not closed in $W^{k,p}$ for any $k \in \mathbb{N}$. These results suggest that training a network to approximate a target function in Sobolev norm does not prevent parameter explosion. Finally, we present experimental results demonstrating that parameter explosion occurs in stochastic training regardless of the norm under which the network is trained. However, the network is still capable of closely approximating a non-network target function with network parameters that grow at a manageable rate.

Keywords—Fixed-architecture neural networks, Neural network expressivity, Closedness, Sobolev space

1 Introduction

From an approximation theory perspective, neural networks use observed training data to approximate an unknown target function. Studying topological properties of the set of neural networks will reveal what kinds of functions can be approximated by neural networks.

*Department of Mathematics, University of California San Diego. 9500 Gilman Drive #0112, La Jolla, CA 92037.
†Corresponding author. E-mail: scmahan@ucsd.edu
‡Department of Mathematics, Colorado State University. 1874 Campus Delivery, Fort Collins, CO 80523.
§Halicioglu Data Science Institute, University of California San Diego. 9500 Gilman Drive #0112, La Jolla, CA 92037.
In particular, closedness of the set of networks is a topological property of interest. If this set is closed with respect to some norm, then one can construct a sequence of neural networks converging to a target function if and only if that target function is itself a network. On the other hand, nonclosedness would mean that neural networks can approximate target functions that are not networks themselves.

To allow neural networks to approximate a wider class of functions, the number of nodes in the network can be increased. Alternatively, the parameters of the network can be allowed to grow without bound. However, networks are often trained with a fixed size and some regularization of the parameters. We will study the closedness properties of the set of neural networks of a fixed size, but we enhance the expressiveness of the networks by not bounding the parameters.

Hornik’s Universal Approximation Theorem shows that neural networks with only one hidden layer can approximate any $p$-integrable function to arbitrary accuracy, as long as the number of hidden nodes is allowed to grow without bound (Hornik, 1991). Other approximation theorems show that neural networks are dense in other function classes, depending on the properties of the activation function, but most of these results allow the depth or width of the network to vary (Cybenko, 1989; Hornik et al., 1989; Krizhevsky et al., 2012). Hornik’s result suggests that the set of realized neural networks is not closed, since not every $p$-integrable function can be represented exactly by a neural network. However, this is a result about the set of networks of any width.

In practice, the architecture of a neural network is fixed before the learning process begins. Hence, it is compelling to consider properties of the set of neural networks with a fixed architecture. Petersen, Raslan, and Voigtlaender discuss the topological properties of this set (Petersen et al., 2019, 2020). Among other results, they prove that for most commonly used activation functions, the set is not closed with respect to $L^p$ norms. However, the set of all neural networks with a fixed architecture and uniformly bounded parameters is closed, indicating that learning a non-network target function requires the parameters to explode. They speculate in (Petersen et al., 2019) that the set of neural networks may be closed in Sobolev space, where convergence is stronger. However, we show that this is not true, extending their nonclosedness results to convergence in Sobolev norm under additional smoothness assumptions on the activation function.

In some cases, such as network compression or distillation, we have data about the derivatives of the target function in addition to the training data. In these instances, one can train a network to learn the target function and its derivatives. This approach, introduced in (Czarnecki et al., 2017) as Sobolev training, often requires less training data and performs better on testing data. Hence, it is natural to consider the theoretical properties of neural networks in Sobolev space, as we do in this paper. We also provide some experimental results using Sobolev training, where we are able to approximate non-network target functions in Sobolev norm. This result indicates that the set of realized neural networks is indeed not closed in Sobolev space, but also that Sobolev training does not prevent parameter explosion and allows us to approximate functions on the boundary of the set of realized networks. Our experiment exhibits slow parameter growth relative to a fast decrease in approximation error, but there may be target functions that require much faster parameter growth to approximate.
1.1 Contributions of this Work

Our work considers the set of realizations of neural networks with a fixed architecture and a fixed nonlinear transformation (called the activation function). In particular, we study the closedness of this set of realized neural networks in Sobolev space. Our main contributions are:

1) We establish in Theorem 3.1 that for an \( m \)-times differentiable activation function, the set of realized neural networks is not closed in order-\((m-1)\) Sobolev space \( W^{m-1,p} \). We prove this result by constructing a sequence of neural networks that converges in Sobolev norm to a target function that is not a neural network.

2) We extend the nonclosedness result of Theorem 3.1 to \( W^{m,p} \) under an additional assumption on the activation function.

3) For real analytic activation functions, Theorem 3.3 shows that the set of realized neural networks in not closed in any order Sobolev space.

4) We conduct some experiments in Section 5 demonstrating that neural networks can approximate target functions that require increasingly large parameters. Our example achieves a fast decay in approximation error with a relatively slow growth in the network parameters, which may not be the case for other non-network target functions.

Our nonclosedness results all indicate that neural networks can be trained to approximate non-network target functions in Sobolev norm. However, we will see that doing so will necessarily cause an explosion of network parameters. Thus, the training process may be difficult in practice, or regularization techniques may prevent a network from approximating a non-network target function. In our experiments, we train a sequence of networks to approximate a non-network target function in Sobolev norm. The networks are able to closely approximate this target function, providing further evidence that the set of realized neural networks is not closed in Sobolev space. Moreover, the ability to numerically approximate functions on the boundary of the set of realized neural networks speaks to the expressiveness of neural networks in practice. An interesting further question is whether some target functions can only be closely approximated by parameters so large that that the network training process fails.

1.2 Outline of this Paper

Our work first provides background material on neural networks, then discusses the closedness of realized neural networks in Sobolev space, and finally presents some related numerical results. Section 2 lays out the definitions and notation required for the rest of the paper. In Section 3 we state our main results that the set of realized neural networks is not closed in Sobolev space under reasonable conditions on the activation function. On the other hand, Section 4 studies realizations of networks with bounded parameters, and presents a result that the set of these realizations is closed in Sobolev space. We provide experimental results in Section 5 that demonstrate the nonclosedness of the set of realized neural networks.
and show that a particular non-network target function can indeed be approximated by a
sequence of networks with slowly increasing parameters. Section 6 concludes the paper.

2 Notation and Definitions

We first define a neural network. Every network has an architecture which specifies the
input dimension, the number of layers, and the number of nodes in each layer. In addition,
the network consists of matrix-vector pairs that determine the affine transformation between
consecutive layers.

Definition 2.1. \textit{Petersen et al.} \cite{Petersen2020} Let \( d, L \in \mathbb{N} \). A neural network \( \Phi \) with input
dimension \( d \) and \( L \) layers is a sequence of matrix-vector pairs
\[
\Phi = ((A_1, b_1), \ldots, (A_L, b_L)),
\]
where \( N_0 = d \) and \( N_1, \ldots, N_L \in \mathbb{N} \), and where each \( A_l \) is an \( N_l \times N_{l-1} \) matrix, and
\( b_l \in \mathbb{R}^{N_l} \). We call \( (d, N_1, \ldots, N_L) \) the architecture of \( \Phi \). \( N_L \) is the output dimension.
Define \( \mathcal{NN}(d, N_1, \ldots, N_L) \) to be the set of all neural networks \( \Phi \) with architecture
\( (d, N_1, \ldots, N_L) \).

To emphasize the role of the activation function, we distinguish between a neural network
and a realization of a neural network. A realization of a network is a function defined by
alternately applying the affine transformations of the network and the activation function.
We also define the set of all realizations of neural networks with a fixed architecture and
the same activation function.

Definition 2.2. \textit{Petersen et al.} \cite{Petersen2020} Let \( \Phi \) be a neural network, \( \Omega \subset \mathbb{R}^d \), and \( \rho : \mathbb{R} \to \mathbb{R} \).
The realization of \( \Phi \) with activation function \( \rho \) over \( \Omega \) is the function \( R^\Omega_{\rho}(\Phi) : \Omega \to \mathbb{R}^{N_L} \) defined by
\[
R^\Omega_{\rho}(\Phi)(x) = W_L(\rho(W_{L-1}(\cdots \rho(W_1(x)))))
\]
where the affine transformation \( W_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l} \) is defined by \( W_l(x) = A_l x + b_l \) and \( \rho \) is
evaluated componentwise. Define \( R^\Omega_{\rho} \) to be the realization map \( \Phi \mapsto R^\Omega_{\rho}(\Phi) \), and let
\[
\mathcal{RN}^\Omega_{\rho}(d, N_1, \ldots, N_L) := R^\Omega_{\rho}(\mathcal{NN}(d, N_1, \ldots, N_L)).
\]
We call \( \mathcal{RN}^\Omega_{\rho}(d, N_1, \ldots, N_L) \) the set of \( \rho \)-realizations of networks with architecture
\( (d, N_1, \ldots, N_L) \) over \( \Omega \).

We will sometimes need to concatenate networks, which creates a new neural network
consisting of the matrix-vector pairs of the first network followed by the pairs of the second
network.

Definition 2.3. \textit{Petersen et al.} \cite{Petersen2020} Let \( \Phi_1 = ((A_1^1, b_1^1), \ldots, (A_{L_1}^1, b_{L_1}^1)) \) and \( \Phi_2 =
((A_1^2, b_1^2), \ldots, (A_{L_2}^2, b_{L_2}^2)) \) be two neural networks such that the input dimension of \( \Phi_1 \) equals
the output dimension of \( \Phi_2 \). Then
\[
\Phi_1 \bullet \Phi_2 := ((A_1^1, b_1^1), \ldots, (A_{L_2}^2, b_{L_2}^2), (A_1^1 A_2^2, A_1^1 b_{L_2}^2 + b_1^1), (A_2^1, b_2^1), \ldots, (A_{L_1}^1, b_{L_1}^1))
\]
defines a neural network with \( L_1 + L_2 - 1 \) layers. We call \( \Phi_1 \bullet \Phi_2 \) the concatenation of
\( \Phi_1 \) and \( \Phi_2 \).
Note that for any activation function \( \rho : \mathbb{R} \to \mathbb{R} \) and any \( \Omega \subset \mathbb{R}^{d_2} \), we have \( R^\Omega_\rho(\Phi_1 \circ \Phi_2) = R^\Omega_\rho(\Phi_1) \circ R^\Omega_\rho(\Phi_2) \), where \( d_i \) is the input dimension of \( \Phi_i \). That is, concatenation of neural networks corresponds to function composition of the realizations of those networks.

For a fixed network architecture \((d,N_1, \ldots, N_L)\), we want to consider the closedness of the set \( \mathcal{RNN}^\Omega_\rho(d,N_1, \ldots, N_L) \) in Sobolev space. We define Sobolev space below.

**Definition 2.4.** Let \( k \in \mathbb{N} \), let \( \Omega \subset \mathbb{R}^k \) be measurable with non-empty interior, and let \( 1 \leq p \leq \infty \). The **Sobolev space** \( W^{k,p}(\Omega) \) consists of all functions \( f \) on \( \Omega \) such that for all multi-indices \( \alpha \) with \( |\alpha| \leq k \), the mixed partial derivative \( f^{(\alpha)} := D^\alpha f \) exists in the weak sense and belongs to \( L^p(\Omega) \). That is,

\[
W^{k,p}(\Omega) = \{ f \in L^p(\Omega) : D^\alpha f \in L^p(\Omega) \text{ for all } |\alpha| \leq k \}.
\]

The number \( k \) is the order of the Sobolev space. The norm

\[
\| f \|_{W^{k,p}(\Omega)} := \sum_{|\alpha| \leq k} \| D^\alpha f \|_{L^p(\Omega)}
\]

makes \( W^{k,p}(\Omega) \) a Banach space for any \( k \in \mathbb{N} \) and any \( p \in [1, \infty] \). Note that \( W^{0,p}(\Omega) = L^p(\Omega) \).

## 3 Nonclosedness in Sobolev Space

In [Petersen et al., 2020] it is shown that \( \mathcal{RNN}^{[-B,B]^d}_\rho(d,N_1, \ldots, N_{L-1},1) \) is *not* closed in \( L^p([-B,B]^d) \) for any \( p \in (0,\infty) \), under mild assumptions satisfied by most commonly used activation functions (including ReLU, the rectified linear unit). Moreover, this set of realizations of neural networks of a fixed architecture is not closed in \( C([-B,B]^d) \) with respect to the \( L^\infty \) norm for most commonly used activation functions. However, the set of ReLU-realizations of two-layer networks is closed in \( C([-B,B]^d) \). These results are shown for \([-B,B]^d \), but generalize to any compact set \( \Omega \subset \mathbb{R}^d \) with non-empty interior.

In this work, we investigate the closedness of the set of realized neural networks in Sobolev space. Since convergence in Sobolev norm is stronger than \( L^p \) convergence, Petersen, Raslan, and Voigtlaender anticipate that \( \mathcal{RNN}^{[-B,B]^d}_\rho(d,N_1, \ldots, N_{L-1},1) \) may be closed in Sobolev space [Petersen et al., 2019]. We prove that this is often not the case. Provided that the activation function \( \rho \) is \( m \)-times differentiable with bounded derivatives, the set is not closed in \( W^{m-1,p}([-B,B]^d) \) for any \( p \in [1,\infty) \).

**Theorem 3.1.** Let \( m, d \in \mathbb{N} \), \( 1 \leq p < \infty \), and \( B > 0 \). Define \( \Omega = [-B,B]^d \). Consider a network architecture \((d,N_1, \ldots, N_{L-1},1)\) with \( L \geq 2 \) and \( N_{L-1} \geq 2 \). Suppose that \( \rho \in C^m(\mathbb{R}) \setminus C^{m+1}(\mathbb{R}) \) and all derivatives of \( \rho \) up to order \( m \) are \( p \)-integrable and bounded. Then:

- The set \( \mathcal{RNN}^{\Omega}_\rho(d,N_1, \ldots, N_{L-1},1) \) is *not* closed in \( W^{m-1,p}(\Omega) \).

- If additionally \( \rho^{(m)} \) is absolutely continuous and the weak derivative \( \rho^{(m+1)} \) exists and is in \( L^p(\Omega) \), then \( \mathcal{RNN}^{\Omega}_\rho(d,N_1, \ldots, N_{L-1},1) \) is *not* closed in \( W^{m,p}(\Omega) \).
Proof. See Appendix A.1. We construct a sequence of networks whose \( \rho \)-realizations converge in order-\((m - 1)\) (or order-\(m\)) Sobolev norm to a function that is not \(m\)-times differentiable, and hence not a \(\rho\)-realization of some network.

Section 3.1 lists several commonly used activation functions and whether they satisfy the assumptions of Theorem 3.1 for some value of \(m\).

Of course, convergence in order-\((m - 1)\) Sobolev norm is stronger than convergence in lower-order Sobolev norm, so \(\mathcal{RNN}_\rho(d, N_1, \ldots, N_{L-1}, 1)\) is not closed in lower-order Sobolev spaces either.

**Corollary 3.2.** Let \(d \in \mathbb{N}\) and \(1 \leq p < \infty\). Suppose \(\rho \in C^m(\mathbb{R}) \setminus C^{m+1}(\mathbb{R})\) with bounded derivatives up to order \(m\). Then \(\mathcal{RNN}_\rho^{(d)}(d, N_1, \ldots, N_{L-1}, 1)\) is not closed in \(W^{k,p}(\Omega)\) for any \(k \in \{0, \ldots, m - 1\}\), where \(\Omega = [-B, B]^d\).

Proof. Note that convergence in \(W^{m-1,p}(\Omega)\) implies convergence in \(W^{k,p}(\Omega)\) for all \(k \in \{0, \ldots, m - 1\}\). So we still have \(f_n \to f\) in \(W^{k,p}(\Omega)\) in the proof of Theorem 3.1 but \(f \notin \mathcal{RNN}_\rho^{(d)}(d, N_1, \ldots, N_{L-1}, 1)\).

In Theorem 3.1 we show that \(\mathcal{RNN}_\rho^{(d)}(d, N_1, \ldots, N_{L-1}, 1)\) is not closed in order-\((m - 1)\) Sobolev space for \(\rho \in C^m(\mathbb{R})\). For an analytic, bounded, and non-constant activation function \(\rho\), we extend this result and prove that \(\mathcal{RNN}_\rho^{(d)}(d, N_1, \ldots, N_{L-1}, 1)\) is not closed in any order Sobolev space.

**Theorem 3.3.** Let \(d \in \mathbb{N}\), \(1 \leq p < \infty\), and \(B > 0\). Suppose that \(\rho : \mathbb{R} \to \mathbb{R}\) is real analytic, bounded, and not constant, and that all derivatives \(\rho^{(n)}\) of \(\rho\) are bounded. Then for all possible neural network architectures \((d, N_1, \ldots, N_{L-1}, 1)\) with \(L \geq 2\) and \(N_{L-1} \geq 2\) and all \(k \in \mathbb{N}\), the set \(\mathcal{RNN}_\rho^{[-B,B]^d}(d, N_1, \ldots, N_{L-1}, 1)\) is not closed in \(W^{k,p}([-B, B]^d)\).

Proof. See Appendix A.2. We construct a sequence of networks whose \(\rho\)-realizations converge in any order Sobolev norm to an unbounded function, which cannot be a \(\rho\)-realization of some network since \(\rho\) is bounded. Lemma A.2 in the proof is interesting in its own right, as it states that realizations of neural networks with analytic activation functions can approximate the coordinate projection maps to arbitrary accuracy in Sobolev norm.

Theorems 3.1 and 3.3 show that \(\mathcal{RNN}_\rho^{[-B,B]^d}(d, N_1, \ldots, N_{L-1}, 1)\) is not closed in Sobolev space, although the smoothness of the activation function dictates the order of the Sobolev space in question. In particular, for \(\rho \in C^m(\mathbb{R}) \setminus C^{m+1}(\mathbb{R})\), Theorem 3.1 shows nonclosedness in order-\((m - 1)\) or order-\(m\) Sobolev space. Since \(\rho \in C^\infty(\mathbb{R} \setminus \{0\})\) for most commonly used activation functions (that is, the high-order differentiability typically only fails at 0), we can consider weak derivatives and ask whether \(\mathcal{RNN}_\rho^{[-B,B]^d}(d, N_1, \ldots, N_{L-1}, 1)\) may be closed in Sobolev spaces of order greater than \(m\). However, the techniques used in the proof of Theorem 3.1 no longer apply since the derivatives of order \(m + 1\) and higher are not continuous. Even so, we speculate that the nonclosedness of \(\mathcal{RNN}_\rho^{[-B,B]^d}(d, N_1, \ldots, N_{L-1}, 1)\) still holds in Sobolev spaces of order greater than \(m\).
3.1 Commonly Used Activation Functions

Though some of the assumptions on the activation function required by Theorems 3.1 and 3.3 seem strong, they are satisfied by many commonly used activation functions. Thus, the results of these theorems apply, and the set of neural networks is not closed in various orders of Sobolev space for $\Omega$ compact and $p \geq 1$. Table 1 summarizes which results apply to several common activation functions.

| Name                                      | $\rho(x)$                                      | Smoothness/Boundedness               | $\mathcal{RNN}^\Omega_p$ not closed in |
|------------------------------------------|-----------------------------------------------|--------------------------------------|----------------------------------------|
| Rectified Linear Unit (ReLU)             | $\max\{0, x\}$                                | $C(\mathbb{R})$, abs. cont., $\rho' \in L^p(\Omega)$ | $L^p(\Omega)$ (Petersen et al., 2020) |
| Exponential Linear Unit                  | $x \cdot \chi_{x \geq 0} + (e^x - 1) \cdot \chi_{x < 0}$ | $C^1(\mathbb{R})$, $\rho'$ abs. cont., $\rho'' \in L^p(\Omega)$ | $W^{1,p}(\Omega)$ |
| Softsign                                 | $\frac{x}{1+|x|}$                             | $C^1(\mathbb{R})$, $\rho'$ abs. cont., $\rho'' \in L^p(\Omega)$ | $W^{1,p}(\Omega)$ |
| Inverse Square Root Linear Unit ($a > 0$) | $x \cdot \chi_{x \geq 0} + \frac{x}{\sqrt{1+ax^2}} \cdot \chi_{x < 0}$ | $C^2(\mathbb{R})$, $\rho''$ abs. cont., $\rho''' \in L^p(\Omega)$ | $W^{2,p}(\Omega)$ |
| Inverse Square Root Unit ($a > 0$)       | $\frac{x}{\sqrt{1+ax^2}}$                    | analytic, all derivatives bounded    | $W^{k,p}(\Omega)$ for all $k$         |
| Sigmoid                                  | $\frac{1}{1+e^{-x}}$                          | analytic, all derivatives bounded    | $W^{k,p}(\Omega)$ for all $k$         |
| tanh                                     | $\frac{e^x-e^{-x}}{e^x+e^{-x}}$               | analytic, all derivatives bounded    | $W^{k,p}(\Omega)$ for all $k$         |
| arctan                                    | $\arctan(x)$                                 | analytic, all derivatives bounded    | $W^{k,p}(\Omega)$ for all $k$         |

Table 1: Many activation functions used in practice satisfy some smoothness and boundedness properties so that Theorems 3.1 and 3.3 apply. Thus, $\mathcal{RNN}^\Omega_p$ is not closed in various orders of Sobolev space for $\Omega$ compact and $p \in [1, \infty)$.

The smoothness and boundedness assumptions for the ReLU, exponential linear unit, softsign, and inverse square root linear unit can be checked by hand. The analyticity of the other activation functions is established by the following remark.

**Remark 3.4.** The inverse square root unit, sigmoid, tanh, and arctan activation functions in Table 1 are real analytic.

**Proof.** See Appendix A.3

7
For these activation functions, the set of realized neural networks in not closed in Sobolev space of certain orders. Thus, using Sobolev training still allows these networks to learn non-network target functions, although doing so will cause an explosion of parameters, which follows from Corollary 4.3 in the next section.

4 Closedness of the Set of Networks with Bounded Parameters

The nonclosedness of the set of realized neural networks is undesirable if we only want to learn network target functions or if we want to prevent explosion of network parameters. If we desire closedness, then we must modify the set of neural networks under consideration in some way. However, requiring closedness will necessarily constrain the set of target functions that we can approximate.

Modifications to enforce closedness of the set of realized neural networks may include relaxing some assumptions on the activation function or placing restrictions on the network parameters. In this section, we examine the closedness of the set of realizations of neural networks whose parameters are all bounded by the same constant. We define a norm on $\mathcal{N}(d, N_1, \ldots, N_L)$ and the set of realized neural networks with bounded norm.

**Definition 4.1.** (Petersen et al., 2020) Let $C > 0$. Define $\mathcal{N}^C(d, N_1, \ldots, N_L) = \{ \Phi \in \mathcal{N}(d, N_1, \ldots, N_L) : \| \Phi \|_{\text{total}} \leq C \}$, as the set of neural networks with uniformly bounded weights, where

$$\| \Phi \|_{\text{total}} = \max_{\ell=1,\ldots,L} \| A_{\ell} \|_{\text{max}} + \max_{\ell=1,\ldots,L} \| b_{\ell} \|_{\text{max}}$$

and $\| \cdot \|_{\text{max}}$ equals the absolute value of the entry of largest magnitude from a matrix or vector. For $\Omega \subset \mathbb{R}^d$ and $\rho : \mathbb{R} \to \mathbb{R}$, also define

$$\mathcal{R}\mathcal{N}\mathcal{A}^{\rho,C}(d, N_1, \ldots, N_L) := \mathcal{R}_\rho^\Omega (\mathcal{N}^C(d, N_1, \ldots, N_L))$$

as the set of realizations of neural networks with uniformly bounded weights and biases.

Petersen, Raslan, and Voigtlaender show that $\mathcal{R}\mathcal{N}\mathcal{A}^{[-B,B]^d}(d, N_1, \ldots, N_L-1, 1)$ is not closed in $L^p([-B, B]^d)$ or $C([-B, B]^d)$ with respect to the $L^\infty$ norm. However, the set of realizations of neural networks with uniformly bounded parameters is closed (in fact, compact) in these spaces (Petersen et al., 2020).

**Proposition 4.2.** (Petersen et al., 2020) Let $\Omega \subset \mathbb{R}^d$ be compact, $C > 0$, $p \in (0, \infty)$, and $\rho : \mathbb{R} \to \mathbb{R}$ be continuous. Then the set $\mathcal{R}\mathcal{N}\mathcal{A}^{\rho,C}(d, N_1, \ldots, N_L)$ of realized neural networks with uniformly bounded weights is compact (and hence closed) in $L^p(\Omega)$ and $C(\Omega)$ (with respect to the $L^\infty$ norm) for any architecture $(d, N_1, \ldots, N_L)$. 

8
Proof. See Proposition 3.5 in (Petersen et al., 2020). The compactness of $\mathcal{N}^\rho_\omega(d, N_1, \ldots, N_L)$ follows from the Heine-Borel Theorem. Since the realization map

$$R^\Omega_\rho : \mathcal{N}^\rho(d, N_1, \ldots, N_L) \to C(\Omega)$$

is continuous (also shown in (Petersen et al., 2020)), the image

$$\mathcal{RNN}^\rho_\omega(d, N_1, \ldots, N_L) = R^\Omega_\rho(\mathcal{N}^\rho(d, N_1, \ldots, N_L))$$

is compact in $C(\Omega)$. Since $\Omega$ is compact, $C(\Omega)$ is continuously embedded in $L^p(\Omega)$ for any $p \in (0, \infty)$, and thus $\mathcal{RNN}^\rho_\omega(d, N_1, \ldots, N_L)$ is compact in $L^p(\Omega)$ as well. 

Since Sobolev convergence is stronger than $L^p$ convergence, $\mathcal{RNN}^\rho_\omega(d, N_1, \ldots, N_L)$ is also closed in Sobolev space.

Corollary 4.3. Let $\Omega \subset \mathbb{R}^d$ be compact, $C > 0$, $p \in (0, \infty)$, $k \in \mathbb{N}$, and $\rho : \mathbb{R} \to \mathbb{R}$ be continuous. Then $\mathcal{RNN}^\rho_\omega(d, N_1, \ldots, N_L)$ is closed in $W^{k,p}(\Omega)$.

Proof. If $(f_n)_{n \in \mathbb{N}} \subset \mathcal{RNN}^\rho_\omega(d, N_1, \ldots, N_L)$ satisfies $\|f_n - f\|_{W^{k,p}(\Omega)} \to 0$ for some $f$, then $f_n \to f$ in $L^p$ norm. Thus, $f \in \mathcal{RNN}^\rho_\omega(d, N_1, \ldots, N_L)$ because this set is closed in $L^p(\Omega)$.

The nonclosedness of $\mathcal{RNN}^\rho_\omega$ in Sobolev space has significant consequences for approximating functions using neural networks. Indeed, it says that for any architecture $S = (d, N_1, \ldots, N_{L-1}, 1)$ with $L \geq 2$ and $N_{L-1} \geq 2$, there is a non-network target function $f \in \mathcal{RNN}^\rho_\omega(S) \setminus \mathcal{RNN}^\rho_\omega(S)$, where the closure can be taken with respect to Sobolev norm of the appropriate order. Combined with the closedness of the set $\mathcal{RNN}^\rho_\omega(S)$ with uniformly bounded weights, this means that if $\|R_\rho(\Omega)(\Phi_n) - f\|_{W^{k,p}(\Omega)} \to 0$ for some sequence of networks $(\Phi_n)_{n \in \mathbb{N}}$ with architecture $S$, then $\|\Phi_n\|_{\text{total}} \to \infty$. This may explain the phenomenon of exploding weights that sometimes occurs when training neural networks, and it indicates that using Sobolev training for neural networks can still lead to an explosion of parameters for some target functions.

5 Experimental Results

We now show some experimental results that demonstrate the nonclosedness of the set of realized neural networks in Sobolev space. Specifically, we use Sobolev training to produce a sequence of neural networks that approximates a non-network target function. We know from the proof of Theorem 3.1 that there is a sequence of networks that converges in Sobolev norm to the derivative of the activation function, which is not a realized neural network. In our experiments, the trained sequence of networks escapes the set of realized neural networks along a very similar path as the construction in that proof. These results demonstrate the nonclosedness of the set of realized neural networks, but also that Sobolev training can cause an explosion of parameters in a predictable way.

The activation function $\rho$ is taken to be the softsign function

$$\rho(x) = \frac{x}{1 + |x|}$$
so that $\rho \in C^1(\mathbb{R}) \setminus C^2(\mathbb{R})$. Moreover, one can verify that $\rho'$ is absolutely continuous and the weak derivative $\rho''$ is in $L^p$ for $p \in [1, \infty)$. Thus, $\mathcal{RNN}_{\rho}([-B, B]^d)(d, N_1, \ldots, N_{L-1}, 1)$ is not closed in $W^{1,p}([-B, B]^d)$ by Theorem 3.1. Specifically, we will provide numerical evidence for the nonclosedness of $\mathcal{RNN}_{\rho}([-5, 5]^1)(1, 2, 1)$ in $W^{1,2}([-5, 5])$ (that is, we take $d = 1, L = 2, p = 2, \text{ and } B = 5$) and by approximating a function on the boundary of the set of realized neural networks. Our code can be found here.

To demonstrate the nonclosedness, we train a sequence of $\rho$-realizations of networks to learn the target function

$$f(x) = \rho'(x) = \frac{1}{(1 + |x|)^2}$$

which is not $C^1$ and hence not a $\rho$-realization of some network. If we let $h_n = R_{\rho}^\mathbb{R}(\Phi_n)$ be defined as in the proof of Theorem 3.1 (see Appendix A.1), then $\|h_n - f\|_{W^{1, p}} \to 0$ as $n \to \infty$. However, by Corollary 4.3, we expect to see exploding weights. For the network $\Phi_n = (\Phi_n^1, \Phi_n^2)$, the weights from the hidden layer to the output layer are $A_n^2 = (n - n)$. Since $\|\Phi_n\|_{\text{total}} \to \infty$ as $n \to \infty$, it may be difficult to approximate the non-network target function. However, our experiments show that Sobolev training can approximate such functions.

To train the network, we run through 10000 epochs, and in each epoch we generate 1000 training points $f(x)$ with $x$ drawn uniformly from $[-5, 5]$. The network parameters are initialized as Gaussian random variables, and then updated at each epoch using an Adam optimizer (Kingma and Ba, 2017) with learning rate $0.005$. We train one network $\Phi$ to minimize $\|\Phi - f\|_{L^2}$ and another network $\Phi_S$ to minimize $\|\Phi_S - f\|_{W^{1, 2}}$ (i.e., Sobolev training). Note that Sobolev training requires us to generate training derivatives $f'(x)$ as well. Results from these training processes are shown below.

![Figure 1](image_url)

(a) Training loss. (b) The network and target function.

Figure 1: A network $\Phi$ is trained to minimize $\|\Phi - f\|_{L^2}$. Left: The $L^2$ training loss gets very close to 0, with a mean squared error less than $10^{-4}$. Right: The trained network’s output appears to closely match that of the target function.

Figure 1 provides two visualizations showing that we can train a network to learn $f$ in $L^2$ norm. The mean $L^2$ training loss between the network and the target function approaches
0, and the network output matches the target function very closely. Since we are training a \( \rho \)-realization of a neural network to learn \( f \not\in C^1 \), this supports the conclusion of Corollary 3.2 that \( \mathcal{RNN}_{\rho}^{[-5,5]}(1, 2, 1) \) is not closed in \( L^2([-5, 5]) \). Moreover, it demonstrates that we can approximate functions on the boundary of the set of realized neural networks, even though does so requires an explosion of parameters.

Figure 2 similarly provides two visualizations showing that we can train a network to approximate \( f \) in Sobolev space \( W^{1,2} \). Since we are now training a \( \rho \)-realization of a neural network to learn \( f \not\in C^1 \) in Sobolev norm, this supports the conclusion of Theorem 3.1 that \( \mathcal{RNN}_{\rho}^{[-5,5]}(1, 2, 1) \) is not closed in \( W^{1,2}([-5, 5]) \). Additionally, we see that neural networks are still expressive enough in practice to approximate non-network target functions in Sobolev norm, despite these approximations requiring parameter explosion.

Finally, figure 3 shows the explosion of parameters that is expected when we train a network to learn a non-network target function. By Corollary 4.3, convergence of the neural network to a target function which is not a realized neural network implies explosion of the parameters. The Frobenius norm of the weight matrix from the hidden layer to the output layer is plotted versus the number of training epochs.

We observe that the training loss decreases at a much faster rate than the growth of the network parameters, at least at the beginning of the training process. Indeed, both the \( L^2 \) and Sobolev losses reach a value close to their minimum within a few epochs. On the other hand, we observe approximately linear growth in the Frobenius norm of the networks, though this growth does eventually slow down due to the convergence of the training process. An interesting question for further research would be to characterize the relationship between the decrease of the training loss and the explosion of network parameters when approximating non-network target functions.

Another interesting phenomenon in Figures 1 and 2 is that Sobolev training produces a network that does not visually approximate the target function as well as \( L^2 \) training does. This is likely because Sobolev training also considers the target derivatives, and hence is
less concerned with the value of the target function itself. Thus, we observe areas where the network output lies slightly above the target function and areas where it lies slightly below. However, it may interesting to further explore the question of how well or how quickly Sobolev training approximates a target function.

6 Conclusion

In this work, we prove that the set of realized neural networks of a fixed architecture is not closed in Sobolev space under certain smoothness conditions on the activation function $\rho$. More specifically, this set is not closed in order-$(m-1)$ Sobolev space when $\rho$ is $m$-times differentiable (with bounded derivatives). Moreover, the set is not closed in any order Sobolev space when $\rho$ is smooth with bounded derivatives of all orders. We present experimental results demonstrating this nonclosedness for the softsign activation function.

The nonclosedness of the set of neural networks in Sobolev space has significant consequences for using Sobolev training to train neural networks. Most importantly, neural networks can be trained to approximate a non-network target function in Sobolev norm, but doing so requires an explosion of the network’s parameters. Our experiments also show that it is also possible to approximate a non-network target function numerically. This can be considered good or bad depending on the goal of the learning process: Sobolev training does not limit the expressivity of the network, but it also does not regularize the parameters of the network.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650112. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

SM is funded by grant NSF DGE GRFP #1650112. AC is funded by grants NSF DMS #1819222, #2012266, and Russell Sage Foundation grant 2196.

References

Adegoke, K. and Layeni, O. (2010). The higher derivatives of the inverse tangent function and rapidly convergent BBP-type formulas for pi. *Applied Mathematics E-Notes*, 10:70–75.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314.

Czarnecki, W. M., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. (2017). Sobolev training for neural networks. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 1–10.

Folland, G. (1999). *Real Analysis*. Pure and Applied Mathematics. John Wiley & Sons, New York, 2 edition.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Kingma, D. P. and Ba, J. L. (2017). Adam: A method for stochastic optimization. arXiv:1412.6980v9.

Krantz, S. G. and Parks, H. R. (2002). *A Primer of Real Analytic Functions*. Birkhäuser Advanced Texts. Birkhäuser, Boston, MA, 2 edition.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25(2):1097–1105.

Minai, A. A. and Williams, R. D. (1993). On the derivatives of the sigmoid. *Neural Networks*, 6(6):845–853.

Petersen, P., Raslan, M., and Voigtlaender, F. (2019). Unfavorable structural properties of the set of neural networks with fixed architecture. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–4.
A Proofs of Results from Section 3

A.1 Proof of Theorem 3.1

The proof of Theorem 3.1 uses the following lemma.

Lemma A.1. Let $d \in \mathbb{N}$, $L \geq 2$, $B > 0$, and $D > 0$. Set $\Omega = [-B, B]^d$. Suppose $\rho \in C^m(\mathbb{R}) \setminus C^{m+1}(\mathbb{R})$ for some $m \in \mathbb{N}$. Then there exists $\Phi \in NN(d, 1, \ldots, 1)$ with $L - 1$ layers such that $R^\Omega_\rho(\Phi)(\Omega) \supset [-D, D]$.

Proof. Since $\rho \in C^m(\mathbb{R}) \setminus C^{m+1}(\mathbb{R})$ for some $m \in \mathbb{N}$, $\rho$ is not a constant function. Moreover, since $\rho \in C^1(\mathbb{R})$, there exists an open interval $U \subset \mathbb{R}$ such that $0 \not\in \rho'(U)$. Hence, $\rho(U)$ is not a single point and must be an interval with non-empty interior. Next, choose $A_1 \in \mathbb{R}^{1 \times d}$ such that $A_1 \Omega \supset U$.

Since $U$ has non-empty interior, we can iteratively choose $A_j \in \mathbb{R}^{1 \times 1}$ such that

$$A_j \rho(A_{j-1} \rho(\cdots \rho(A_1 \Omega) \cdots)) \supset U$$

for $j = 2, \ldots, L - 2$. Again since $U$ has non-empty interior, we can choose $A_{L-1} \in \mathbb{R}^{1 \times 1}$ and $b_{L-1} \in \mathbb{R}$ such that

$$A_{L-1} \rho(A_{L-2} \rho(\cdots \rho(A_1 \Omega) \cdots)) + b_{L-1} \supset [-D, D].$$

Finally, if we let $\Phi = ((A_1, 0), \ldots, (A_{L-2}, 0), (A_{L-1}, b_{L-1}))$ then $R^\Omega_\rho(\Phi)(\Omega) \supset [-D, D]$ by construction. \qed

Proof of Theorem 3.1. Given $\rho \in C^m(\mathbb{R}) \setminus C^{m+1}(\mathbb{R})$ as in the statement of the theorem, we will construct a sequence of functions $(f_n)_{n=1}^\infty$ in $R^K_\rho(d, 1, \ldots, 1, 2, 1)$ whose $W^{m-1,p}(\Omega)$-limit $f$ is not in $C^m(\Omega)$. Since $\rho \in C^m(\mathbb{R})$ implies

$$\mathcal{N}\mathcal{N}_\rho^d(d, 1, \ldots, 1, 2, 1) \subset C^m(\Omega),$$

we have $f \not\in \mathcal{N}\mathcal{N}_\rho^d(d, 1, \ldots, 1, 2, 1)$ and hence $\mathcal{N}\mathcal{N}_\rho^d(d, 1, \ldots, 1, 2, 1)$ is not closed in $W^{m-1,p}(\Omega)$. Since $N_\ell \geq N_\ell$ for $\ell = 1, \ldots, L - 1$ implies

$$\mathcal{N}\mathcal{N}_\rho^d(d, N_1, \ldots, N_{L-1}, N_L) \subset \mathcal{N}\mathcal{N}_\rho^d(d, N_1, \ldots, N_{L-1}, N_L)$$

by Lemma 2.5 in (Petersen et al., 2020), the nonclosedness of $\mathcal{N}\mathcal{N}_\rho^d(d, N_1, \ldots, N_{L-1}, 1)$ holds for any architecture $\rho(d, N_1, \ldots, N_{L-1}, 1)$ with $L \geq 2$ and $N_{L-1} \geq 2$.  

14
To construct \((f_n)_{n=1}^\infty\), first note that \(\rho \notin C^{m+1}(\mathbb{R})\) implies \(\rho \notin C^{m+1}([-C,C])\) for some \(C > 0\). Let \(\Phi \in \mathcal{N}(d,1,\ldots,1)\) have \(L-1\) layers such that \(J(x) := R^0_\rho(\Phi)(x)\) satisfies \(J(\Omega) \supset [-C,C]\) as in Lemma [A.1]. Define a sequence of neural networks \(\Phi_n = ((A^1_1, b^1_1), (A^2_2, b^2_2)) \in \mathcal{N}(1,2,1)\) by

\[
A^1_1 = \left(\begin{array}{c} 1 \\ 1 \end{array}\right) \in \mathbb{R}^{2 \times 1}, \quad b^1_1 = \left(\begin{array}{c} 1/n \\ 0 \end{array}\right) \in \mathbb{R}^2, \quad A^2_2 = (n-n) \in \mathbb{R}^{1 \times 2}, \quad b^2_2 = 0 \in \mathbb{R}^1.
\]

Next define \(h_n : \mathbb{R} \to \mathbb{R}\) by

\[
h_n(x) = R^0_\rho(\Phi_n)(x) = n\rho(x + 1/n) - n\rho(x) = \rho(x + 1/n) - \rho(x)
\]

and let \(f_n = h_n \circ J = R^0_\rho(\Phi_n \bullet \Phi)\) so that

\[
f_n(x) = \frac{\rho(J(x) + 1/n) - \rho(J(x))}{1/n}
\]

for \(n \in \mathbb{N}\). Notice that \(f_n \in C^m([\mathbb{R}^d])\) for all \(n \in \mathbb{N}\), with

\[
\frac{\partial^l}{\partial x_1 \cdots \partial x_l} f_n(x) = \sum_{\pi \in \Pi} \left(\frac{\rho(|\pi|)(J(x) + 1/n) - \rho(|\pi|)(J(x))}{1/n}\right) \cdot \prod_{B \in \pi} \frac{\partial^{|B|} J(x)}{\partial x_j} = \frac{\partial^l}{\partial x_1 \cdots \partial x_l} \rho'(J(x))
\]

for \(l \leq m\) by Faà di Bruno's formula (here the sum is taken over all set partitions \(\pi\) of the set \(\{1,\ldots,l\}\) and the product is taken over all blocks \(B\) in the partition \(\pi\)). We then have

\[
\lim_{n \to \infty} \frac{\partial^l}{\partial x_1 \cdots \partial x_l} f_n(x) = \sum_{\pi \in \Pi} \left(\frac{\rho(|\pi|)(J(x))}{\rho(|\pi|)(J(x))}\right) \cdot \prod_{B \in \pi} \frac{\partial^{|B|} J(x)}{\partial x_j} = \frac{\partial^l}{\partial x_1 \cdots \partial x_l} \rho'(J(x))
\]

pointwise for \(l < m\). For \(l = 0,\ldots,m - 1\), \(\rho^{(l+1)}\) is continuous and bounded, so \(\rho^{(l)}\) is Lipschitz continuous and therefore absolutely continuous on every bounded interval. Moreover, \(\rho^{(l+1)} \in L^p(\mathbb{R})\) by assumption. Hence, by Exercise 8.9 in [Folland 1999], the pointwise derivative of \(\rho^{(l)}\) agrees with its strong \(L^p\) derivative. Thus, we have

\[
\lim_{n \to \infty} \left\| \frac{\rho^{(l)}(J(x) + 1/n) - \rho^{(l)}(J(x))}{1/n} - \rho^{(l+1)}(J(x)) \right\|_{L^p(\Omega)} = 0
\]

for \(l = 1,\ldots,m - 1\). Since all terms in Equation [1] are bounded, it follows that

\[
\lim_{n \to \infty} \left\| f_n - \rho'(J(\cdot)) \right\|_{W^{m-1,p}(\Omega)} = \lim_{n \to \infty} \sum_{|\alpha| \leq m-1} \left\| D^\alpha (f_n - \rho'(J(\cdot))) \right\|_{L^p(\Omega)} = 0.
\]

That is, the sequence \((f_n)_{n=1}^\infty\) converges in \(W^{m-1,p}(\Omega)\) to the function \(\Omega \to \mathbb{R}\) given by \(x \mapsto f(x) := \rho'(J(x))\). Since \([-C,C] \subset J(\Omega)\), we have \(f \notin C^m(\Omega)\) and therefore \(f \notin R^0_\rho(d,1,\ldots,1,2,1)\). Hence, \(R^0_\rho(d,1,\ldots,1,2,1)\) is not closed in \(W^{m-1,p}(\Omega)\).
If in addition $\rho^{(m)}$ is absolutely continuous and the weak derivative $\rho^{(m+1)}$ exists and is in $L^p(\mathbb{R})$, then Exercise 8.9 in [Folland 1999] gives
\[
\lim_{n \to \infty} \left\| \frac{\rho^{(m)}(J(\cdot) + 1/n) - \rho^{(m)}(J(\cdot))}{1/n} - \rho^{(m+1)}(J(\cdot)) \right\|_{L^p(\Omega)} = 0
\]
and therefore
\[
\lim_{n \to \infty} \left\| f_n - \rho'(J(\cdot)) \right\|_{W^{m,p}(\Omega)} = \lim_{n \to \infty} \sum_{|\alpha| \leq m} \| D^\alpha (f_n - \rho'(J(\cdot))) \|_{L^p(\Omega)} = 0.
\]
Hence, $R^\Omega_\rho(d,1,\ldots,1,2,1)$ is not closed in $W^{m,p}(\Omega)$ under this additional assumption on $\rho$.

\[\square\]

### A.2 Proof of Theorem 3.3

The proof of Theorem 3.3 uses an intermediate result which states that neural networks can approximate coordinate projection maps in Sobolev norm under certain conditions on the activation function.

**Lemma A.2.** Let $\rho : \mathbb{R} \to \mathbb{R}$ be real analytic, bounded, and not constant. Suppose that all derivatives $\rho^{(n)}$ of $\rho$ are bounded. Then for every $d$, $L$, $k \in \mathbb{N}$, $p \in [1, \infty)$, $\epsilon > 0$, $B > 0$, and every $i \in \{1, \ldots, d\}$, one can construct a neural network $\tilde{\Phi}^B_{\epsilon,i,k,p} \in \mathcal{NN}(d,1,\ldots,1)$ with $L$ layers such that
\[
\left\| R^\Omega_\rho[-B,B]^d(\tilde{\Phi}^B_{\epsilon,i,k,p}) - P_i \right\|_{W^{k,p}([-B,B]^d)} \leq \epsilon,
\]
where $P_i(x) = x_i$.

**Proof.** Without loss of generality, we only consider the case $\epsilon \leq 1$. Set $\Omega := [-B,B]^d$ and $\epsilon' := \epsilon / ((k + 1)(2B)^d L)$. Since $\rho$ is in $C^\infty(\mathbb{R})$ and is not constant, there exists $x_0 \in \mathbb{R}$ such that $\rho'(x_0) \neq 0$. For each $C > 0$, define $\Phi^C_1 := ((A_1, b_1), (A_2, b_2)) \in \mathcal{NN}(d,1,1)$ by
\[
A_1 = \left( \frac{1}{C} \ 0 \ \cdots \ 0 \right) \in \mathbb{R}^{1 \times d}, \quad b_1 = x_0 \in \mathbb{R}, \quad A_2 = \frac{C}{\rho'(x_0)} \in \mathbb{R}^{1 \times 1}, \quad b_2 = -\frac{C \rho(x_0)}{\rho'(x_0)} \in \mathbb{R}
\]
so that
\[
R^\Omega_\rho(\Phi^C_1)(x) = \frac{C}{\rho'(x_0)} \cdot \rho \left( \frac{x_1}{C} + x_0 \right) - \frac{C \rho(x_0)}{\rho'(x_0)}.
\]
where $x = (x_1, \ldots, x_d)$. Notice that
\[
\lim_{C \to \infty} R^\Omega_\rho(\Phi^C_1)(x) = \lim_{C \to \infty} x_1 \cdot \frac{1}{\rho'(x_0)} \cdot \frac{\rho(x_0 + x_1/C) - \rho(x_0)}{x_1/C} = x_1
\]
pointwise. In fact, there exists some $C_0 > 0$ such that $|R^\Omega_\rho(\Phi^C_1)(x) - x_1| \leq \epsilon'$ for all $x \in (-B - L \epsilon, B + L \epsilon)^d$ and all $C \geq C_0$. To see why, notice that by the definition of the derivative there exists $\delta > 0$ such that
\[
\left| \frac{\rho(x_0 + t) - \rho(x_0)}{t} - \rho'(x_0) \right| \leq \frac{\|\rho'(x_0)\| \cdot \epsilon'}{1 + B + L}
\]
(2)
for all \( t \in \mathbb{R} \) with \( |t| \leq \delta \). Set \( C_0 := (B + L)/\delta \) and let \( C \geq C_0 \) be arbitrary. Since \( \epsilon \leq 1 \), every \( x \in (-B - L_\epsilon, B + L_\epsilon)^d \) satisfies \( |x_1| \leq B + L \). If we set \( t = x_1/C \), then
\[
|t| = |x_1|/C \leq (B + L)/C \leq (B + L)/C_0 = \delta.
\]
It follows that
\[
|R^\Omega_\rho(\tilde{\Phi}^C_1)(x) - x_1| = \left| \frac{C}{\rho'(x_0)} \cdot \rho \left( \frac{x_1}{C} + x_0 \right) - \frac{C \rho(x_0)}{\rho'(x_0)} - x_1 \right|
\]
\[
= \frac{C}{\rho'(x_0)} \left| \rho \left( x_0 + \frac{x_1}{C} \right) - \rho(x_0) - \rho'(x_0) \cdot \frac{x_1}{C} \right|
\]
\[
= \frac{C}{\rho'(x_0)} \left| \rho(x_0 + t) - \rho(x_0) - \rho'(x_0)t \right|
\]
\[
\leq \frac{C}{\rho'(x_0)} \cdot \left| \rho'(x_0) \right| \cdot \epsilon' \cdot |t|
\]
\[
= \frac{|x_1|}{1 + B + L} \cdot \epsilon' \leq \epsilon'
\]
for all \( x \in (-B - L_\epsilon, B + L_\epsilon)^d \) and all \( C \geq C_0 \). We also have
\[
\frac{\partial}{\partial x_1} R^\Omega_\rho(\tilde{\Phi}^C_1)(x) = \frac{1}{\rho'(x_0)} \cdot \rho' \left( \frac{x_1}{C} + x_0 \right) \xrightarrow{C \to \infty} 1
\]
pointwise. Note that \( \rho' \) is Lipschitz because \( \rho'' \) is bounded. Hence, if we define \( C_1 := B \cdot \operatorname{Lip}(\rho')/\left(\|\rho'(x_0)\| \cdot \hat{\epsilon}_1\right) \) (where \( \hat{\epsilon}_1 > 0 \) will be chosen later) then
\[
\left| \frac{\partial}{\partial x_1} R^\Omega_\rho(\tilde{\Phi}^C_1)(x) - 1 \right| = \left| \frac{1}{\rho'(x_0)} \cdot \rho' \left( \frac{x_1}{C} + x_0 \right) - 1 \right|
\]
\[
\leq \frac{1}{\rho'(x_0)} \left| \rho' \left( x_0 + \frac{x_1}{C} \right) - \rho'(x_0) \right|
\]
\[
\leq \frac{1}{\rho'(x_0)} \operatorname{Lip}(\rho') \frac{x_1}{C}
\]
\[
\leq \frac{B \cdot \operatorname{Lip}(\rho')}{|\rho'(x_0)|} \cdot \frac{1}{C_1}
\]
\[
= \hat{\epsilon}_1
\]
for all \( x \in [-B, B]^d \) and all \( C \geq C_1 \). We also have
\[
\frac{\partial^n}{\partial x_1^n} R^\Omega_\rho(\tilde{\Phi}^C_1)(x) = \frac{1}{C^{n-1}} \cdot \frac{1}{\rho'(x_0)} \cdot \rho^{(n)} \left( \frac{x_1}{C} + x_0 \right) \xrightarrow{C \to \infty} 0
\]
pointwise for \( n \geq 2 \). In fact, if we define \( C_n := \max(1, \|\rho^{(n)}\|_{\infty}/(\|\rho'(x_0)\| \cdot \hat{\epsilon}_n)) \) (where \( \hat{\epsilon}_n > 0 \) will be chosen later) then
\[
\left| \frac{\partial^n}{\partial x_1^n} R^\Omega_\rho(\tilde{\Phi}^C_1)(x) - 0 \right| = \left| \frac{1}{C^{n-1}} \cdot \frac{1}{\rho'(x_0)} \cdot \rho^{(n)} \left( \frac{x_1}{C} + x_0 \right) - 0 \right|
\]
where $\rho$ for all $x \in L$ to get and define $\tilde{\Phi}$ for all $x$ for all $C \geq C_n$. Then

$$
\left| R^\Omega_\rho(\tilde{\Phi}^{C^*})(x) - x_1 \right| \leq \epsilon'
$$

$$
\left| \frac{\partial}{\partial x_1} R^\Omega_\rho(\tilde{\Phi}^{C^*})(x) - 1 \right| \leq \tilde{\epsilon}_1
$$

$$
\left| \frac{\partial^n}{\partial x_1^n} R^\Omega_\rho(\tilde{\Phi}^{C^*})(x) - 0 \right| \leq \tilde{\epsilon}_n
$$

for all $x \in \Omega$. Furthermore, any partial derivatives involving $x_2, \ldots, x_n$ are identically zero. Next define a network $\tilde{\Phi}^{C^*} = ((A'_1, b'_1), (A'_2, b'_2)) \in \mathcal{NN}(1, 1, 1)$ by

$$
A'_1 = \frac{1}{\epsilon^2} \in \mathbb{R}^{1 \times 1}, \quad b'_1 = x_0 \in \mathbb{R}^1, \quad A'_2 = \frac{C^*}{\rho'(x_0)} \in \mathbb{R}^{1 \times 1}, \quad b'_2 = -\frac{C^* \rho(x_0)}{\rho'(x_0)} \in \mathbb{R}^1
$$

and define $\tilde{\Phi}^{C^*} = \tilde{\Phi}_2^{C^*} \cdots \tilde{\Phi}_1^{C^*} \in \mathcal{NN}(d, 1, \ldots, 1)$, where we take $L-2$ concatenations to get $L$ layers. Then

$$
R^\Omega_\rho(\tilde{\Phi}^{C^*})(x) = (\rho_{C^*} \circ \cdots \circ \rho_{C^*})(x_1)
$$

where $\rho_{C^*}(x) = \frac{C^*}{\rho'(x_0)} \cdot \rho\left(\frac{x_0}{\rho'(x_0)} + x_0\right) - \frac{C^* \rho(x_0)}{\rho'(x_0)}$ is applied $L$ times. Inductively, we have

$$
\left| R^\Omega_\rho(\tilde{\Phi}^{C^*})(x) - x_1 \right| \leq L \epsilon' = \frac{\epsilon}{(k + 1)(2B)^{d/p}}
$$

by applying $\rho_C$ $L$ times. Working with the derivatives $\frac{\partial^n}{\partial x_1^n}(\rho_C \circ \cdots \circ \rho_C)(x_1)$ is not as simple because the derivatives of this composition of functions will involve applications of the chain rule and product rule. However, we only apply the chain rule and product rule finitely many times, so Equation (3) and the boundedness of all derivatives of $\rho$ guarantee that $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_k$ can be chosen so that

$$
\left| \frac{\partial}{\partial x_1} R^\Omega_\rho(\tilde{\Phi}^{C^*})(x) - 1 \right| \leq \frac{\epsilon}{(k + 1)(2B)^{d/p}}
$$

$$
\left| \frac{\partial^n}{\partial x_1^n} R^\Omega_\rho(\tilde{\Phi}^{C^*})(x) - 0 \right| \leq \frac{\epsilon}{(k + 1)(2B)^{d/p}} \quad n = 2, \ldots, k.
$$

Finally, it follows that

$$
\left\| R^\Omega_\rho(\tilde{\Phi}^{C^*}) - P_1 \right\|_{W^{k,p}(\Omega)} = \sum_{|\alpha| \leq k} \left\| D^\alpha R^\Omega_\rho(\tilde{\Phi}^{C^*}) - D^\alpha P_1 \right\|_{L^p(\Omega)}
$$
Lemma A.3. Let the following lemma will be useful.

as desired. For \( i = 2, \ldots, n \) we can just permute the columns of \( A_1 \) accordingly.

We will also need to consider Sobolev convergence of compositions of functions, for which the following lemma will be useful.

**Lemma A.3.** Let \( k \in \mathbb{N} \) and \( p \in [1, \infty) \), and let \( \Omega \subset \mathbb{R}^d \) be bounded and compact. Suppose \( (g_n)_{n \in \mathbb{N}} \subset C^\infty(\Omega) \) is a sequence of functions such that \( \|g_n - h\|_{W^{k,p}(\Omega)} \to 0 \) for some \( h \in C^\infty(\Omega) \). If \( (f_n)_{n \in \mathbb{N}} \subset C^\infty(\Omega) \) is a sequence of functions whose derivatives are all bounded independent of \( n \), then

\[
\lim_{n \to \infty} \|f_n \circ g_n - f_n \circ h\|_{W^{k,p}(\Omega)} = 0.
\]

That is, if \( g_n \to h \) in Sobolev norm, then \( f_n \circ g_n \to f_n \circ h \) in Sobolev norm.

**Proof.** For any \( \ell \leq k \), we have

\[
\frac{\partial^\ell}{\partial x_1 \cdots \partial x_\ell} (f_n \circ g_n)(x) = \sum_{\pi \in \Pi} f_n^{(|\pi|)}(g_n(x)) \cdot \prod_{B \in \pi} \frac{\partial |B| g_n(x)}{\prod_{j \in B} \partial x_j}
\]

and

\[
\frac{\partial^\ell}{\partial x_1 \cdots \partial x_\ell} (f_n \circ h)(x) = \sum_{\pi \in \Pi} f_n^{(|\pi|)}(h(x)) \cdot \prod_{B \in \pi} \frac{\partial |B| h(x)}{\prod_{j \in B} \partial x_j}
\]

by Faà di Bruno’s formula, where the sum is taken over all set partitions \( \pi \) of the set \( \{1, \ldots, l\} \) and the product is taken over all blocks \( B \) in the partition \( \pi \). Subtracting these two expressions, we get a sum over \( \pi \in \Pi \) of terms of the form

\[
f_n^{(|\pi|)}(g_n(x)) \cdot \prod_{B \in \pi} \frac{\partial |B| g_n(x)}{\prod_{j \in B} \partial x_j} - f_n^{(|\pi|)}(h(x)) \cdot \prod_{B \in \pi} \frac{\partial |B| h(x)}{\prod_{j \in B} \partial x_j} = f_n^{(|\pi|)}(g_n(x)) \cdot \prod_{B \in \pi} \frac{\partial |B| g_n(x)}{\prod_{j \in B} \partial x_j} - f_n^{(|\pi|)}(g_n(x)) \cdot \prod_{B \in \pi} \frac{\partial |B| h(x)}{\prod_{j \in B} \partial x_j}
\]

(4)
\[
+ f_n^{(\pi)}(g_n(x)) \cdot \prod_{B \in \pi} \frac{\partial B | h(x) - f_n^{(\pi)}(h(x))}{\prod_{j \in B} \partial x_j} \cdot \prod_{B \in \pi} \frac{\partial B | h(x)}{\prod_{j \in B} \partial x_j}
\]

(note that we added and subtracted a cross term). For the first term \(\text{[4]}\), we have

\[
\left\| f_n^{(\pi)}(g_n(\cdot)) \prod_{B \in \pi} \frac{\partial B | g_n(\cdot) - f_n^{(\pi)}(g_n(\cdot))}{\prod_{j \in B} \partial x_j} - f_n^{(\pi)}(\cdot) \prod_{B \in \pi} \frac{\partial B | h(\cdot)}{\prod_{j \in B} \partial x_j} \right\|_{L^p(\Omega)} \leq C \left\| \prod_{B \in \pi} \frac{\partial B | g_n(\cdot)}{\prod_{j \in B} \partial x_j} - \prod_{B \in \pi} \frac{\partial B | h(\cdot)}{\prod_{j \in B} \partial x_j} \right\|_{L^p(\Omega)}
\]

for some constant \(C\) since \(f_n^{(\pi)}\) is bounded independent of \(n\). Note that for each factor in the product over \(B \in \pi\), we have

\[
\lim_{n \to \infty} \left( \frac{\partial B | g_n(\cdot)}{\prod_{j \in B} \partial x_j} - \frac{\partial B | h(\cdot)}{\prod_{j \in B} \partial x_j} \right) = 0
\]

since \(\lim_{n \to \infty} \|g_n - h\|_{W^{k,p}(\Omega)} = 0\) and \(|B| \leq \ell \leq k\). It follows that the entire product converges to 0 in \(L^p\) norm as \(n \to \infty\). Indeed, for \(\phi_1, \phi_2, \psi_1, \psi_2\) bounded, we have

\[
\|\phi_1 \phi_2 - \psi_1 \psi_2\|_{L^p(\Omega)}^p = \int_{\Omega} |\phi_1(x) \phi_2(x) - \psi_1(x) \psi_2(x)|^p dx
\]

\[
= \int_{\Omega} |\phi_1(x) \phi_2(x) - \psi_1(x) \phi_2(x) + \psi_1(x) \phi_2(x) - \psi_1(x) \psi_2(x)|^p dx
\]

\[
\leq 2^{p-1} \left( \int_{\Omega} |\phi_1(x) \phi_2(x) - \psi_1(x) \phi_2(x)|^p dx + \int_{\Omega} |\psi_1(x) \phi_2(x) - \psi_1(x) \psi_2(x)|^p dx \right)
\]

\[
\leq 2^{p-1} C_{\phi_1, \phi_2} \left( \|\phi_1 - \psi_1\|_{L^p(\Omega)}^p + \|\phi_2 - \psi_2\|_{L^p(\Omega)}^p \right)
\]

by Minkowski’s inequality, where \(C_{\phi_1, \phi_2}\) bounds both \(|\phi_1|^p\) and \(|\phi_2|^p\) on \(\Omega\). Thus, if \(\phi_{1,n} \to \psi_1\) and \(\phi_{2,n} \to \psi_2\) in \(L^p\) norm as \(n \to \infty\) with \(C_{\phi_{1,n}, \phi_{2,n}}\) independent of \(n\), then \(\phi_{1,n} \phi_{2,n} \to \psi_1 \psi_2\) in \(L^p\) norm. For the second term \(\text{[5]}\), we have

\[
\left\| f_n^{(\pi)}(g_n(\cdot)) \prod_{B \in \pi} \frac{\partial B | h(-)}{\prod_{j \in B} \partial x_j} - f_n^{(\pi)}(\cdot) \prod_{B \in \pi} \frac{\partial B | h(\cdot)}{\prod_{j \in B} \partial x_j} \right\|_{L^p(\Omega)} \leq C_h \left\| f_n^{(\pi)}(g_n(\cdot)) - f_n^{(\pi)}(\cdot) \right\|_{L^p(\Omega)}
\]

for some constant \(C_h\) since all derivatives of \(h\) are bounded because \(h \in C^\infty(\Omega)\) with \(\Omega\) compact. Since all derivatives of \(f_n\) are bounded independent of \(n\), \(f_n^{(\pi)}\) is Lipschitz with constant \(M\) independent of \(n\). Thus,

\[
\left\| f_n^{(\pi)}(g_n(\cdot)) - f_n^{(\pi)}(h(\cdot)) \right\|_{L^p(\Omega)}^p = \int_{\Omega} |f_n^{(\pi)}(g_n(x)) - f_n^{(\pi)}(h(x))|^p dx
\]

\[
\leq M^p \int_{\Omega} |g_n(x) - h(x)|^p dx
\]

20
\[ = M^p \|g_n - h\|_{L^p(\Omega)}^p \]

so the second term converges to 0 as \( n \to \infty \). It follows that

\[
\lim_{n \to \infty} \|D^\alpha (f_n \circ g_n - f_n \circ h)\|_{L^p(\Omega)} = 0
\]

for any multi-index \( \alpha \) with \( |\alpha| \leq k \), so \( \lim_{n \to \infty} \|f_n \circ g_n - f_n \circ h\|_{W^{k,p}(\Omega)} = 0. \]

**Proof of Theorem 3.3.** Set \( \Omega := [-B,B]^d \). Since \( \rho \in C^\infty(\mathbb{R}) \) is not constant, there exists \( x_0 \in \mathbb{R} \) such that \( \rho'(x_0) \neq 0 \). For each \( n \in \mathbb{N} \), define \( \Phi_n := ((A_1^n, b_1^n), (A_2^n, b_2^n)) \in \mathcal{NN}(1,2,1) \) by

\[
A_1^n = \left( \frac{1}{1/n} \right) \in \mathbb{R}^{2 \times 1}, \quad b_1^n = \begin{pmatrix} 0 \\ x_0 \end{pmatrix} \in \mathbb{R}^2, \quad A_2^n = (1, n) \in \mathbb{R}^{1 \times 2}, \quad b_2^n = -n \rho(x_0) \in \mathbb{R}
\]

so that

\[
R_{\rho}^\infty(\Phi_n^1)(x) = \rho(x) + n \rho(x/n + x_0) - n \rho(x_0) = \rho(x) + x \cdot \frac{\rho(x_0 + x/n) - \rho(x_0)}{x/n}
\]

for all \( x \in \mathbb{R} \). By Lemma A.2, there exists a sequence of neural networks \( (\Phi_n^2)_{n \in \mathbb{N}} \subset \mathcal{NN}(d,1,\ldots,1) \) with \( L - 1 \) layers such that

\[
\|R_{\rho}^\infty(\Phi_n^2) - P_1\|_{W^{k,p}(\Omega)} \leq \frac{1}{n}
\]

for each \( n \in \mathbb{N} \). Define \( \Phi_n := \Phi_n^1 \bullet \Phi_n^2 \in \mathcal{NN}(d,1,\ldots,1,2,1) \). Consider the map \( F : \mathbb{R}^d \to \mathbb{R} \) defined by \( F(x) = \rho(x_1) + \rho'(x_0)x_1 \). We have

\[
\|R_{\rho}^\Omega(\Phi_n) - F\|_{W^{k,p}((-B,B)^d)} = \left\| R_{\rho}^\Omega(\Phi_n) - R_{\rho}^\infty(\Phi_n^1) \circ P_1 + R_{\rho}^\infty(\Phi_n^1) \circ P_1 - F \right\|_{W^{k,p}(\Omega)}
\]

\[
\leq \left\| R_{\rho}^\infty(\Phi_n^1) \circ R_{\rho}^\infty(\Phi_n^2) - R_{\rho}^\infty(\Phi_n^1) \circ P_1 \right\|_{W^{k,p}(\Omega)} + \left\| R_{\rho}^\infty(\Phi_n^1) \circ P_1 - F \right\|_{W^{k,p}(\Omega)}.
\]

Since \( \frac{d^\ell}{dx^\ell} R_{\rho}^\infty(\Phi_n^1)(x) = \rho^{(\ell)}(x) + \rho^{(\ell)}(x_n + x_0)/n^{\ell-1} \), each derivative of \( R_{\rho}^\infty(\Phi_n^1) \) is bounded independent of \( n \). Combined with equation (6) and the boundedness of all derivatives of \( \rho \), this implies that

\[
\left\| R_{\rho}^\infty(\Phi_n^1) \circ R_{\rho}^\infty(\Phi_n^2) - R_{\rho}^\infty(\Phi_n^1) \circ P_1 \right\|_{W^{k,p}(\Omega)} \xrightarrow{n \to \infty} 0
\]

by Lemma A.3. Next note that

\[
R_{\rho}^\infty(\Phi_n^1) \circ P_1(x) = \rho(x_1) + x_1 \cdot \frac{\rho(x_0 + x_1/n) - \rho(x_0)}{x_1/n} \xrightarrow{n \to \infty} \rho(x_1) + \rho'(x_0)x_1 = F(x)
\]

pointwise. An argument similar to the one in the proof of Theorem 3.1 shows that

\[
\left\| R_{\rho}^\infty(\Phi_n^1) \circ P_1 - F \right\|_{W^{k,p}(\Omega)} \xrightarrow{n \to \infty} 0
\]

21
since the pointwise derivative agrees with the strong $L^p$ derivative. In total, we have

$$\|R^{\Omega}_\rho(\Phi_n) - F\|_{W^{k,p}(\Omega)} \xrightarrow{n \to \infty} 0.$$  

However, $F|_{\Omega} \notin \mathcal{RNN}(d, 1, \ldots, 1, 2, 1)$. To see why suppose we had $F|_{\Omega} = R^{\Omega}_\rho(\Psi)$ for some neural network $\Psi$. Since $\rho$ is analytic, $F$ and $R^{\mathbb{R}^d}_\rho(\Psi)$ are both analytic and coincide on $\Omega = [-B, B]^d$. Hence, we must have $F = R^{\mathbb{R}^d}_\rho(\Psi)$ on all of $\mathbb{R}^d$. Note that $F(x) = \rho(x_1) + \rho'(x_0)x_1$ is unbounded because $\rho$ is bounded and $\rho'(x_0) \neq 0$. However, $R^{\mathbb{R}^d}_\rho(\Psi)$ is bounded because $\rho$ is. This contradicts $F = R^{\mathbb{R}^d}_\rho(\Psi)$, so $F|_{\Omega} \neq R^{\Omega}_\rho(\Psi)$. Since

$$\|R^{\Omega}_\rho(\Phi_n) - F\|_{W^{k,p}((-B,B)^d)} \xrightarrow{n \to \infty} 0$$

with $R^{\Omega}_\rho(\Phi_n) \in \mathcal{RNN}(d, 1, \ldots, 1, 2, 1)$ but $F|_{\Omega} \notin \mathcal{RNN}(d, 1, \ldots, 1, 2, 1)$, we have shown that $\mathcal{RNN}(d, 1, \ldots, 1, 2, 1)$ is not closed. \hfill $\square$

### A.3 Activation Function Assumptions

We can use simple properties of real analytic functions to see that the last four activation functions in Table 1 are analytic.

**Proof.** We will use several properties of real analytic functions from (Krantz and Parks, 2002). First, note that all polynomials are real analytic, and the square root function is real analytic on $(0, \infty)$. Also, compositions of real analytic functions, where the range of the inner function is a subset of the domain of the outer function, are again real analytic. Moreover, quotients of real analytic functions are real analytic, provided that the denominator never vanishes. It follows that the inverse square root unit

$$\rho(x) = \frac{x}{\sqrt{1 + ax^2}}$$

is real analytic. Next, note that $e^x$ is analytic because it can be expressed as a power series on all of $\mathbb{R}$. Thus, the sigmoid $\rho(x) = \frac{1}{1 + e^{-x}}$ and the hyperbolic tangent function $\rho(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ are real analytic, since the denominators never vanish. Finally, if a function has an analytic derivative, then the function is analytic itself. Since the derivative of $\rho(x) = \arctan(x)$ is $\rho'(x) = \frac{1}{1+x^2}$ and this is analytic, the inverse tangent function is real analytic. \hfill $\square$