Topological Expressivity of ReLU Neural Networks

Ekin Ergen ERGEN@MATH.TU-BERLIN.DE and **Moritz Grillo** GRILLO@MATH.TU-BERLIN.DE
Technische Universität Berlin, Berlin, Germany

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study the expressivity of ReLU neural networks in the setting of a binary classification problem from a topological perspective. Recently, empirical studies showed that neural networks operate by changing topology, transforming a topologically complicated data set into a topologically simpler one as it passes through the layers. This topological simplification has been measured by Betti numbers, which are algebraic invariants of a topological space. We use the same measure to establish lower and upper bounds on the topological simplification a ReLU neural network can achieve with a given architecture. We therefore contribute to a better understanding of the expressivity of ReLU neural networks in the context of binary classification problems by shedding light on their ability to capture the underlying topological structure of the data. In particular the results show that deep ReLU neural networks are exponentially more powerful than shallow ones in terms of topological simplification. This provides a mathematically rigorous explanation why deeper networks are better equipped to handle complex and topologically rich data sets.

Keywords: ReLU neural networks, Betti numbers, expressivity, topology

1. Introduction

Neural networks are at the core of many AI applications. A crucial task when working with neural networks is selecting the appropriate architecture to effectively tackle a given problem. Therefore, it is of fundamental interest to understand the range of problems that can be solved by neural networks with a given architecture, i.e., its *expressivity*.

In recent years, many theoretical findings have shed light on the expressivity of neural networks. Universal approximation theorems (Cybenko, 1989; Hornik, 1991) state that one hidden layer is already sufficient to approximate any continuous function with arbitrary accuracy. On the other hand, it is known that deep networks can represent more complex functions than their shallow counterparts (see e.g., Telgarsky (2016); Eldan and Shamir (2016); Arora et al. (2018)).

The measure of expressivity of a neural network should always be related to the problem it has to solve. A common scenario in which neural networks are employed is the binary classification problem, where the network serves as a classifier for a binary labeled data set. Since topological data analysis has revealed that data often has nontrivial topology, it is important to consider the topological structure of the data when dealing with a binary classification problem. Naitzat et al. (2020) show through empirical methods that neural networks operate topologically, transforming a topologically complicated data set into a topologically simple one as it passes through the layers. Given a binary labeled data set, they assume that the positively labeled and the negatively labeled points are sampled from topological spaces X_a and X_b respectively that are entangled with each other in a nontrivial way. Their experiments show that a well-trained neural network gradually disentangles the topological spaces until they are linearly separable in the end, i.e, the space X_b is mapped to the positive real line and X_a to the negative real line. From a theoretical point of view, it is of interest to determine the extent of "topological change" that can be achieved by neural networks of a particular architecture. The topological expressivity of a neural network can therefore be measured by the complexity of the most complex topological spaces it can separate and is directly related to the complexity of binary classification problems.

In this paper we investigate the topological expressivity of ReLU neural networks, which are one of the most commonly used types of neural networks (Glorot et al., 2011; Goodfellow et al., 2016). A $(L+1)$-*layer neural network (NN)* is defined by $L+1$ affine transformations $T_\ell\colon \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$, $x \mapsto A_\ell x + b_\ell$ for $A_\ell \in \mathbb{R}^{n_{\ell-1}\times n_\ell}, b_\ell \in \mathbb{R}^{n_\ell}$ and $\ell = 1,\dots,L+1$. The tuple $(n_0, n_1, \dots, n_L, n_{L+1})$ is called the *architecture*, $L+1$ the *depth*, n_ℓ the *width of the ℓ-layer*, $\max\{n_1,\dots,n_L\}$ the *width* of the NN and $\sum_{\ell=1}^{L} n_\ell$ the *size* of the NN. The entries of A_ℓ and b_ℓ for $\ell = 1, ..., L+1$ are called *weights* of the NN and the vector space of all possible weights is called the *parameter space* of an architecture. A ReLU neural network computes the function

$$F = T_{L+1} \circ \sigma_{n_L} \circ T_L \circ \cdots \circ \sigma_{n_1} \circ T_1,$$

where $\sigma_n\colon \mathbb{R}^n \to \mathbb{R}^n$ is the *ReLU function* given by $\sigma_n(x) = (\max(0, x_1), \dots, \max(0, x_n))$.

Note that the function F is piecewise linear and continuous. In fact, it is known that any continuous piecewise linear function F can be computed by a ReLU neural network (Arora et al., 2018). However, for a fixed architecture A, the class $\mathcal{F}_A$ of piecewise linear functions that is representable by this architecture is not known (Hertrich et al., 2021; Haase et al., 2023). Conveniently, in the setting of a binary classification problem we are merely interested in the *decision regions*, i.e., $F^{-1}((-\infty, 0])$ and $F^{-1}((0,\infty))$ rather than the continuous piecewise linear function F itself.

A common choice to measure the complexity of a topological space X is the use of algebraic invariants. Homology groups are the essential algebraic structures with which topological data analysis analyzes data (Dey and Wang, 2022) and hence Betti numbers as the ranks of these groups are a natural measure of topological expressivity. Intuitively, the k-th Betti number $\beta_k(X)$ corresponds to the number of $(k+1)$-dimensional holes in the space X for $k > 0$ and $\beta_0(X)$ corresponds to the number of path-connected components of X. Thus, one can argue that when a space (the support of one class of the data) has many connected components and higher dimensional holes, it is more difficult to separate this space from the rest of the ambient space, e.g., mapping it to the negative line. In Appendix A.2, we present a brief introduction to homology groups. For an in-depth discussion of the aforementioned concepts, we refer to textbooks on algebraic topology (e.g., Hatcher (2002)).

In order to properly separate X_a and X_b, the sublevel set $F^{-1}((-\infty, 0])$ of the function F computed by the neural network should have the same topological complexity as X_a. Bianchini and Scarselli (2014) measured the topological complexity of the decision region $F^{-1}((-\infty, 0])$ with the sum of all its Betti numbers. This notion of topological expressivity does not differentiate between connected components and higher dimensional holes. On the other hand, if an architecture is not capable of expressing the Betti numbers of different dimensions of the underlying topological space of the data set, then for every $F \in \mathcal{F}_A$ there is a set of data points U such that F misclassifies every $x \in U$ (Guss and Salakhutdinov, 2018). Therefore it is of fundamental interest to understand each Betti number of the decision regions, by which we define our notion of expressivity:

Definition 1 *The* topological expressivity *of a ReLU neural network $F\colon \mathbb{R}^d \to \mathbb{R}$ is defined as the vector $\beta(F) = (\beta_k(F))_{k=0,\dots,d-1} = (\beta_k(F^{-1}((-\infty, 0]))_{k=0,\dots,d-1}$.*

1.1. Main Results

Our main contribution consists of lower and upper bounds on the topological expressivity for ReLU NNs. These bounds demonstrate that the growth of Betti numbers in neural networks depends on their depth. With an unbounded depth, Betti numbers in every dimension can exhibit exponential growth as the network size increases. However, in the case of a shallow neural network, where the depth remains constant, the Betti numbers of the sublevel set are polynomially bounded in size. This implies that increasing the width of a network while keeping the depth constant prevents exponential growth in the Betti numbers. Consequently, if a data set possesses exponentially high Betti numbers (parameterized by some parameter p), accurate modeling of the data set requires a deep neural network when the size of the neural network is constrained to be polynomial in parameter p since the topological expressivity serves, as discussed above, as a bottleneck measure for effective data representation.

In Theorem 13, the lower bounds for the topological expressivity are given by an explicit formula, from which we can derive the following asymptotic lower bounds:

Corollary 2 *Let $A = (d, n_1, \dots, n_L, 1)$ with $n_L \geq 4d$ and $M = \prod_{\ell=1}^{L-1} 2 \cdot \lfloor \frac{n_\ell}{2d} \rfloor$, then there is a ReLU NN $F \colon \mathbb{R}^d \mapsto \mathbb{R}$ with architecture A such that*

(i) $\beta_0(F) \in \Omega(M^d \cdot n_L)$

(ii) $\beta_k(F) \in \Omega(M^k \cdot n_L)$ *for* $0 < k < d$.

In particular, given $\boldsymbol{v} = (v_1, \dots, v_d) \in \mathbb{N}^{d-1}$, there is a ReLU NN F of size $O\left(\log\left(\sum_{k=1}^{d-1} v_k\right)\right)$ such that $\beta_k(F) \geq v_{k+1}$ for all $k \in \{0, \dots, d-1\}$.

Corollary 2 provides a proof for a conjecture on lower bounds for the zeroth Betti number of the decision region given in the paper of Guss and Salakhutdinov (2018); in fact, it generalizes the statement to arbitrary dimensions. Furthermore, we observe that $L = 2$ hidden layers are already sufficient to increase the topological expressivity as much as we want at the expense of an increased width due to the above lower bound.

Corollary 3 *Given $v \in \mathbb{N}^d$, there exists an NN $F \colon \mathbb{R}^d \to \mathbb{R}$ of depth 2 such that $\beta_k(F) \geq v_{k+1}$ for all $k \in \{0, \dots, d-1\}$.*

We obtain the lower bound by making choices for the weights of the NN, nevertheless, we show that our construction is robust with respect to small perturbations. In fact, in Proposition 17 we prove that we actually have an open set in the parameter space in which the respective functions all have the same topological expressivity.

Using an upper bound on the number of linear regions (Serra et al., 2017), we obtain an explicit formula for an upper bound on $\beta_k(F)$ for an arbitrary architecture in Proposition 18. This gives rise to the following asymptotic upper bounds:

Corollary 4 *Let $F \colon \mathbb{R}^d \to \mathbb{R}$ be a neural network of architecture $(d, n_1, \dots, n_L, 1)$. Then it holds that $\beta_k(F) \in O\left(\left(\prod_{i=1}^{L} n_i\right)^{d^2}\right)$ for $k \in [d-2]$ and $\beta_0, \beta_{d-1} \in O\left(\left(\prod_{i=1}^{L} n_i\right)^{d}\right)$.*

By combining Corollary 2 and Corollary 4, we can conclude that there is an exponential gap in the topological expressivity between shallow and deep neural networks. This aligns with other popular measures of expressivity, such as the number of linear regions, where similar exponential gaps are known (Serra et al., 2017; Montúfar et al., 2014; Montúfar, 2017).

1.2. Related Work

1.2.1. TOPOLOGY AND NEURAL NETWORKS

Vast streams of research studying neural networks by means of topology using empirical methods (Petri and Leitão, 2020; Guss and Salakhutdinov, 2018; Naitzat et al., 2020; Li et al., 2020) as well as from a theoretical perspective (Basri and Jacobs, 2017; Melodia and Lenz, 2020; Grigsby and Lindsey, 2022; Bianchini and Scarselli, 2014; Grigsby et al., 2022a; Hajij and Istvan, 2020) have emerged recently. Bianchini and Scarselli (2014) were the first that used Betti numbers as a complexity measure for decision regions of neural networks. Their work studies NNs with sigmoidal activation functions and shows that there is an exponential gap with respect to the sum of Betti numbers between deep neural networks and neural networks with one hidden layer. However, there are no insights about distinct Betti numbers. In Guss and Salakhutdinov (2018), the decision regions of ReLU neural networks ares studied with empirical methods and an exponential gap for the zeroth Betti number is conjectured. Our results prove the conjecture and extend the results of Bianchini and Scarselli (2014) for the ReLU case (see Section 4 and Appendix). Furthermore, topological characteristics such as connectivity or boundedness of the decision regions are also investigated in (Fawzi et al., 2018; Grigsby and Lindsey, 2022; Grigsby et al., 2022a; Nguyen et al., 2018).

1.2.2. EXPRESSIVITY OF (RELU) NEURAL NETWORKS

In addition to the universal approximation theorems (Cybenko, 1989; Hornik, 1991), there is a significant amount of research on the expressivity of neural networks, e.g., indicating that deep neural networks can be exponentially smaller in size than shallow ones. For ReLU neural networks, the number of linear regions is often used as a measure of complexity for the continuous piecewise linear (CPWL) function computed by the network. It is well established that deep ReLU neural networks can compute CPWL functions with exponentially more linear regions than shallow ones, based on various results such as lower and upper bounds on the number of linear regions for a given architecture (Montúfar, 2017; Serra et al., 2017; Montúfar et al., 2014; Arora et al., 2018). We partially use techniques from aforementioned works to establish our bounds on topological expressivity, which offers the advantage of being directly related to the complexity of binary classification problems.

1.3. Notation and Definitions

A function $F\colon \mathbb{R}^d \to \mathbb{R}^d$ is continuous piecewise linear (CPWL) if there is a polyhedral complex covering $\mathbb{R}^d$, such that F is affine linear over each polyhedron of this complex. A linear region of f is a maximal connected convex subspace R such that f is affine linear on R, i.e., a full-dimensional polyhedron of the complex.[1]

For a survey on polyhedral theory in deep learning see Huchette et al. (2023), and for a general introduction to polyhedra we refer to Schrijver (1986).

We denote by $[n]$ the set $\{1, \dots, n\}$ and by $[n]_0$ the set $\{0, \dots, n\}$. We denote by $\pi_j\colon \mathbb{R}^d \to \mathbb{R}$ the projection onto the j-th component of $\mathbb{R}^d$ and by $p_j\colon \mathbb{R}^d \to \mathbb{R}^j$ the projection onto the first j components.

1. In the literature there exists also a slightly different definition of a linear region leaving out the necessity of the region being convex, but the bounds we use are all applicable to this definition of a linear region.

A crucial part of our construction is decomposing a unit cube into a varying number of small cubes. Thereby, given $\mathbf{m} = (m_1, \dots, m_L) \in \mathbb{N}^L$ and $M = \left(\prod_{\ell=1}^{L} m_\ell\right)$, the set $W^{(L,\mathbf{m},d)}_{i_1,\dots,i_d}$ is defined as the cube of volume $\frac{1}{M^d}$ with "upper right point" $\frac{1}{M} \cdot (i_1, \dots, i_d)$, i.e., the cube $\prod_{k=1}^{d}[\frac{(i_k-1)}{M}, \frac{i_k}{M}] \subset [0,1]^d$. The indices $(L, \mathbf{m}, d)$ are omitted whenever they are clear from the context.

We denote by $D^k = \{x \in \mathbb{R}^k \colon \|x\| < 1\}$ the k-dimensional standard (open) disk and by $S^k = \{x \in \mathbb{R}^{k+1} \colon \|x\| = 1\}$ the k-dimensional standard sphere. We consider these sets as "independent" topological spaces. Therefore, it is justified to abstain from picking a specific norm, since all norms on $\mathbb{R}^k$ are equivalent.

For $k, m \in \mathbb{N}$ with $m \leq k$, the *(j-dimensional open) k-annulus* is the product space $S^k \times D^{j-k}$. Note that since S^k has one connected component and a $(k+1)$-dimensional hole, it holds that $\beta_0(S^k) = \beta_k(S^k) = 1$ and the remaining Betti numbers equal zero. The j-dimensional k-annulus is a j-dimensional manifold that can be thought as a thickened k-sphere and hence its Betti numbers coincide with the ones from the k-sphere. In Appendix A.2 the reader can find a more formal treatment of the latter fact.

In contrast to D^k and S^k, which are only seen as spaces equipped with a topology, we also consider neighborhoods around certain points $x \in \mathbb{R}^d$ as subsets of $\mathbb{R}^d$. To make a clear distinction, we define the space $B^d_r(x)$ as the *d-dimensional open r-ball around x with respect to the 1-norm*, i.e., the space $\{y \in \mathbb{R}^d \colon \|x-y\|_1 < r\}$. Note that for $r < r'$, the set $B^k_r(x) \setminus \overline{B^k_{r'}(x)}$ is homeomorphic to a k-dimensional $(k-1)$-annulus and we will refer to them as $(k-1)$-annuli as well. These annuli will be the building blocks of our construction for the lower bound.

The rest of the paper is devoted to proving the lower and upper bounds. Most of the statements come with an explanation or an illustration. In addition, formal proofs for these statements are also provided in the appendix.

2. Lower Bound

In this section, our aim is to construct a family of neural networks $F \colon \mathbb{R}^d \to \mathbb{R}$ of depth $L+2$ for $L \in \mathbb{N}$ such that $\beta_k(F)$ grows exponentially in the size of the neural network for all $k \in [d-1]_0$.

We propose a construction that is restricted to architectures where the widths $n_1, \dots, n_{L+1}$ of all hidden layers but the last one are divisible by $2d$. This construction, however, is generalized for any architecture where the dimension of all hidden layers is at least $2d$ by inserting at most $2d$ auxiliary neurons at each layer at which a zero map is computed. Correspondingly, one obtains a lower bound by rounding down the width n_ℓ at each layer to the largest possible multiple of $2d$. In particular, a reduction to the case in Theorem 12 in the appendix does not have an effect on the asymptotic size of the NN.

The key idea is to construct $F = f \circ h$ as a consecutive execution of two neural networks f and h, where the map $h \colon \mathbb{R}^d \to \mathbb{R}^d$ is an ReLU NN with L hidden layers that identifies exponentially many regions with each other. More precisely, h cuts the unit cube of $\mathbb{R}^d$ into exponentially many small cubes $W_{i_1,\dots,i_d} \in [0,1]^d$ and maps each of these cubes to the whole unit cube by scaling and mirroring. The one hidden layer ReLU NN f then cuts the unit cube into pieces so that f on the pieces alternatingly takes exclusively positive respectively negative values (cf. Figures 3 and 4). Since h maps all $W_{i_1,\dots,i_d}$ to $[0,1]^d$ by scaling and mirroring, every $W_{i_1,\dots,i_d}$ is cut into positive-valued and negative-valued regions by the composition $f \circ h$ in the same way as $[0,1]^d$ is mapped

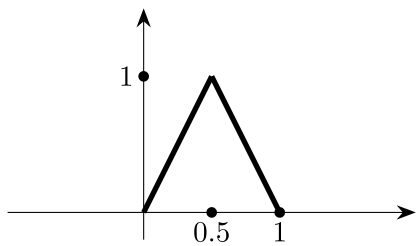


Figure 1: The graph of the function $\pi_j \circ h^{(1,2,d)}$ that folds the unit interval, i.e., mapping the interval $[0, 0.5]$ and $[0.5, 1]$ to the unit interval. This function is realised by a hidden layer with 2 hidden neurons.

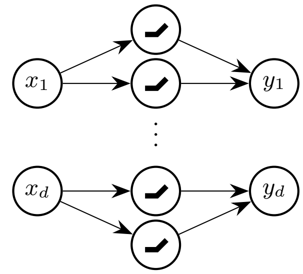


Figure 2: The architecture of the one hidden layer neural network $h^{(1,2,d)}$ that folds the d-dimensional unit cube by folding every component of the cube as described in Figure 1.

by f, up to mirroring. The cutting of the unit cube and the mirroring of the small cubes in the map to $[0,1]^d$ are chosen in such a way that the subspaces on which F takes negative values assemble at the corners (each corner belongs to 2^d small cubes by the cutting by h) of the cubes so that they form k-annuli for every $k \in [d-1]$. Since h cuts the unit cube into exponentially many small cubes, we obtain exponentially many k-annuli for every $k \in [d-1]$ in $F^{-1}((-\infty, 0))$ (cf. Figures5 and 6). Some technical adjustments will then yield the result for $F^{-1}((-\infty, 0])$.

The idea of constructing a ReLU neural network that folds the input space goes back to Montúfar et al. (2014), where the construction was used to show that a deep neural network with ReLU activation function can have exponentially many linear regions. For our purposes, we explicitly state the continuous piecewise linear map that arises from the construction instead of proving only the existence of such a neural network. Using their techniques, we first build a 1-hidden layer NN $h^{(1,m,d)} \colon \mathbb{R}^d \to \mathbb{R}^d$ for $m \in \mathbb{N}$ even that folds the input space, mapping m^d many small cubes $W^{(1,m,d)}_{i_1,\ldots,i_d} \subset [0,1]^d$ by scaling and mirroring to $[0,1]^d$. More precisely, the NN $h^{(1,m,d)}$ has $m \cdot d$ many neurons in the single hidden layer which are partitioned into m groups. The weights are chosen such that the output of the neurons in one group depends only on one input variable and divides the interval $[0,1]$ into m subintervals of equal length, each of which is then mapped to the unit interval $[0,1]$ by the output neuron. Figure 1 illustrates this construction. In Appendix B.1 or in Montúfar et al. (2014), the reader can find an explicit construction of $h^{(1,m,d)}$.

The map $h^{(1,m,d)}$ identifies only $O(m^d)$ many cubes with each other. To subdivide the input space into exponentially many cubes and map them to the unit cube, we need a deep neural network. For this purpose, we utilize a vector $\mathbf{m}$ of folding factors instead of a single number m. Let $\mathbf{m} = (m_1, \ldots, m_L) \in \mathbb{N}^L$ with m_ℓ even for all $\ell \in [L]$ and define the neural network $h^{(L,\mathbf{m},d)}$ with L hidden layers as $h^{(L,\mathbf{m},d)} = h^{(1,m_L,d)} \circ \cdots \circ h^{(1,m_1,d)}$. Since each of the m_1^d cubes that results from the subdivision by the first layer is mapped back to $[0,1]^d$, each cube is subdivided again into m_2^d cubes by the subsequent layer. Thus, after L such layers, we obtain a subdivision of the input space into $\left(\prod_{\ell=1}^L m_\ell\right)^d$ cubes.

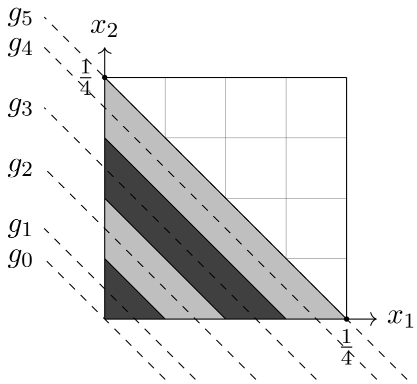


Figure 3: Illustration of the preimage of the map $\hat{g}^{(4,2)}$ in $[0, \frac{1}{4}]^2$, where the dark gray regions correspond to $\hat{g}^{-1}((0,\infty))$ and the light gray regions to $\hat{g}^{-1}((-\infty, 0))$.

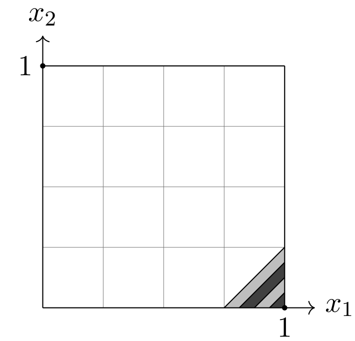


Figure 4: Illustration of the preimage of the map $g^{(4,2)}$ in $[0,1]^2$.

In the following, we define fixed but arbitrary variables: $L \in \mathbb{N}$, $\mathbf{m} = (m_1, \ldots, m_L) \in \mathbb{N}^L$ and $M = \left(\prod_{\ell=1}^{L} m_\ell\right)$ with $m_\ell > 1$ even for all $\ell \in [L]$. The following lemma recollects the existence of a map $h^{(L,\mathbf{m},d)}$ that actually enjoys the aforementioned properties.

Lemma 5 *(cf. (Montúfar et al., 2014)) Let $d \in \mathbb{N}$. Then there exists a map $h^{(L,\mathbf{m},d)} \colon \mathbb{R}^d \to \mathbb{R}^d$ such that for all $(i_1, \ldots, i_d) \in [M]^d$, the following hold:*

1. $h^{(L,\mathbf{m},d)}(W^{(L,\mathbf{m},d)}_{(i_1,\ldots,i_d)}) = [0,1]^d$

2. $\pi_j \circ h^{(L,\mathbf{m},d)}_{|W^{(L,\mathbf{m},d)}_{(i_1,\ldots,i_d)}}(x_1, \ldots, x_d) = \begin{cases} M \cdot x_j - (i_j - 1) & i_j \text{ odd} \\ -M \cdot x_j + i_j & i_j \text{ even} \end{cases}$

We now define cutting points as the points that are mapped to the point $(1, 1, 1, \ldots, 1, 0)$ by the map $h^{(L,\mathbf{m},d)}$. They will play a central role in counting the annuli in the sublevel set of F.

Definition 6 *We call a point $x \in [0,1]^d$ a* cutting point *if it has coordinates of the form $x_i = \frac{x_i'}{M}$ for all $i \in \{1, \ldots, d\}$, where the x_i' are odd integers for $1 \leq i \leq d-1$ and x_d' is an even integer.*

Next, for $w \geq 2$, we build a 1-hidden layer neural network $\hat{g}^{(w,d)} \colon \mathbb{R}^d \to \mathbb{R}$ that cuts the d-dimensional unit cube into w pieces such that $\hat{g}^{(w,d)}$ maps the pieces alternatingly to $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{\leq 0}$, respectively. We omit the indices w and d whenever they are clear from the context.

In order to build the neural network, we fix w and d and define the maps $\hat{g}_q \colon \mathbb{R}^d \to \mathbb{R}$, $q = 0, \ldots, w+1$ by

$$\hat{g}_q(x) = \begin{cases} \max\{0, \mathbf{1}^T x\} & q = 0 \\ \max\{0, \mathbf{1}^T x - \frac{1}{4}\} & q = w+1 \\ \max\{0, 2(\mathbf{1}^T x - (2q-1)/8w)\} & \text{else} \end{cases}$$

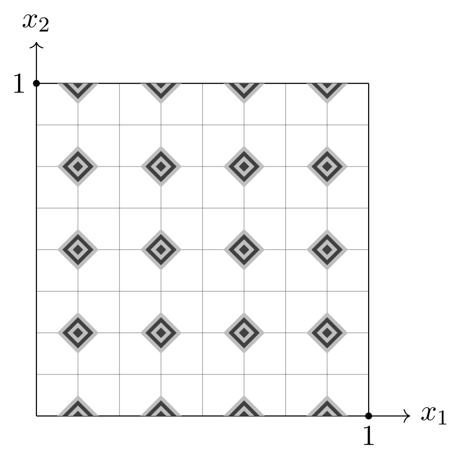


Figure 5: Illustration of the preimage of the composition $g^{(4,2)} \circ h^{(3,2,2)}$.

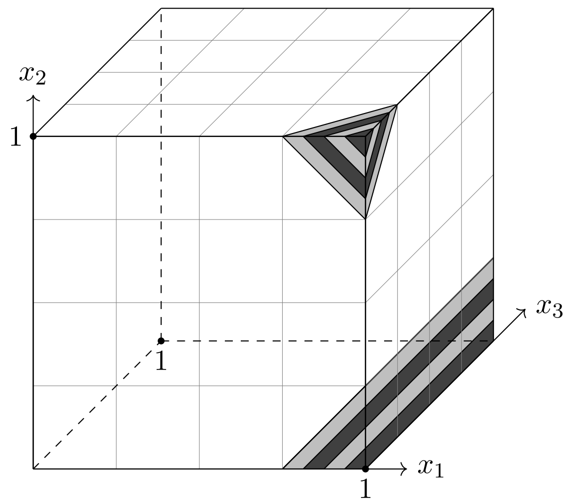


Figure 6: Illustration of the preimage of $f^{(4,4)} = g^{(4,3)} + g^{(4,2)} \circ p_2$.

Later in this section, we will iteratively construct k-annuli in the sublevel set of F for all $k \in [d-1]$. In order to ensure that these annuli are disjoint, it is convenient to place them around the cutting points. To achieve this, we mirror the map $\hat{g}$ before precomposing it with h. The mirroring transformation that maps the origin to the point $(1, \ldots, 1, 0)$ is an affine map $t : [0,1]^d \to [0,1]^d$ defined by $t(x_1, x_2, \ldots, x_d) = (1 - x_1, 1 - x_2, \ldots, 1 - x_{d-1}, x_d)$. We define the neural network $g = \hat{g} \circ t$ as the consecutive execution of $\hat{g}$ and t.

Lemma 7 *Let $d, w \in \mathbb{N}$ with w odd and*

$$R_q = \{x \in [0,1]^d : \frac{q}{4w} < \|(1, 1, \ldots, 1, 0) - x\|_1 < \frac{q+1}{4w}\}.$$

Then there exists a 1-hidden layer neural network $g^{(w,d)} \colon \mathbb{R}^d \to \mathbb{R}$ of width $w + 2$ such that $g^{(w,d)}(R_q) \subseteq (-\infty, 0)$ for all odd $\in [w-1]_0$, $g^{(w,d)}(R_q) \subseteq (0, \infty)$ for all even $q \in [w-1]_0$ and $g^{(w,d)}(x) = 0$ for all $x \in [0,1]^d$ with $\|(1, 1, \ldots, 1, 0) - x\|_1 \geq \frac{1}{4}$.

Lemma 37 in the appendix characterizes the regions around cutting points that admit positive respectively negative values under the map $g^{(w,d)} \circ h^{(L,\mathbf{m},d)}$. We focus on the regions that admit negative values, i.e., the space $Y_{d,w} \coloneqq (g^{(w,d)} \circ h^{(L,\mathbf{m},d)})^{-1}((-\infty, 0))$ and observe that we obtain d-dimensional $(d-1)$-annuli around each cutting point.

In order to count the annuli we count the cutting points.

Observation 8 *Cutting points lie on a grid in the unit cube, with $\frac{M}{2}$ many cutting points into dimensions $1, \ldots, d-1$ and $\frac{M}{2} + 1$ many in dimension d. Thus, there are $\frac{M^{(d-1)}}{2^{d-1}} \cdot \left(\frac{M}{2} + 1\right)$ cutting points. Note that since M is an even number, these points cannot lie on the boundary unless the last coordinate is 0 or M. This means, $2 \cdot \frac{M^{(d-1)}}{2^{d-1}} = \frac{M^{(d-1)}}{2^{d-2}}$ of the cutting points are located on the boundary of the unit cube and the remaining $\frac{M^{(d-1)}}{2^{d-1}} \cdot \left(\frac{M}{2} - 1\right)$ are in the interior.*

Combining Lemma 37 and Observation 8, we can finally describe $Y_{d,w}$ as a topological space.

Proposition 9 *The space $Y_{d,w}$ is homeomorphic to the disjoint union of $p_d = \frac{M^{(d-1)}}{2^{d-1}} \cdot \left(\frac{M}{2} - 1\right) \cdot \left\lceil \frac{w}{2} \right\rceil$ many $(d-1)$-annuli and $p'_d = \frac{M^{(d-1)}}{2^{d-2}} \cdot \left\lceil \frac{w}{2} \right\rceil$ many disks, that is,*

$$Y_{d,w} \cong \bigsqcup_{k=1}^{p_d} (S^{d-1} \times [0,1]) \sqcup \bigsqcup_{k=1}^{p'_d} D^d.$$

In order to obtain exponentially many k-annuli for all $k \in [d-1]$, we follow a recursive approach: At each step, we start with a k-dimensional space that has exponentially many j-annuli for all $j \in [k-1]$. We then cross this space with the interval $[0,1]$, transforming the k-dimensional j-annuli into $(k+1)$-dimensional j-annuli. Finally, we "carve" $(k+1)$-dimensional k-annuli in this newly formed product space. To allow us flexibility with respect to the numbers of annuli carved in different dimensions, we fix an arbitrary vector $\mathbf{w} = (w_1, \dots, w_{d-1}) \in \mathbb{N}^{d-1}$ such that $\sum_{i=1}^{d-1}(w_i + 2) = n_{L+1}$. We iteratively define the 1-hidden layer neural network

$$f^{(w_1,\dots,w_{k-1})} \colon \mathbb{R}^k \to \mathbb{R}$$

of width n_{L+1} by $f^{(w_1)} = g^{(w_1,2)}$ and

$$f^{(w_1,\dots,w_{k-1})} = f^{(w_1,\dots,w_{k-2})} \circ p_{k-1} + g^{(w_{k-1},k)}$$

for $k \le d$. Roughly speaking, the following lemma states that the carving map does not interfere with the other maps, i.e., there is enough space in the unit cubes to place the k-annuli after having placed all k'-annuli ($k' < k$) in the same, inductive manner.

Lemma 10 *For $k \le d$ and $\mathbf{w} = (w_1, \dots, w_{d-1}) \in \mathbb{N}^{d-1}$ it holds that*

1. $f^{(w_1,\dots,w_{k-2})} \circ p_{k-1}(x) \neq 0 \implies g^{(w_{k-1},k)}(x) = 0$ *and*
2. $g^{(w_{k-1},k)}(x) \neq 0 \implies f^{(w_1,\dots,w_{k-2})} \circ p_{k-1}(x) = 0$

for all $x \in [0,1]^k$.

Using Lemma 10 and the fact that the folding maps $h^{(L,\mathbf{m},k)}$ are compatible with projections (cf. Lemma 39 in Appendix), we can make sure that we can construct the cuts iteratively so that we obtain k-annuli for every $k \in [d-1]$, which is stated in the following lemma.

Lemma 11 *For $2 \le k \le d$, the space $X_k \coloneqq (f^{(w_1,\dots,w_{k-1})} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0))$ satisfies*

$$X_k = (X_{k-1} \times [0,1]) \sqcup Y_{k,w_{k-1}}$$

with $X_1 \coloneqq \emptyset$.

Lemma 11, Proposition 9 and the disjoint union axiom of singular homology (Proposition 27 in Appendix A.2) allow us to compute the Betti numbers of the decision region of $F \coloneqq f^{(w_1,\dots,w_{d-1})} \circ h^{(L,\mathbf{m},d)}$ as stated in Theorem 12 in the appendix. One can easily generalize this statement by rounding down the widths $n_1, \dots, n_L$ to the nearest even multiple of d:

Theorem 12 *Given an architecture* $A = (d, n_1, \dots, n_L, 1)$ *with* $n_\ell \geq 2d$ *for all* $\ell \in [L]$ *and numbers* $w_1, \dots, w_{d-1} \in \mathbb{N}$ *such that* $\sum_{k=1}^{d-1}(w_k + 2) = n_L$*, there is a neural network* $F \in \mathcal{F}_A$ *with weights bounded from above by* $\max_{\ell=1,\dots L} 2\frac{n_\ell}{d}$ *such that*

(i) $\beta_0(F^{-1}((-\infty, 0))) = \sum_{k=2}^{d} \frac{M^{(k-1)}}{2^{k-1}} \cdot \left(\frac{M}{2} + 1\right) \cdot \left\lceil \frac{w_k}{2} \right\rceil$

(ii) $\beta_k(F^{-1}((-\infty, 0))) = \frac{M^{(k-1)}}{2^{k-1}} \cdot \left(\frac{M}{2} - 1\right) \cdot \left\lceil \frac{w_{k-1}}{2} \right\rceil$ *for* $0 < k < d$*,*

where $M = \prod_{\ell=1}^{L-1} 2 \cdot \lfloor \frac{n_\ell}{2d} \rfloor$*.*

In order to obtain lower bounds for $\beta_k(F)$, we modify the construction slightly by adding a small constant b to the output layer, which yields a neural network F' such that there is no full-dimensional region R such that $F'(R) = \{0\}$. The construction then yields that $\overline{F'^{-1}((-\infty, 0))} = F'^{-1}((-\infty, 0])$. Since adding a small constant b only makes the annuli in the sublevel set $F^{-1}((-\infty, 0))$ thinner, the spaces $F^{-1}((-\infty, 0))$ and $F'^{-1}((-\infty, 0))$ are homeomorphic. Furthermore, since annuli are homotopy equivalent to their closures, the sublevel set $F'^{-1}((-\infty, 0))$ is homotopy equivalent to its closure $\overline{F'^{-1}((-\infty, 0))} = F'^{-1}((-\infty, 0])$. Since Betti numbers are invariant under homotopy equivalences, it follows that $\beta_k(F') = \beta_k(F^{-1}((-\infty, 0))$ for all $k \in [d-1]_0$, resulting in the following theorem.

Theorem 13 *Given an architecture* $A = (d, n_1, \dots, n_L, 1)$ *with* $n_\ell \geq 2d$ *for all* $\ell \in [L]$ *and numbers* $w_1, \dots, w_{d-1} \in \mathbb{N}$ *such that* $\sum_{k=1}^{d-1}(w_k + 2) = n_L$*, there is a neural network* $F \in \mathcal{F}_A$ *with weights bounded from above by* $\max_{\ell=1,\dots L} 2\frac{n_\ell}{d}$ *such that*

(i) $\beta_0(F) = \sum_{k=2}^{d} \frac{M^{(k-1)}}{2^{k-1}} \cdot \left(\frac{M}{2} + 1\right) \cdot \left\lceil \frac{w_k}{2} \right\rceil$

(ii) $\beta_k(F) = \frac{M^{(k-1)}}{2^{k-1}} \cdot \left(\frac{M}{2} - 1\right) \cdot \left\lceil \frac{w_{k-1}}{2} \right\rceil$ *for* $0 < k < d$*,*

where $M = \prod_{\ell=1}^{L-1} 2 \cdot \lfloor \frac{n_\ell}{2d} \rfloor$*.*

The special case $\left\lfloor \frac{w_1}{2} \right\rfloor = \ldots = \left\lfloor \frac{w_d}{2} \right\rfloor$ then corresponds precisely to Corollary 2.

As mentioned previously, the sum of Betti numbers, the notion of topological expressivity used in Bianchini and Scarselli (2014), does not provide us with an understanding of holes of different dimensions. On the other hand, our bounds are an extension of this result. In addition, the dimension-wise lower bound allows further implications, one of them being a lower bound on the *Euler characteristic*, which is the alternating sum $\chi(X) = \sum_{k=1}^{d} \beta_k(X)$ of the Betti numbers.

Corollary 14 *Let* A *be the architecture as in Theorem 13 , then there is a ReLU NN* $F \colon \mathbb{R}^d \mapsto \mathbb{R}$ *with architecture* A *such that* $\chi(F^{-1}((-\infty, 0])) \in \Omega\left(M^d \cdot \sum_{i=1}^{d-1} w_i\right)$*, where* $\chi(F^{-1}((-\infty, 0]))$ *denotes the Euler characteristic of the space* $F^{-1}((-\infty, 0])$*.*

Proof The Euler characteristic of a finite CW complex X is given by the alternating sum of its Betti numbers, i.e., by the sum $\sum_{k \in \mathbb{N}} (-1)^k \beta_k(X)$. By Theorem 12, this term is dominated by the zeroth Betti number, from which the claim follows. ■

The Euler characteristic is an invariant used widely in differential geometry in addition to algebraic topology. For instance, it can also be defined by means of the index of a vector field on a compact smooth manifold.

3. Topologically Stable ReLU Neural Networks

In the following, we establish a sufficient criterion for the parameters of a neural network such that the topological expressivity of the corresponding neural networks is constant in an open neighbourhood of this parameter. The neural network constructed explicitly in Section 2 to obtain the lower bound fulfills this criterion, so that the lower bound is attained in an open subset of the parameter space.

We denote by $\Phi\colon \mathbb{R}^D \to C(\mathbb{R}^d)$ the map that assigns a vector of weights in the parameter space $\mathbb{R}^D \cong \bigoplus_{\ell=1}^{L+1} \mathbb{R}^{(n_{\ell-1}+1)\times n_\ell}$ to the function computed by the ReLU neural network with this weights, i.e.,

$$\Phi(p) \coloneqq T_{L+1}(p) \circ \sigma_{n_L} \circ T_L(p) \circ \cdots \circ \sigma_{n_1} \circ T_1(p)$$

(c.f. Definition 42 in the appendix) and by

$$\Phi^{(i,\ell)}(p) \coloneqq \pi_i \circ T_\ell(p) \circ \cdots \circ \sigma_{n_1} \circ T_0(p),$$

the map computed at the i-th neuron in layer ℓ. Any neuron (i, ℓ) defines a hyperplane

$$H_{i,\ell}(p) = \{x \in \mathbb{R}^{n_{\ell-1}} \mid \pi_i \circ (T_\ell(p))\,(x) = 0\}$$

in the output space of the previous layer. One can iteratively define a sequence of polyhedral complexes $\mathcal{P}^{(i,\ell)}(p)$ such that $\Phi^{(k,j)}(p)$ is affine linear on the polyhedra of $\mathcal{P}^{(i-1,\ell)}(p)$ for all (k, j) lexigrophically smaller than (i, ℓ) by intersecting all the polyhedra in $\mathcal{P}^{(i,\ell)}(p)$ with the pullback of the hyperplane $H_{i,\ell}(p)$ to the input space i.e., $\{x \in \mathbb{R}^d \mid \left(\Phi^{(i,\ell)}(p)\right)(x) = 0\}$ and the pullbacks of the corresponding half-spaces to the input space i.e., $\{x \in \mathbb{R}^d \mid \left(\Phi^{(i,\ell)}(p)\right)(x) \leq 0\}$ and $\{x \in \mathbb{R}^d \mid \left(\Phi^{(i,\ell)}(p)\right)(x) \geq 0\}$ (c.f. Definition 43 in the appendix). For a polytope K, let $\hat{K} \coloneqq \{F \mid F \text{ is a face of } K\}$ be the polyhedral complex consisting of the faces of K. The canonical polyhedral complex (with respect to K) is then defined as $\mathcal{P}(p, K) \coloneqq \mathcal{P}^{(n_L,L)}(p) \cap \hat{K}$ (c.f. Grigsby and Lindsey (2022)). Furthermore, $\Phi(p)$ is affine linear and non-positive respectively non-negative on all polyhedra of $\mathcal{P}^{(n_{L+1},1)}(p) \cap \hat{K}$, which is a refinement of the canonical polyhedral complex.

We call a neural network $\Phi(p)$ *topologically stable with respect to* K if the pullback of the hyperplane $H_{i,\ell}(p)$ does not intersect any vertices (i.e., faces of dimension 0) of $\mathcal{P}^{(i,\ell)}(p) \cap \hat{K}$ for all neurons (i, ℓ) (c.f. Definition 46 in the appendix). One can perturb the weights (and hence the hyperplanes as well as their pullbacks) of a topologically stable neural network within a small enough magnitude such that the combinatorial structure of the refinement $\mathcal{P}^{(n_{L+1},L+1)}(p) \cap \hat{K}$ is preserved.

Proposition 15 *(c.f. Proposition 45 in the appendix). Let K be a polytope and $\Phi(p)\colon K \to \mathbb{R}$ be a topologically stable (w.r.t to K) ReLU neural network of architecture $(n_0, \ldots n_{L+1})$. Then for every $\delta > 0$, there is an open set $U \subseteq \mathbb{R}^D$ such that for every $u \in U$ there is an isomorphism $\varphi_u\colon \mathcal{P}(p, K) \to \mathcal{P}(u, K)$ with $\|v - \varphi_u(v)\|_2 < \delta$ for every vertex $v \in \mathcal{P}(p, K)$.*

Since the isomorphisms extend to the refinements $\mathcal{P}^{(n_{L+1},1)}(p) \cap \hat{K}$ and $\mathcal{P}^{(n_{L+1},1)}(u) \cap \hat{K}$ and an isomorphism of polyhedral complexes $\varphi\colon \mathcal{P} \to \mathcal{Q}$ yield a PL-homeomorphism between the respective supports $|\varphi|\colon |\mathcal{P}| \to |\mathcal{Q}|$, we obtain the following result.

Proposition 16 *Let K be a polytope and $\Phi(p)$ a topologically stable ReLU neural network with respect to K, then there is a $\delta > 0$ such that for all $u \in B_\delta(p)$ it holds that $K \cap \Phi(p)^{-1}((-\infty, 0])$ is homeomorphic to $K \cap \Phi(u)^{-1}((-\infty, 0])$.*

Lastly, since our construction for the lower bound yield topologically stable ReLU NN with respect to the unit cube, Proposition 16 implies that our results are stable with respect to small perturbations.

Proposition 17 *There is an open set $U \subseteq \mathbb{R}^D$ in the parameter space of the architecture $A = (d, n_1, \ldots, n_L, 1)$ such that $\Phi(u)$ restricted to the unit cube has at least the same topological expressivity as F in Theorem 13 for all $u \in U$.*

4. Upper Bound

In this section we derive an upper bound for $\beta_k(F)$ for a ReLU neural network $F\colon \mathbb{R}^d \to \mathbb{R}$ for all $k \in [d-1]$, showing that they are polynomially bounded in the width using an upper bound on the linear regions of F. A linear region R of F contains at most one maximal convex polyhedral subspace where F takes on exclusively non-negative function values. Intuitively, every such polyhedral subspace can be in the interior of at most one d-dimensional hole of the sublevel set $F^{-1}((-\infty, 0])$ and thus the number of linear regions is an upper bound for $\beta_{d-1}(F)$. In the following proposition we will formalize this intuition and generalize it to $\beta_k(F)$ for all $k \in [d-1]_0$.

Proposition 18 *Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be a neural network of architecture $(d, n_1, \ldots, n_L, 1)$. Then it holds that $\beta_0(F) \leq \sum_{(j_1,\ldots,j_L)\in J} \quad \prod_{\ell=1}^{L} \binom{n_\ell}{j_\ell}$ and for all $k \in [d-1]$ that*

$$\beta_k(F) \leq \binom{\sum_{(j_1,\ldots,j_L)\in J} \quad \prod_{\ell=1}^{L} \binom{n_\ell}{j_\ell}}{d-k-s},$$

where $J = \left\{(j_1, \ldots, j_L) \in \mathbb{Z}^L \colon 0 \leq j_\ell \leq \min\{d, n_1 - j_1, \ldots, n_{\ell-1} - j_{\ell-1}\} \text{ for all } \ell = 1, \ldots, L\right\}$ and $s \in [d]$ is the dimension of the lineality space of a refinement of the canonical polyhedral complex of F.

Proof Sketch Theorem 1 in (Serra et al., 2017) states that F has at most $r := \sum_{(j_1,\ldots,j_L)\in J} \prod_{l=1}^{L} \binom{n_l}{j_l}$ linear regions. In Section D we will provide a formal proof for the statement that we sketch here. Let $\mathcal{P}$ be the canonical polyhedral complex of F, i.e, F is affine linear on all polyhedra in $\mathcal{P}$ (c.f Definition 43 in the appendix) and $\mathcal{P}^-$ be a subcomplex of a refinement of $\mathcal{P}$ such that F takes on exclusively non-positive values on all polyhedra in $\mathcal{P}^-$. Therefore, the support $|\mathcal{P}^-|$ of $\mathcal{P}^-$ equals $F^{-1}((-\infty, 0])$ and we then proceed by showing the chain of inequalities

$$\beta_k(F) = \beta_k(|\mathcal{P}^-|) \leq \#\mathcal{P}_{k+1-\ell} \leq \binom{r}{d-k-\ell}$$

using cellular homology and polyhedral geometry, where $\mathcal{P}_{k+1} \subseteq \mathcal{P}$ is the set of $(k+1)$-dimensional polyhedra in $\mathcal{P}$. This concludes the proof, since it also holds that $\beta_0(|\mathcal{P}^-|) \leq \#\mathcal{P}_d = r$.

■

This implies that we obtain an upper bound that is polynomial in the width:

Corollary 4 *Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be a neural network of architecture $(d, n_1, \ldots, n_L, 1)$. Then it holds that $\beta_k(F) \in O\left(\left(\prod_{i=1}^{L} n_i\right)^{d^2}\right)$ for $k \in [d-2]$ and $\beta_0, \beta_{d-1} \in O\left(\left(\prod_{i=1}^{L} n_i\right)^{d}\right)$.*

5. Extensions of the Results

So far, we have demonstrated an exponential gap for the Betti numbers of decision regions computable by deep and shallow networks, where our lower bound was derived using a special construction. However, this construction can be adapted to extend our results to a broader class of classification tasks and neural network architectures.

Multi-categorical Classification The bounds for binary classification can be achieved for multi-categorical classification as well: Assume that we have k classes that are classified by the intervals $(-\frac{p}{k}, -\frac{p-1}{k}]$ for $p \in [k]$. In our constructions, this corresponds to subdividing every annulus into $2k$ annuli (two for each class) and therefore it holds that $2\beta_j(F^{-1}((-\infty, 0]) = \beta_j(F^{-1}((-\frac{p}{k}, -\frac{p-1}{k}])$ for all $p \in [k]$. To have decision intervals of the form $(0, 1], (1, 2], \ldots, (k-1, k]$, one can scale this accordingly. Hence, if we have k classes, the Betti numbers of the decision regions of all these classes can be exponentially high simultaneously for deep neural networks. The upper bounds can be achieved in the same way as for the binary case.

Recurrent Neural Networks For $\mathbf{m} = (m, \ldots, m)$, our construction can be seen as a recurrent neural network (RNN) followed by a dense layer: Since we concatenate the same layer map (the folding map $h^{(1,m,d)}$) L times, this can be seen as an RNN and the last layer map (the cutting map f) as the dense layer.

If we restrict ourselves to consider RNNs without an additional dense layer, one can define the two-hidden-layer NN $F^{(w_1,\ldots,w_{d-1},m)} = (\lambda \cdot f^{(w_1,\ldots,w_{d-1})}) \circ h^{(1,m,d)}$, where $\lambda = d \cdot \max_{x\in[0,1]^d} f(x)$. Then, for every cutting point c, there exist 2^d many cubes who share the vertex c and are mapped back to the unit cube under the map $F^{(w_1,\ldots,w_{d-1},m)}$. Therefore, after every subsequent application of $F^{(w_1,\ldots,w_{d-1},m)}$, this behaviour is replicated inside every such cube. Furthermore, there are $\Omega(m^d \cdot w_{k-1})$ many k-annuli inside every such cube, resulting in $\Omega(m^{dL} \cdot w_{k-1})$ many k-annuli after L applications of $F^{(w_1,\ldots,w_{d-1},m)}$. Since the upper bounds apply to RNNs regarded as unfolded DNNs as well, we obtain the same exponential gap.

6. Conclusion, Limitations and Outlook

Since it is widely accepted that data sets often have nontrivial topologies, investigating a neural network's ability to capture topological properties, as characterized by all Betti numbers, is an exciting and essential question that yields insight into the nature of ReLU networks. In an attempt to shed light on this question, we have proven lower and upper bounds for the topological expressivity of ReLU neural networks with a given architecture. Our bounds give a rough estimate on how the architecture needs to be in order to be at least theoretically able to capture the topological complexity of the data set.

Even though our lower bounds apply under certain restrictions of neural network architecture, this does not pose a big limitation for our purposes. Since our results are of a theoretical and mostly asymptotic nature, a constant factor (in the hidden layers resp. in the last hidden layer) is negligible. Besides, since our layers merely consists of many small layers put in parallel, one could also concatenate the layers in order to achieve a smaller width maintaining all the asymptotic results.

As a byproduct of our analysis we have seen that two hidden layers are sufficient to increase the topological expressivity as much as we want at the expense of an increased width. Although there are finer topological invariants such as cohomology rings or homotopy groups; from a computational point of view, Betti numbers are a good trade-off between the ability to capture differences of spaces and tractability. Nevertheless, it might be interesting to employ further topological or geometrical invariants to investigate the expressivity of neural networks in the setting of classification tasks.

Besides, there is no CPWL function known for which the non-existence of a two hidden layer ReLU neural network computing this function has been proven (Hertrich et al., 2021; Haase et al., 2023). Therefore, a further research goal might also be to find topological spaces for which one can show that two hidden layers are not sufficient to have them as a decision region.

It seems straightforward that the construction in Section 2 can be adapted to neural networks with sigmoidal activation functions in a "smoothed" way. Therefore we conjecture that the same lower bound holds for the topological expressivity of neural networks with sigmoidal activation function, which would generalise the lower bound for the zeroth Betti number given in Bianchini and Scarselli (2014) to all Betti numbers.

Another interesting follow-up research topic would be to investigate the distribution of Betti numbers over the parameter space of the neural network. To that end, it is of interest to understand the realization map extensively. Some work in this direction has been done by Grigsby et al. (2022b).

Acknowledgments

The authors would like to thank Christoph Hertrich and Martin Skutella for many valuable discussions and their careful proofreading. Moreover, we thank the various anonymous referees for their remarks. Moritz Grillo and Ekin Ergen are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

References

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=B1J_rgWRW.

Ronen Basri and David W. Jacobs. Efficient representation of low-dimensional manifolds using deep networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJ3filKll.

Monica Bianchini and Franco Scarselli. On the complexity of shallow and deep neural network classifiers. In *The European Symposium on Artificial Neural Networks*, 2014.

George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

Tamal Krishna Dey and Yusu Wang. *Complexes and Homology Groups*, page 26–59. Cambridge University Press, 2022. doi: 10.1017/9781009099950.003.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on*

Learning Theory, volume 49 of *Proceedings of Machine Learning Research*, pages 907–940, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/eldan16.html.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2018. doi: 10.1109/CVPR.2018.00396.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/glorot11a.html.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

J Elisenda Grigsby and Kathryn Lindsey. On transversality of bent hyperplane arrangements and the topological expressiveness of relu neural networks. *SIAM Journal on Applied Algebra and Geometry*, 6(2), 2022.

J. Elisenda Grigsby, Kathryn Lindsey, and Marissa Masden. Local and global topological complexity measures of relu neural network functions, 2022a.

J. Elisenda Grigsby, Kathryn Lindsey, Robert Meyerhoff, and Chenxi Wu. Functional dimension of feedforward relu neural networks, 2022b.

William H. Guss and Ruslan Salakhutdinov. On characterizing the capacity of neural networks using algebraic topology, 2018. URL https://arxiv.org/abs/1802.04443.

Christian Haase, Christoph Hertrich, and Georg Loho. Lower bounds on the depth of integral relu neural networks via lattice polytopes, 2023.

Mustafa Hajij and Kyle Istvan. A topological framework for deep learning, 2020. URL https://arxiv.org/abs/2008.13697.

Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-79160-X; 0-521-79540-0.

Christoph Hertrich, Amitabh Basu, Marco Di Summa, and Martin Skutella. Towards lower bounds on the depth of reLU neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=0OWwNh-4in1.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4: 251–257, 1991.

Joey Huchette, Gonzalo Muñoz, Thiago Serra, and Calvin Tsay. When deep learning meets polyhedral theory: A survey, 2023.

Weizhi Li, Gautam Dasarathy, Karthikeyan Natesan Ramamurthy, and Visar Berisha. Finding the homology of decision boundaries with active learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, page 8355–8365. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5f14615696649541a025d3d0f8e0447f-Paper.pdf.

Luciano Melodia and Richard Lenz. Estimate of the neural network dimension using algebraic topology and lie theory. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part V*, volume 12665 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2020. doi: 10.1007/978-3-030-68821-9_2. URL https://doi.org/10.1007/978-3-030-68821-9_2.

Guido Montúfar. Notes on the number of linear regions of deep neural networks. 03 2017.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2924–2932, Cambridge, MA, USA, 2014. MIT Press.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020. URL http://jmlr.org/papers/v21/20-345.html.

Quynh Nguyen, Mahesh Chandra Mukkamala, and Matthias Hein. Neural networks should be wide enough to learn disconnected decision regions. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3740–3749. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/nguyen18b.html.

Giovanni Petri and António Leitão. On the topological expressive power of neural networks. In *TDA & Beyond*, 2020. URL https://openreview.net/forum?id=I44kJPuvqPD.

Alexander Schrijver. Theory of linear and integer programming. In *Wiley-Interscience series in discrete mathematics and optimization*, 1986.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. *CoRR*, abs/1711.02114, 2017. URL http://arxiv.org/abs/1711.02114.

Matus Telgarsky. Benefits of depth in neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/telgarsky16.html.

Appendix A. Mathematical Background

A.1. Polyhedral Geometry

In this section we recall the definition of a polyhedral complex and introduce some notation related to them. For an introduction to polyhedra, we refer to Schrijver (1986). Furthermore, we prove two lemmata that we apply in Section C.

Definition 19 (Polyhedral complex) *A collection of polyhedra $\mathcal{P}$ is called a* polyhedral complex *if*

1. *Every face F of any polyhedra $P \in \mathcal{P}$ is also in $\mathcal{P}$ and*
2. *it holds that $P \cap Q \in \mathcal{P}$ for all $P, Q \in \mathcal{P}$.*

There is a poset structure given on $\mathcal{P}$ by $Q \preceq P :\iff Q$ is a face of P and we call $(\mathcal{P}, \preceq)$ the face poset *of the polyhedral complex. Furthermore we denote $\mathcal{P}_k$ the set of k-dimensional polyhedra in $\mathcal{P}$ and by $\operatorname{sk}_k(\mathcal{P})$ the k-skeleton of $\mathcal{P}$. Note that for any polyhedron, the set of all its faces forms a polyhedral complex.*

Definition 20 (Isomorphisms of polyhedral complexes and polytopes) *Let $\mathcal{P}$ and $\mathcal{Q}$ be polyhedral complexes. A map $\varphi\colon \mathcal{P} \to \mathcal{Q}$ is called an isomorphism if it is an isomorphism of the face posets of $\mathcal{P}$ and $\mathcal{Q}$ and it holds that $\dim(\varphi(P)) = \dim(P)$ for all $P \in \mathcal{P}$.*

If P and Q are polytopes we call a map $\varphi\colon P \to Q$ an isomorphism if it is an isomorphism when considering P and Q as polyhedral complexes.

Definition 21 *We call $\varphi\colon \mathcal{P} \to \mathcal{Q}$ an ε-*isomorphism *if it is an isomorphism (of polyhedral complexes) and it holds that $\|\varphi(v) - v\|_2 < \varepsilon$ for all $v \in \mathcal{P}_0$.*

Definition 22 *Let $x \mapsto a^T x + b$ be an affine linear map and $H(a, b) \coloneqq \{x \in \mathbb{R}^d \mid a^T x + b = 0\}$ the hyperplane given by the kernel. Then we denote the corresponding half-spaces by*

$$H^1(a, b) \coloneqq \{x \in \mathbb{R}^d \mid a^T x \geq b\},$$

$$H^{-1}(a, b) \coloneqq \{x \in \mathbb{R}^d \mid a^T x \leq b\}.$$

We will also use the notation $H^0(a, b) \coloneqq H(a, b)$. We will simply write H^s for $H^s(a, b)$ whenever a and b are clear from the context.

Lemma 23 *Let $P \subseteq \mathbb{R}^d$ be a polytope, $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $P_0 \cap H(a, b) = \emptyset$. Then for all $\varepsilon > 0$, there is a $\delta > 0$ such that for all $(a', b') \in B_\delta^{d+1}((a, b))$ there are ε-isomorphisms*

$$\psi^s\colon P \cap H^s(a, b) \to P \cap H^s(a', b')$$

for $s \in \{-1, 0, 1\}$, and it holds that $P_0 \cap H(a', b') = \emptyset$.

Proof Let $e \in P_1$ and $\mathbb{R}_e \coloneqq \operatorname{Aff}(e)$ be the affine space spanned by e. First, assume that

$$\mathbb{R}_e \cap H(a, b) \neq \emptyset.$$

Since $H(a,b) \cap P_0 = \emptyset$ we know that $\mathbb{R}_e \cap H(a,b) = \{v_e^{(a,b)}\}$ with $v_e^{(a,b)} \in e^\circ$ or $v_e^{(a,b)} \in \mathbb{R}_e \setminus e$, where e° denotes the relative interior of e. Let

$$\varepsilon_e := \begin{cases} \min\{\varepsilon, \frac{1}{2} \inf_{y \in e^\circ} \|y - v_e^{(a,b)}\|_\infty\} & v_e^{(a,b)} \in e^\circ \\ \min\{\varepsilon, \frac{1}{2} \inf_{y \in \mathbb{R}_e \setminus e} \|y - v_e^{(a,b)}\|_\infty\} & v_e^{(a,b)} \in \mathbb{R}_e \setminus e \end{cases}$$

It is easily verified that the map $(c,d) \mapsto H(c,d) \cap \mathbb{R}_e$ is locally continuous around (a,b) and hence there is a $\delta_e > 0$ such that $\|(a,b) - (a',b')\| < \delta_e$ implies that $\|v_e^{(a,b)} - v_e^{(a',b')}\|_\infty < \varepsilon_e$ for all $e \in P_1$. On the other hand, if $\mathbb{R}_e \cap H(a,b) = \emptyset$, then there is a $\delta_e > 0$ such that $e^\circ \cap H(a',b') = \emptyset$. Let $\delta := \min_{e \in P_1} \delta_e$. Note that

$$(P \cap H(a',b'))_0 = \{v_e^{(a',b')} \mid v_e^{(a',b')} \in e^\circ\}$$

and hence $f(v_e^{(a,b)}) := v_e^{(a',b')}$ defines a bijection

$$f\colon (P \cap H)_0 \to (P \cap H')_0$$

for $(a',b') \in B_\delta^{d+1}((a,b))$. Let F be a face of $P \cap H(a,b)$, then $F = F' \cap H(a,b)$ for some face F' of P and furthermore $F = \operatorname{conv}(\{v_e^{(a,b)} \cap e \mid e \preceq F\}$. It now easily follows by induction on the dimension of the face F that F is isomorphic to $\operatorname{conv}(\{v_e^{(a',b')} \cap e \mid e \preceq F\}$ and therefore in particular that $P \cap H(a,b)$ is isomorphic to $P \cap H(a',b')$. We can extend f to a bijection $f\colon (P \cap H^s(a,b))_0 \to (P \cap H^s(a',b'))_0$ by $f(v) = v$ for all $v \in P_0 \cap H^s(a,b)$ and by the same arguments we obtain that $P \cap H^s(a,b)$ is isomorphic to $P \cap H^s(a',b')$ for $s \in \{-1,1\}$. ∎

Lemma 24 *Let $P \subseteq \mathbb{R}^d$ be a polytope and let $H = \{x \in \mathbb{R}^d \mid a^T x = b\}$ be a hyperplane. If $P_0 \cap H = \emptyset$ then for all $\varepsilon > 0$ there is a $\delta > 0$ such that for all polytopes $Q \subseteq \mathbb{R}^d$ with δ-isomorphisms $\varphi\colon P \to Q$ there are ε-isomorphisms*

$$\gamma^s\colon P \cap H^s \to Q \cap H^s$$

for $s \in \{-1,0,1\}$ and furthermore it holds that $Q_0 \cap H = \emptyset$.

Proof Let $e \in P_1$ and let $\partial e = \{u,v\}$. We adopt the notation $e = uv$ and define

$$\delta_e := \min\{\varepsilon, \frac{1}{2} \inf_{y \in H} \|y - v\|_\infty, \frac{1}{2} \inf_{y \in H} \|y - u\|_\infty\}$$

and $\delta := \min_{e \in P_1} \delta_e$. Since $P_0 \cap H = \emptyset$ it holds that $\delta > 0$. Let $\varphi\colon P \to Q$ be a δ-isomorphism. Then it holds $H \cap uv = \{v_e\} \neq \emptyset$ if and only if $H \cap \varphi(u)\varphi(v) = \{v_{\varphi(e)}\} \neq \emptyset$. Note that $(P \cap H)_0 = \{v_e \mid H \cap e \neq \emptyset\}$ and $(Q \cap H)_0 = \{v_{\varphi(e)} \mid H \cap \varphi(e) \neq \emptyset\}$ and hence $f(v_e) := v_{\varphi(e)}$ defines a bijection $f\colon (P \cap H)_0 \to (Q \cap H)_0$. The remainder of the proof follows analogously to the proof of Lemma 23.

∎

A.2. Topology

In the following, we summarize background knowledge necessary for our purposes that the reader may not have been acquainted with. The content of this subsection can also be found in many algebraic topology textbooks such as Hatcher (2002, Chapter 2).

First, we recall two well-known constructions in topology that yield well-behaved, yet more complex topological spaces.

Definition 25 *For two topological spaces X and Y, the space $X \sqcup Y$ denotes the* disjoint union *of X and Y endowed with the disjoint union topology. Similarly for an arbitrary index set I, the set $\bigsqcup_{i \in I} X_i$ denotes the disjoint union of the topological spaces X_i for $i \in I$. If I is a finite set, i.e., $I = \{1, \ldots, q\}$ for a suitable $q \in N$, we also denote this space by $\bigsqcup_{i=1}^{q} X_i$.*

We also create *product spaces*: For two topological spaces X and Y, the product space is the Cartesian product $X \times Y$ endowed with the product topology. Even though it is possible to extend this definition to infinite families of topological spaces as well, this will not be needed for our purposes.

To relate topological spaces X, Y with each other, one often defines *maps*, i.e., continuous functions $f\colon X \to Y$, and investigates properties of such maps. In our work, we use the following special cases of maps:

Definition 26 *Let X, Y be topological spaces.*

(i) *A map $f\colon X \to Y$ is called a* homeomorphism *if it is open and bijective. In this case, we call X and Y* homeomorphic.

(ii) *A map $F\colon X \times [0,1] \to Y$ is a* homotopy between functions $f\colon X \to Y$ and $g\colon X \to Y$ *if $F(x,0) = f(x)$ and $F(x,1) = g(x)$ for all $x \in X$. In this case, the maps f and g are called* homotopic*, denoted by $f \sim g$.*

(iii) *A map $f\colon X \to Y$ is a* homotopy equivalence *if there exists a map $g\colon Y \to X$ such that $g \circ f$ resp. $f \circ g$ are homotopic to id_X resp. id_Y. The spaces X and Y are called* homotopy equivalent *in this case.*

(iv) *Let $A \subseteq X$ be a subspace of X. A map $r\colon X \to A$ is called a* retraction *if $r(a) = a$ for all $a \in A$.*

(v) *A homotopy between a retraction $r\colon X \to A$ (or more precisely, a map $r\colon X \to X$ with $r(X) \subseteq A$ and $r|_A = \mathrm{id}_A$) and the identity map id_X is called a* deformation retraction.

Next, we introduce the notion of homology by giving a sketch of the construction of homology groups.

Let X be a topological space and

$$\Delta_n = \left\{ \sum_{i=0}^{n} \theta_i x_i \colon x \in \mathbb{R}^n, \sum_{i=0}^{n} \theta_i = 1, \theta_i \geq 0 \text{ for all } i = 0, \ldots, n \right\}$$

denote the standard n-simplex. Note that the standard n-simplex is the convex combination of $n+1$ points $\{p_0, \ldots, p_n\}$. Taking the convex combination of an n-subset $\{p_0, \ldots, p_n\} \setminus \{p_i\}$ of these points, one obtains a subspace homeomorphic to the standard $n-1$-simplex, which we call an *i-th n-face* of the simplex.

The $\mathbb{Z}$-module C_n, the group of *n-chains*, is defined as the free abelian group generated by continuous maps $\sigma\colon \Delta_n \to X$, called *simplices*. The inclusion $\iota_i\colon \Delta_{n-1} \hookrightarrow \Delta_n$ induces $n-1$-simplices $\sigma_i := \sigma \circ \iota_i\colon \Delta_{n-1} \to X$ by the inclusion of the i-th n-face into the standard simplex.

The map $\partial_n\colon C_n \to C_{n-1}$, which we call the *boundary map*, is constructed as $\sigma \mapsto \sum_{i=1}^{n}(-1)^i\sigma_i$ on the generators and by linear extension elsewhere. This yields a *chain complex*

$$\ldots \to C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \to \ldots,$$

where we have $\partial_{n-1} \circ \partial_n = 0$ for all n. Therefore, we can define the *n-th singular homology group* as

$$H_n(X) := \ker(\partial_n) / \operatorname{im}(\partial_{n+1}).$$

We list some well-known properties of (singular) homology groups that will be used in our constructions.

Proposition 27 *Let $n \in \mathbb{N}$.*

1. *(Disjoint union axiom, implication) For any index set I and topological spaces X_i for $i \in I$, it holds that $H_n\left(\bigsqcup_{i\in I} X_i\right) \cong \bigoplus_{i\in I} H_n(X_i)$.*
2. *(Homotopy invariance axiom) Let X, Y be topological spaces and $r\colon X \to Y$ a homotopy equivalence. Then the map $H_n(r)$ is an isomorphism for all $n \in \mathbb{N}$. In particular, it holds that $H_n(Y) \cong H_n(X)$.*
3. *(Dimension axiom)* $H_n(D^d) = \begin{cases} \mathbb{Z}, & n = 0 \\ 0 & \text{else} \end{cases}$
4. $H_n(S^d) = \begin{cases} \mathbb{Z}\oplus\mathbb{Z} & n = d = 0 \\ \mathbb{Z}, & n = d \neq 0 \text{ or } d \neq n = 0 \\ 0 & \text{else} \end{cases}$

Observation 28 *Using Proposition 27 and given definitions, one can immediately calculate the homology groups of a d-dimensional k-annuli:*

$$H_n(S^k \times D^{d-k}) = H_n(S^k) = \begin{cases} \mathbb{Z}\oplus\mathbb{Z} & n = k = 0 \\ \mathbb{Z}, & n = k \neq 0 \text{ or } k \neq n = 0 \\ 0 & \text{else} \end{cases}$$

To ease our computations for upper bounds, we deviate to another homology theory called *cellular homology* which is defined on a special class of topological spaces called *CW-complexes*.

Definition 29 *A Hausdorff space X with a filtration $\emptyset = X_{-1} \subseteq X_0 \subseteq \ldots \subseteq \bigcup_{i=1}^{d} X_d = X$ is a* d-dimensional finite CW complex *if the following axioms hold:*

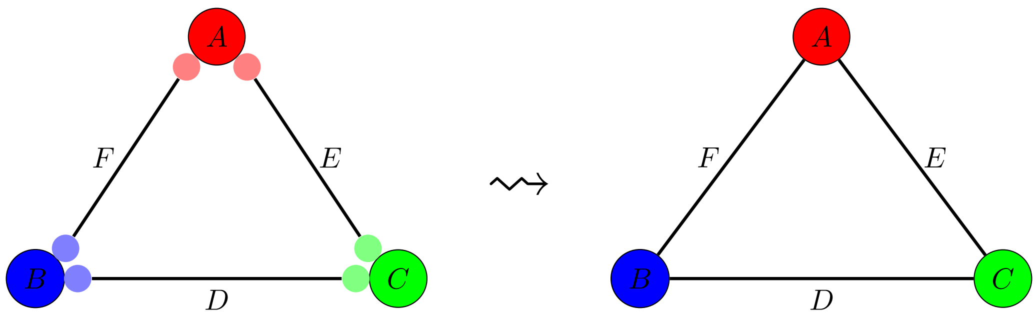


Figure 7: A CW-complex X of dimension 1 that is homeomorphic to S^1. The darker shaded points constitute the 0-cells, i.e., we have $X_0 = \{A, B, C\}$. The line segments on the left are the 1-cells. The triangle on the right illustrates $X = X_1$. The lighter shades of colors indicate the attaching maps $q_1^j \colon S^0 \to \{A, B, C\}$ for $j \in \{D, E, F\}$.

(i) A subset $A \subseteq X$ is closed in X if and only if $A \cap X_i$ is closed in X_i for all $i \in [d]_0$.

(ii) The spaces X_i in the filtration are each called i-skeleton. The i-skeleton is recursively obtained from X_{i-1} by attaching cells, i.e. we have pushout maps of the form

$$\begin{array}{ccc} \bigsqcup_{j\in I_i} S^{i-1} & \xrightarrow{\bigsqcup_{j\in I_i} q_i^j} & X_{i-1} \\ \big\downarrow & & \big\downarrow \\ \bigsqcup_{j\in I_i} D^{i} & \xrightarrow{\bigsqcup_{j\in I_i} Q_i^j} & X_{i} \end{array}$$

for finite index sets I_i for $i \in [d]_0$. The maps q_i^j are called attaching maps *and the maps Q_i^j are called* characteristic maps.

For $i \in [d]_0$, the set of path components of $X_i \setminus X_{i-1}$ is called the set of open i-cells. *The set of closures of open i-cells are called* closed i-cells. *We almost always make use of closed i-cells and therefore refer to them simply as* i-cells.

The non-expert can understand a pushout map as one that simply glues the boundary of an i-dimensional cell (that is, a topological disk/polyhedron of dimension i) onto the $(i-1)$-skeleton X_i. Here, the choice of the attaching maps define the commutative pushout diagram above, while the characteristic maps are those that are "uniquely" defined by the attaching maps "in a natural way". Figure 7 illustrates the above definition.

A more general definition of CW complexes allow an infinte dimension of cells as well as an (arbitrarily indexed) infinite number of cells in every dimension. For our results, it is sufficient to restrict to our definition, as we will restrict to a CW complex that is a compact topological space without loss of generality:

Lemma 30 *A CW-complex is finite if and only if it is compact.*

One can naturally endow finite polyhedral complexes with CW-structures by defining the i-cells as the i-facets (therefore uniquely defining the filtration), and the attaching maps as those that include each face into the i-skeleton of the CW complex for each i. This way, the poset structure is compatible with the characteristic maps as well (because faces of polyhedra lie on their topological boundary and the attaching maps q_i^j are injective in this case). One can also observe that any polyhedral complex is homotopy equivalent to a finite CW-complex, see 51.

Our motivation to endow polyhedral complexes with CW-structures is to use *cellular homology*, which, given the natural CW-structure of a polyhedral complex, allows to compute homology groups conveniently. Among other advantages, we will rely on cellular homology for an induction on the number of polyhedra.

Definition 31 *Let X be a finite CW-complex. The* cellular chain complex $(C_i)_{i\in\mathbb{N}}$ *of X is given by free abelian groups $C_i \cong \mathbb{Z}^{|I_i|}$ that are generated by the i-cells of X. The boundary maps, which are given by a construction using attachment maps in the general case, can be greatly simplified for our purposes: The boundary map $\partial_i \colon C_i \to C_{i-1}$ is defined by the incidence matrix $\Delta \in \mathbb{Z}^{|I_{i-1}|\times|I_i|}$ between i and $(i-1)$-cells, that is, it is given by entries*

$$\delta_{jk} = \begin{cases} 1 & \text{the } (i-1)\text{-cell } j \text{ lies in the boundary of the } i\text{-cell } k \\ 0 & \text{else.} \end{cases}$$

The i-th cellular homology $H_i^{cell}(X)$ *of X is defined by the homology of the cellular chain complex $(C_i)_{i\in\mathbb{N}}$, that is, we have*

$$H_i^{cell}(X) = {}^{\ker(\partial_i)}\!\big/\!_{\operatorname{im}(\partial_{i-1})}.$$

It is well-known that on CW-complexes, cellular homology groups coincide with singular homology groups. Therefore, we may make use of cellular homology groups in order to compute Betti numbers.

Appendix B. Lower Bound

In this section we provide formal proofs for the statements made in sections 2 and additional lemmata that are used in these proofs. For the sake of completeness, we also recall the statements we prove.

B.1. Proof of Lemma 5

Definition 32 *We define the one hidden layer ReLU neural network $h^{(1,m,d)}$ in the following way: The neurons $\{v_{i,j}\}_{i=0,\dots,m-1,j=1,\dots,d}$ in the hidden layer are given by:*

- $v_{0,j}(x) = \max\{0, mx_j\}, j = 1,\dots,d$
- $v_{i,j}(x) = \max\{0, 2m(x_j - i/m)\}, j = 1,\dots,d\; i = 1,\dots,m-1$

and the output neurons by: $h_j^{(1,m,d)}(x) = \sum_{i=0}^{m-1}(-1)^i \cdot v_{i,j}(x)$.

Lemma 33 *Let $d, m \in \mathbb{N}$ with $m > 1$. Then*

1. $h^{(1,m,d)}(W^{(1,m,d)}_{(i_1,\dots,i_d)}) = [0,1]^d$

2. $\pi_j \circ h^{(1,m,d)}_{|W^{(1,m,d)}_{(i_1,\ldots,i_d)}}(x_1,\ldots,x_d) = \begin{cases} m \cdot x_j - (i_j - 1) & i_j \text{ odd} \\ -m \cdot x_j + i_j & i_j \text{ even} \end{cases}$

for all $(i_1,\ldots,i_d) \in [m]^d$.

Proof Throughout this proof we denote $W^{(1,m,d)}_{(i_1,\ldots,i_d)}$ by $W_{(i_1,\ldots,i_d)}$ and $h^{(1,m,d)}$ by h. We prove that h satisfies the second property. The first property then follows immediately from the second since

$$\pi_j \circ h_{|W_{(i_1,\ldots,i_d)}}(W_{(i_1,\ldots,i_d)}) = \left[m \cdot \frac{(i_j-1)}{m} - (i_j-1), m \cdot \frac{i_j}{m} - (i_j-1)\right] = [0,1]$$

if i_j is odd and

$$\pi_j \circ h_{|W_{(i_1,\ldots,i_d)}}(W_{(i_1,\ldots,i_d)}) = \left[m \cdot \frac{(i_j-1)}{m} + i_j, m \cdot \frac{i_j}{m} + i_j\right] = [0,1]$$

if i_j is even.

Let $j \in \{1,\ldots,d\}$ and $x \in W_{(i_1,\ldots,i_d)}$, so in particular $x_j \in \left[\frac{(i_j-1)}{m}, \frac{i_j}{m}\right]$. Since $i \geq i_j$ implies $2m(x_j - i/m) \leq 0$, it follows that $v_{i,j}(x) = 0$ for all $i \geq i_j$. Similarly, $i < i_j$ implies $2m(x_j - i/m) \geq 0$, and therefore it follows that $v_{i,j}(x) = 2m(x_j - i/m)$ for all $i < i_j$. Hence

$$h_j(x) = \sum_{i=0}^{i_j-1} (-1)^i \cdot v_{i,j}(x) = mx_j + \sum_{i=1}^{i_j-1} (-1)^i \cdot 2m(x_j - i/m).$$

If i_j is even, then

$$\begin{aligned} h_j(x) &= mx_j + \sum_{i=1}^{i_j/2-1} 2m(x_j - 2i/m) - \sum_{i=1}^{i_j/2} 2m(x_j - (2i-1)/m) \\ &= mx_j - 2(i_j/2 - 1) - 2m(x_j - (2i_j/2 - 1)/m) \\ &= mx_j - i_j + 2 - 2mx_j + 2i_j - 2 \\ &= -mx_j + i_j \end{aligned}$$

If i_j is odd, then

$$\begin{aligned} h_j(x) &= mx_j + \sum_{i=1}^{(i_j-1)/2} 2m(x_j - 2i/m) - \sum_{i=1}^{(i_j-1)/2} 2m(x_j - (2i-1)/m) \\ &= mx_j - 2(i_j - 1/2) \\ &= mx_j - (i_j - 1). \end{aligned}$$

■

Lemma 34 *(cf. (Montúfar et al., 2014)) Let* $d \in \mathbb{N}$. *Then there exists a map* $h^{(L,\mathbf{m},d)}: \mathbb{R}^d \to \mathbb{R}^d$ *such that for all* $(i_1,\ldots,i_d) \in [M]^d$*, the following hold:*

1. $h^{(L,\mathbf{m},d)}(W^{(L,\mathbf{m},d)}_{(i_1,\dots,i_d)}) = [0,1]^d$

2. $\pi_j \circ h^{(L,\mathbf{m},d)}_{|W^{(L,\mathbf{m},d)}_{(i_1,\dots,i_d)}}(x_1,\dots,x_d) = \begin{cases} M \cdot x_j - (i_j - 1) & i_j \text{ odd} \\ -M \cdot x_j + i_j & i_j \text{ even} \end{cases}$

Proof

We apply induction over L. The base case has already been covered by Lemma 33. Assume that there exists a NN $h^{(L-1,(m_1,\dots,m_{L-1}),d)}$ that satisfies the desired properties and define $h^{(L,\mathbf{m},d)} = h^{(1,m_L,d)} \circ h^{(L-1,(m_1,\dots,m_{L-1}),d)}$. Let $j \in \{1,\dots,d\}$ and $x \in W^{(L,\mathbf{m},d)}_{(i_1,\dots,i_d)}$. Define $i_j^{(1)} := \left\lfloor \frac{m_L \cdot (i_j - 1)}{M} \right\rfloor + 1$. It holds that $\left[\frac{(i_j-1)}{M}, \frac{i_j}{M}\right] \subset \left[\frac{(i_j^{(1)}-1)}{m_L}, \frac{i_j^{(1)}}{m_L}\right]$.

Case 1: $i_j^{(1)}$ odd. Then by Lemma 33:

$$\begin{aligned} h_j^{(1,m_L,d)}\left(\tfrac{(i_j-1)}{M}\right) &= m_L \cdot \frac{(i_j-1)}{M} - (i_j^{(1)} - 1) = \frac{m_L \cdot (i_j-1)}{M} - \left\lfloor \frac{m_L \cdot (i_j-1)}{M} \right\rfloor \\ &= \frac{(i_j-1) \bmod (M/m_L)}{(M/m_L)} \end{aligned}$$

and

$$\begin{aligned} h_j^{(1,m_L,d)}\left(\tfrac{i_j}{M}\right) &= m_L \cdot \frac{i_j}{M} - (i_j^{(1)} - 1) = \frac{m_L \cdot (i_j)}{M} - \left\lfloor \frac{m_L \cdot (i_j-1)}{M} \right\rfloor \\ &= \frac{(i_j-1) \bmod (M/m_L) + 1}{(M/m_L)}. \end{aligned}$$

Define $i_j^{(L-1)} := ((i_j - 1) \bmod (M/m_L)) + 1$. Then it holds that

$$h_j^{(1,m_L,d)}(x) \in \left[\frac{m_L \cdot (i_j^{(L-1)} - 1)}{M}, \frac{m_L \cdot i_j^{(L-1)}}{M}\right].$$

Moreover, $i_j^{(L-1)} = ((i_j - 1) \bmod (M/m_L)) + 1$ is odd if and only if i_j is odd because $\frac{M}{m_L}$ is an even number.

Case 1.i: $i_j^{(L-1)}$ (and therefore i_j) is odd. Then follows with the induction hypothesis:

$$\begin{aligned} h_j^{(L,\mathbf{m},d)}(x) &= h_j^{(L-1,(m_1,\dots,m_{L-1}),d)}(h_j^{(1,m_L,d)}(x)) \\ &= \frac{M}{m_L} \cdot (h_j^{(1,m_L,d)}(x)) - (i_j^{(L-1)} - 1) \\ &= \frac{M}{m_L} \cdot (m_L x - (i_j^{(1)} - 1)) - (i_j^{(L-1)} - 1) \\ &= Mx - \frac{M}{m_L} \cdot \left\lfloor \frac{m_L \cdot (i_j-1)}{M} \right\rfloor - ((i_j - 1) \bmod (M/m_L)) \\ &= Mx - (i_j - 1). \end{aligned}$$

Case 1.ii: $i_j^{(L-1)}$ (and therefore i_j) is even. Then follows with the induction hypothesis:

$$\begin{aligned}
h_j^{(L,\mathbf{m},d)}(x) &= h_j^{(L-1,(m_1,\ldots,m_{L-1},d)}(h_j^{(1,m_L,d)}(x)) \\
&= -\frac{M}{m_L}\cdot(h_j^{(1,m_L,d)}(x)) + i_j^{(L-1)} \\
&= -\frac{M}{m_L}\cdot(m_L x - (i_j^{(1)}-1)) + i_j^{(L-1)} \\
&= -Mx + \frac{M}{m_L}\cdot\left\lfloor\frac{m_L\cdot(i_j-1)}{M}\right\rfloor + ((i_j-1) \bmod (M/m_L)+1) \\
&= -Mx + i_j - 1 + 1 \\
&= -Mx + i_j.
\end{aligned}$$

Case 2: $i_j^{(1)}$ even. Then by Lemma 33:

$$\begin{aligned}
h_j^{(1,m_L,d)}(\tfrac{(i_j-1)}{M}) &= -m_L\cdot\frac{(i_j-1)}{M} + i_j^{(1)} = -\frac{m_L\cdot(i_j-1)}{M} + \left\lfloor\frac{m_L\cdot(i_j-1)}{M}\right\rfloor + 1 \\
&= 1 - \frac{(i_j-1) \bmod (M/m_L)}{M/m_L}
\end{aligned}$$

and

$$\begin{aligned}
h_j^{(1,m_L,d)}(\tfrac{i_j}{M}) &= -m_L\cdot\frac{i_j}{M} + i_j^{(1)} = -m_L\cdot\frac{i_j}{M} + \left\lfloor m_L\cdot\frac{(i_j-1)}{M}\right\rfloor + 1 \\
&= 1 - \frac{(i_j-1) \bmod (M/m_L))-1}{M/m_L}
\end{aligned}$$

Define $i_j^{(L-1)} := \frac{M}{m_L} - ((i_j-1) \bmod (M/m_L))$. Then it holds that

$$h_j^{(1,m_L,d)}(x) \in \left[\frac{m_L\cdot(i_j^{(L-1)}-1)}{M}, \frac{m_L\cdot(i_j^{(L-1)})}{M}\right].$$

Moreover, $i_j^{(L-1)}$ is even if and only if i_j is odd, once more because $\frac{M}{m_L}$ is an even number.

Case 2.i: $i_j^{(L-1)}$ is odd (i.e., i_j even). Then follows with the induction hypothesis:

$$\begin{aligned}
h_j^{(L,\mathbf{m},d)}(x) &= h_j^{(L-1,(m_1,\ldots,m_{L-1}),d)}(h_j^{(1,m_L,d)}(x)) \\
&= \frac{M}{m_L}\cdot(h_j^{(1,m_L,d)}(x)) - (i_j^{(L-1)}-1) \\
&= \frac{M}{m_L}\cdot(-m_L x + i_j^{(1)}) - (i_j^{(L-1)}-1) \\
&= -Mx + \frac{M}{m_L}\cdot\left\lfloor\frac{m_L\cdot(i_j-1)}{M}\right\rfloor + \frac{M}{m_L} - \left(\frac{M}{m_L} - ((i_j-1) \bmod M/m_L) - 1\right) \\
&= -Mx + i_j
\end{aligned}$$

Case 2.ii: $i_j^{(L-1)}$ is even (i.e., i_j odd). Then follows with the induction hypothesis:

$$\begin{aligned}
h_j^{(L,\mathbf{m},d)}(x) &= h_j^{(L-1,(m_1,\ldots,m_{L-1}),d)}(h_j^{(1,m_L,d)}(x)) \\
&= -\frac{M}{m_L}\cdot(h_j^{(1,m_L,d)}(x)) + i_j^{(L-1)} \\
&= -\frac{M}{m_L}\cdot(-m_L x + i_j^{(1)}) + i_j^{(L-1)} \\
&= Mx - \frac{M}{m_L}\cdot\left(\left\lfloor\frac{m^L(i_j-1)}{M}\right\rfloor + 1\right) + \frac{M}{m_L} - ((i_j-1) \bmod (M/m_L) + 1) + 1 \\
&= Mx - (i_j - 1).
\end{aligned}$$

This concludes the proof for all cases.

■

B.2. Proof of Lemma 7

Lemma 35 *Let $d, w \in \mathbb{N}$ and*

$$R_q = \{x \in \mathbb{R}^d : x_1,\ldots,x_d > 0, \frac{q}{4w} < \|x\|_1 < \frac{q+1}{4w}\}.$$

Then $\operatorname{sgn}(\hat{g}(R_q)) = (-1)^q$ *for all* $q = 0,\ldots,w-1$ *and* $\hat{g}(x) = 0$ *for all* $x \in [0,1]^d$ *with* $\|x\|_1 \geq \frac{1}{2}$.

Proof Let $q \in \{0,\ldots,w-1\}$ and $x \in R_q$. Note that $\hat{g}_0(x) = \mathbf{1}^T x$ for all $q \in \{0,\ldots,w-1\}$.

Case 1: $\mathbf{1}^T x < (2q+1)/4w$. This implies $\hat{g}_i(x) = 0 \; \forall q > i$ and $g_i(x) = 2(\mathbf{1}^T x - ((2i-1)/4w))$ for all $1 < i \leq q$ and therefore

$$\hat{g}(x) = \sum_{i=0}^{q}(-1)^i \hat{g}_i(x) = x_1 + \sum_{i=1}^{q}(-1)^i 2(\mathbf{1}^T x - ((2i-1)/4w)).$$

Case 1.i: If q is even, then it holds:

$$\begin{aligned}
\hat{g}(x) &= \mathbf{1}^T x + \sum_{i=1}^{q/2} 2(\mathbf{1}^T x - ((2(2i)-1)/4w)) - \sum_{i=1}^{q/2} 2(\mathbf{1}^T x - ((2(2i-1)-1)/4w)) \\
&= \mathbf{1}^T x + \sum_{i=1}^{q/2} 2(\mathbf{1}^T x - ((4i-1)/4w)) - \sum_{i=1}^{q/2} 2(\mathbf{1}^T x - ((4i-3)/4w)) \\
&= \mathbf{1}^T x - q/2w > 0
\end{aligned}$$

Case 1.ii: If q is odd, then it holds:

$$\begin{aligned}
\hat{g}(x) &= \mathbf{1}^T x + \sum_{i=1}^{(q-1)/2} 2(\mathbf{1}^T x - ((4i-1)/4w)) - \sum_{i=1}^{(q+1)/2} 2(\mathbf{1}^T x - ((4i-3)/4w)) \\
&= \mathbf{1}^T x - 2(q-1)/4w - 2(\mathbf{1}^T x - ((4(q+1)/2) - 3)/4w)) \\
&= -(\mathbf{1}^T x) - 2(q-1)/4w + (4(q+1)-6)/4w \\
&= -(\mathbf{1}^T x) + q/2w < 0
\end{aligned}$$

Case 2: $\mathbf{1}^T x \geq (2q+1)/4w$. This implies $\hat{g}_i(x) = 0\ \forall i > q+1$ and

$$g_i(x) = 2(\mathbf{1}^T x - ((2i-1)/4w))$$

for all $1 < i \leq q+1$ and therefore

$$\hat{g}(x) = \sum_{i=0}^{q+1} (-1)^q \hat{g}_i(x) = x_1 + \sum_{i=1}^{q+1} (-1)^i 2(\mathbf{1}^T x - ((2i-1)/4w)).$$

Case 2.i: If q is even, then it holds:

$$\begin{aligned}
\hat{g}(x) &= \mathbf{1}^T x + \sum_{i=1}^{q/2} 2(\mathbf{1}^T x - ((4i-1)/4w)) - \sum_{i=1}^{q/2+1} 2(\mathbf{1}^T x - ((4i-3)/4w)) \\
&= \mathbf{1}^T x - 2q/4w - 2(\mathbf{1}^T x - ((4(q/2+1)-3)/4w)) \\
&= -(\mathbf{1}^T x) - q/w + 2(2q+1)/4w \\
&= -(\mathbf{1}^T x) + (q+1)/2w > 0
\end{aligned}$$

Case 2.ii: If q is odd, then it holds:

$$\begin{aligned}
\hat{g}(x) &= \mathbf{1}^T x + \sum_{i=1}^{(q+1)/2} 2(\mathbf{1}^T x - ((4i-1)/4w)) - \sum_{i=1}^{(q+1)/2} 2(\mathbf{1}^T x - ((4i-3)/4w)) \\
&= \mathbf{1}^T x - (q+1)/2w < 0
\end{aligned}$$

and hence $\operatorname{sgn}(\hat{g}(x)) = (-1)^q\ \forall x \in R_q\ \forall q = 1, \ldots, w-1$.

Let $x \in [0,1]^d$ with $\mathbf{1}^T x \geq \frac{1}{2}$.

Case 1: w even. Then

$$\begin{aligned}
\hat{g}(x) &= \sum_{q=0}^{w+1} (-1)^q \cdot \hat{g}_q(x) \\
&= \mathbf{1}^T x - (\mathbf{1}^T x - \frac{1}{2}) + \sum_{q=1}^{w/2} \hat{g}_{2_q}(x) - \hat{g}_{2_{q-1}}(x) \\
&= \frac{1}{2} + \sum_{q=1}^{w/2} 2(\mathbf{1}^T x - (2 \cdot 2q - 1)/4w) - 2(\mathbf{1}^T x - (2 \cdot (2q-1)) - 1)/4w)) \\
&= \frac{1}{2} + \sum_{q=1}^{w/2} 2(-(4q-1)/4w + (4q-3)/4w) \\
&= \frac{1}{2} + \sum_{q=1}^{w/1} -1/2w \\
&= 0
\end{aligned}$$

Case 2: w odd. Then

$$\begin{aligned}\hat{g}(x) &= \sum_{q=0}^{w+1}(-1)^q \cdot \hat{g}_q(x) \\ &= \hat{g}_0(x) - \hat{g}_w(x) + \hat{g}_{w+1}(x) + \sum_{q=1}^{w-1/2} \hat{g}_{2_q}(x) - \hat{g}_{2_{q-1}}(x) \\ &= \mathbf{1}^T x - 2(\mathbf{1}^T x - (2w-1)/4w) + \left(\mathbf{1}^T x - \frac{1}{2}\right) + \sum_{q=1}^{w-1/2} -1/w \\ &= (2w-1)/2w - \frac{1}{2} + \sum_{q=1}^{w-1} -1/2w \\ &= 1 - 1/2w - \frac{1}{2} - \left(\frac{1}{2} - 1/2w\right) \\ &= 0\end{aligned}$$

■

Lemma 36 *Let $d, w \in \mathbb{N}$ with w odd and*

$$R_q = \{x \in [0,1]^d : \frac{q}{4w} < \|(1,1,\dots,1,0) - x\|_1 < \frac{q+1}{4w}\}.$$

Then there exists a 1-hidden layer neural network $g^{(w,d)} \colon \mathbb{R}^d \to \mathbb{R}$ of width $w+2$ such that $g^{(w,d)}(R_q) \subseteq (-\infty, 0)$ for all odd $\in [w-1]_0$, $g^{(w,d)}(R_q) \subseteq (0,\infty)$ for all even $q \in [w-1]_0$ and $g^{(w,d)}(x) = 0$ for all $x \in [0,1]^d$ with $\|(1,1,\dots,1,0) - x\|_1 \geq \frac{1}{4}$.

Proof Let the affine map $t \colon [0,1]^d \to [0,1]^d$ be given by $x \mapsto (1-x_1, \dots, 1-x_{d-1}, x_d)$ and let $\hat{g}$ be the 1-hidden layer neural network from Lemma 35. We prove that the neural network $g := \hat{g} \circ t$ satisfies the assumptions. Let $q \in \{0, \dots, w-1\}$ and $x \in R_q$. Then

$$\|(1,1,\dots,1,0) - t(x)\|_1 = \|(1,1,\dots,1,0) - (1-x_1,\dots,1-x_{d-1},x_d)\|_1 = \|x\|_1.$$

Since $g(x) = g \circ t(x) = \hat{g}(t(x))$, Lemma 35 implies that $\operatorname{sgn}(g(R_q)) = (-1)^q$. Analogously follows that $g(x) = 0$ for all $x \in [0,1]^d$ with $\|(1,1,\dots,1,0) - x\|_1 \geq \frac{1}{4}$. ■

B.3. Proof of Proposition 9

Lemma 37 *Let $g^{(w,d)}$ be the NN from Lemma 7 and C the set of cutting points. Define*

$$R_{q,c} := B^d_{(q+1)/(4w\cdot M)}(c) \setminus \overline{B^d_{q/(4w\cdot M)}(c)}$$

for a cutting point $c \in C$ and $q \in \{1, \dots, w\}$. Then

1. *$g^{(w,d)} \circ h^{(L,\mathbf{m},d)}(R_{q,c}) \subseteq (-\infty, 0]$ for q odd,*

2. *$g^{(w,d)} \circ h^{(L,\mathbf{m},d)}(R_{q,c}) \subseteq [0, \infty)$ for q even*

3. *$x \notin \bigcup_{q\in[w],c\in C} R_{q,c}$ implies $g^{(w,d)} \circ h^{(L,\mathbf{m},d)}(x) = 0$.*

In particular, $g^{(w,d)} \circ h^{(L,\mathbf{m},d)}(x) = 0$ for all $x \in \partial R_{q,c}$.

Proof By definition of c being a cutting point, there exist odd numbers $i_1, \dots, i_{d-1} \in [M]$ and an even number $i_d \in [M]$ such that $c = (\frac{i_1-1}{M}, \dots, \frac{i_d-1}{M})$. Let $x \in [0,1]^d$ with $\|x - c\|_\infty \leq \frac{1}{M}$, then either $x_j \in \left[\frac{i_j-2}{M}, \frac{i_j-1}{M}\right]$ or $x_j \in \left[\frac{i_j-1}{M}, \frac{i_j}{M}\right]$. Let $J^+ := \{j \in [d] : x_j - c_j \leq 0\}$ be the set of indices j such that $x_j \in \left[\frac{i_j-1}{M}, \frac{i_j}{M}\right]$ and $J^- := [d] \setminus J^+$. Let $y = h^{(L,\mathbf{m},d)}(x) \in [0,1]^d$. Then it follows with Lemma 5 that

$$\begin{aligned}
\mathbf{1}^T y &= \sum_{j\in J^+} M \cdot x_j - (i_j - 1) + \sum_{j\in J^-} -M \cdot x_j + (i_j - 1) \\
&= \sum_{j\in J^+} M \cdot (x_j - c_j + c_j) - (i_j - 1) + \sum_{j\in J^-} -M \cdot (x_j - c_j + c_j) + (i_j - 1) \\
&= \sum_{j\in J^+} M \cdot (x_j - c_j) + \sum_{j\in J^+} M \cdot c_j - (i_j - 1) \\
&+ \sum_{j\in J^-} -M \cdot (x_j - c_j) + \sum_{j\in J^-} -M \cdot c_j + (i_j - 1) \\
&= \sum_{j\in J^+} M \cdot (x_j - c_j) + \sum_{j\in J^-} -M \cdot (x_j - c_j) \\
&= M \cdot \sum_{j=1}^{d} |x_j - c_j| \\
&= M \cdot \|x - c\|_1
\end{aligned}$$

If $x \in R_{q,c}$, then in particular $\|x - c\|_\infty \leq \frac{1}{M}$ and thus:

$$\mathbf{1}^T y = M \cdot \|x - c\|_1 < M \cdot \frac{q+1}{M \cdot 4w} = \frac{q+1}{4w}$$

and

$$\mathbf{1}^T y = M \cdot \|x - c\|_1 > M \cdot \frac{q}{M \cdot 4w} = \frac{q}{4w}.$$

If q is even, with Lemma 7 it follows that $g(y) \in (0, \infty)$ and therefore $g \circ h^{(L,\mathbf{m},d)}(x) = g(y)$. The case where q is odd follows analogously and this concludes the first two cases.
If x is not in any $R_{q,c}$, If x is not in any $R_{q,c}$, then either $x \in \partial R_{q,c}$ for some cutting point c and some q or it holds that $\|x - c\|_1 \geq \frac{1}{2M}$ for every cutting point c. In the first case it follows directly from the above shown that $g \circ h^{(L,m,d)}(x)) = 0$, since that $g \circ h^{(L,m,d)}$ is continuous. In the second case there exists a cutting point c such that $\|x - c\|_\infty \leq \frac{1}{M}$, since for every x_j either $\lfloor M \cdot x_j \rfloor$ or $\lceil M \cdot x_j \rceil$ is even. Thus $\mathbf{1}^T h^{(L,\mathbf{m},d)}(x) = M \cdot \|x - c\|_1 \geq M \cdot \frac{1}{4 \cdot M} = \frac{1}{4}$ and therefore it follows with Lemma 7 that $g \circ h^{(L,\mathbf{m},d)}(x) = 0$, which concludes the proof.

■

Proposition 9 *The space $Y_{d,w}$ is homeomorphic to the disjoint union of $p_d = \frac{M^{(d-1)}}{2^{d-1}} \cdot \left(\frac{M}{2} - 1\right) \cdot \lceil \frac{w}{2} \rceil$ many $(d-1)$-annuli and $p'_d = \frac{M^{(d-1)}}{2^{d-2}} \cdot \lceil \frac{w}{2} \rceil$ many disks, that is,*

$$Y_{d,w} \cong \bigsqcup_{k=1}^{p_d} (S^{d-1} \times [0,1]) \sqcup \bigsqcup_{k=1}^{p'_d} D^d.$$

Proof We observe that the sets $Y_{d,w} \cap \overline{B^d}_{1/4M}(c)$ are disjoint for cutting points c because we have $||c - c'||_1 \geq \frac{2}{M}$ for any two distinct cutting points c, c'. Therefore, the sets $Y_{d,w} \cap \overline{B}^d_{1/4M}(c)$ are pairwise disjoint for $c \in C$. Since

$$\bigsqcup_{c \in C} Y_{d,w} \cap \overline{B^d_{1/4M}(c)} = Y_{d,w},$$

the number of cutting points of the interior is $\frac{M^{\cdot(d-1)}}{2^{d-1}} \cdot \left(\frac{M}{2} - 1\right)$ and the number of the cutting points on the boundary is $\frac{M^{\cdot(d-1)}}{2^{d-2}}$ by Observation 8, it suffices to show that

$$Y_{d,w} \cap \overline{B^d_{1/4M}(c)} \cong \bigsqcup_{i=1}^{\lceil \frac{nw}{2} \rceil} S^{d-1} \times D^1$$

for every $c \in C \cap \mathrm{int}([0,1]^d)$ and $Y_{d,w} \cap \overline{B^d_{1/4M}}(c) \cong \bigsqcup_{i=1}^{\lceil \frac{w}{2} \rceil} D^d$ for every $c \in C \cap \partial[0,1]^d$.

By Lemma 37, we can see that for every $c \in C \cap \mathrm{int}([0,1]^d)$, we have

$$\begin{aligned} Y_{d,w} \cap \overline{B^d_{1/4M}(c)} &= \bigsqcup_{1 \leq q \leq w \text{ odd}} B^d_{q/(w\cdot 4M)}(c) \setminus \overline{B^d_{(q-1)/(w\cdot 4M)}(c)} \\ &\cong \bigsqcup_{1 \leq q \leq w \text{ odd}} S^{d-1} \times [0,1] \\ &= \bigsqcup_{q=1}^{\lceil \frac{w}{2} \rceil} S^{d-1} \times [0,1], \end{aligned}$$

as well as for every $c \in C \cap \partial([0,1]^d)$, we have

$$Y_{d,w} \cap \overline{B^d_{1/4M}(c)} \cong \bigsqcup_{1 \leq q \leq w \text{ odd}} \left(B^d_{q/(w\cdot 4M)}(c) \setminus \overline{B^d_{(q-1)/(w\cdot 4M)}(c)} \right) \cap [0,1]^d \cong \bigsqcup_{q=1}^{\lceil \frac{w}{2} \rceil} D^d,$$

proving the claim. ■

B.4. Proof of Lemma 10

Lemma 38 *For $k \leq d$ and $\mathbf{w} = (w_1, \ldots, w_{d-1}) \in \mathbb{N}^{d-1}$ it holds that*

1. $f^{(w_1,\ldots,w_{k-2})} \circ p_{k-1}(x) \neq 0 \implies g^{(w_{k-1},k)}(x) = 0$ *and*

2. $g^{(w_{k-1},k)}(x) \neq 0 \implies f^{(w_1,\dots,w_{k-2})} \circ p_{k-1}(x) = 0$

for all $x \in [0,1]^k$.

Proof We adopt the notation $c^{(k)} := (1,1,\dots,1,0) \in \mathbb{R}^k$ throughout.

We first show that for all $x \in [0,1]^k$,

$$\|x - c^{(k)}\|_1 \leq \frac{1}{4} \Rightarrow g^{(w_{k-2},k-1)} \circ p_{k-1}(x) = 0. \tag{1}$$

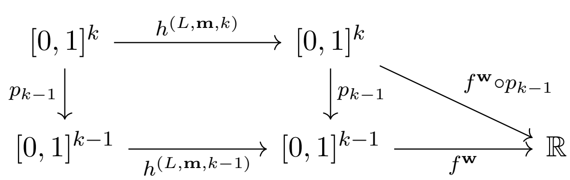
Let $x \in [0,1]^k$ with $g^{(w_{k-2},k-1)} \circ p_{k-1}(x) \neq 0$. Lemma 7 implies that $\|p_{k-1}(x) - c^{(k-1)}\|_1 < \frac{1}{4}$. Therefore we have $\frac{1}{4} > |\pi_{k-1} \circ p_{k-1}(x) - 0| = |\pi_{k-1}(x) - 0| = x_{k-1}$ which also means $|x_{k-1} - 1| > \frac{1}{4}$ and thus $\|x - c^{(k)}\|_1 > \frac{1}{4}$.
Note that by Lemma 7 it suffices to show that $f^{(w_1,\dots,w_{k-1})} \circ p_{k-1}(x) = 0$ for all x with $\|x - c^{(k)}\|_1 \leq \frac{1}{4}$. We prove this by induction over k. The base case has already been covered since $g^{(w_1,2)} = f^{w_1}$. Furthermore

$$\begin{aligned} f^{(w_1,\dots,w_{k-2})} \circ p_{k-1} &= (f^{(w_1,\dots,w_{k-3})} \circ p_{k-2} + g^{(w_{k-2},k-1)}) \circ p_{k-1} \\ &= f^{(w_1,\dots,w_{k-3})} \circ p_{k-2} \circ p_{k-1} + g^{(w_{k-2},k-1)} \circ p_{k-1} \\ &= f^{(w_1,\dots,w_{k-3})} \circ p_{k-2} + g^{(w_{k-2},k-1)} \circ p_{k-1} \end{aligned}$$

and thus the induction hypothesis and (1) imply that $f^{(\mathbf{w})} \circ p_{k-1}(x) = 0$ for x with $\|x - c^{(d)}\|_1 \leq \frac{1}{4}$.
■

B.5. Proof of Lemma 11

Lemma 39 *The following diagram commutes:*



Proof In order to show that the left half of the diagram commutes, we prove that

$$(\pi_j \circ h^{(L,\mathbf{m},k-1)} \circ p_{k-1})(x) = (\pi_j \circ p_{k-1} \circ h^{(L,\mathbf{m},k)})(x)$$

for every $j \in \{1,\dots,k-1\}$ and $x \in [0,1]^k$. For any $x \in [0,1]^k$, there exist indices $i_1,\dots,i_k$ such that $x = (x_1,\dots,x_k) \in W^{(L,\mathbf{m},k)}_{(i_1,\dots,i_k)}$. Moreover, if $x \in W^{(L,\mathbf{m},k)}_{(i_1,\dots,i_k)}$, we have $p_{k-1}(x) \in W^{(L,\mathbf{m},k-1)}_{(i_1,\dots,i_{k-1})}$ because

$$p_{k-1}\left(W^{(L,\mathbf{m},k)}_{(i_1,\dots,i_k)}\right) = p_{k-1}\left(\prod_{j=1}^{k}\left[\frac{(i_j-1)}{M}, \frac{i_j}{M}\right]\right) = \prod_{j=1}^{k-1}\left[\frac{(i_j-1)}{M}, \frac{i_j}{M}\right].$$

We use this observation combined with Lemma 5, assuming that i_j is odd:

$$\begin{aligned}(\pi_j \circ h^{(L,\mathbf{m},k-1)} \circ p_{k-1})(x) &= M \cdot (p_{k-1}(x))_j - (i_j - 1) \\ &= M \cdot x_j - (i_j - 1) \\ &= (\pi_j \circ h^{(L,\mathbf{m},k)})(x) \\ &= (\pi_j \circ p_{k-1} \circ h^{(L,\mathbf{m},k)})(x),\end{aligned}$$

as claimed. The case where i_j is even follows analogously. ■

Lemma 40 *For $2 \le k \le d$, the space $X_k \coloneqq (f^{(w_1,\dots,w_{k-1})} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0))$ satisfies*

$$X_k = (X_{k-1} \times [0,1]) \sqcup Y_{k,w_{k-1}}$$

with $X_1 \coloneqq \emptyset$.

Proof For $k = 2$ it holds that $f^{w_1} = g^{(w_1,2)}$ and therefore the claim holds trivially. Now let $k \ge 3$. Since $f^{(w_1,\dots,w_{k-1})} = f^{(w_1,\dots,w_{k-2})} \circ p_{k-1} + g^{(w_{k-1},k)}$ and the spaces

$$(f^{(w_1,\dots,w_{k-2})} \circ p_{k-1} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0])$$

and $(g^{(w_{k-1},k)} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0])$ are disjoint by Lemma 10, it follows that

$$\begin{aligned}&(f^{(w_1,\dots,w_{k-1})} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0]) \\ &= ((f^{(w_1,\dots,w_{k-2})} \circ p_{k-1} + g^{(w_{k-1},k)}) \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0]) \\ &= (f^{(w_1,\dots,w_{k-2})} \circ p_{k-1} \circ h^{(L,\mathbf{m},k)} + g^{(w_{k-1},k)} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0]) \\ &= (f^{(w_1,\dots,w_{k-2})} \circ p_{k-1} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0]) \sqcup (g^{(w_{k-1},k)} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0]) \\ &= (f^{(w_1,\dots,w_{k-2})} \circ h^{(L,\mathbf{m},k-1)} \circ p_{k-1})^{-1}((-\infty, 0]) \sqcup (g^{(w_{k-1},k)} \circ h^{(L,\mathbf{m},k)})^{-1}((-\infty, 0]) \\ &= X_{k-1} \times [0,1] \sqcup Y_{k,w},\end{aligned}$$

where the second last equality is due to Lemma 39. ■

B.6. Proof of Theorem 12

Lemma 41 *The space $Y_{d,w} \coloneqq (g^{(w,d)} \circ h^{(L,\mathbf{m},d)})^{-1}((-\infty, 0])$ satisfies*

(i) $H_0(Y_{d,w}) \cong \mathbb{Z}^{p+p'}$,

(ii) $H_{d-1}(Y_{d,w}) \cong \mathbb{Z}^p$,

(iii) $H_k(Y_{d,w}) = 0$ *for* $k \ge d$

with $p = \frac{M^{(d-1)}}{2^{d-1}} \cdot \left(\frac{M}{2} - 1\right) \cdot \left\lceil \frac{w}{2} \right\rceil$ *and* $p' = \frac{M^{(d-1)}}{2^{d-2}} \cdot \left\lceil \frac{w}{2} \right\rceil$

Proof Follows directly from Observation 28 and Proposition 9 and the disjoint union axiom (Proposition 27). ■

Theorem 12 *Given an architecture* $A = (d, n_1, \dots, n_L, 1)$ *with* $n_\ell \geq 2d$ *for all* $\ell \in [L]$ *and numbers* $w_1, \dots, w_{d-1} \in \mathbb{N}$ *such that* $\sum_{k=1}^{d-1}(w_k + 2) = n_L$*, there is a neural network* $F \in \mathcal{F}_A$ *with weights bounded from above by* $\max_{\ell=1,\dots L} 2\frac{n_\ell}{d}$ *such that*

(i) $\beta_0(F^{-1}((-\infty, 0))) = \sum_{k=2}^{d} \frac{M^{(k-1)}}{2^{k-1}} \cdot \left(\frac{M}{2} + 1\right) \cdot \lceil \frac{w_k}{2} \rceil$

(ii) $\beta_k(F^{-1}((-\infty, 0))) = \frac{M^{(k-1)}}{2^{k-1}} \cdot \left(\frac{M}{2} - 1\right) \cdot \lceil \frac{w_{k-1}}{2} \rceil$ *for* $0 < k < d$*,*

where $M = \prod_{\ell=1}^{L-1} 2 \cdot \lfloor \frac{n_\ell}{2d} \rfloor$*.*

Proof We consider the map $F := f^{(\mathbf{w})} \circ h^{(L,\mathbf{m},d)}$ that was previously constructed (Lemma 11) and let $X_d = F^{-1}((-\infty, 0))$. For $d = 2$, the statement is identical to Lemma 41. Indeed, we have

$$2 \cdot \frac{\left(\frac{M}{2} + 1\right)^3 - 1}{M} - \frac{M}{2} - 2 = \left(\frac{M}{2} + 1\right)^2 + \frac{M}{2} + 1 - \frac{M}{2} - 2$$
$$= \frac{M}{2}\left(\frac{M}{2} + 1\right).$$

Let $d \geq 3$. Using Proposition 9, we see that

$$H_k(X_d) \cong H_k(X_{d-1} \sqcup Y_{d,w_{d-1}}) \cong H_k(X_{d-1}) \oplus \prod_{i=1}^{p_d} H_k(S^{d-1}) \oplus \prod_{i=1}^{p'_d} H_k(D^d) \tag{2}$$

and therefore

$$\beta_k(X_d) = \beta_k(X_{d-1}) + \sum_{i=1}^{p_d} \left(\beta_k(S^{d-1})\right) + \sum_{i=1}^{p'_d} \beta_k(D^d) \tag{3}$$

where $p_d = \frac{M^{d-1}}{2^{d-1}} \cdot \left(\frac{M}{2} - 1\right) \cdot \lceil \frac{w_{d-1}}{2} \rceil$ and $p'_d = \frac{M^{d-1}}{2^{d-2}} \cdot \lceil \frac{w_{d-1}}{2} \rceil$. Fix some $k \in \mathbb{N}$. For different values of k, we obtain the claims:

(i) For $k = 0$, equation (3) implies

$$\beta_0(X_d) = \beta_0(X_{d-1}) + p_d + p'_d = \sum_{i=2}^{d} \frac{M^{(i-1)}}{2^{i-1}} \cdot \left(\frac{M}{2} + 1\right) \cdot \left\lceil \frac{w_{i-1}}{2} \right\rceil$$

(ii) For $k \leq d - 1$, we have $\beta_{d-1}(X_d) = 0$ and therefore

$$\beta_{d-1}(X_d) = p_d = \left(\frac{M}{2} - 1\right) \cdot \frac{M^{d-1}}{2^{d-1}} \cdot \left\lceil \frac{w_{d-1}}{2} \right\rceil.$$

For $0 < k < d - 1$, we have $\beta_k(X_d) = \beta_k(X_{d-1})$, i.e., the claim is satisfied by induction.

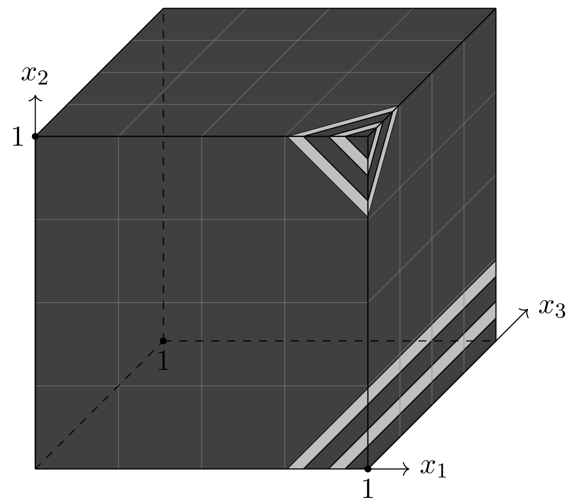


Figure 8: The last hidden layer of the modified F in the proof of Theorem 13; Adding a small constant b in the output layer in order to not have full dimensional "0-regions".

(iii) Finally for $k \geq d$, we observe that all summands of (3) vanish.

Lastly, the norm of the weights of the map $h^{(L,\mathbf{m},d)}$ are bounded from above by $\max_{\ell=1,\ldots L} 2\frac{n_\ell}{d}$ and the norm of the weights in the output layer $f^{(\mathbf{w})}$ are bounded from above by $1 \leq \max_{\ell=1,\ldots L} 2\frac{n_\ell}{d}$.

■

B.7. Closure

In this section we modify the construction slightly to obtain a lower bound for $\beta_k(F^{-1}((-\infty, 0])$ additionally to the lower bound for $\beta_k(F^{-1}((-\infty, 0))$

Theorem 13 *Given an architecture $A = (d, n_1, \ldots, n_L, 1)$ with $n_\ell \geq 2d$ for all $\ell \in [L]$ and numbers $w_1, \ldots, w_{d-1} \in \mathbb{N}$ such that $\sum_{k=1}^{d-1}(w_k + 2) = n_L$, there is a neural network $F \in \mathcal{F}_A$ with weights bounded from above by $\max_{\ell=1,\ldots L} 2\frac{n_\ell}{d}$ such that*

(i) $\beta_0(F) = \sum_{k=2}^{d} \frac{M^{(k-1)}}{2^{k-1}} \cdot \left(\frac{M}{2} + 1\right) \cdot \left\lceil \frac{w_k}{2} \right\rceil$

(ii) $\beta_k(F) = \frac{M^{(k-1)}}{2^{k-1}} \cdot \left(\frac{M}{2} - 1\right) \cdot \left\lceil \frac{w_{k-1}}{2} \right\rceil$ *for* $0 < k < d$,

where $M = \prod_{\ell=1}^{L-1} 2 \cdot \lfloor \frac{n_\ell}{2d} \rfloor$.

Proof Let F' be the neural network constructed in Theorem 12. We want to modify this neural network to obtain a neural network F such that it holds that

- $F^{-1}((-\infty, 0))$ and $F'^{-1}((-\infty, 0))$ are homeomorphic,
- $F^{-1}((-\infty, 0))$ is homotopy equivalent to its closure $\overline{F^{-1}((-\infty, 0))}$ and

- $\overline{F^{-1}((-\infty, 0))} = F^{-1}((-\infty, 0])$.

In order to guarantee the third property, we need to ensure that $F^{-1}(\{0\}) = \partial F^{-1}((-\infty, 0))$. Therefore, we transform the full-dimensional $0-$regions into slightly positive, but still constant regions by adding a small constant b, i.e., by simply setting $F(x) = F'(x) + b$, where $b = \min_{k=2,...,d}\{\frac{1}{8w_k \cdot M}\}$.

Next, we argue that $F^{-1}((-\infty, 0))$ and $F'^{-1}((-\infty, 0))$ are homeomorphic, since adding b also just makes the annuli in $F'^{-1}((-\infty, 0))$ thinner. Let $k \in [d]$ and A be an $k-$annuli in $F'^{-1}((-\infty, 0))$ and $A_k = p_k(A)$ be its projection onto the first k coordinates. It follows that there is a cutting point $c \in \mathbb{R}^k$ such that $A_k = B^k_{q/(4w_k \cdot M)}(c) \setminus \overline{B^k_{(q-1)/(4w_k \cdot M)}(c)}$ for a suitable $q = 1, ..., w_k$. Since $b \leq \{\frac{1}{8w_k \cdot M}\}$ it follows that

$$A'_{k,c,q} := B^k_{(q-2b)/(4w_k \cdot M)}(c) \setminus \overline{B^k_{(q-1+2b)/(4w_k \cdot M)}(c)}$$

is also a k-annuli and revisiting the proof of Lemma 35 reveals that $F(A'_{k,c,q} \times \mathbb{R}^{d-k}) = (-\frac{1}{4w_k}, -b)$ and hence $F(A'_{k,c,q} \times \mathbb{R}^{d-k}) = (-\frac{1}{4w_k} + b, 0)$ since $\frac{1}{8w_k} \geq b$. We conclude that for every k-annuli in $F'^{-1}((-\infty, 0))$ there is a "shrunk" k-annuli in $F^{-1}((-\infty, 0))$ and the "shrinking" for every annuli is an homeomorphism. The case for the disks at the boundary of the unit cube follows analogously. Since $F^{-1}((-\infty, 0))$ is the disjoint union of these disks and annuli and it clearly holds that $F(x) > 0$ for all $x \in F'^{-1}((0, \infty))$, it follows that $F^{-1}((-\infty, 0))$ and $F'^{-1}((-\infty, 0))$ are homeomorphic. Since disks and annuli are homotopy equivalent to their closures, it follows that $F^{-1}((-\infty, 0))$ is homotopy equivalent to $\overline{F^{-1}((-\infty, 0))}$, proving the claim.

■

Appendix C. Stability

In this section we aim to prove Proposition 17. Before we prove the stability of our construction, we prove stability for a wider range of neural networks. Throughout this section we will use definitions and statements from Section A.1. First, we define the realization map, that maps, for a given architecture, a vector of weights to its corresponding neural network.

Definition 42 (The realization map) *Let* $(n_0, \ldots n_{L+1})$ *be an architecture. Its* corresponding parameter space *is given by* $\mathbb{R}^D \cong \bigoplus_{\ell=1}^{L+1} \mathbb{R}^{(n_{\ell-1}+1) \times n_\ell}$*, where the vector space isomorphism is given by* $p \mapsto (A^{(\ell)}(p), b^{(\ell)}(p))_{\ell=1,...,L+1}$ *for* $A^{(\ell)}(p) \in \mathbb{R}^{n_{\ell-1} \times n_\ell}, b^{(\ell)}(p) \in \mathbb{R}^{n_\ell}$ *and* $\ell = 1, \ldots, L+1$*. For a polyhedron* $K \subseteq \mathbb{R}^d$*, we define* $\Phi \colon \mathbb{R}^D \to C(K)$ *to be the* realization map *that assigns to a vector of weights the function the corresponding neural network computes, i.e.,*

$$\Phi(p) := T_{L+1}(p) \circ \sigma_{n_L} \circ T_L(p) \circ \cdots \circ \sigma_{n_1} \circ T_1(p)$$

where $T_\ell(p) \colon \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$*,* $x \mapsto A^{(\ell)}(p)x + b^{(\ell)}(p)$*.*

Furthermore let

$$\Phi^{(\ell)}(p) := T_\ell(p) \circ \sigma_{n_\ell} \circ \cdots \circ \sigma_{n_1} \circ T_0(p)$$

and

$$\Phi^{(i,\ell)}(p) := \pi_i \circ T_\ell(p) \circ \cdots \circ \sigma_{n_1} \circ T_0(p),$$

which we call the map computed at neuron (i, ℓ). We denote the points of non-linearity introduced by the i-th neuron in the ℓ-th layer by

$$\tilde{H}_{i,\ell}(p) \coloneqq H\left(A_i^{(\ell)}(p), b_i^{(\ell)}(p)\right)$$

Now we define a sequence of polyhedral complexes associated to a neural network. Every polyhedral complex $\mathcal{P}^{(i,\ell)}$ in this sequence corresponds to a refinement of the input space K such that the CPWL map computed at a neuron (i, ℓ) in the neural network as well as the CPWL map computed at any neuron (j, k) with (j, k) lexicographically smaller than (i, ℓ) is affine linear on all polyhedra in $\mathcal{P}^{(i,\ell)}$. Moreover, it is a refinement of the previous polyhedral complexes $\mathcal{P}^{(i-1,\ell)}$ in this sequence that results by intersecting $\mathcal{P}^{(i-1,\ell)}$ with the pullbacks of $\tilde{H}^s_{i,\ell}(p), s \in \{-1, 0, 1\}$ by $\Phi^{(\ell)}(p)$.

Definition 43 (Canonical polyhedral complex (Grigsby et al. (2022a))) *Let $p \in \mathbb{R}^D$ be a vector of weights. Recall that $\Phi(p)$ is the corresponding neural network.*

We recursively define polyhedral complexes $\mathcal{P}^{(i,\ell)}(p, K)$ by

$$\mathcal{P}^{(0,0)}(p, K) \coloneqq \{F \mid F \textit{ is a face of } K\}$$

and

$$\mathcal{P}^{(i,\ell)}(p, K) \coloneqq \{R \cap (\Phi^{(\ell-1)}(p))^{-1}(\tilde{H}^s_{i,\ell}(p)) \mid R \in \mathcal{P}^{(i-1,\ell)}(p, K), s \in \{-1, 0, 1\}\}$$

for $i = 2, \ldots n_\ell, \ell = 1, \ldots L$ and

$$\mathcal{P}^{(1,\ell)}(p, K) \coloneqq \{R \cap (\Phi^{(\ell-1)}(p))^{-1}(\tilde{H}^s_{i,\ell}(p, K)) \mid R \in \mathcal{P}^{(n_{\ell-1},\ell-1)}(p, K), s \in \{-1, 0, 1\}\}$$

for $\ell = 1, \ldots L$.

Note that for all $j \leq i$, it holds that $\Phi^{(j,\ell)}(p)$ is affine linear on R for each $R \in \mathcal{P}^{(i,\ell)}(p)$ and we denote this affine linear map by $\Phi^{(j,\ell)}_{|R}(p)$. For $\ell \in [L], i \in [n_\ell]$ and $R \in \mathcal{P}^{(i,\ell)}_d(p, K)$ we denote the points of non-linearity in the region R with respect to the first $\ell - 1$ layer map introduced by the i-th neuron in the ℓ-th layer by

$$H_{i,\ell,R}(p) \coloneqq (\Phi^{(\ell-1)}_{|R}(p))^{-1}(\tilde{H}_{i,\ell}(p)) = H\left(A_i^{(\ell)}(p)\left(\Phi^{(\ell-1)}_{|R}(p)(x)\right), b_i^{(\ell)}(p)\right).$$

For the sake of simplification we set $\mathcal{P}^{(0,\ell)}(p, K) \coloneqq \mathcal{P}^{(n_{\ell-1},\ell-1)}(p, K)$. Furthermore, since for $R \in \mathcal{P}^{(i-1,\ell)}_d(p, K), F \in \mathcal{P}^{(i-1,\ell)}(p, K), F \preceq R$ it holds that

$$F \cap H^s_{i,\ell,R}(p) = F \cap (\Phi^{(\ell-1)}(p))^{-1}(\tilde{H}^s_{i,\ell}(p))$$

due to the continuity of the function $\Phi(u)$, we have that

$$\mathcal{P}^{(i,\ell)}(p) = \{F \cap H^s_{i,\ell,R}(p) \mid R \in \mathcal{P}^{(i-1,\ell)}_d(p), F \in \mathcal{P}^{(i-1,\ell)}(p), F \preceq R, s \in \{-1, 0, 1\}\}$$

We call $\mathcal{P}(p, K) \coloneqq \mathcal{P}^{(n_L,L)}(p, K)$ the *canonical polyhedral complex* of $\Phi(u)$ (with respect to K). We omit p respectively K whenever p respectively K is clear from the context.

In the following, we find a sufficient condition for a neural network, which is, that the points of non-linearity introduced at any neuron (i, ℓ) do not intersect the vertices of the polyhedral complex $\mathcal{P}^{(i,\ell)}(p, K)$, to compute a "similar looking map on a polytope K" $\Phi(u)$ for "close enough" parameters u. Note that the boundedness of the polyhedron is required because any perturbation of two parallel unbounded $(d - 1)-$faces results in new intersection patterns in the corresponding polyhedral complex.

Definition 44 *Let $K \subseteq \mathbb{R}^d$ be a polytope and $\Phi(p)\colon K \to \mathbb{R}$ be a ReLU neural network of architecture $(n_0, \dots n_{L+1})$. Then we call $\Phi(p)$* combinatorially stable *(with respect to K) if for every $\ell \in [L+1], i \in [n_\ell]$ and all $R \in \mathcal{P}_d^{(i-1,\ell)}(p, K)$ it holds*

(i) $\dim(H_{i,\ell,R}(p)) = d - 1$ *and*

(ii) $H_{i,\ell,R}(p) \cap R_0 = \emptyset$.

We will now prove that this condition is indeed sufficient.

Proposition 45 *Let K be a polytope and $\Phi(p)\colon K \to \mathbb{R}$ be a stable ReLU neural network of architecture $(n_0, \dots n_{L+1})$. Then for every $\varepsilon > 0$, there is a an open set $U \subseteq \mathbb{R}^D$ such that for every $u \in U$ there is an ε-isomorphism $\varphi_u\colon \mathcal{P}(p, K) \to \mathcal{P}(u, K)$.*

Proof We will prove the following stronger statement by induction on the indexing of the neurons.

Claim *For every $\ell \in [L+1]$, $i \in [n_\ell]$ and every $\varepsilon > 0$, there is a $\delta > 0$ such that for all $u \in B_\delta^{\|\cdot\|_\infty}(p)$ there is an ε-isomorphism $\varphi_u^{(i,\ell)}\colon \mathcal{P}^{(i,\ell)}(p) \to \mathcal{P}^{(i,\ell)}(u)$.*

The induction base is trivially satisfied.
So we assume that the statement holds for $(i-1, \ell)$. For simpler notation we denote $\varphi_u^{(i-1,\ell)}$ by φ_u and $H_{i,\ell,R}(p)$ by $H_R(p)$. Let $\varepsilon > 0$ and $F \in \mathcal{P}^{(i-1,\ell)}(p)$. There is an $R \in \mathcal{P}_d^{(i-1,\ell)}(p)$ such that $F \preceq R$. In the following we wish to find a $\delta_F > 0$ such that there are ε-isomorphisms

$$\varphi_{(u,R,s)}^{(i,\ell)}\colon F \cap H_R^s(p) \to \varphi_u(F) \cap H_{\varphi_u(R)}^s(u)$$

for $s \in \{-1, 0, 1\}$ and all $u \in B_{\delta_F}^{\|\cdot\|_\infty}(p)$.

Since $\Phi(p)$ is stable, we obtain by Lemma 24 a $\delta_2 > 0$ such that for all δ_2-isomorphisms $\varphi\colon F \to Q$ there are $\frac{\varepsilon}{3}$-isomorphisms $\gamma^s\colon F \cap H_R^s(p) \to \varphi(F) \cap H_R^s(p)$. By the induction hypothesis we obtain $\delta_1 > 0$ such that for all $u \in B_{\delta_1}^{\|\cdot\|_\infty}(p)$ there is an δ_2-isomorphism

$$\varphi_u\colon \mathcal{P}(p)^{(i-1,\ell)} \to \mathcal{P}(u)^{(i-1,\ell)}$$

and hence we obtain $\frac{\varepsilon}{3}$-isomormorphisms

$$\gamma^{(s,F)}\colon F \cap H_R^s(p) \to \varphi_u(F) \cap H_R^s(p).$$

Let $H_{\varphi_u(R)}(u, p) \coloneqq H_R(u_{1,1}, \dots u_{i-1,\ell}, p_{i,\ell}, \dots p_{n_{L+1},L+1})$ with $u_{j,k}, p_{j,k} \in \mathbb{R}^{n_k}$ being the parameters associated to the j-th neuron in the k-th layer. Again, for simpler notation, let the affine maps $\Phi_{|R}^{(\ell-1)}(p)$ be given by $x \mapsto Mx + c$ and $\Phi_{|\varphi_u(R)}^{(\ell-1)}(u)$ by $x \mapsto Nx + d$ and the non-linearity points introduced by the i-th neuron in the ℓ-th layer by $\tilde{H}_{i,\ell}(p) = H(a, b)$. Then we have that

$$H_R(p) = H(a^T M, a^T c + b)$$

and

$$H_{\varphi_u(R)}(u, p) = H(a^T N, a^T d + b).$$

By Lemma 24 we know that $(\varphi_u(F))_0 \cap H_R(p) = \emptyset$ and hence by Lemma 23 there is a $\delta_3 > 0$ such that there are $\frac{\varepsilon}{3}$-isomorphisms $\psi^s\colon \varphi_u(F) \cap H_R^s(p) \to \varphi_u(F) \cap H^s(y, z)$ for all

$(y,z) \in B^{d+1}_{\delta_3}((a^T M, a^T c + b))$. Let $C \coloneqq n_{\ell-1} \cdot \max_{i=1,\dots,n_{\ell-1}} \{a_i\}$ and $u \in \mathbb{R}^D$ with $\|u-p\|_\infty < \frac{\delta_3}{C}$. Then we have that

$$\begin{aligned}
\|(a^T M, a^T c + b) - (a^T N, a^T d + b)\|_\infty &= \max_{i=1,\dots,d} \left\{ \sum_{j=1}^{n_{\ell-1}} a_j (m_{ij} - n_{ij}), \sum_{j=1}^{n_{\ell-1}} a_j (c_j - d_j) \right\} \\
&< \max_{i=1,\dots,d} \left\{ \sum_{j=1}^{n_{\ell-1}} a_j \frac{\delta_3}{C}, \sum_{j=1}^{n_{\ell-1}} a_j \frac{\delta_3}{C} \right\} \\
&< \delta_3
\end{aligned}$$

and hence there are $\frac{\varepsilon}{3}$-isomorphisms

$$\psi^{(s,F)} \colon \varphi_u(F) \cap H^s_R(p) \to \varphi_u(F) \cap H^s_{\varphi_u(R)}(u,p)$$

By Lemma 23 we know that $(\varphi_u(F))_0 \cap H_{\varphi(R)}(u,p) = \emptyset$ and hence by the same lemma there is a $\delta_4 > 0$ such that there are $\frac{\varepsilon}{3}$-isomorphisms $\alpha^s \colon \varphi_u(F) \cap H^s_{\varphi(R)}(u,p) \to \varphi_u(F) \cap H^s(y,z)$ for all $(y,z) \in B^{d+1}_{\delta_4}((a^T N, a^T d + b))$. Let $a' \in \mathbb{R}^{n_{\ell-1}}, b' \in \mathbb{R}$ such that $\tilde{H}_{i,\ell}(u) = H(a'^T, b')$. Then we have that

$$H_{\varphi_u(R)}(u) = H(a'^T N, a'^T d + b').$$

Let $E \coloneqq n_{\ell-1} \cdot \max_{i,j=1,\dots,n_{\ell-1}} \{n_{ij}, d_j\}$ and $u \in \mathbb{R}^D$ with $\|u-p\|_\infty < \frac{\delta_5}{E}$. Then we have that

$$\begin{aligned}
&\|(a'^T N, a'^T d + b') - (a^T N, a^T d + b)\|_\infty \\
&\qquad = \max_{i=1,\dots,d} \left\{ \sum_{j=1}^{n_{\ell-1}} n_{ij} (a'_j - a_j), \Big(\sum_{j=1}^{n_{\ell-1}} d_j (a'_j - a_j) \Big) + (b'_j - b_j) \right\} \\
&\qquad < \max_{i=1,\dots,d} \left\{ \sum_{j=1}^{n_{\ell-1}} n_{ij} \frac{\delta_5}{E}, \sum_{j=1}^{n_{\ell-1}} n_{ij} \frac{\delta_5}{E} \right\} \\
&\qquad < \delta_4
\end{aligned}$$

and hence there are $\frac{\varepsilon}{3}$-isomorphisms

$$\alpha^{(s,F)} \colon \varphi_u(F) \cap H^s_{\varphi(R)}(u,p) \to \varphi_u(F) \cap H^s_{\varphi_u(R)}(u).$$

Let $\delta_F \coloneqq \min\{\delta_2, \frac{\delta_4}{C}, \frac{\delta_5}{E}\}$, then for all $u \in B^{D,\|\cdot\|_\infty}_{\delta_F}(p)$ there is an ε-isomorphism

$$\varphi^{(i,\ell)}_{(u,F,s)} \colon F \cap H^s_R(p) \to \varphi_u(F) \cap H^s_{\varphi_u(R)}(u)$$

given by

$$\varphi^{(i,\ell)}_{(u,F,s)} = \alpha^{(s,F)} \circ \psi^{(s,F)} \circ \gamma^{(s,F)}.$$

Lastly, let $\delta = \min\{\delta_F \mid F \in \mathcal{P}^{(i-1,\ell)}(p)\}$. Since every element of $\mathcal{P}^{(i,\ell)}(p)$ is of the form $F \cap H^s_R(p)$, it now remains to show that the map $\varphi^{(i,\ell)}_u \colon \mathcal{P}^{(i,\ell)}(p) \to \mathcal{P}^{(i,\ell)}(u)$ defined by

$$\varphi^{(i,\ell)}_u(F \cap H^s_R(p)) \coloneqq \varphi_u(F) \cap H^s_{\varphi_u(R)}(u)$$

is an ε-isomorphism for all $u \in B_\delta^{\|\cdot\|_\infty}(p)$. Since φ_u and $\varphi_{(u,F,s)}^{(i,\ell)}$ are bijections, the same holds for $\varphi_u^{(i,\ell)}$. Furthermore let $G \preceq F \cap H_R^s(p)$, then there is a $G' \preceq F$ and a $s' \in \{0, s\}$ such that $G = G' \cap H_R^{s'}(p)$. Since φ_u is an isomorphism by the induction hypothesis, it follows that

$$\varphi_u^{(i,\ell)}(G' \cap H_R^{s'}(p)) = \varphi_u(G') \cap H_{\varphi_u(R)}^{s'}(u) \preceq \varphi_u(F) \cap H_{\varphi_u(R)}^{s}(u)$$

and hence $\varphi_u^{(i,\ell)}$ is an ε-isomorphism as claimed. ■

Taking the hyperplanes where the output layer equals zero into account, we define a topological stable parameter.

Definition 46 *Let K be a polytope and $\Phi(p)\colon K \to \mathbb{R}$ be a ReLU neural network of architecture $(n_0, \ldots, n_L, 1)$. Then we call $\Phi(p)$* topologically stable *if it is combinatorially stable (with respect to K), and for all $R \in \mathcal{P}_d^{(n_L,L)}(p, K)$ it holds that*

(i) $\dim(H_{1,L+1,R}(p)) = d - 1$ *and*

(ii) $H_{1,L+1,R}(p) \cap R_0 = \emptyset$.

We now prove that topologically stable is the right definition for our purposes, that is, finding an open set $U \subseteq \mathbb{R}^d$ with $p \in U$ such that the sublevel set of $\Phi(u)$ is homeomorphic to the sublevel set $\Phi(p)$ for all $u \in U$.

Proposition 16 *Let K be a polytope and $\Phi(p)$ a topologically stable ReLU neural network with respect to K, then there is a $\delta > 0$ such that for all $u \in B_\delta(p)$ it holds that $K \cap \Phi(p)^{-1}((-\infty, 0])$ is homeomorphic to $K \cap \Phi(u)^{-1}((-\infty, 0])$.*

Proof Let

$$\begin{aligned} \mathcal{P}^-(p) &:= \{F \cap H_{1,L+1,R}^s(p) \mid R \in \mathcal{P}_d(p), F \in \mathcal{P}(p), F \preceq R, s \in \{-1, 0\}\} \\ &= \{P \cap F^{-1}((-\infty, 0]) \mid P \in \mathcal{P}\} \end{aligned}$$

be the polyhedral complex consisting of all maximal subpolyhedron of $\mathcal{P}(p)$ where $\Phi(p)$ takes on non-negative values. Analogously to the proof of Proposition 45 we obtain a $\delta > 0$ such that $\mathcal{P}^-(p, K)$ and $\mathcal{P}^-(u, K)$ are isomorphic as polyhedral complexes and hence in particular there is a homeomorphism $\varphi\colon |\mathcal{P}^-(p, K)| \to |\mathcal{P}^-(u, K)|$ for all $u \in B_\delta(p)$, where $|\mathcal{P}^-(p, K)|$ denotes the support of $\mathcal{P}^-(p, K)$. This concludes the proof since $|\mathcal{P}^-(p, K)| = K \cap \Phi(p)^{-1}((-\infty, 0])$ and $|\mathcal{P}^-(u, K)| = K \cap \Phi(u)^{-1}((-\infty, 0])$. ■

Having this at hand, we can finally show the stability of the constructed neural network in Theorem 13 for the lower bound of the topological expressive power.

Proposition 17 *There is an open set $U \subseteq \mathbb{R}^D$ in the parameter space of the architecture $A = (d, n_1, \ldots, n_L, 1)$ such that $\Phi(u)$ restricted to the unit cube has at least the same topological expressivity as F in Theorem 13 for all $u \in U$.*

Proof Let $p \in \mathbb{R}^D$ such that $\Phi(p) = F$ from Theorem 13. Then, since $\Phi(p)$ is topologically stable with respect to any cube it follows by Proposition 16 that there is an open set in $\mathbb{R}^D$ containing u such that $\Phi(u)^{-1}((-\infty, 0)) \cap K$ is homeomorphic to $\Phi(p)^{-1}((-\infty, 0)) \cap K$ for all $u \in U$, where K is the unit cube. ■

Using the results from above, we can even show that if p is topologically stable, then also $\Phi(p)^{-1}((-\infty, 0))$ is homeomorphic to $\Phi(u)^{-1}((-\infty, 0))$ for all u in an open set $U \subseteq \mathbb{R}^D$.

Proposition 47 *For every topologically stable network $\Phi(p)$, there is a $\delta > 0$ such that for all $u \in B_\delta(p)$, it holds that $K \cap \Phi(p)^{-1}((-\infty, 0))$ is homeomorphic to $K \cap \Phi(u)^{-1}((-\infty, 0))$.*

Proof We adapt the notation from the proof of Proposition 16 and we know that $|\mathcal{P}^-(p, K)|$ and $|\mathcal{P}^-(u, K)|$ are homeomorphic.

We wish to show that $|\mathcal{P}^-(p, K)|^\circ = K^\circ \cap \Phi(p)^{-1}((-\infty, 0])$. Due to the continuity of $\Phi(u)$ it holds that $K^\circ \cap \Phi(p)^{-1}((-\infty, 0)) \subseteq |\mathcal{P}^-(p)|^\circ$. Let x be chosen such that $\Phi(p)(x) = 0$, i.e., $x \in |\mathcal{P}^-(p)| \setminus \left(K^\circ \cap \Phi(p)^{-1}((-\infty, 0))\right)$. Since $\mathcal{P}(u)$ is a pure polyhedral complex, there is a full-dimensional polyhedron $R \in \mathcal{P}_d(u)$ such that $x \in R$. It follows that $x \in H_{1,L+1,R}(p) \cap R$ with $\dim(H_{1,L+1,R}(p)) = d - 1$.

Assume for sake of contradiction that $\dim(H^1_{1,L+1,R}(p) \cap R) < d$, then there would be a face $F \preceq R$ such that $F \subseteq H_{1,L+1,R}(p)$, which is a contradiction to $H_{1,L+1,R}(p) \cap R_0 = \emptyset$. Therefore, the space $H^1_{1,L+1,R}(p) \cap R$ is full-dimensional. As a result, $\Phi(u)$ takes on exclusively positive values on $(H^1_{1,L+1,R}(p) \cap R)^\circ \neq \emptyset$ and hence for every open subset $U \subseteq \mathbb{R}^d$ with $x \in U$, it holds that $U \cap \Phi(p)^{-1}((0, \infty)) \neq \emptyset$. Thus, $x \notin |\mathcal{P}^-(p)|^\circ$ and hence

$$|\mathcal{P}^-(p)|^\circ = K^\circ \cap \Phi(p)^{-1}((-\infty, 0)).$$

Since $\mathcal{P}^-(p)$ and $\mathcal{P}^-(u)$ are isomorphic and $\Phi(u)$ is also topologically stable due to Lemma 24 and Lemma 23, the same arguments can be applied in order to show

$$|\mathcal{P}^-(u)|^\circ = K^\circ \cap \Phi(u)^{-1}((-\infty, 0)).$$

Observing that the restriction of φ to the interiors $\varphi_{|\mathcal{P}^-(p)|^\circ} : |\mathcal{P}^-(p)|^\circ \to |\mathcal{P}^-(u)|^\circ$ is a homeomorphism as well, we conclude that $K^\circ \cap \Phi(u)^{-1}((-\infty, 0))$ and $K^\circ \cap \Phi(p)^{-1}((-\infty, 0))$ are homeomorphic.

Let F now be any face of K with $\dim(F) \neq 0$, then by the same arguments it follows that $F^\circ \cap \Phi(u)^{-1}((-\infty, 0))$ and $F^\circ \cap \Phi(p)^{-1}((-\infty, 0))$ are homeomorphic. Furthermore, due to the fact that $\Phi(p)$ is topologically stable and the choice of u, if $\dim(F) = 0$, the fact that

$$F \subseteq K \cap \Phi(p)^{-1}((-\infty, 0))$$

implies that $F \subseteq K \cap \Phi(u)^{-1}((-\infty,0))$ and hence

$$\begin{aligned}
\partial K \cap \Phi(p)^{-1}((-\infty,0)) &= \left(\bigsqcup_{\substack{F \preceq K, F \neq K \\ \dim(F) \neq 0}} F^\circ \sqcup \bigsqcup_{F \in K_0} F \right) \cap \Phi(p)^{-1}((-\infty,0)) \\
&\cong \left(\bigsqcup_{\substack{F \preceq K, F \neq K \\ \dim(F) \neq 0}} F^\circ \sqcup \bigsqcup_{F \in K_0} F \right) \cap \Phi(u)^{-1}((-\infty,0)) \\
&= \partial K \cap \Phi(u)^{-1}((-\infty,0))
\end{aligned}$$

Altogether, we conclude that $K \cap \Phi(p)^{-1}((-\infty,0))$ is homeomorphic to $K \cap \Phi(u)^{-1}((-\infty,0))$. ■

Appendix D. Upper Bound

In this section we will provide a formal proof for the upper bounds on the Betti numbers of $F^{-1}((-\infty,0])$. For the sake of simplicity, we compute $\beta_k(F^{-1}((-\infty,0])$ using cellular homology. Ideally, we would like to equip $F^{-1}((-\infty,0])$ with a canonical CW-complex structure, i.e., the k-cells of the CW-complexes precisely correspond to the k-faces of the respective polyhedral complex, and attachment maps are given by face incidences (c.f. Appendix A.2). However, $F^{-1}((-\infty,0])$ may contain unbounded polyhedra. In particular, an unbounded polyhedron cannot correspond to a CW-cell. To sidestep this issue, we construct a bounded polyhedral complex $\mathcal{Q}$ that is homotopy equivalent to $F^{-1}((-\infty,0])$.

The lineality space $L(P)$ of a polyhedron P is defined as the vector space V such that $p+v \in P$ for all $p \in P$ and $v \in V$. If $\mathcal{P}$ is a complete d-dimensional polyhedral complex(i.e., $|\mathcal{P}| = \mathbb{R}^d$), then all polyhedra in $\mathcal{P}$ have the same lineality space and hence the lineality space of the polyhedral complex is well-defined in this case.

Lemma 48 *Let $\mathcal{P}$ be a d-dimensional polyhedral complex and let $\mathcal{P}' \subseteq \mathcal{P}$ be a subcomplex with $\#\mathcal{P}'_k \leq \binom{r}{d-k+1}$. Then there is a polyhedral complex $\mathcal{Q}$ such that*

1. *all polyhedra in $Q \in \mathcal{Q}$ contain a vertex,*
2. *$|\mathcal{Q}|$ is a deformation retract of $|\mathcal{P}'|$ and*
3. *the number of k-faces $\#\mathcal{Q}_k$ is bounded by $\binom{r}{d-k-\ell+1}$, where ℓ is the dimension of the lineality space of $\mathcal{P}$.*

Proof *Let V be the lineality space of $\mathcal{P}$, $W \subseteq \mathbb{R}^d$ the subspace orthogonal to V and $\pi\colon \mathbb{R}^d \to W$ the orthogonal projection. Then it holds that $\pi(P)$ is a face of $\pi(P')$ iff $P \preceq P'$ for all $P, P' \in \mathcal{P}'$ and $\mathcal{Q} = \{\pi(P) \mid P \in \mathcal{P}'\}$ is a polyhedral complex. Furthermore, it holds that $\dim(P) = \dim(\pi(P)) + \ell$, where ℓ is the dimension of V and therefore*

$$\#\mathcal{Q}_k = \#\mathcal{P}'_{k+l} \leq \#\mathcal{P}_{k+l} \leq \binom{r}{d-k-\ell+1}.$$

Since W is the lineality space of $\mathcal{P}$ and hence also of $\mathcal{P}'$, it follows that the map

$$R\colon |\mathcal{P}'| \times [0,1] \to |\mathcal{Q}|$$

given by $R(w+v,t) = w + (1-t)v$ with $v \in V, w \in W, w+v \in |\mathcal{P}'|$ is continuous and therefore a deformation retraction, proving the claim. ■

Lemma 49 *Let $P \subseteq \mathbb{R}^d$ be a d-dimensional unbounded pointed polyhedron for some $d > 0$. Then we have $P \cong \mathbb{R}^{d-1} \times [0,\infty)$.*

Proof Without loss of generality, we may assume that $0 \in P^\circ$ (else we apply translation). There exists a radius r such that all vertices of the polyhedron P are in the interior of the open r-ball $B_r(0) := \{x : \|x\|_2 < r\}$. We show that the space $P \cap B_r(0)$ is homeomorphic to $\mathbb{R}^{d-1} \times [0,\infty)$. Finally, we observe that $P \cap B_r(0) \cong P$ by scaling points in the polyhedra along the extreme rays.

Now consider $P \cap B_r(0) \subset D_r^d$ as a subset of the closed d-dimensional disk of radius r, i.e., $D_r^d := \{x : \|x\|_2 \leq r\}$. Because P and $B_r(0)$ are both convex, so is their intersection, therefore, radial projection from the origin $0 \in P$, which fixes the origin and maps the boundary of P to the boundary of D_r^d is the desired homeomorphism. ■

Lemma 50 *Let $\mathcal{P}$ be a finite polyhedral complex such that each $P \in \mathcal{P}$ contains a vertex. Then $\mathcal{P}$ is homotopy equivalent to a bounded polyhedral complex $\mathcal{Q}$ such that the number of k-faces $\#\mathcal{Q}_k$ is at most the number of k-faces $\#\mathcal{P}_k$.*

Proof We prove the statement by induction on the number $\ell_{\mathcal{P}}$ of unbounded faces of $\mathcal{P}$. If $\ell_{\mathcal{P}} = 0$, we can pick $\mathcal{Q} = \mathcal{P}$ because $\mathcal{P}$ is bounded.

In the following, we construct a polyhedral complex $\mathcal{Q}' \simeq \mathcal{P}$ with $\ell_{\mathcal{Q}'} = \ell_{\mathcal{P}} - 1$ and $\#\mathcal{Q}'_k \leq \#\mathcal{P}_k$. By induction hypothesis, we then obtain $\mathcal{P} \simeq \mathcal{Q}' \simeq \mathcal{Q}$ and $\#\mathcal{Q}_k \leq \#\mathcal{Q}'_k \leq \#\mathcal{P}_k$ for each $k \in \mathbb{N}$, proving the statement.

Let $P \in \mathcal{P}$ be an unbounded n-dimensional polyhedron that is maximal with respect to inclusion. By Lemma 49, P is homeomorphic to $\mathbb{R}^{n-1} \times [0,\infty)$. It is easy to observe that any homeomorphism (in particular the map $\phi\colon P \to \mathbb{R}^{n-1} \times [0,\infty)$ described in the proof of Lemma 49) maps the (topological) boundary ∂P of P precisely to the boundary $\mathbb{R}^{n-1} \times \{0\} \subset \mathbb{R}^{n-1} \times [0,\infty)$ of the codomain. Hence we have the following commutative diagram:

$$\begin{array}{ccc} P & \xleftarrow{\quad\iota\quad} & \partial P \\ {\scriptstyle\phi}\big\downarrow{\scriptstyle\cong} & & {\scriptstyle\phi|_{\partial P}}\big\downarrow{\scriptstyle\cong} \\ \mathbb{R}^{n-1} \times [0,\infty) & \xrightarrow{\ \mathrm{pr}\ } & \mathbb{R}^{n-1} \times \{0\} \end{array}$$

where $\iota\colon \partial P \hookrightarrow P$ is the canonical inclusion and pr is the canonical projection onto the first $d-1$ components.

Since P is maximal with respect to inclusion, $\mathcal{Q}' := \mathcal{P} \setminus \{P\}$ is a polyhedral complex with support $|\mathcal{P}| \setminus P^{\mathrm{o}}$. Moreover, we have $\ell_{\mathcal{Q}'} = \ell_{\mathcal{P}} - 1$ and $\#\mathcal{Q}'_k \leq \#\mathcal{P}_k$ for all $k \in \mathbb{N}$.

Using $\Psi := (\phi|_{\partial P})^{-1}\mathrm{pr} \circ \phi$, we define the following map $\Sigma\colon |\mathcal{P}| \to |\mathcal{P}| \setminus P^{\mathrm{o}}$:

$$\Sigma(x) = \begin{cases} \Psi(x), & x \in P \\ x, & \text{else} \end{cases}$$

Note that this map is well-defined and continuous because one can easily observe that $\Psi(x) = x$ for all $x \in \partial P$. Moreover, it is a retraction, in particular a homotopy equivalence. ■

Lemma 51 *Let $\mathcal{P}$ be a subcomplex of a d-dimensional complete polyhedral complex. Then there is a CW-complex X, such that*

1. *$|\mathcal{P}|$ is homotopy equivalent to X and*
2. *and the number of k-cells of X is bounded by $\#\mathcal{P}_{k+\ell}$,*

where ℓ is the dimension of the lineality space of $\mathcal{P}$.

Proof Follows by Lemma 48, Lemma 50 and the fact that bounded polyhedral complexes are canonically homeormorphic to CW-complexes(c.f. Appendix A.2). ■

Lemma 52 *Let C be a bounded d-dimensional polyhedral complex such that $|C|$ is contractible and X a subcomplex of C. Then it holds that*

$$\beta_k(X) \leq \#\{(k+1) - \textit{dimensional polyhedra in } C \setminus X\}$$

for all $k \in [d-1]$.

Proof Since C only contains bounded polyhedra, we can equip C with a CW-complex structure (c.f. Appendix A.2) and compute the Betti numbers using cellular homology.

Let $Y = C \setminus X$, then Y is a set of polytopes, but not necessarily a polyhedral complex. By abuse of notation, we denote by Y_k the k-dimensional polytopes in Y and by $\mathrm{sk}_k(Y) = \{\ell-$dimensional polytopes in $Y \mid \ell \leq k\}$. By the definition of cellular homology it holds that $\beta_k(\mathrm{sk}_{k+1}(X)) = \beta_k(X)$ and hence, if $\#Y_{k+1} = 0$, then

$$\beta_k(X) = \beta_k(\mathrm{sk}_{k+1}(C \setminus Y)) = \beta_k(\mathrm{sk}_{k+1}(C) \setminus (\mathrm{sk}_k(Y) \cup Y_{k+1})) \leq \beta_k(\mathrm{sk}_{k+1}(C)) = 0$$

settling the induction hypothesis. To show the induction step, let $\mathrm{sk}_{k+1}(Y) = D \cup \hat{P}$, where $\hat{P}$ consists of one $(k+1)$-polytopes P and all its faces in Y. Therefore, $\#Y_{k+1} = \#D_{k+1} + 1$. Furthermore, let $B = C \setminus D$ and $A = C \setminus Y$. By the induction hypothesis we know that $\beta_k(B) \leq \#D_{k+1}$.

Our goal is to embed the cellular homology group $H_k(B)$ into $\mathbb{Z} \oplus H_k(A)$. Such an embedding readily implies that $\beta_k(B) \leq 1 + \beta_k(A)$. From this, the induction step follows:

$$\beta_k(B) \leq 1 + \beta_k(A) \leq 1$$

We first delete the $(k+1)$-dimensional polytope P itself (that is, without deleting the redundant faces, resulting in a polyhedral complex whose support we denote by B'), and observe by the elaborate definition of cellular homology groups that this induces a map $\phi_1 \colon H_k(B') \to \mathbb{Z} \oplus H_k(A)$. One can additionally observe that this map is an embedding: Notice that the homology arises from the (relative) homologies of the chain complex

$$\ldots \to C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \to \ldots$$

Deleting P decreases the image of the boundary map ∂_{k+1} and hence increases the k-th homology; however, it is straightforward to observe that

$$\phi_1 \colon H_k(B') \cong {}^{\ker \partial_k}\!/\!_{\operatorname{im} \partial_{k+1}} \to [\sigma] \oplus {}^{\ker \partial_k}\!/\!_{\operatorname{im} \partial_{k+1} \cup [\sigma]} \cong \mathbb{Z} \oplus H_k(A)$$

which maps a homology class $[\sum_{\tau \in C_k} c_\tau \tau]$ from the domain to $(c_\sigma, [\sum_{\tau \in C_k \setminus [\sigma]} c_\tau \tau])$ is an embedding, where σ is the generator corresponding to P.

To finish off the construction of the embedding, we finally define $\phi_2 \colon H_k(B) \to H_k(B')$, i.e., the map induced by deleting all faces of P in $\hat{P}$. This operation might reduce the kernel of the k-th boundary map, and hence potentially decrease the k-th Betti number. It is, however, again straightforward to observe that the map is injective in any case, in a similar fashion as above.

The claimed embedding is now $\phi_1 \circ \phi_2$, finishing the proof for $k \in [d-1]$. ■

Proposition 18 *Let $F \colon \mathbb{R}^d \to \mathbb{R}$ be a neural network of architecture $(d, n_1, \ldots, n_L, 1)$. Then it holds that $\beta_0(F) \leq \sum_{(j_1,\ldots,j_L) \in J} \quad \prod_{\ell=1}^{L} \binom{n_\ell}{j_\ell}$ and for all $k \in [d-1]$ that*

$$\beta_k(F) \leq \binom{\sum_{(j_1,\ldots,j_L) \in J} \quad \prod_{\ell=1}^{L} \binom{n_\ell}{j_\ell}}{d-k-s},$$

where $J = \big\{(j_1, \ldots, j_L) \in \mathbb{Z}^L \colon 0 \leq j_\ell \leq \min\{d, n_1 - j_1, \ldots, n_{\ell-1} - j_{\ell-1}\}$ for all $\ell = 1, \ldots, L\big\}$ and $s \in [d]$ is the dimension of the lineality space of a refinement of the canonical polyhedral complex of F.

Proof Theorem 1 in Serra et al. (2017) states that F has at most $r \coloneqq \sum_{(j_1,\ldots,j_L) \in J} \prod_{l=1}^{L} \binom{n_l}{j_l}$ linear regions. Let $\mathcal{P}$ be the canonical polyhedral complex of F. Since $\mathcal{P}$ is complete, it follows that $\#\mathcal{P}_k \leq \binom{r}{d-k+1}$. Furthermore, for $P \in \mathcal{P}$ it holds that $F_{|P}$ is affine linear and hence $F_{|P}^{-1}((-\infty, 0])$ is a half-space and therefore $P \cap F^{-1}((-\infty, 0]) = P \cap F_{|P}^{-1}((-\infty, 0])$ is a polyhedron. We define $\mathcal{P}^- = \{P \cap F^{-1}((-\infty, 0]) \mid P \in \mathcal{P}\}$ and $\mathcal{P}^+ = \{P \cap F^{-1}([0, \infty)) \mid P \in \mathcal{P}\}$, which due to the continuity of F are subcomplexes of the refinement of $\mathcal{P}$ where F takes on exclusively non-positive respectively non-negative values on all polyhedra. It follows immediately from the definition that $\#\mathcal{P}_k^+ \leq \#\mathcal{P}_k \leq \binom{r}{d-k+1}$. Let s be the dimension of the lineality space of the complete polyhedral complex $\mathcal{P}^- \cup \mathcal{P}^+$. By Lemma 51 we obtain a CW-complex C that is homotopy equivalent to $\mathcal{P}^- \cup \mathcal{P}^+$, and therefore in particular contractible. Furthermore, since $\mathcal{P}^-$ and $\mathcal{P}^-$ are subcomplex of $\mathcal{P}^- \cup \mathcal{P}^+$, we obtain subcomplexes X and Y of C such that X is homotopy equivalent to $F^{-1}((-\infty, 0])$ and Y is homotopy equivalent to $F^{-1}([0, \infty))$. It clearly holds that $C \setminus X \subseteq Y$ and hence by Lemma 52 it follows that

$$\beta_k(F^{-1}((-\infty, 0])) = \beta_k(X) \leq \#\{(k+1)\text{-dimensional polyhedra in } (C \setminus X)\} \leq \binom{r}{d-k-s},$$

proving the claim. ■