Journal of Computational and Applied Mathematics 441 (2024) 115667

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

On the number of regions of piecewise linear neural networks☆

Alexis Goujon *, Arian Etemadi, Michael Unser

Biomedical Imaging Group, École polytechnique fédérale de Lausanne (EPFL), Station 17, 1015 Lausanne, Switzerland

ARTICLE INFO

Keywords:
Deep learning
Expressivity
Activation functions
Continuous and piecewise-linear functions
Splines
Convex partitions

ABSTRACT

Many feedforward neural networks (NNs) generate continuous and piecewise-linear (CPWL) mappings. Specifically, they partition the input domain into regions on which the mapping is affine. The number of these so-called linear regions offers a natural metric to characterize the expressiveness of CPWL NNs. The precise determination of this quantity is often out of reach in practice, and bounds have been proposed for specific architectures, including for ReLU and Maxout NNs. In this work, we generalize these bounds to NNs with arbitrary and possibly multivariate CPWL activation functions. We first provide upper and lower bounds on the maximal number of linear regions of a CPWL NN given its depth, width, and the number of linear regions of its activation functions. Our results rely on the combinatorial structure of convex partitions and confirm the distinctive role of depth which, on its own, is able to exponentially increase the number of regions. We then introduce a complementary stochastic framework to estimate the average number of linear regions produced by a CPWL NN. Under reasonable assumptions, the expected density of linear regions along any 1D path is bounded by the product of depth, width, and a measure of activation complexity (up to a scaling factor). This yields an identical role to the three sources of expressiveness: no exponential growth with depth is observed anymore.

1. Introduction

The ability to train deep parametric models has enabled dramatic advances in a wide variety of fields, ranging from computer vision to natural-language processing [1,2]. Many popular deep models belong to the family of feedforward neural networks (NNs), for which the input–output mapping takes the form[1]

$$\mathbf{x} \mapsto (\boldsymbol{\sigma}_L \circ \boldsymbol{f}_{\theta_L} \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\sigma}_2 \circ \boldsymbol{f}_{\theta_2} \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_{\theta_1})(\mathbf{x}), \tag{1}$$

where L is the number of layers of the NN (referred to as the depth of the NN), $\boldsymbol{f}_{\theta_k} : \mathbb{R}^{d_k} \to \mathbb{R}^{d_{k+1}}$ is an affine function parameterized by θ_k, and $\boldsymbol{\sigma}_k$ is a non-affine activation function. One of the most widespread activation functions in deep learning is the rectified linear unit $\mathrm{ReLU}(x) = \max(x, 0)$ [3–5]. With this choice, the mapping is a composition of continuous and piecewise-linear (CPWL) functions, which yields a map that is CPWL too [6]. Remarkably, the reverse also holds true: any CPWL function $\mathbb{R}^d \to \mathbb{R}$ can be parameterized by a ReLU NN with at most $\lceil \log_2(d+1) \rceil$ hidden layers [7]. The family of NNs generating CPWL functions (referred to as CPWL NNs in the sequel) is broad. It benefits from a large choice of effective activation functions that includes ReLU [5], leaky ReLU [4], PReLU [8], CReLU [9], Maxout [10], linear splines [11,12], GroupSort [13], Householder [14] as well as other

☆ This work was supported by the Swiss National Science Foundation, Grant 200020_184646/1.

* Corresponding author.
E-mail addresses: alexis.goujon@epfl.ch (A. Goujon), arietemadi@gmail.com (A. Etemadi), michael.unser@epfl.ch (M. Unser).

[1] In practice, the last layer of the NN is not necessarily an activation function. This case is also covered by our framework since $\boldsymbol{\sigma}_L$ can be any continuous and piecewise-linear function, including the identity function.

https://doi.org/10.1016/j.cam.2023.115667
Received 7 May 2023; Received in revised form 1 November 2023
Available online 10 November 2023
0377-0427/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

components such as convolutional layers, max- and average-pooling, skip connections [15], and batch normalization [16](once the model is trained). While the depth of the architecture is instrumental to overcome the curse of dimensionality [17–19], it concurrently deters our understanding of the parameterization when compared to simpler models [20].

The observation that a ReLU NN produces a CPWL function sheds light on its behavior. In effect, a ReLU NN partitions the input domain into affine regions [21,22]. The characteristics of the regions are therefore fundamental to grasp the structure of the learnt mapping and there exist different approaches to define them [23,24]. The regions can be described as polyhedrons or union of polyhedrons, which results from the continuity and the piecewise-affine property of the mapping. In the case of ReLU NNs, it is common to define activation regions, which are sets of points that fire the same group of neurons. On each activation region, the mapping is affine and these sets are convex [25]. Unfortunately, the linear regions in deep NNs are only indirectly specified. While they can be locally described [26] their global delimitation becomes computationally less and less tractable as the dimension increases, which compromises the interpretability of deep NNs. Yet, it is entangled with their ability to overcome the curse of dimensionality.

The successive compositions inherent in deep models prevent us from attributing a specific role to each parameter. The size and the expressiveness of the function space $\mathcal{H}_{\mathcal{N}}$ generated by a given architecture $\mathcal{N}$ is consequently remotely connected to the number of trainable parameters. With their remarkable structure, CPWL NNs benefit from another meaningful descriptor: the distribution of counts of regions of all the mappings that the architecture can produce. Two approaches have been proposed to give a better understanding of this descriptor.

- *Upper and lower bound the maximum number of regions of the CPWL mappings generated by a given architecture.* The first bounds, given in [6], showed that the maximum number of regions that can be produced by ReLU NNs increases exponentially with their depth. This revealed that deep models have the ability to generate much more complex functions than shallow ones do. The bounds for ReLU NNs have since been refined, for example in [27] and then in [28], and also extended to other NN architectures. For instance, [29] specifies bounds for the maximum number of regions of convolutional NNs (CNNs). It is shown that CNNs produce more regions per parameter than fully connected NNs do. For Maxout NNs, bounds can be derived directly from the ones on ReLU NNs [6,27]. However, this approach usually yields loose bounds, as recently shown in [30]. The derivation of sharp bounds for Maxout NNs, as proposed in [30], requires to take into account the specificities of the Maxout unit, and it was handled via the use of tropical geometry.
  The available bounds show that the maximum number of regions in ReLU and Maxout NNs increases exponentially with their depth. It suggests that deep models have the ability to generate more complex functions than shallow ones do [6,7,27,28,30].
- *Upper bound the average number of regions of the mappings generated by ReLU and Maxout NNs.* The available bound for ReLU NNs depends on the number of neurons, regardless of whether the NN is deep or wide, and depth does not produce exponentially more regions on average [25,31]. In other words, this behavior drastically differs from the maximum number of regions. This new perspective was then recently extended to Maxout NNs, with a similar qualitative conclusion [32].

The existing toolbox of CPWL NNs is broad and likely not complete yet, as hinted by recent works on the MaxMin or more generally GroupSort activation function, in the field of Lipschitz-constrained NNs [13,33], or with the Piecewise Linear Unit (PWLU) [34]. Previous studies on the count of linear regions have provided insights on some specific CPWL NNs only, mostly ReLU and Maxout NNs. Their qualitative outcomes turn out to hold true for CPWL NNs in general. We intend to prove this claim with quantitative results in this paper. We want to improve the understanding of the role of the three main ways to increase the expressiveness of CPWL NNs (Fig. 1), namely,

- **depth**, which is the number of composed CPWL functions;
- **width**, which relates the input and output dimensions of the composed CPWL layers;
- **activation complexity**, the rationale there being that the expressiveness of a CPWL NN can be heightened by increasing the complexity of its activation functions. This strategy is used with both univariate and multivariate activation functions, and it gave rise to deepspline, Maxout, GroupSort, and PWLU NNs for example. In the remainder of the paper, the complexity of an activation will refer to its number of linear regions. For example, a rank-k Maxout unit has a complexity of k (see Fig. 6 for visual examples).

Our contributions are as follows.

(i) Generalization of the notion of arrangement of hyperplanes to arrangement of convex partitions with analogous tight bounds on the number of regions.
(ii) Determination of precise bounds on the maximal number of linear convex regions generated by the primary operations of the space of CPWL functions (sum, vectorization, and composition). The compositional upper and lower bound grow exponentially with depth and polynomially with the width and the activation complexity.
(iii) Demonstration that, under reasonable assumptions, the expected number of regions along a 1D path for random CPWL NNs is at most linear with the product of the depth, the width, and the activation complexity (up to an independent factor), which yields equivalent roles to the three descriptors in terms of expressiveness.

The paper is organized as follows: In Section 2, we present the relevant mathematical concepts. In Section 3, we bound from below and from above the maximal number of regions produced by CPWL NNs and, in Section 4, we present a stochastic framework to quantify the average expressiveness of CPWL NNs with random parameters.

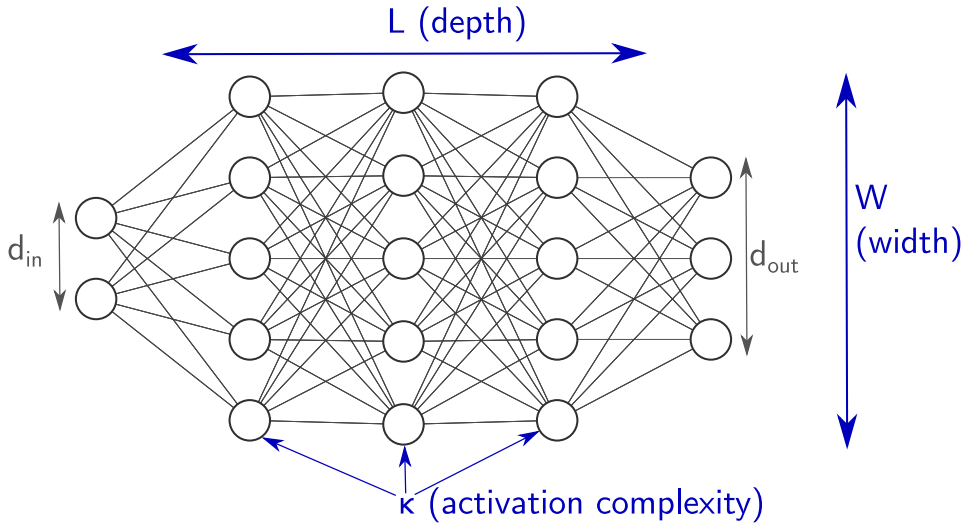

Fig. 1. The three sources of complexity of CPWL NNs.

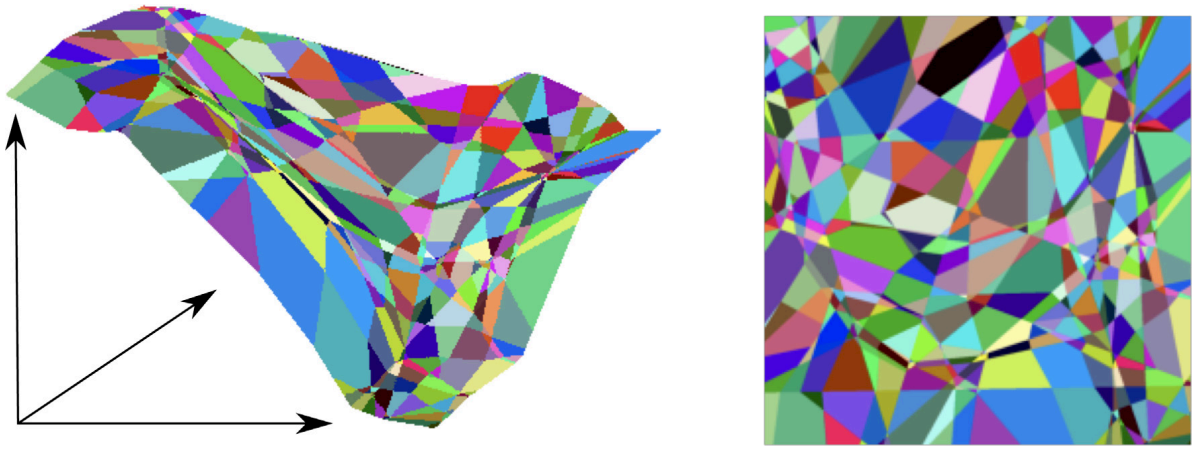

Fig. 2. An $\mathbb{R}^2 \to \mathbb{R}$ CPWL function and its corresponding partition of the input space.

2. Mathematical preliminaries

2.1. CPWL functions

Definition 1. A function $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ is continuous and piecewise-linear (CPWL) if it is continuous and if there exists a set $\{f^k : k \in \{1, \dots, K\}\}$ of affine functions and closed subsets $(\Omega_k)_{k=1}^K$ of $\mathbb{R}^d$ with nonempty and pairwise disjoint interiors such that $\cup_{k=1}^K \Omega_k = \mathbb{R}^d$ and $f_{|\Omega_k} = f^k$ on Ω_k. The f^k are called the affine pieces of f, and the Ω_k the corresponding projection regions.

An example of a CPWL function and of its partition is given in Fig. 2. The kth component of a vector-valued CPWL function f_ℓ, which is necessarily CPWL as well, will be denoted by $f_{\ell,k}$. The space of CPWL functions has the following remarkable properties:

- it is closed under compatible compositions;
- it is closed under compatible linear combinations;
- it is closed under compatible vectorization.

Since the function $\mathbf{x} \mapsto \max(\mathbf{x}) = \max(x_1, \dots, x_d)$ is CPWL (with d regions), the space of CPWL functions is also closed under max-pooling.

2.2. Regions of CPWL functions and convex partitions

The term *linear region* is frequently used in an ambiguous way and may refer to different mathematical definitions. In the sequel, we shortly present some relevant definitions and discuss them in the context of CPWL NNs.

Fig. 3. Convex and projection regions of the CPWL function $(x, y) \mapsto \mathrm{ReLU}(\min(x + 1, -x + 1))$.

2.2.1. Projection regions

We recall that a polyhedron is the intersection of finitely many half-spaces, and that a polytope is a bounded polyhedron. The subsets Ω_k in Definition 1 are commonly referred to as projection regions [23,24]. The affine pieces of different projection regions are distinct and, since the overall function is continuous, the common points of two neighboring regions lie in a hyperplane. This implies that the Ω_k are polyhedrons or unions of polyhedrons. These projection regions might, however, not be connected (Fig. 3).

2.2.2. Convex regions

It is usually preferred to work with (connected) convex regions because of their simpler geometrical structure. We now precisely define convex linear regions of CPWL functions.

Definition 2 (*Convex Partitions of $\mathbb{R}^d$, Adapted From [35]*)**.** Let n and d be two positive integers. A convex partition of $\mathbb{R}^d$ is a collection $\Pi = (P_1, P_2, \dots, P_n)$ of convex and closed subsets of $\mathbb{R}^d$ with nonempty and pairwise-disjoint interiors so that the union $\bigcup_{k=1}^{n} P_k = \mathbb{R}^d$. Each of the sets P_k is called a region of Π. Convex partitions with n regions are called n-partitions.

Definition 3 (*Linear Convex Partition*)**.** A convex partition Π of $\mathbb{R}^d$ is said to be a linear convex partition of a CPWL function $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ if f is affine on each region of Π.

The existence of a linear convex partition is guaranteed for any CPWL function but not its unicity. This motivates Definition 4, which gives a precise meaning to the *number of convex linear regions* for CPWL functions.

Definition 4 (*Number of Convex Linear Regions*)**.** The number κ_f of convex linear regions of f is defined as the minimal cardinality of all linear convex partitions of f.

A special instance of the linear convex regions for scalar-valued CPWL functions are the uniquely-ordered regions. Each of these regions has the same ordering of the values of the affine pieces f^k of f in all its points [24]. Uniquely-ordered regions are used to build the lattice representation of a CPWL function [23] and are tightly connected to the GroupSort activation function [13].

2.2.3. Projection vs. convex linear regions

In the remainder of the paper we shall keep in mind the following connections between projection and convex linear regions.

- Projection regions can always be partitioned into convex regions so that any upper bound on the number of convex regions also applies to the number of projection regions. Conversely, the number of convex regions can also be upper bounded by the number of projection regions (Proposition 1).
- The majority of commonly used parameterizations have typically convex projection regions. The local parameterization with hat basis functions produces simplicial linear splines whose natural regions are simplices [20] and, therefore, are convex. Other known linear expansions, such as the generalized hinging-hyperplanes model [36], use nonlocal CPWL basis functions that partition the input domain into convex regions. The generated function will produce projection regions that are convex for all sets of parameters except for some specific values that are usually encountered with zero probability in a learning framework. The convex regions are also naturally adapted to compositional models such as ReLU and Maxout NNs as explained it [31].

Proposition 1. *Let $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a CPWL function with ρ projection regions. The number κ of linear convex regions of f is no larger than the number of convex regions formed by the arrangement of $\rho(\rho-1)/2$ hyperplanes*

$$\rho \leq \kappa \leq \begin{cases} 2^{\rho(\rho-1)/2}, & \rho(\rho-1)/2 \leq d \\ \sum_{k=0}^{d} \binom{\rho(\rho-1)/2}{k}, & \textit{otherwise.} \end{cases} \tag{2}$$

The proof of Proposition 1 is given in Appendix A.

2.2.4. Useful properties of convex partitions

We now give a series of lemmas on convex partitions that are used in the proofs of Section 3. The proofs are given in Appendix A.

For convenience, we extend the definition of convex partitions of $\mathbb{R}^d$ to convex partitions of affine subspaces of $\mathbb{R}^d$. In particular, a convex partition of an affine subspace E of $\mathbb{R}^d$ consists of convex and closed subsets of E of dimension $\dim(E)$ whose pairwise intersection is of dimension smaller than $\dim(E)$ and whose union is E.

Lemma 1 (*Projection of a Convex Partition*). *Let E be an affine subspace of $\mathbb{R}^d$ and Π an n-partition of $\mathbb{R}^d$. Then, there exists a convex partition Π_E of E in $\mathbb{R}^d$ with no more than n regions such that, for $P_E \in \Pi_E$, there is $P \in \Pi$ with $P_E \subset P$.*

Lemma 2 (*Preimage of a Convex Partition Under Affine Maps*). *Let $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^{d'}$ be an affine function and Π be an N-partition of the affine space $\boldsymbol{f}(\mathbb{R}^d)$ in $\mathbb{R}^{d'}$. Then, $\boldsymbol{f}^{-1}(\Pi) = \{\boldsymbol{f}^{-1}(P) : P \in \Pi\}$ is an N-partition of $\mathbb{R}^d$.*

Lemma 3. *Let $(\boldsymbol{f}_\ell)_{\ell\in[L]}$ be a collection of affine functions with $\boldsymbol{f}_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_{\ell+1}}$. Then,*

$$\dim((\boldsymbol{f}_L \circ \cdots \circ \boldsymbol{f}_1)(\mathbb{R}^{d_1})) \leq \min(d_1, \dots, d_{L+1}). \tag{3}$$

2.3. Arrangement of convex partitions

The known results on the number of convex regions of ReLU NNs are built upon the theory of hyperplane arrangements. In combinatorial geometry, an arrangement of hyperplanes refers to a set of hyperplanes. It is known that the number of connected regions formed by an arrangement of N hyperplanes in $\mathbb{R}^d$ is at most $\sum_{k=0}^{\min(d,N)} \binom{N}{k}$ [37]. This bound is reached when the hyperplanes are in general position: any collection of k of them intersect in a $(d-k)$-dimensional plane for $1 \leq k \leq d$ and have an empty intersection for $k > d$. Although this positioning seems very specific, it is qualified as "general" because it almost surely happens when the hyperplanes are randomly generated (with a "reasonable" notion of randomness). When it comes to the study of generic CPWL NNs, the concept of arrangement of hyperplanes lacks precision since only a small fraction of all convex partitions can be seen as arrangement of hyperplanes. We thus introduce the notion of arrangement of convex partitions (Definition 5 and Fig. 4) as a generalization, which will prove to be necessary to find the precise bounds given in Section 3. Note that, in the case of an arrangement of N hyperplanes, our terminology differs. Instead of considering the hyperplanes, we rather consider the N 2-partitions they form, which consist of pairs of closed half-spaces separated by the hyperplanes.

Definition 5 (*Arrangement of Convex Partitions*). Let $(\Pi_k)_{k\in[N]}$ be a collection of N convex partitions, with $\Pi_k = (P_1^k, \dots, P_{n_k}^k)$, for $k \in [N]$. The arrangement $\mathcal{A}(\Pi_1, \dots, \Pi_N)$ of these partitions is the convex partition whose regions are the $A_{m_1,\dots,m_N}$ that have nonempty interiors, where

$$A_{m_1,\dots,m_N} = \bigcap_{k=1}^{N} P_{m_k}^k,$$

for $(m_1, \dots, m_N) \in \{1, \dots, n_1\} \times \cdots \times \{1, \dots, n_N\}$.

3. Maximum number of regions produced by CPWL NNs

In this section, we characterize the largest number of regions that can be generated by simple operations with CPWL functions, including sums, vectorizations, and compositions. In particular, we strictly generalize the known upper and lower bounds on the number of regions of ReLU NNs [27] and Maxout NNs [30] to NNs activated by generic CPWL activation functions.

3.1. Upper bound on the number of regions of arrangements

Operations with CPWL functions imply arrangement of convex partitions, either explicitly, for sums and vectorizations, or implicitly, for compositions. It is straightforward to see that an arrangement $\mathcal{A}(\Pi_1, \dots, \Pi_N)$ of N convex partitions $\Pi_1, \dots, \Pi_N$ of $\mathbb{R}^d$ with $n_1, \dots, n_N$ regions cannot yield more than $n_1 n_2 \cdots n_N$ regions. This naive bound is a polynomial of degree N in $n_1, \dots, n_N$. In dimension $d = 1$ one can, however, check that the bound is not sharp: the number of regions is no more than $1+(n_1-1)+\cdots+(n_N-1)$. More generally, the number of regions of the arrangement is bounded by a polynomial in the cardinality of the partitions $n_1, \dots, n_N$ of degree $\min(d, N)$ (Theorem 1), which highlights the role played by the dimension of the ambient space.

Theorem 1 (*Arrangements' Upper Bound*). *The maximum cardinality $\beta^d(n_1, \dots, n_N)$ of the arrangement $\mathcal{A}(\Pi_1, \dots, \Pi_N)$ of N convex partitions $\Pi_1, \dots, \Pi_N$ of $\mathbb{R}^d$ with cardinality $n_1, \dots, n_N$ is a polynomial in $n_1, \dots, n_N$ of degree $\min(d, N)$. It is given by*

$$\beta^d(n_1, \dots, n_N) = 1 + \sum_{k=1}^{\min(d,N)} \sum_{1\leq \ell_1 < \cdots < \ell_k \leq N} \prod_{q=1}^{k} (n_{\ell_q} - 1). \tag{4}$$

Moreover, this bound satisfies

$$\begin{aligned} \beta^d(n_1, \dots, n_N) &= \textstyle\prod_{k=1}^{N} n_k, && \text{if } N \leq d \\ \beta^d(n_1, \dots, n_N) &\leq \left(1 + \textstyle\sum_{k=1}^{N}(n_k - 1)\right)^d \leq \left(\textstyle\sum_{k=1}^{N} n_k\right)^d, && \text{otherwise.} \end{aligned} \tag{5}$$

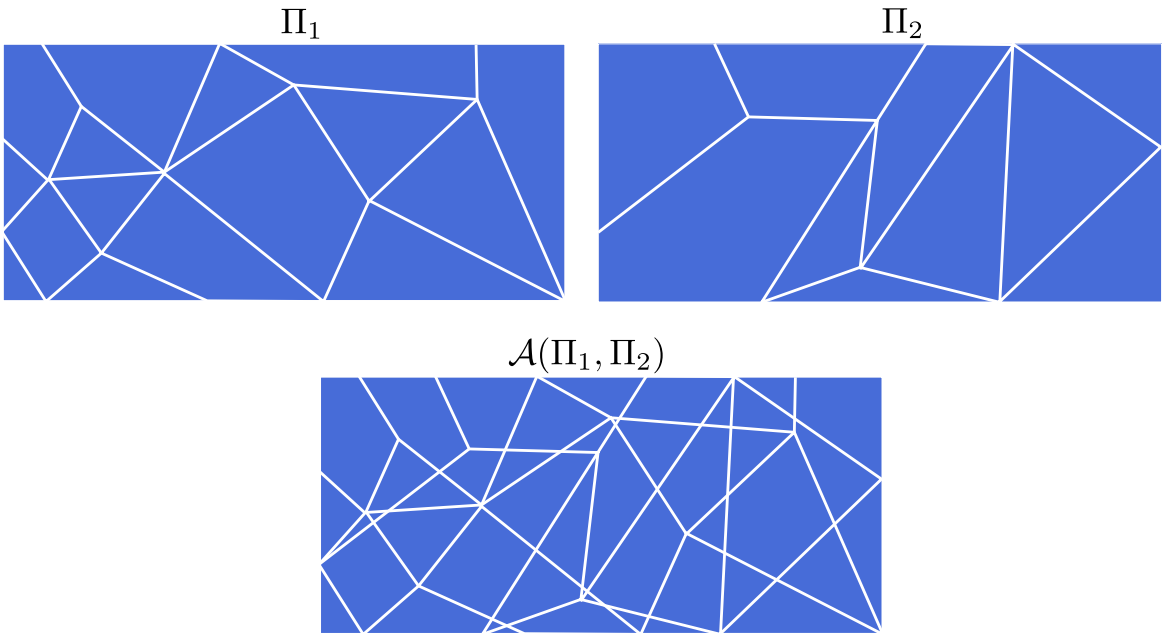

Fig. 4. Arrangement of two convex partitions of $\mathbb{R}^2$.

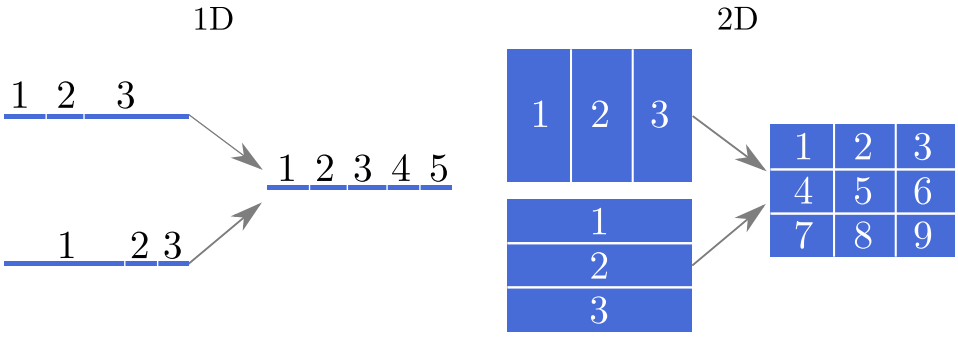

Fig. 5. Arrangement of two convex partitions with 3 regions each. While in 2D the maximal number of regions is $3 \times 3 = 9$, this cannot be reached in 1D, for which the maximum is 5.

The expression of the bound in Theorem 1 is based on a broad result of discrete geometry [38]. We then relied on Zaslavsky's Theorem [37] and Whitney's formula to construct a specific arrangement of convex partitions for any set of parameters $d, N, n_1, \dots, n_N \in \mathbb{N}\backslash\{0\}$ that achieves the bound. The proof of Theorem 1 is given in Appendix A. We now discuss the result and its implications.

- Theorem 1 is a generalization of the hyperplane-arrangement bound. Indeed, let us consider the number of regions generated by an arrangement of N hyperplanes: each hyperplane defines a 2-partition of $\mathbb{R}^d$ and the bound yields $\beta^d(2, \dots, 2) = 1 + \sum_{k=1}^{\min(d,N)} \sum_{1 \le \ell_1 < \dots < \ell_k \le N} 1 = 1 + \sum_{k=1}^{\min(d,N)} \binom{N}{k} = \sum_{k=0}^{\min(d,N)} \binom{N}{k}$, which is known to be exactly the number of convex regions generated by an arrangement of N hyperplanes in general position [37].
- The naive upper bound can be rewritten as $\prod_{k=1}^{N} n_k = \prod_{k=1}^{N}((n_k - 1) + 1) = 1 + \sum_{k=1}^{N} \sum_{1 \le \ell_1 < \dots < \ell_k \le N} (n_{\ell_1} - 1) \cdots (n_{\ell_k} - 1)$. This shows that when $N \le d$, the naive bound is optimal. By contrast, when $N > d$, the dimension enforces the existence of one or more empty intersections between regions of different partitions. This is illustrated in Fig. 5 with a simple example.
- For partitions with the same number n of regions, we introduce the simpler notation $\beta_N^d(n) := \beta^d(n, \dots, n) = 1 + \sum_{k=1}^{\min(d,N)} \sum_{1 \le \ell_1 < \dots < \ell_k \le N} (n-1)^k = \sum_{k=1}^{\min(d,N)} \binom{N}{k} (n-1)^k \le \min(n^N, (1 + N(n-1))^d)$.
- The bound is reached when the partitions Π_k are made of the regions of the arrangement of $(n_k - 1)$ distinct parallel hyperplanes, where the hyperplanes are in general position when only one per partition is selected (more detailed in the proof in Appendix A).

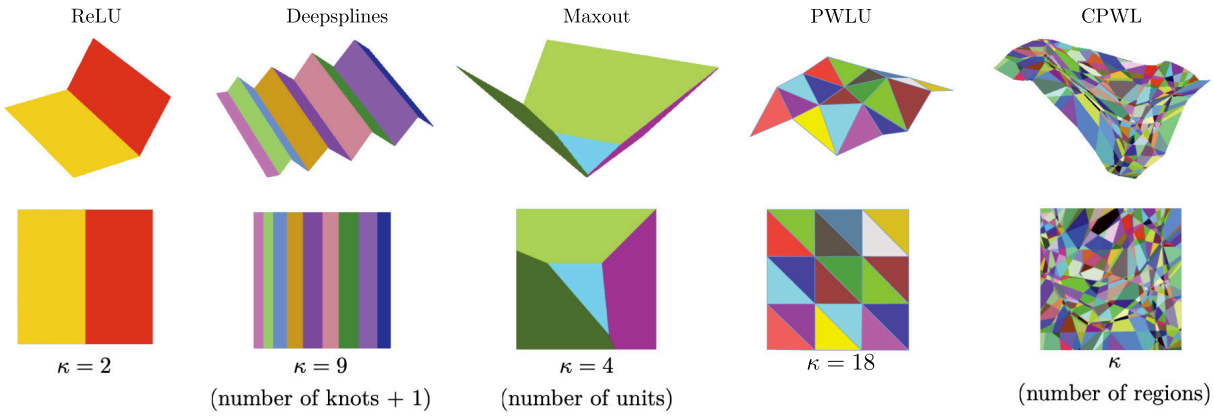

Fig. 6. Partition and complexity of some CPWL components.

Remark 1. After the disclosure of our work on arXiv, we became aware of [30], which contains highly relevant results on the complexity of Maxout NNs. From Theorem 1, we can directly recover the sharp bound on the number of regions of a shallow Maxout NN recently given in [30, Theorem 3.7]. Regarding the converse, i.e. inferring Theorem 1 from [30, Theorem 3.7], we believe that it could perhaps be done but it is not immediate. Indeed, [30, Theorem 3.7] is specific to convex partitions that are the linear partitions of the maximum of affine functions, i.e. only specific CPWL functions. Note that the proofs in [30] rely on tropical geometry, which gives an overall perspective very different from ours.

3.2. Single hidden-layer: Bound for the sum and vectorization operations

The sum and vectorization of CPWL functions both yield the same bound on the number of linear convex regions. We now give a novel optimal bound in Proposition 2. The proof is given in Appendix A.

Proposition 2. *Let $f_1, \dots, f_N : \mathbb{R}^d \to \mathbb{R}$ be CPWL functions with $\kappa_1, \dots, \kappa_N$ convex linear regions. The number of convex linear regions of the sum $(f_1 + \cdots + f_N)$ and of the vector-valued function $(f_1, \dots, f_N)$ can be bounded by a polynomial in $\kappa_1, \dots, \kappa_N$ of degree $\min(d, N)$, namely*

$$\kappa_{f_1+\cdots+f_N} \leq \beta^d(\kappa_1, \dots, \kappa_N), \tag{6}$$

$$\kappa_{(f_1,\dots,f_N)} \leq \beta^d(\kappa_1, \dots, \kappa_N), \tag{7}$$

and these bounds are sharp.

Remark 2. Bounds similar to the ones given in Proposition 2 have recently been derived for one hidden-layer Maxout NNs [30]. The latter work is a specific instance of our setting, in which the CPWL functions considered are the maximum of a finite set of affine functions.

As an illustration of Proposition 2 and Theorem 1, we give some direct implications on the number of regions of some building blocks of CPWL NNs before going deeper.

Ridge functions. Consider the ridge expansion $f_R : \mathbf{x} \mapsto \sum_{k=1}^N \lambda_k \mathrm{ReLU}(\mathbf{w}_k^T \mathbf{x} + b_k)$, where $\mathbf{w}_k \in \mathbb{R}^d$ and $b_k \in \mathbb{R}$. The number κ_{Ridge} of linear convex regions of f_R is upper-bounded as

$$\kappa_{\mathrm{Ridge}} \leq \beta_N^d(2) = \sum_{k=0}^{\min(d,N)} \binom{N}{k} \leq \min(2^N, (N+1)^d), \tag{8}$$

and the bound is tight.

Max-pooling. The kth component of the max-pooling operation $\boldsymbol{f}_{\mathrm{mp}} : \mathbb{R}^d \to \mathbb{R}^{d'}$ takes the form $\boldsymbol{f}_{\mathrm{mp}}^k(x_1, \dots, x_d) = \max_{p \in I_k}(x_p)$, where I_k is a set of chosen cardinality N of "neighboring" coordinate indices. The number κ_{mp} of convex linear regions of the max-pooling operation is upper-bounded as

$$\kappa_{\mathrm{mp}} \leq \beta_{d'}^d(N) = \sum_{k=0}^{\min(d,d')} \binom{d'}{k} (N-1)^k. \tag{9}$$

Generalized Hinging Hyperplanes (GHH). Consider the GHH expansion $\mathbb{R}^d \to \mathbb{R}$ $f_G = \sum_{k=1}^{N} \epsilon_k \max(f_1^k, \dots, f_{d+1}^k)$, where f_p^k are affine functions and $\epsilon_k = \pm 1$ [36]. The number κ_{GHH} of convex linear regions of f_G is upper-bounded as

$$\kappa_{\text{GHH}} \leq \beta_N^d(d+1) = \sum_{k=0}^{\min(d,N)} \binom{N}{k} d^k \leq \min\left((d+1)^N, (Nd+1)^d\right). \tag{10}$$

GroupSort layer. The sort operation takes as input a vector $\mathbf{x} \in \mathbb{R}^d$ and simply sorts its components. For any permutation σ of the set $\{1, \dots, d\}$, we define the uniquely-ordered region $P_\sigma = \{\mathbf{x} \in \mathbb{R}^d : x_{\sigma(1)} \leq \cdots \leq x_{\sigma(d)}\}$, where x_k is the kth component of $\mathbf{x}$. These regions are convex as intersections of half-spaces and the sort operation agrees on them with distinct affine functions, namely, permutations. We infer that the sort operation has exactly $d!$ linear convex regions and the same number of projection regions.

The GroupSort activation was recently introduced and shown to be beneficial in the context of Lipschitz-constrained learning [13]. It generalizes the minmax and sort activations: it splits the pre-activation into a chosen number n_g of groups of size g_s (with $n_g g_s = d$), sorts each pre-activation of each group in ascending order, and outputs the combined sorted groups. Each group produces $g_s!$ linear convex regions which are invariant along the coordinates that are not in the group. We infer the number of linear convex regions of the GroupSort activation to be $\kappa_{\text{GS}} = (g_s!)^{n_g}$, which can be bounded as

$$(g_s/2)^{d/2} \leq \kappa_{\text{GS}} \leq (g_s^{g_s})^{n_g} = g_s^d, \tag{11}$$

where we have used the known inequalities $(n/2)^{n/2} \leq n! \leq n^n$. The bounds support the intuition that larger group sizes generate more regions than smaller ones. Note, however, that they simultaneously increase the computational complexity of the layer.

PWLU. The PWLU [34] is a learnable CPWL activation function with control points placed on a grid and with fixed linear regions (namely simplices whose vertices are control points). In its 2D version, a PWLU $\varphi_{\text{PWLU}} : \mathbb{R}^2 \to \mathbb{R}$ with M^2 control points has $2(M-1)^2$ linear regions that are triangles, see Fig. 6 for an illustration with $M = 4$, and see [34, Figure 5] for a more generic representation of PWLUs. Consider the one-hidden layer $\mathbb{R}^d \to \mathbb{R}$ PWLU NN $f_{\text{PWLU}}(\mathbf{x}) = \sum_{k=1}^{N} \varphi_{\text{PWLU}}^k(\mathbf{W}_k \mathbf{x})$ with 2D PWLU activations φ_{PWLU}^k with M^2 control points and corresponding weight matrices $\mathbf{W}_k \in \mathbb{R}^{2\times d}$. The number κ_{PWLU} of convex linear regions of this PWLU NN is upper-bounded as

$$\kappa_{\text{PWLU}} \leq \beta_N^d\left(2(M-1)^2\right) = \sum_{k=1}^{\min(d,N)} \binom{N}{k} (2(M-1)^2 - 1)^k \tag{12}$$

$$\leq \min((2(M-1)^2)^N, (1 + N(2(M-1)^2 - 1))^d) \tag{13}$$

$$\leq \min((2M^2)^N, (1 + N(2M^2))^d). \tag{14}$$

Our framework also allows one to derive bounds for NNs activated with higher dimensional PWLUs, but we are not aware of their use in practice.

3.3. Multiple hidden-layers: Compositional bounds

The architecture of a CPWL NN $\mathbb{R}^{d_1} \to \mathbb{R}^{d_{L+1}}$ is specified by its depth L, its layer dimensions $(d_1, \dots, d_{L+1})$, and its activation complexity $\kappa_{\ell,k}$ at each node (ℓ, k), which is naturally depicted by the number of linear convex regions of the kth component of the ℓth composed function (Fig. 6). Theorem 2 below yields precise bounds on the maximal number of convex linear regions of any CPWL NN. It is complemented by Corollary 1 which tackles the following question: given a CPWL NN with fixed input and output dimensions, how is the maximal number of regions related to depth, width, and activation complexities? Our results confirm and generalize the following qualitative intuitions:

- (i) depth can exponentially increase the complexity of the generated function;
- (ii) width and activation complexity, on the contrary, can only increase the number of linear convex regions of the generated function polynomially;
- (iii) layers with small dimensions reduce the maximal number of regions produced by the NN, especially if they are located toward the input of the NN. This bottleneck effect stems from the upper bound given in Theorem 1.

Note that (i) is well known and was first proven in [6], (ii) is in agreement with the recent results in [30] obtained for the particular instance of Maxout NNs, and (iii) was observed for ReLU NNs in [27,39].

Theorem 2. *The maximal number $\kappa_{\max}$ of convex linear regions of a CPWL NN with depth L, layer dimensions $(d_1, \dots, d_{L+1})$, and activation complexities $\kappa_{\ell,k}$ for $k = 1, \dots, d_{\ell+1}$ and $\ell = 1, \dots, L$, is bounded as*

$$\prod_{\ell=1}^{L} \alpha^{\min(d_1,\dots,d_{L+1})}(\kappa_{\ell,1}, \dots, \kappa_{\ell,d_{\ell+1}}) \leq \kappa_{\max} \leq \prod_{\ell=1}^{L} \beta^{\min(d_1,\dots,d_\ell)}(\kappa_{\ell,1}, \dots, \kappa_{\ell,d_{\ell+1}}), \tag{15}$$

where $\beta^{\cdot}(\cdot)$ is the upper bound on the number of regions of an arrangement of convex partitions (Theorem 1) and where

$$\alpha^{\min(d_1,\dots,d_{L+1})}(\kappa_{\ell,1}, \dots, \kappa_{\ell,d_{\ell+1}}) = \max_{\tau \in \mathcal{T}_{d_\ell}} \prod_{r=1}^{\min(d_1,\dots,d_{L+1})} \sum_{k \in \tau^{-1}(\{r\})} \kappa_{\ell,k}. \tag{16}$$

There, $\mathcal{T}_{d_\ell}$ denotes the set of all mappings from $\{k \in \mathbb{N} : 1 \leq k \leq d_{\ell+1}\}$ to $\{k \in \mathbb{N} : 1 \leq k \leq \min(d_1, \dots, d_{L+1})\}$, and $\tau^{-1}(\{r\})$ denotes the preimage of $\{r\}$ under τ.

Corollary 1. *The maximal number $\kappa_{\max}$ of convex linear regions of a CPWL NN with L layers, layer dimensions $(d_{\text{in}}, W, \dots, W, d_{\text{out}})$, with $d_{\text{in}}, W, d_{\text{out}} \in \mathbb{N}\backslash\{0\}$ and $W \geq d_{\text{in}}$, where each component of the composed functions has κ linear convex regions is bounded as*

$$(\kappa \lfloor W/d^* \rfloor)^{Ld^*} \leq \kappa_{\max} \leq (\kappa W)^{Ld_{\text{in}}}, \tag{17}$$

where $d^* = \min(d_{\text{in}}, d_{\text{out}})$.

Corollary 2. *The bounds given in Theorem 2 and Corollary 1 also apply to the maximal number of projection regions of a CPWL NN and, equivalently, to its maximal number of distinct affine pieces.*

The proofs of Theorem 2 and of its corollaries can be found in Appendix A.4.

3.4. Application to some popular CPWL NNs

In the sequel, we consider the CPWL NN $\boldsymbol{f}_L \circ \cdots \circ \boldsymbol{f}_1$ where $\boldsymbol{f}_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_{\ell+1}}$. We now apply Theorem 2 to bound the maximal number of convex linear regions produced by the most popular architectures. Note that the lower bound given in Theorem 2 only applies to CPWL NNs with pointwise activation functions. This includes ReLU and, more generally, deepspline NNs. The reason is that the lower bound of Theorem 2 was found by building a deepspline NN.

ReLU/PReLU/Leaky ReLU NNs. In a ReLU NN, the kth component $f_{\ell,k}$ of $\boldsymbol{f}_\ell$ takes the form $f_{\ell,k} : \mathbf{x} \mapsto \text{ReLU}(\mathbf{w}_{\ell,k}\mathbf{x} + b_{\ell,k})$ and has two convex linear regions (half-spaces). Theorem 2 then yields

$$\kappa_{\text{ReLU}} \leq \prod_{\ell=1}^{L} \sum_{k=0}^{\min(d_1,\dots,d_\ell)} \binom{d_{\ell+1}}{k}, \tag{18}$$

which is the bound proposed in [6]. However, it is not the tightest upper bound known [27]. The reason is that the ReLU function is only a very specific instance of 1D CPWL functions with 2 linear regions: the image of the half real line $(-\infty, 0]$ by the ReLU function is only the singleton $\{0\}$. This reduces the apparent dimension of the problem for any region that would not fire all neurons. This observation was exploited in [27] to get a better estimate. In that sense, (18) is better tailored to PReLU and Leaky ReLU NNs, which have activations with two nonzero slope regions.

Deepspline NN. Deepspline NNs have learnable pointwise 1D CPWL activation functions [11,12,40]. Given activation functions with $(\kappa - 1)$ knots (at most κ linear convex regions), the number of linear convex regions of the NN is bounded as

$$\kappa_{\text{Deepspline}} \leq \prod_{\ell=1}^{L} \sum_{k=0}^{\min(d_1,\dots,d_\ell)} \binom{d_{\ell+1}}{k} (\kappa - 1)^k. \tag{19}$$

Maxout NN. In a Maxout NN with κ units, the kth component f_ℓ^k of $\boldsymbol{f}_\ell$ takes the form $f_\ell^k : \mathbf{x} \mapsto \max(h_{\ell,k}^1, \dots, h_{\ell,k}^\kappa)$, where $h_{\ell,k}^1, \dots, h_{\ell,k}^\kappa$ are learnable affine functions [10]. Theorem 2 yields that

$$\kappa_{\text{Maxout}} \leq \prod_{\ell=1}^{L} \sum_{k=0}^{\min(d_1,\dots,d_\ell)} \binom{d_{\ell+1}}{k} (\kappa - 1)^k. \tag{20}$$

This bound is an improvement over [27]. In their work they plug $d_\ell = d$ for $\ell = 1, \dots, L$ and obtain the bound $2^{\frac{\kappa(\kappa-1)}{2} dL}$, to be compared to κ^{dL} for (20).

GroupSort NNs. To bound the number of linear convex regions of a GroupSort NN [13] with the same group size g_s in each layer, we consider for each composition the arrangement of $d_{\ell+1}/g_s$ convex partitions (one per group) with $g_s!$ regions each and obtain that

$$\kappa_{\text{GroupSort}} \leq \prod_{\ell=1}^{L} \sum_{k=0}^{\min(d_1,\dots,d_\ell)} \binom{d/g_s}{k} (g_s! - 1)^k. \tag{21}$$

These bounds provide an intuition of the role of the hyperparameters of CPWL NNs in terms of expressiveness. For instance, the number of units in Maxout NNs plays a role in the bound that is analogous to that of the number of knots of the activation functions in deepspline NNs. However, these two architectures do not induce the same implementation complexity. To increase the activation complexity by one unit, Maxout requires the inclusion of an additional learnable multidimensional affine function, whereas deepspline simply requires the insertion of one more knot to a 1D CPWL function.

While it is tempting to compare architectures on the sole basis of their expressiveness, it can be very delicate to draw generic practical conclusions from this comparison. The final choice of an architecture is guided by a tradeoff between expressiveness, computation complexity, memory usage, and ability to learn over the functional space. For instance, an increase in the group size of a GroupSort activation function increases the expressiveness with no additional parameters, but usually small group sizes are favored to keep the computational impact limited.

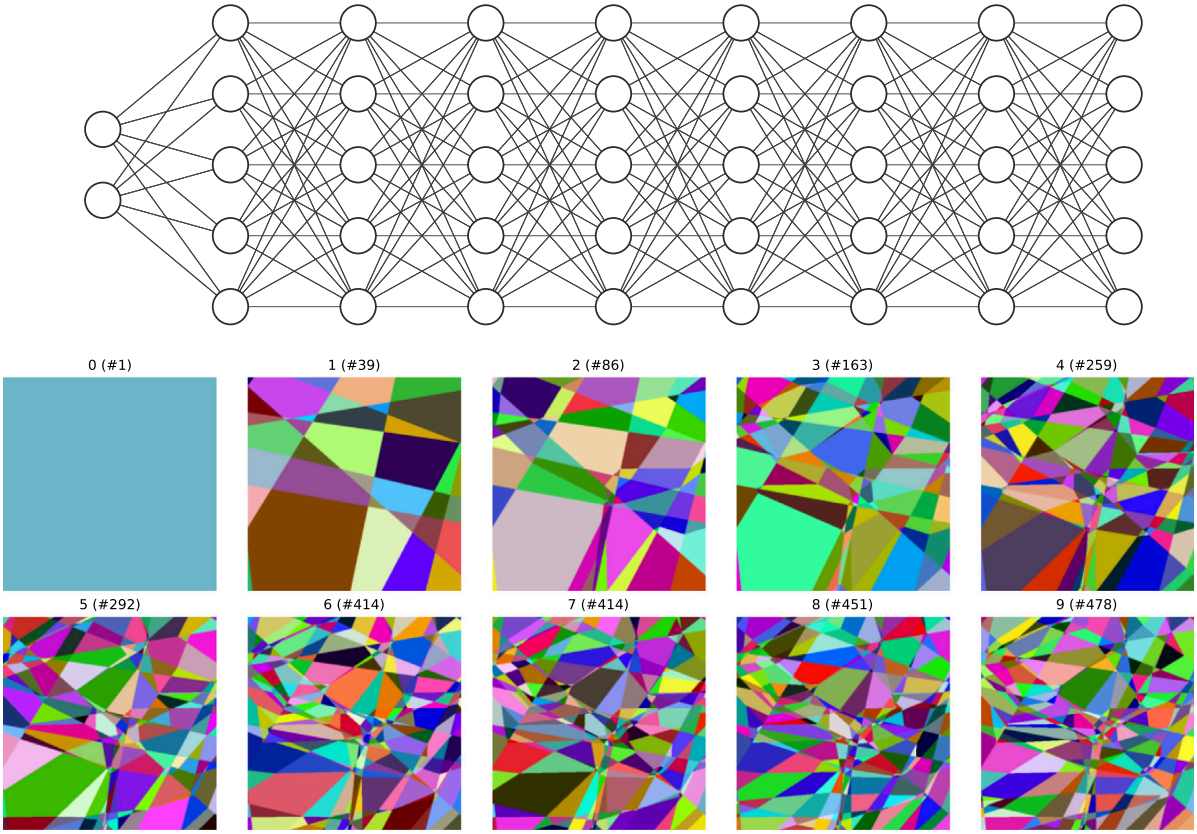

Fig. 7. Linear region-splitting process for a CPWL NN with absolute-value activation function and randomly generated parameters. The figure shows the linear regions of the mapping after k activation layers, for $k = 0, \dots, 9$. From one layer to the next, the regions are partitioned into smaller pieces. The number of linear regions is indicated in parentheses and suggests that the splitting process saturates with depth. The regions were numerically identified by evaluating the Jacobian of the mapping on a very fine grid.

4. Expected number of regions produced by CPWL NNs along 1D paths

In Section 3, we found that depth increases the expressiveness of the model exponentially when the corresponding metric is the maximal number of regions. However, the compositions that achieve the lower bound of Theorem 2 could be very specific and hard to reach in practice.

The composition $(\boldsymbol{f}_2 \circ \boldsymbol{f}_1)$ of two CPWL functions results in the partitioning of each linear region of $\boldsymbol{f}_1$ into smaller linear pieces. The successive compositions $(\boldsymbol{f}_\ell \circ \cdots \circ \boldsymbol{f}_1)$ have regions that are obtained from splitting of the regions of the previous compositions (Fig. 7). As such, we expect the image of each region of the composition to shrink when depth increases, at least for compositions with reasonable gradients magnitude (~ 1). The extent of the split should therefore depend on the depth of the composition. The more there are regions produced by the first compositions, the fewer splits each region will undergo after the next compositions. This intuition rules out an exponential growth of the average number of regions with ℓ. This effect has already been revealed for ReLU NNs in [25] and recently extended to Maxout NNs in [32]. We now aim to prove that it is universal to NNs with any type of CPWL activations under reasonable assumptions.

Throughout this section, we consider CPWL functions $\boldsymbol{f}_\theta$ parameterized by random parameters θ. We shall specify the parameterization and characteristics of the underlying stochastic model whenever needed. The natural extension of Section 3 is to estimate the expected number of regions of compositions of randomly generated CPWL functions. This task seems unfortunately very complex as it mixes stochasticity and combinatorial geometry. It would involve an overly heavy framework with the risk to lose focus on the high-level intuition. Instead, we propose a simpler but closely related metric: the expected density of regions along 1D paths. This quantity is valuable in practice since it gives the expected number of linear regions that are found in-between two locations of the input space that are 1 unit distance apart. In addition, the inverse of the density gives a rough measure of the average size of a linear region along one direction.

4.1. Knot density

The characterization of the density of linear regions of CPWL NNs along one-dimensional paths requires the introduction of some mathematical concepts.

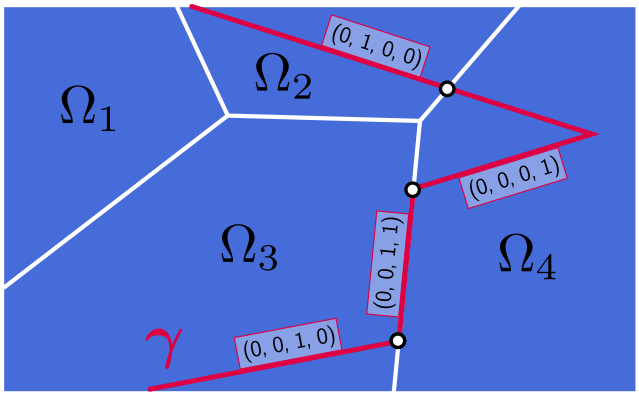

Fig. 8. Example of a 1D CPWL path $\gamma : \mathbb{R} \to \mathbb{R}^2$. The value of the characteristic function φ_f^γ along γ is given as a 4D vector and allow one to identify the 3 knots along γ.

1D CPWL path. 1D CPWL path denotes in the sequel any CPWL mapping $\gamma : S \to \mathbb{R}^d$ on a closed segment $S = [a, b]$ $(a, b \in \mathbb{R})$ with finitely many knots. This path will serve to "navigate" within the input domain of CPWL NNs for counting the linear regions. The length of γ is computed as $\mathrm{Len}(\gamma) := \int_{t\in S} \|\frac{\mathrm{d}\gamma}{\mathrm{d}t}\|_2 \mathrm{d}t$. Note that γ is a parameterization of what is often referred to as a polygonal chain.

In this Section we only study the density of linear regions along CPWL paths because of their simplicity and connections with CPWL NNs, e.g. the composition of a CPWL path and a CPWL NN is again a CPWL path. This choice is, however, not very restrictive since CPWL paths can approximate any continuous path arbitrarily close.

Knot density along a path. Given a 1D CPWL path, the goal is to characterize the complexity of a CPWL NN along it. Informally, the number of knots of a CPWL NN along the path is the number of times the path crosses regions. This intuitive definition is unfortunately not sufficiently precise since it does not specify how to count knots when some nonzero-length portion of the path γ is contained in a face of a linear region, see in Fig. 8 for an example. To avoid any ambiguity, we introduce the characteristic function

$$\varphi_f^\gamma : \mathbb{R} \to \mathbb{R}^K \tag{22}$$

$$t \mapsto (\mathbb{1}_{\Omega_1}(\gamma(t)), \dots, \mathbb{1}_{\Omega_K}(\gamma(t))) \tag{23}$$

of a CPWL f along γ, where the sets Ω_k are the projection regions of f and $\mathbb{1}_{\Omega_k}(\gamma(t)) = 1$ if $\gamma(t) \in \Omega_k$ and $\mathbb{1}_{\Omega_k}(\gamma(t)) = 0$ otherwise. Since γ is continuous with finitely many knots, and since the projection regions are unions of polyhedrons, φ_f^γ is a binary function with finitely many jumps, see Fig. 8. Note that φ_f^γ uniquely identifies the supporting affine function active at location $\gamma(t)$. Hence, in practice, the knowledge of $f(\gamma(t))$ and $\nabla f(\gamma(t))$, which is computable in any deep-learning library, suffice to identify $\varphi_f^\gamma(t)$.

Definition 6 (*Knot Density Along 1D CPWL Curves*)**.** Let $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a CPWL function, γ a 1D CPWL path, and φ_f^γ the characteristic function of f along γ. The number kt_f^γ of knots of f along γ is the number of discontinuous points of the piecewise-constant function φ_f^γ. The knot density λ_f^γ of f along γ is defined as

$$\lambda_f^\gamma = \mathrm{kt}_f^\gamma / \mathrm{Len}(\gamma), \tag{24}$$

where $\mathrm{Len}(\gamma)$ is the length of γ.

We stress that alternative definitions of the knot density that correspond to the same informal intuition are possible, but they would differ when the path γ follows the boundaries of some projection regions. In the sequel, this will not matter since, in any reasonable stochastic framework, the path does not follow some boundaries almost surely.

The knot density along a path is subadditive for the sum and vectorization of CPWL functions, and can be bounded for the composition of CPWL functions, see Propositions 3 and 4 and Appendix B for the corresponding proofs.

Proposition 3. *Let $\gamma : S \to \mathbb{R}^d$ be a 1D CPWL path on the segment $S \subset \mathbb{R}$ and let $f_1 : \mathbb{R}^d \to \mathbb{R}^{d'}$ and $f_2 : \mathbb{R}^d \to \mathbb{R}^{d'}$ be two CPWL functions. The knot density along γ of either the sum $f_1 + f_2$ or of the vectorized function (f_1, f_2) is bounded as*

$$\lambda_{f_1+f_2}^\gamma \le \lambda_{f_1}^\gamma + \lambda_{f_2}^\gamma, \tag{25}$$

$$\lambda_{(f_1,f_2)}^\gamma \le \lambda_{f_1}^\gamma + \lambda_{f_2}^\gamma, \tag{26}$$

where $\lambda_{f_1}^\gamma$ and $\lambda_{f_2}^\gamma$ are the knot density of f_1 and f_2 along γ, respectively.

Proposition 4. *Let $\gamma : S \to \mathbb{R}^{d_1}$ be a 1D CPWL path on $S \subset \mathbb{R}$ and let $f_1 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ and $f_2 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ be two CPWL functions. Then, the knot density of $f_2 \circ f_1$ on γ is bounded as*

$$\lambda^{\gamma}_{f_2 \circ f_1} \leq \lambda^{\gamma}_{f_1} + \left(\frac{\mathrm{Len}(f_1 \circ \gamma)}{\mathrm{Len}(\gamma)} \right) \lambda^{f_1 \circ \gamma}_{f_2}, \tag{27}$$

where λ^{γ}_1 is the knot density of f_1 along γ and λ^{γ}_2 the one of f_2 along $f_1 \circ \gamma$.

4.2. Knot density of CPWL layers

The goal of this subsection is to show that the knot density is well behaved for classical CPWL NN layers, which justifies the assumption (i) of Theorem 3 and Corollary 3. The proofs can be found in Appendix C.

Proposition 5 (*Knot Density - ReLU*)**.** *Let $(\boldsymbol{w}, b) \in \mathbb{R}^d \times \mathbb{R}$ be independent random variables with bounded probability density functions ρ_b for b and ρ_w for the components of $\boldsymbol{w}$, which are i.i.d. Then, the expected knot density of the ReLU CPWL component $\mathbf{x} \mapsto \mathrm{ReLU}(\boldsymbol{w}^T \mathbf{x} + b)$ along any 1D CPWL path γ is bounded as*

$$\mathbb{E}\left[\lambda^{\gamma}_{\mathrm{ReLU}}\right] \leq \sqrt{\mathbb{E}[w_1^2]} \sup_{t \in \mathbb{R}} \rho_b(t). \tag{28}$$

In particular, when b and the components of $\boldsymbol{w}$ are normally distributed with zero mean and standard deviation σ_b and σ_w, respectively, the following tighter bound holds true

$$\mathbb{E}\left[\lambda^{\gamma}_{\mathrm{ReLU}}\right] \leq \frac{\sigma_w}{\pi \sigma_b}. \tag{29}$$

When the ReLU activation function is replaced by a 1D CPWL function with a given number K of knots, we conjecture that the bounds can simply be multiplied by K.

Proposition 6 (*Knot Density - Maxout*)**.** *Let $((w_{k1}, \dots, w_{kd}), b_k) \in \mathbb{R}^d \times \mathbb{R}$ for $k = 1, \dots, K$ be independent random variables with bounded probability density functions ρ_b for any b_k and ρ_w for all components w_{kl} of $\boldsymbol{w}_k$, which are i.i.d. over both $k \in [K]$ and $l \in [d]$. Then, the expected knot density of the rank K Maxout unit $f : \mathbf{x} \mapsto \max_{k=1,\dots,K}(\boldsymbol{w}_k^T \mathbf{x} + b_k)$ along any 1D CPWL path γ is bounded as*

$$\mathbb{E}\left[\lambda^{\gamma}_{\mathrm{Maxout}}\right] \leq \sqrt{2} \binom{K}{2} \sigma_w \sup_{t \in \mathbb{R}} \rho_b(t), \tag{30}$$

where σ_w is the standard deviation of any w_{kl}. In particular, when b_k and w_{kl} are normally distributed with zero mean and standard deviation σ_b and σ_w, respectively, a tighter bound holds true, according to

$$\mathbb{E}\left[\lambda^{\gamma}_{\mathrm{Maxout}}\right] \leq \sqrt{2} \binom{K}{2} \frac{\sigma_w}{\pi \sigma_b}. \tag{31}$$

The bounds provided in Proposition 6 grow quadratically in terms of the number of Maxout units; we conjecture the existence of a tighter linear bound.

Proposition 7 (*Knot Density - Groupsort*)**.** *Let $(\boldsymbol{w}_k, b_k)$ be as in Proposition 6. Then, the expected knot density λ^{γ}_f of the GroupSort layer $f : \mathbb{R}^d \to \mathbb{R}^d : \mathbf{x} \mapsto \mathrm{GS}_{n_g, g_s}(\boldsymbol{W}\mathbf{x})$, where GS_{n_g, g_s} is the GroupSort activation with n_g groups of size g_s, is bounded along any 1D CPWL path γ as*

$$\mathbb{E}\left[\lambda^{\gamma}_{\mathrm{GroupSort}}\right] \leq \frac{\sqrt{2}}{2} d(g_s - 1) \sigma_w \sup_{t \in \mathbb{R}} \rho_b(t), \tag{32}$$

where σ_w is the standard deviation of any $w_{k,l}$. In particular, when b_k and w_{kl} are normally distributed with zero mean and standard deviation σ_b and σ_w, respectively, a tighter bound can be given as

$$\mathbb{E}\left[\lambda^{\gamma}_{\mathrm{GroupSort}}\right] \leq \frac{\sqrt{2}}{2} d(g_s - 1) \frac{\sigma_w}{\pi \sigma_b}. \tag{33}$$

For ReLU and Maxout layers with multidimensional outputs, the bounds given in Propositions 5 and 6 are simply multiplied by the output dimension (see Proposition 3). We note that all bounds proposed take the form $(\kappa W \sigma_w \sup_{t \in \mathbb{R}} \rho_b(t))$, where the prefactor κ only depends on the activation function and W is the number of outputs of the layer. The learnable parameters are typically initialized by sampling a uniform or normal distribution with the same characteristics for the biases and the weights of a same layer. In this case, although the characteristics of the distribution usually depend on the input and output dimensions of the layer [8], the quantity $\sigma_w \sup_{t \in \mathbb{R}} \rho_b(t)$ is determined only by the distribution: normal or uniform (since, for these distributions, the supremum of the probability density function is proportional to the standard deviation). All in all, it should be reminded that

- the expected knot density is well defined for learnable CPWL layers;
- with standard initialization methods, it is reasonable to assume that the expected knot density of the components of a CPWL layer depends neither on its width nor on the total depth of the NN (at least at initialization stage).

It is tempting to take advantage of the previous results to adjust the distributions of the weights and biases at initialization in the hope to increase the upper bound and, possibly, the knot density of a NN. The effect is, however, subtle: for instance, if one narrows the distribution of the biases, the bound increases as $\sup_{t\in\mathbb{R}} \rho_b(t)$ increases. While this may increase the average knot density at some specific locations, it will inevitably decrease it elsewhere.

4.3. Bounds on the expected knot density of CPWL NNs

In Theorem 3 and Corollary 3, we introduce two different settings to bound the expected knot density of CPWL NNs. Theorem 3 highlights the role played by the gradients of the composed layers: larger gradients allow for a more intense splitting process within the composition and should lead to a greater knot density. With Corollary 3, we propose a more practical analysis: given a learning task that dictates the input and output dimensions, how does the expected density of linear regions along 1D curves relate to the depth, width, and activation complexity of the CPWL NN? In accordance with the intuition given in Fig. 7, depth cannot provide exponentially more linear regions on average. This key result relies mainly on the assumption (ii), which is discussed in Section 4.3.1.

The directional derivative of the function f in the direction $\mathbf{u}$ is denoted by $D_{\mathbf{u}}\{f\}$, and the proofs of the results proposed in this section can be found in Appendix D.

Theorem 3. *Let $f_{\theta_1}, \dots, f_{\theta_L}$, with $f_{\theta_\ell} : \mathbb{R}^W \to \mathbb{R}^W$, be CPWL functions parameterized by the independent and identically distributed random variables $\theta_1, \dots, \theta_L$. Suppose that there exist $\lambda_0, D_0 \in \mathbb{R}$ such that*

(i) for any 1D CPWL path γ, $\mathbb{E}[\lambda^{\gamma}_{f_{\theta_\ell,k}}] \leq \lambda_0$, where $f_{\theta_\ell,k}$ is the kth component of f_{θ_ℓ} (bounded expected knot density of the components);

(ii) for any $\mathbf{x}, \mathbf{u} \in \mathbb{R}^W$ with $\|\mathbf{u}\|_2 = 1$, $\mathbb{E}[D_{\mathbf{u}}\{f_\theta\}(\mathbf{x})] \leq D_0$ (bounded expected directional derivative).

Then, on any 1D CPWL path γ, the expected knot density of the CPWL NN is bounded as

$$\mathbb{E}[\lambda^{\gamma}_{f_{\theta_L}\circ\cdots\circ f_{\theta_1}}] \leq \begin{cases} \lambda_0\ W \left(\frac{1-D_0^L}{1-D_0}\right), & D_0 \neq 1 \\ \lambda_0\ W L, & D_0 = 1. \end{cases} \tag{34}$$

Corollary 3. *Let $f_{\theta_1}, \dots, f_{\theta_L}$, with $f_{\theta_\ell} : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_{\ell+1}}$, be CPWL functions parameterized by the independent and identically distributed random variables $\theta_1, \dots, \theta_L$ and $d_2 = \cdots = d_L = W > d_{L+1}$. Suppose that there exist $\lambda_0, D_0 \in \mathbb{R}$ such that*

(i) for any 1D CPWL path γ, $\mathbb{E}[\lambda^{\gamma}_{f_{\theta_\ell,k}}] \leq \lambda_0$, where $f_{\theta_\ell,k}$ is the kth component of f_{θ_ℓ} (bounded expected knot density of the components),

(ii) for any $\mathbf{x}, \mathbf{u} \in \mathbb{R}^d$ with $\|\mathbf{u}\|_2 = 1$, $\mathbb{E}[D_{\mathbf{u}}\{f_{\theta_\ell} \circ \cdots \circ f_{\theta_1}\}(\mathbf{x})] \leq D_0$, for $1 \leq \ell \leq L$ (bounded expected directional derivative within the composition).

Then, on any 1D CPWL path γ, the expected knot density of the CPWL NN is bounded as

$$\mathbb{E}[\lambda^{\gamma}_{f_{\theta_L}\circ\cdots\circ f_{\theta_1}}] \leq D_0^*(\lambda_0\ W L), \tag{35}$$

where $D_0^* = \max(D_0, 1)$.

The proof of Theorem 3 relies on Lemma 4. In the bound presented in this lemma, the expected value is evaluated before taking the supremum, whilst a switch of the order of the operators would yield a much looser bound.

Lemma 4. *Let $\gamma : S \to \mathbb{R}^d$ be a 1D CPWL path and $f_\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$ a CPWL function parameterized by the random variable θ such that, for any $\mathbf{x}, \mathbf{u} \in \mathbb{R}^d$, f_θ is differentiable at $\mathbf{x}$ in direction $\mathbf{u}$ with probability 1. Then, the expected length of the 1D CPWL path $f_\theta \circ \gamma : S \to \mathbb{R}^{d'}$ is bounded as*

$$\mathbb{E}[\mathrm{Len}(f_\theta \circ \gamma)] \leq \mathrm{Len}(\gamma) \sup_{\substack{\mathbf{x},\mathbf{u}\in\mathbb{R}^d \\ \|\mathbf{u}\|_2=1}} \mathbb{E}[\|D_{\mathbf{u}}\{f_\theta\}(\mathbf{x})\|_2]. \tag{36}$$

4.3.1. Discussion of the compositional bounds

Our approach relies on the independence of the randomly generated CPWL functions. It usually holds at initialization stage, but it is not true anymore in the learning stage. While this can be regarded as a limitation, it is a legitimate and convenient way to explore and depict the whole function space that a given architecture gives access to.

Assumption (i) of Theorem 3 and its corollary (bounded expected knot density of the learnable CPWL components) have been discussed in details in Section 4.2, where it was remarked that it is reasonable to assume that λ_0 is independent of W and L.

Theorem 3 and Corollary 3 differ on Assumption (ii) (well behaved gradients). While the assumption of the theorem seems more natural at first sight (gradient controlled for each layer), the one of the corollary is closer to practical observations. Assumption (ii) of Corollary 3 was invoked to bound the expected length of the image of any finite-length 1D CPWL path, independently of the depth of the composition. While early works suggested that this expected length grows exponentially with depth [41], it was recently shown otherwise in a more realistic setup, both theoretically and experimentally [42]. For instance, for ReLU NNs, with the usual 2/fan-in weight variance, depth typically does not affect the expected length [42]. More generally, a control of the magnitude

of the directional derivatives that is independent of the depth is highly desirable in the learning stage for a stable back-propagation algorithm [8] and, in the inference stage, to produce robust models [43]. In short, it is also reasonable to assume that the parameter D_0 depends neither on W nor on L.

The previous discussion suggests a simple and important bound on the density of regions of CPWL NNs. It attributes an identical role to the three sources of complexity, namely depth, width, and activation complexity.

The quality of the proposed bounds seems to be completely determined by the tightness of the bounds in Assumptions (i) and (ii). Based on the proofs of Theorem 3 and Corollary 3, we believe that the compositional bounds are sharp provided that the expected knot density is uniform (i.e., the same for any 1D CPWL curve) and that the expected norm of the directional derivative is uniform and isotropic within the NN.

5. Conclusion

In this work, we have investigated the role of depth, width, and activation complexity in the expressiveness of CPWL NNs. By invoking results from combinatorial geometry, we have found that depth has a predominant role over width and activation complexity: it is the only descriptor able to increase the number of linear regions exponentially. However, this exponential growth is only observed for the maximal number of regions. Indeed, when exploring the whole function space produced by a given CPWL NN, we have found that, on average, the number of regions along a line is bounded by a quantity that only depends on the product of the three descriptors. In that perspective, the three complexity parameters have an identical role: no exponential behavior with depth is observed anymore.

The ability to train deeper and deeper NNs has led to major improvements in machine learning. However, depth comes at a price in applications where the NN needs to be stable, for instance by constraining its global Lipschitz constant. In such settings, we therefore believe that complex learnable activations should always be regarded as a valuable opportunity to increase substantially the expressiveness of the model without resorting to deeper NNs [44,45].

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors are thankful to Shayan Aziznejad for helpful discussions.

Appendix A. Proofs for Section 3

A.1. Number of convex vs. projection regions

Proof of Proposition 1. The first inequality follows from the fact that there cannot be fewer linear convex regions than affine pieces. Consider two neighboring projection regions Ω_k and Ω_p of $\boldsymbol{f}$, where $1 \leq k < p \leq \rho$, with corresponding affine pieces $\boldsymbol{f}^k : \mathbf{x} \mapsto \mathbf{W}_k^T\mathbf{x} + \mathbf{b}_k$ and $\boldsymbol{f}^p : \mathbf{x} \mapsto \mathbf{W}_p^T\mathbf{x} + \mathbf{b}_p$, where $\mathbf{W}_k, \mathbf{W}_p \in \mathbb{R}^{d' \times d}$ and $\mathbf{b}_k, \mathbf{b}_p \in \mathbb{R}^{d'}$. Since $\boldsymbol{f}$ is continuous, any $\mathbf{x} \in \Omega_k \cap \Omega_p$ satisfies that $(\mathbf{W}_k - \mathbf{W}_p)^T\mathbf{x} + (\mathbf{b}_k - \mathbf{b}_p) = 0$. The set of all boundary points of $\boldsymbol{f}$ is therefore included in $\cup_{1\leq k<p\leq\rho} H_{kp}$, where $H_{kp} = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{W}_k - \mathbf{W}_p)^T\mathbf{x} + (\mathbf{b}_k - \mathbf{b}_p) = 0\}$ is an affine subspace of dimension at most $(d-1)$ since $k \neq p$. The arrangement of the $\rho(\rho-1)/2$ hyperplanes H_{kq} with $k \neq p \in [\rho]$ yields convex regions on which $\boldsymbol{f}$ is affine since these regions do not contain boundary points. The number of such regions is, therefore, an upper bound on the number of convex regions of $\boldsymbol{f}$. It is known from [37] that the number of convex regions formed by an arrangement of N hyperplanes in $\mathbb{R}^d$ is at most $\sum_{k=0}^{\min(d,N)} \binom{N}{k}$. Hence, for $\rho(\rho-1)/2 > d$, we directly reach the announced result. Otherwise, the bound yields $\sum_{k=0}^{\rho(\rho-1)/2} \binom{\rho(\rho-1)/2}{k} = 2^{\rho(\rho-1)/2}$. □

Proof of Lemma 1. Let $e = \dim(E)$. The natural candidate for Π_E is the partition

$$\Pi' = \left\{P' : P' = P \cap E, P \in \Pi, \text{ and } \operatorname{Int} P' \neq \emptyset\right\}, \tag{A.1}$$

which is unfortunately not necessarily a proper convex partition. Indeed, if E contains an e-face of a region, then some elements of Π' will not have disjoint interiors. Since the regions of Π are polyhedrons, there exist a given number n_H of distinct boundary hyperplanes $H_p = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{a}_p^T\mathbf{x} + b_p = 0\}$ and such that for each $P_k \in \Pi$, there exists a subset $I_k \subset [n_H]$ and $\epsilon_{k,p} \in \{-1, 1\}$ for $p \in I_k$ such that

$$P_k = \{\mathbf{x} \in \mathbb{R}^d : \epsilon_{k,p}(\mathbf{a}_p^T\mathbf{x} + b_p) \geq 0 \quad \forall p \in I_k\}. \tag{A.2}$$

We now consider a mapping ϕ that assigns to each hyperplane H_p a unique region $\phi(p)$ such that $p \in I_{\phi(p)}$. We can now define n new pairwise-disjoint convex regions as

$$P'_k = \left\{ \mathbf{x} \in \mathbb{R}^d : \begin{array}{ll} \epsilon_{k,p}(\mathbf{a}_p^T\mathbf{x} + b_p) \geq 0 & \text{for } p \in [n_k] \text{ and } \phi(p) = k \\ \epsilon_{k,p}(\mathbf{a}_p^T\mathbf{x} + b_p) > 0 & \text{for } p \in [n_k] \text{ and } \phi(p) \neq k \end{array} \right\}. \tag{A.3}$$

It is clear that $\cup_{k=1}^{n} P'_k = \mathbb{R}^d$. From these new regions, one can eventually build the proper convex partition

$$\Pi_E = \left\{ P_E = \overline{P'_k \cap E} : k \in [n] \text{ and } \operatorname{Int} P_E \neq \emptyset \right\}. \tag{A.4}$$

By construction, all regions of Π_E are closed with nonempty interiors; their union covers E. Let $P_{E,1} = P'_{k_1} \cap E$ and $P_{E,2} = P'_{k_2} \cap E$ be two (nonempty) regions of Π_E. We have that $\operatorname{Int}(P_{E,1}) \cap \operatorname{Int}(P_{E,2}) = \operatorname{Int}(P_{E,1} \cap P_{E,2}) = \operatorname{Int}(P'_{k_1} \cap P'_{k_2} \cap E) = \emptyset$ for $k_1 \neq k_2$. We, therefore, proved that Π_E is a convex partition of E; it has at most n regions and is such that, for any $P_E \in \Pi_E$, there is $P \in \Pi$ with $P_E \subset P$. □

Proof of Lemma 2. Let $P \in \Pi$. Recall that P is a closed and convex subset of the affine space $f(\mathbb{R}^d)$ with dimension $\dim(f(\mathbb{R}^d))$. We first prove that $f^{-1}(P)$ meets the requirements to form a convex partition of $\mathbb{R}^d$.

- The continuity of f implies that $f^{-1}(P)$ is closed.
- The function f is written as $f : \mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{d' \times d}$ and $\mathbf{b} \in \mathbb{R}^{d'}$. For $\mathbf{x}, \mathbf{y} \in f^{-1}(P)$ and $\beta \in [0,1]$, we have that $f(\beta\mathbf{x} + (1-\beta)\mathbf{y}) = \mathbf{A}(\beta\mathbf{x} + (1-\beta)\mathbf{y}) + \mathbf{b} = \beta f(\mathbf{x}) + (1-\beta) f(\mathbf{y}) \in P$ since P is convex. Therefore, $f^{-1}(P)$ is also convex.
- We have that $\cup_{P \in \Pi} f^{-1}(P) = f^{-1}(\cup_{P \in \Pi} P) = f^{-1}(f(\mathbb{R}^d)) = \mathbb{R}^d$.
- For two distinct regions $P_1, P_2 \in \Pi$, we have that $f^{-1}(P_1) \cap f^{-1}(P_2) = f^{-1}(P_1 \cap P_2)$. Since P_1 and P_2 are distinct regions of Π, $\dim(P_1 \cap P_2) < \dim(f(\mathbb{R}^d))$, which implies that $\dim(f^{-1}(P_1) \cap f^{-1}(P_2)) < d$ and proves that P_1 and P_2 have disjoint interiors.
- We decompose the input space as the direct sum $\mathbb{R}^d = \ker(\mathbf{A}) \oplus U$. Note that $f(U) = f(\mathbb{R}^d)$. It is clear that, for any $\mathbf{x} \in f^{-1}(P)$ and $\mathbf{y} \in \ker(\mathbf{A})$, we have that $\mathbf{x} + \mathbf{y} \in f^{-1}(P)$, which implies that $\dim(\operatorname{Proj}_U(f^{-1}(P))) = \ker(\mathbf{A})$. In addition, we use the fact that f restricted to U has full rank and write $\dim(\operatorname{Proj}_{\ker(\mathbf{A})}(f^{-1}(P))) = \dim(P) = \dim(f(\mathbb{R}^d))$. All in all, we have proved that $\dim(f^{-1}(P)) = d$, which implies that the regions of $f^{-1}(\Pi)$ have nonempty interiors. □

Proof of Lemma 3. The result stems from the fact that the rank of a product of matrices is bounded by the smallest rank of these matrices. □

A.2. Upper bound on the number of regions of arrangements

Proof of Theorem 1. First, we prove that the expression given in the theorem is an upper bound. To that end, we need to formalize our problem with the notion of abstract simplicial complex so as to focus solely on the combinatorial structure of the task and be compliant with the formalism of [38]. Let $\Pi_k^* = \{\operatorname{int}(P) : P \in \Pi_k\}$, where $\operatorname{int}(P)$ denotes the interior of P in $\mathbb{R}^d$, and let $\mathcal{F} = \cup_{k=1}^{N} \{P^* : P^* \in \Pi_k^*\}$ be the set that contains the elements of the N sets Π_k^*. The nerve K of $\mathcal{F}$ is defined as

$$\mathrm{K} = \{X \subset \mathcal{F} : \cap X \neq \emptyset\}. \tag{A.5}$$

In simple words, K is made of all the nonempty intersections of sets in any of the Π_k^*. The nerve of an open covering is an abstract simplicial complex which, therefore, applies to K since $\mathcal{F}$ is an open covering of $\mathbb{R}^d$. This more simply follows from the definition of an abstract simplicial complex: it is a family of sets that is closed under taking subsets. In the sequel, we need K to be a d-representable simplicial complex, which is granted because it is the nerve of a finite family of convex sets in $\mathbb{R}^d$ (more details in [38]). In our problem, the faces of dimension 0 of the complex, also known as vertices, are the elements of $\mathcal{F}$. More generally, a face of K of dimension p is a nonempty intersection of $p+1$ elements of $\mathcal{F}$. Each set Π_k^* induces a sub-complex $\mathrm{K}[\Pi_k^*] = \{X \subset \Pi_k^* : \cap X \neq \emptyset\}$ of K. The dimension of this sub-complex, which is the largest dimension of its faces, is 0 because the elements of Π_k^* are disjoint. We note that the interior of the regions of the arrangement of the convex partitions are $(N+1)$-faces of the abstract simplicial complex K, which are also called $\mathbf{1}$-colorful faces, where $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^N$ specifies that each region of the arrangement is built from one region per partition. We are therefore looking to bound the number $f_{\mathbf{1}}(\mathrm{K})$ of $\mathbf{1}$-colorful faces of the complex K. Since we have now fully translated our problem into the framework of [38], we can apply [38, Theorem 10] to $\mathcal{F}$. The parameter $\mathbf{r} = (r_1, \dots, r_N)$ can be chosen so that $\dim(\mathrm{K}[\Pi_k^*]) \leq (r_k - 1)$. Therefore, we simply choose $\mathbf{r} = \mathbf{1}$ and obtain that

$$\beta^d(n_1, \dots, n_N) = f_{\mathbf{1}}(\mathrm{K}) \leq p_{\mathbf{1}}(\mathbf{n}, d, \mathbf{1}), \tag{A.6}$$

where

$$p_{\mathbf{k}}(\mathbf{n}, d, \mathbf{r}) = \sum_{\ell = (\ell_1, \dots, \ell_N) \in L_{\mathbf{k}}(d)} \prod_{i=1}^{N} \binom{n_i - r_i}{\ell_i} \binom{r_i}{k_i - \ell_i} \tag{A.7}$$

and

$$L_{\mathbf{k}}(d) = \{\ell = (\ell_1, \dots \ell_N) \in \mathbb{N}^N : \ell_1 + \cdots + \ell_N \leq d \text{ and } \ell_i \leq k_i \text{ for } i \in [N]\}. \tag{A.8}$$

In our problem, $\mathbf{k} = \mathbf{1}$ and

$$L_{\mathbf{1}}(d) = \{\ell = (\ell_1, \dots \ell_N) \in \mathbb{N}^N : \ell_1 + \cdots + \ell_N \leq d \text{ and } \ell_i \in \{0, 1\} \text{ for } i \in [N]\}. \tag{A.9}$$

With $\mathbf{r} = \mathbf{1}$, we have that

$$\beta^d(n_1, \dots, n_N) \leq p_{\mathbf{1}}(\mathbf{n}, d, \mathbf{1}) = \sum_{\ell = (\ell_1, \dots, \ell_N) \in L_{\mathbf{1}}(d)} \prod_{i=1}^{N} \binom{n_i - 1}{\ell_i} \binom{1}{1 - \ell_i}$$

$$
\begin{aligned}
&= \sum_{\ell=(\ell_1,\dots,\ell_N)\in L_{\mathbf{1}}(d)} \prod_{i=1}^{N} \binom{n_i-1}{\ell_i} \\
&= \sum_{k=0}^{d} \sum_{\substack{\ell_1,\dots,\ell_N\in\{0,1\} \\ \ell_1+\cdots+\ell_N=k}} \prod_{i=1}^{N} \binom{n_i-1}{\ell_i} \\
&= \sum_{k=0}^{d} \sum_{\substack{\ell_1,\dots,\ell_N\in\{0,1\} \\ \ell_1+\cdots+\ell_N=k}} \prod_{\substack{i=1 \\ \ell_i=1}}^{N} \binom{n_i-1}{\ell_i} \prod_{\substack{i=1 \\ \ell_i=0}}^{N} \binom{n_i-r_i}{\ell_i} \\
&= \sum_{k=0}^{d} \sum_{\substack{\ell_1,\dots,\ell_N\in\{0,1\} \\ \ell_1+\cdots+\ell_N=k}} \prod_{\substack{i=1 \\ \ell_i=1}}^{N} (n_i-1) \\
&= 1+\sum_{k=1}^{d} \sum_{1\le\ell_1<\cdots<\ell_k\le N} \prod_{i=1}^{k} (n_{\ell_i}-1),
\end{aligned} \tag{A.10}
$$

which proves that the bound given in the Theorem holds true.

Now we show that this upper bound is sharp. To that end, consider that each partition Π_k is made of the regions of the arrangement of the (n_k-1) distinct parallel hyperplanes H_q^k for $q=1,\dots,(n_k-1)$ so that the hyperplanes are in general position when only one per partition is selected. Recall that N hyperplanes are in general position if any collection of k of them intersect in a $(d-k)$-dimensional plane for $1\le k\le d$ and have empty intersection for $k>d$. The number of regions of the arrangement $\mathcal{A}(\Pi_1,\dots,\Pi_N)$ is exactly the number of regions of the arrangement of all the hyperplanes H_q^k for $q=1,\dots,(n_k-1)$ and $k=1,\dots,N$. Following Zavalasky's Theorem, the number of regions can be computed by

$$
\#\mathcal{R}(\mathcal{A}) = (-1)^d \chi_{\mathcal{A}}(-1), \tag{A.11}
$$

where $\chi_{\mathcal{A}}$ is the characteristic polynomial of the arrangement. There is no need here to define the characteristic polynomial in detail since Whitney's formula provides a direct way to evaluate it as

$$
\chi_{\mathcal{A}}(-1) = \sum_{\substack{\mathcal{B}\subset\mathcal{A} \\ \cap_{H\in\mathcal{B}}H\neq\emptyset}} (-1)^{\#\mathcal{B}} (-1)^{\dim(\cap_{H\in\mathcal{B}}H)}. \tag{A.12}
$$

The subsets $\mathcal{B}\subset\mathcal{A}$ that have a nonempty intersection can be written as $\mathcal{B}=\{H_{q_{k_1}}^{k_1},\dots,H_{q_{k_p}}^{k_p}\}$ with $1\le k_1<\cdots<k_p\le N$, $q_{k_i}\in[n_{k_i}-1]$ where $i=1,\dots,p$ and $0\le p\le d$. This holds because, for $q\neq q'$, $H_q^k\cap H_{q'}^k=\emptyset$. Note that, by convention, the set $\mathcal{B}=\emptyset$ is also considered in the sum. Because of the particular choice of the hyperplanes, for a given p, $\mathcal{B}$ is the nonempty intersection of p hyperplanes and there are $\sum_{1\le\ell_1<\cdots<\ell_p\le N}\prod_{i=1}^{p}(n_{\ell_i}-1)$ such subsets of $\mathcal{A}$. The intersection of the elements of $\mathcal{B}$ has dimension $(d-p)$ (recall that the hyperplanes of $\mathcal{B}$ are in general position). All in all, we have that

$$
\begin{aligned}
\#\mathcal{R}(\mathcal{A}) &= (-1)^d \sum_{\substack{\mathcal{B}\subset\mathcal{A}: \\ \cap_{H\in\mathcal{B}}H\neq\emptyset}} (-1)^{\#\mathcal{B}} (-1)^{\dim(\cap_{H\in\mathcal{B}}H)} \\
&= 1+(-1)^d \sum_{k=1}^{d} \sum_{1\le\ell_1<\cdots<\ell_k\le N} (-1)^k (-1)^{d-k} \prod_{i=1}^{k} (n_{\ell_i}-1) \\
&= 1+\sum_{k=1}^{d} \sum_{1\le\ell_1<\cdots<\ell_k\le N} \prod_{i=1}^{k} (n_{\ell_i}-1),
\end{aligned} \tag{A.13}
$$

which is the upper bound given in the theorem.

When $N\le d$, we readily check that the bound is giving $n_1\cdots n_N$. To prove the second additional bound for $N>d$ given in the theorem, we invoke the binomial theorem and remark that

$$
\begin{aligned}
\left(1+\sum_{p=1}^{N}(n_p-1)\right)^d &= 1+\sum_{k=1}^{d}\binom{d}{k}\left(\sum_{p=1}^{N}(n_p-1)\right)^k = 1+\sum_{k=1}^{d}\binom{d}{k}\sum_{1\le\ell_1,\dots,\ell_k\le N}\prod_{i=1}^{k}(n_{\ell_i}-1) \\
&\ge 1+\sum_{k=1}^{d}\sum_{1\le\ell_1,\dots,\ell_k\le N}\prod_{i=1}^{k}(n_{\ell_i}-1) \\
&\ge 1+\sum_{k=1}^{d}\sum_{1\le\ell_1<\cdots<\ell_k\le N}\prod_{i=1}^{k}(n_{\ell_i}-1). \quad \square
\end{aligned} \tag{A.14}
$$

A.3. Sum and vectorization

Proof of Proposition 2. Let Π_k be a linear convex partition of f_k for $k=1,\dots,N$. On each region of the arrangement $\mathcal{A}(\Pi_1,\dots,\Pi_N)$, the f_k are affine, and so is their sum and their vectorization. This implies that $\mathcal{A}(\Pi_1,\dots,\Pi_N)$ is a linear convex partition of both

the sum and the vectorization of the scalar-valued CPWL functions, which shows that $\beta^d(\kappa_1, \dots, \kappa_N)$ is a valid upper bound on the number of convex linear regions.

We now prove that the bounds are sharp. First, consider N convex partitions Π_k where each Π_k is made of the regions of the arrangement of $(\kappa_k - 1)$ distinct parallel hyperplanes $H_k^p = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}_k^T \mathbf{x} = b_k^p\}$, $p = 1, \dots, (\kappa_k - 1)$, and such that the hyperplanes are in general position when only one per partition is selected. In such a way, the arrangement $\mathcal{A}(\Pi_1, \dots, \Pi_N)$ has exactly $\beta^d(\kappa_1, \dots, \kappa_N)$ convex regions (see proof of Theorem 1). Second, for each partition, we consider a CPWL function $\varphi_k : \mathbb{R} \to \mathbb{R}$ with knots $(b_k^p)_{p=1}^{\kappa_k - 1}$ and κ_k distinct affine pieces $(\varphi_k^p)_{p=1}^{\kappa_k}$. In the sequel, the affine pieces are written $\varphi_k^p : x \to a_k^p x + c_k^p$. The function $f_k : \mathbf{x} \mapsto \varphi(\mathbf{w}_k^T \mathbf{x})$ has exactly n_k linear convex regions and Π_k is a linear convex partition of it. The construction implies that $\mathcal{A}(\Pi_1, \dots, \Pi_N)$ is a linear convex partition of both $(f_1 + \cdots + f_N)$ and $(f_1, \dots, f_N)$. Because the affine pieces of each φ_k are distinct, the vector-valued function $(f_1, \dots, f_N)$ will agree with distinct affine pieces on each region of $\mathcal{A}(\Pi_1, \dots, \Pi_N)$, which proves that this partition has the minimal number of linear convex regions. This yields CPWL functions such that $\kappa_{(f_1, \dots, f_N)} = \beta^d(\kappa_1, \dots, \kappa_N)$. On the contrary, in the case of the sum $(f_1 + \cdots + f_N)$, there is on the contrary no guarantee that $\mathcal{A}(\Pi_1, \dots, \Pi_N)$ is a partition with the minimal number of linear convex regions. To ensure that the regions of this partition have different affine pieces, it is sufficient to choose the pieces (φ_k^p) such that $\sum_{k=1}^N \varphi_k^{p_k^1}(\mathbf{w}_k^T \cdot) \neq \sum_{k=1}^N \varphi_k^{p_k^2}(\mathbf{w}_k^T \cdot)$ for any $1 \leq p_k^1, p_k^2 \leq \kappa_k$ and $(p_1^1, \dots, p_N^1) \neq (p_1^2, \dots, p_N^2)$. An explicit choice is $a_k^p = pm^{k-1}$ with $m = \max(\kappa_k)$. The biases b_k^p are then set such that φ_k^p is continuous. In such a way, the slope of $\sum_{k=1}^N \varphi_k^{p_k}(\mathbf{w}_k^T \cdot)$ is $\sum_{k=1}^N p_k m^k$. This number can be represented in base m as "$(p_N \cdots p_1)_m$", which shows that it is uniquely related to the choice of indices (p_k). Although this choice seems very specific, a random choice of the slopes would also satisfy the condition almost surely. We have therefore found a collection of CPWL functions whose sum has exactly $\beta^d(\kappa_1, \dots, \kappa_N)$ linear convex regions. □

A.4. Compositional bounds

Proof of Theorem 2. We use the notation $m_\ell = \min(d_1, \dots, d_\ell)$ and $\boldsymbol{F}_\ell = \boldsymbol{f}_\ell \circ \cdots \circ \boldsymbol{f}_1$. First, we prove by induction the validity of the proposed upper bound. The initial step is given by Proposition 2. Now suppose that the result holds for $\boldsymbol{F}_{\ell-1}$ with $\ell - 1 > 0$. Let Ω be a linear convex region of $\boldsymbol{F}_{\ell-1}$ and let $\boldsymbol{g}_\Omega$ be the corresponding affine function. The affine space $\boldsymbol{g}_\Omega(\mathbb{R}^{d_1}) \subset \mathbb{R}^{d_\ell}$ is of dimension at most $\min(d_1, \dots, d_\ell)$ (Lemma 3). Each linear convex partition $\Pi_{\ell,k}$ of the components of $\boldsymbol{f}_\ell$ yields a convex partition $\Pi'_{\ell,k}$ of the affine subspace $\boldsymbol{g}_\Omega(\mathbb{R}^{d_1})$ with no more than $\kappa_{\ell,k}$ regions on which $f_{\ell,k}$ is affine (Lemma 1). The arrangement of the partitions $\Pi'_{\ell,1}, \dots, \Pi'_{\ell,d_{\ell+1}}$ results in a convex partition of $\boldsymbol{g}_\Omega(\mathbb{R}^{d_1})$ with no more than $\beta^{m_\ell}(\kappa_{\ell,1}, \dots, \kappa_{\ell,d_\ell+1})$ regions (Theorem 1). Lemma 2 shows that $\boldsymbol{g}_\Omega^{-1}(\mathcal{A}(\Pi'_{\ell,1}, \dots, \Pi'_{\ell,d_{\ell+1}}))$ is a convex partition of $\mathbb{R}^{d_1}$ with $\boldsymbol{F}_\ell$ affine on each of its sets. In short, each linear convex region of $\boldsymbol{F}_{\ell-1}$ is partitioned into no more than $\beta^{m_\ell}(\kappa_{\ell,1}, \dots, \kappa_{\ell,d_{\ell+1}})$ linear convex regions, which concludes the first part of the proof.

Second, we propose a construction inspired from [6] to derive the lower bound given on the maximal number of regions. Let the sawtooth function sw_p of order p be the unique 1D CPWL function with knots located at k/p for $k = 1, \dots, (p-1)$ that satisfies $\mathrm{sw}_p(k/p) = \frac{1}{2}(1 - (-1)^k)$ for $k = 0, \dots, p$. The key properties of the sawtooth function of order p that will prove useful in the sequel are

- it has p projection regions that are also convex linear regions;
- it can be decomposed as

$$\mathrm{sw}_p : x \mapsto \sum_{k=1}^{p} \varphi_{k,p}, \tag{A.15}$$

  where $\varphi_{k,p} = p(x + 2(-1)^p |x - k/p|)$ is a CPWL function with 2 projection regions;
- the composition of sawtooth functions is a sawtooth function whose order is the product of the orders of the composed functions, as in

$$\mathrm{sw}_p \circ \mathrm{sw}_q = \mathrm{sw}_{pq}, \tag{A.16}$$

  for $p, q \in \mathbb{N}$.

The strategy is now to build a CPWL NN which mimics a given NN with independent sawtooth components. Let $\mathbf{e}_{d,k}$ be the kth element of the canonical basis of $\mathbb{R}^d$, $d^* = \min(d_\ell)$ and $\tau_\ell : \{1, \dots, d_{\ell+1}\} \to \{1, \dots, d^*\}$ for $\ell = 1, \dots, L$. Consider the dimension-reduction linear operator $\boldsymbol{u}_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d^*}$ associated to $\tau_{\ell-1}$, which is defined on the canonical basis by

$$\boldsymbol{u}_\ell : \mathbf{e}_{d_\ell,k} \mapsto \mathbf{e}_{d^*,\tau_{\ell-1}(k)}, \tag{A.17}$$

for $\ell = 2, \dots, L$ and

$$\boldsymbol{u}_1(\mathbf{e}_{d_\ell,k}) = \begin{cases} e_{d^*,k}, & k \leq d^* \\ 0, & \text{otherwise.} \end{cases} \tag{A.18}$$

Similarly, let $\boldsymbol{v}_\ell : \mathbb{R}^{d^*} \to \mathbb{R}^{d_{\ell+1}}$ be the dimension-augmentation linear operator

$$\boldsymbol{v}_\ell : \mathbf{e}_{d^*,k} \mapsto \sum_{q \in \tau_\ell^{-1}(\{k\})} \mathbf{e}_{d_{\ell+1},q}, \tag{A.19}$$

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

We now define the nonlinear pointwise function $\boldsymbol{\phi}_\ell : \mathbb{R}^{d_{\ell+1}} \to \mathbb{R}^{d_{\ell+1}}$. For $r \in \{1, \dots, d^*\}$ let $p_{\ell,r} = \sum_{k \in \tau_\ell^{-1}(\{r\})} \kappa_{\ell,k}$ and $\{J_{\ell,\tau_\ell,i}\}_{i=1}^{|\tau_\ell^{-1}(\{r\})|}$ be a partition of the set $\{1, \dots, p_{\ell,r}\}$, where the cardinality of the subsets is in one-to-one correspondence with $\{\kappa_{\ell,q}\}_{q \in \tau_\ell^{-1}(\{r\})}$. In this way, we assign to each $k \in \{1, \dots, d_\ell + 1\}$ a set of indices J_{ℓ,τ_ℓ,i_k} that allows us to define the kth component of $\boldsymbol{\phi}_\ell$ as

$$\phi_{\ell,k} = \sum_{j \in J_{\ell,\tau_\ell,i_k}} \varphi_{|\tau_\ell^{-1}(\{\tau_\ell(k)\})|,j}. \tag{A.20}$$

This ensures that

$$\sum_{k \in \tau_\ell^{-1}(\{r\})} \phi_{\ell,k} = \mathrm{sw}_{p_{\ell,r}}. \tag{A.21}$$

From the pointwise property of $\boldsymbol{\phi}$, we deduce that, for any $t_1, \dots, t_{d^*} \in \mathbb{R}$,

$$\begin{aligned}
(\boldsymbol{u}_{\ell+1} \circ \boldsymbol{\phi} \circ \boldsymbol{v}_\ell)\left(\sum_{r=1}^{d^*} t_r \mathbf{e}_{d^*,r}\right) &= (\boldsymbol{u}_{\ell+1} \circ \boldsymbol{\phi})\left(\sum_{r=1}^{d^*} \sum_{k \in \tau_\ell^{-1}(\{r\})} t_r \mathbf{e}_{d_\ell,k}\right) \\
&= \boldsymbol{u}_{\ell+1}\left(\sum_{r=1}^{d^*} \sum_{k \in \tau_\ell^{-1}(\{r\})} \phi_{\ell,k}(t_r) \mathbf{e}_{d_\ell,k}\right) \\
&= \sum_{r=1}^{d^*} \sum_{k \in \tau_\ell^{-1}(\{r\})} \phi_{\ell,k}(t_r) \boldsymbol{u}_{\ell+1}(\mathbf{e}_{d_\ell,k}) \\
&= \sum_{r=1}^{d^*} \sum_{k \in \tau_\ell^{-1}(\{r\})} \phi_{\ell,k}(t_r) \mathbf{e}_{d^*,\tau_\ell(k)} \\
&= \sum_{r=1}^{d^*} \left(\sum_{k \in \tau_\ell^{-1}(\{r\})} \phi_{\ell,k}(t_r)\right) \mathbf{e}_{d^*,r} \\
&= \sum_{r=1}^{d^*} \mathrm{sw}_{p_{\ell,r}}(t_r) \mathbf{e}_{d^*,r},
\end{aligned} \tag{A.22}$$

which means that $\boldsymbol{u}_{\ell+1} \circ \boldsymbol{\phi} \circ \boldsymbol{v}_\ell$ is a pointwise multivariate function with 1D sawtooth components of order $p_{\ell,r}$ for $r = 1, \dots, d^*$. We denote it by $\mathbf{sw}_{\mathbf{p}_\ell}$ with $\mathbf{p}_\ell = (p_{\ell,1}, \dots, p_{\ell,d^*})$. The function $\boldsymbol{f}_\ell$ of the NN is chosen to be $\boldsymbol{f}_\ell = \boldsymbol{\phi}_\ell \circ \boldsymbol{v}_\ell \circ \boldsymbol{u}_\ell$. Each component $f_{\ell,k}$ can be written in the form of $f_{\ell,k} : \mathbf{x} \mapsto \phi_{\ell,k}(\mathbf{w}_{\ell,k}^T \mathbf{x})$ with $\mathbf{w}_{\ell,k} = \sum_{q \in \tau_{\ell-1}^{-1}(\{\tau_\ell(k)\})} \mathbf{e}_{d_\ell,q}$. This shows that $f_{\ell,k}$ has the same number of projection regions as $\phi_{\ell,k}$ ($\kappa_{\ell,k}$) whenever $\mathbf{w}_{\ell,k} \neq \mathbf{0}$.

All in all, we have that

$$\begin{aligned}
\boldsymbol{f}_L \circ \boldsymbol{f}_{L-1} \circ \cdots \circ \boldsymbol{f}_2 \circ \boldsymbol{f}_1 &= (\boldsymbol{\phi}_L \circ \boldsymbol{v}_L \circ \boldsymbol{u}_L) \circ (\boldsymbol{\phi}_{L-1} \circ \boldsymbol{v}_{L-1} \circ \boldsymbol{u}_{L-1}) \circ \cdots \circ (\boldsymbol{\phi}_2 \circ \boldsymbol{v}_2 \circ \boldsymbol{u}_2) \circ (\boldsymbol{\phi}_1 \circ \boldsymbol{v}_1 \circ \boldsymbol{u}_1) \\
&= \boldsymbol{\phi}_L \circ \boldsymbol{v}_L \circ (\boldsymbol{u}_L \circ \boldsymbol{\phi}_{L-1} \circ \boldsymbol{v}_{L-1}) \circ (\boldsymbol{u}_{L-1} \circ \cdots \circ \boldsymbol{\phi}_2 \circ \boldsymbol{v}_2) \circ (\boldsymbol{u}_2 \circ \boldsymbol{\phi}_1 \circ \boldsymbol{v}_1) \circ \boldsymbol{u}_1 \\
&= \boldsymbol{\phi}_L \circ \boldsymbol{v}_L \circ \mathbf{sw}_{\mathbf{p}_{L-1}} \circ \cdots \circ \mathbf{sw}_{\mathbf{p}_1} \circ \boldsymbol{u}_1.
\end{aligned} \tag{A.23}$$

We now note that there are no fewer projection regions of $\boldsymbol{f}_L \circ \boldsymbol{f}_{L-1} \circ \cdots \circ \boldsymbol{f}_2 \circ \boldsymbol{f}_1$ than the number of projection regions of $\boldsymbol{h} = \boldsymbol{u}_{L+1} \circ \boldsymbol{f}_L \circ \cdots \circ \boldsymbol{f}_1$ because $\boldsymbol{u}_{L+1}$ is a linear mapping. In addition,

$$\begin{aligned}
\boldsymbol{u}_{L+1} \circ \boldsymbol{f}_L \circ \cdots \circ \boldsymbol{f}_1 &= \mathbf{sw}_{\mathbf{p}_L} \circ \cdots \circ \mathbf{sw}_{\mathbf{p}_1} \circ \boldsymbol{u}_1 \\
&= \mathbf{sw}_{\mathbf{q}} \circ \boldsymbol{u}_1,
\end{aligned} \tag{A.24}$$

where $\mathbf{q} = (q_1, \dots, q_{d^*})$ and $q_r = \prod_{\ell=1}^{L} p_{\ell,r}$. The properties of the sawtooth functions and the special form of $\boldsymbol{u}_1$ yields the projection regions for $\boldsymbol{h}$ as

$$\{\mathbf{x} \in \mathbb{R}^{d_1} : \text{for } r = 1, \dots, d^*, \begin{cases} -\infty < x_r \leq 1/q_r, & i_r = 0 \\ 1 - 1/q_r \leq x_r < +\infty, & i_r = q_r - 1 \\ i_r/q_r \leq x_r \leq (i_r + 1)/q_r, & \text{otherwise} \end{cases}\}, \tag{A.25}$$

where $i_r = 0, \dots, (q_r - 1)$ for $r = 1, \dots, d^*$. In summary, the number of projection regions of the constructed CPWL NN is at least

$$\prod_{r=1}^{d^*} q_r = \prod_{\ell=1}^{L} \prod_{r=1}^{d^*} p_{\ell,r} = \prod_{\ell=1}^{L} \prod_{r=1}^{d^*} \sum_{k \in \tau_\ell^{-1}(\{r\})} \kappa_{\ell,k}. \tag{A.26}$$

The conclusion is reached by noticing that the reasoning does not depend on any property of the mappings τ_ℓ: one can therefore pick the ones that yield the largest lower bound. □

Proof of Corollary 1. To get the upper bound, we combine Theorem 2 and the simplified version of the bound given in Theorem 1 with the assumption that $W \geq d_{\text{in}} \geq d^* := \min(d_{\text{in}}, d_{\text{out}})$. To get the lower bound, we have to compute

$$\alpha^{d^*}(\kappa_{\ell,1}, \dots, \kappa_{\ell,d_{\ell+1}}) = \max_{\tau \in \mathcal{T}_{d_\ell}} \prod_{r=1}^{d^*} \sum_{k \in \tau^{-1}(\{r\})} \kappa_{\ell,k}. \tag{A.27}$$

We lower-bound this quantity by selecting an arbitrary mapping $\tau : [W] \to [d^*]$ such that, for $r \in [d^*]$, the cardinality of $\tau^{-1}(\{r\})$ is at least $\lfloor W/d^* \rfloor$. In this way, we obtain that

$$\alpha^{d^*}(\kappa_{\ell,1}, \dots, \kappa_{\ell,d_{\ell+1}}) \geq (\kappa \lfloor W/d^* \rfloor)^{d^*} \tag{A.28}$$

for $\ell = 1, \dots, L$ and reach the given lower bound. □

Proof of Corollary 2. The upper bound is a direct consequence of Proposition 1: the number of convex linear regions is never smaller than the number of projection regions. The CPWL NN built to provide the lower bound of Theorem 2 had exactly as many projection regions as convex linear regions, hence justifying the lower bound. □

Appendix B. Proofs for the knot density of the sum, vectorization, and composition of CPWL functions

Proof of Proposition 3. Consider two CPWL functions f_1, f_2 with characteristic functions $\boldsymbol{\varphi}_1^{\boldsymbol{\gamma}}$ and $\boldsymbol{\varphi}_2^{\boldsymbol{\gamma}}$ along $\boldsymbol{\gamma}$ and projection regions $(\Omega_k^1)_{k=1}^{K_1}$ and $(\Omega_k^2)_{k=1}^{K_2}$. Let $\boldsymbol{\varphi}_{(1,2)}^{\boldsymbol{\gamma}}$ denote the characteristic function of (f_1, f_2) and $\boldsymbol{\varphi}_{1+2}^{\boldsymbol{\gamma}}$ the one of $f_1 + f_2$ along $\boldsymbol{\gamma}$. Consider a subset $R \subset S$ on which $\boldsymbol{\varphi}_1^{\boldsymbol{\gamma}}$ and $\boldsymbol{\varphi}_2^{\boldsymbol{\gamma}}$ are continuous. Following the definition of the characteristic function, any projection region Ω of f_1 or of f_2 either entirely contains $\boldsymbol{\gamma}(R)$ or does not intersect with it; otherwise, $\boldsymbol{\varphi}_1^{\boldsymbol{\gamma}}$ or $\boldsymbol{\varphi}_2^{\boldsymbol{\gamma}}$ would not be continuous on R. Any projection region $\Omega^{(1,2)}$ of (f_1, f_2) is a nonempty intersection of the form $\Omega_p^1 \cap \Omega_q^2$, with $p \in [K_1]$ and $q \in [K_2]$. Since the regions Ω_p^1 and Ω_q^2 either entirely contain $\boldsymbol{\gamma}(R)$ or do not intersect with it, the same holds true for $\Omega^{(1,2)}$, which implies that $\boldsymbol{\varphi}_{(1,2)}^{\boldsymbol{\gamma}}$ must be continuous on R. The same argument holds true for $\boldsymbol{\varphi}_{(1+2)}^{\boldsymbol{\gamma}}$, the only difference being that the projection regions of $f_1 + f_2$ are unions of nonempty subsets of the form $\Omega_p^1 \cap \Omega_q^2$, which also has the same implications. So, on the one hand, we proved that, where $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ are continuous, $\boldsymbol{\varphi}_{1+2}$ and $\boldsymbol{\varphi}_{(1,2)}$ are also continuous. On the other hand, the number of points where either $\boldsymbol{\varphi}_1$ or $\boldsymbol{\varphi}_2$ is discontinuous is no greater than the number of points where $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ are discontinuous, which concludes the proof. □

Proof of Proposition 4. Consider the characteristic function $\boldsymbol{\varphi}_{f_1}^{\boldsymbol{\gamma}}$ of f_1 on $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}_{f_2}^{f_1 \circ \boldsymbol{\gamma}}$ of f_2 on $f_1 \circ \boldsymbol{\gamma}$. Moreover, consider a subset $R \subset S$ on which both $\boldsymbol{\varphi}_{f_1}^{\boldsymbol{\gamma}}$ and $\boldsymbol{\varphi}_{f_2}^{f_1 \circ \boldsymbol{\gamma}}$ are continuous. The projection regions of f_1 and f_2 are denoted by $(\Omega_k^1)_{k=1}^{K_1}$ and $(\Omega_k^2)_{k=1}^{K_2}$. On any region $\Lambda_{p,q} = \Omega_p^1 \cap f_1^{-1}(\Omega_q^2)$, the function $f_2 \circ f_1$ is affine, meaning that each projection region of $f_2 \circ f_1$ is a union of some of the regions $\Lambda_{p,q}$. Each region Ω_p^1 either entirely contains the subset $\boldsymbol{\gamma}(R)$ or does not intersect with it; otherwise, $\boldsymbol{\varphi}_{f_1}^{\boldsymbol{\gamma}}$ would not be continuous on R. We now remark that $\boldsymbol{\varphi}_{f_2}^{f_1 \circ \boldsymbol{\gamma}}(t) = (\mathbb{1}_{\Omega_1^2}(f_1 \circ \boldsymbol{\gamma}(t)), \dots, \mathbb{1}_{\Omega_{K_2}^2}(f_1 \circ \boldsymbol{\gamma}(t))) = (\mathbb{1}_{f_1^{-1}(\Omega_1^2)}(\boldsymbol{\gamma}(t)), \dots, \mathbb{1}_{f_1^{-1}(\Omega_{K_2}^2)}(\boldsymbol{\gamma}(t)))$, which means that each region $f_1^{-1}(\Omega_q^2)$ either entirely contains the subset $\boldsymbol{\gamma}(R)$ or does not intersect with it since, otherwise, $\boldsymbol{\varphi}_{f_2}^{f_1 \circ \boldsymbol{\gamma}}$ would not be continuous on R. We therefore have that each region $\Lambda_{p,q}$ either entirely contains the subset $\boldsymbol{\gamma}(R)$ or does not intersect with it. Consequently, the same holds true for union of regions $\Lambda_{p,q}$ and, as a result, for the projection regions of $f_2 \circ f_1$. This shows that, where $\boldsymbol{\varphi}_{f_1}^{\boldsymbol{\gamma}}$ and $\boldsymbol{\varphi}_{f_2}^{f_1 \circ \boldsymbol{\gamma}}$ are continuous, $\boldsymbol{\varphi}_{f_2 \circ f_1}^{\boldsymbol{\gamma}}$ is also continuous. In short, the number of points of discontinuities of $\boldsymbol{\varphi}_{f_2 \circ f_1}^{\boldsymbol{\gamma}}$ is no greater than the number of points of discontinuity of either $\boldsymbol{\varphi}_{f_1}^{\boldsymbol{\gamma}}$ or $\boldsymbol{\varphi}_{f_2}^{f_1 \circ \boldsymbol{\gamma}}$, which concludes the proof. □

Appendix C. Proof of the bounds on the knot density of classical CPWL components

Proof of Proposition 5. First, we prove the result when $\boldsymbol{\gamma}$ parameterizes a linear segment. Let $\mathbf{x}_0, \mathbf{u} \in \mathbb{R}^d$ with $\|\mathbf{u}\|_2 = 1$ and $\boldsymbol{\gamma} : t \mapsto \mathbf{x}_0 + t\mathbf{u}$ for $t \in S$, where $S = [0, |S|] \subset \mathbb{R}$ is a segment. The first step is to compute the probability $\mathbb{P}(\text{kt}_f^{\boldsymbol{\gamma}} = 1)$ that f has a knot along $\boldsymbol{\gamma}$. The hyperplane $\{\mathbf{x} \in \mathbb{R}^d : \boldsymbol{w}^T\mathbf{x} + b = 0\}$ intersects the line $\{\mathbf{x}_0 + t\mathbf{u} : t \in \mathbb{R}\}$ for t_0 such that $\boldsymbol{w}^T(\mathbf{x}_0 + t_0\mathbf{u}) + b = 0$ or, equivalently, $b = (-\boldsymbol{w}^T(\mathbf{x}_0 + t_0\mathbf{u}))$. In order to have a knot along $\boldsymbol{\gamma}_{|S}$, t_0 has to lie in S. For a given $\boldsymbol{w}$, this implies that b should be in an interval of length $|S \parallel \boldsymbol{w}^T\mathbf{u}|$, more precisely $[-\boldsymbol{w}^T\mathbf{x}_0, |S|\boldsymbol{w}^T\mathbf{u} - \boldsymbol{w}^T\mathbf{x}_0]$ if $\boldsymbol{w}^T\mathbf{u} < 0$ and $[|S|\boldsymbol{w}^T\mathbf{u} - \boldsymbol{w}^T\mathbf{x}_0, -\boldsymbol{w}^T\mathbf{x}_0]$ otherwise. Therefore, $\mathbb{P}(\text{kt}_f^{\boldsymbol{\gamma}} = 1|\boldsymbol{w}) \leq \sup_{t \in \mathbb{R}} \rho_b(t)|S \parallel \boldsymbol{w}^T\mathbf{u}|$. From the independence of the random variables and from the fact that $\text{kt}_f^{\boldsymbol{\gamma}} = 0$ or $\text{kt}_f^{\boldsymbol{\gamma}} = 1$ almost surely, we infer that $\mathbb{E}[\lambda_f^{\boldsymbol{\gamma}|S}] \leq \sup_{t \in \mathbb{R}} \rho_b(t)\mathbb{E}[|\boldsymbol{w}^T\mathbf{u}|] \leq \sup_{t \in \mathbb{R}} \rho_b(t)\sqrt{\mathbb{E}[|\boldsymbol{w}^T\mathbf{u}|^2]} = \sup_{t \in \mathbb{R}} \rho_b(t)\sqrt{\mathbf{u}^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\mathbf{u}} \leq \sup_{t \in \mathbb{R}} \rho_b(t)\sqrt{\mathbb{E}[w^2]}$. In the last step, the assumption that the random variables w_k are i.i.d. has allowed us to infer that $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] = \mathbb{E}[w^2]\mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix.

If b is normally distributed with standard deviation σ_b, then $\sup_{t \in \mathbb{R}} \rho_b(t) = (\sigma_b\sqrt{2\pi})^{-1}$. In addition, suppose that the components w_k are independent and normally distributed with standard deviation σ_w. The random variable $\boldsymbol{w}^T\mathbf{u}$ is also normally distributed with standard deviation σ_w (since $\|\mathbf{u}\|_2 = 1$). We can now compute explicitly $\mathbb{E}[|\boldsymbol{w}^T\mathbf{u}|] = \sigma_w\sqrt{2}/\sqrt{\pi}$ based on the properties of half-normal distributions.

The result is extended to any polygonal chain through the linearity of the expectation operator and by application of the result to the finitely many pieces of the polygonal chain. □

Proof of Proposition 6. A knot of f along a line $\boldsymbol{\gamma}$ must lie on a hyperplane $H_{p,q} = \{\mathbf{x} : (\boldsymbol{w}_p - \boldsymbol{w}_q)^T\mathbf{x} + (b_p - b_q) = 0\}$ with $1 \leq p < q \leq K$, since elsewhere the Maxout unit is affine. Therefore, the expected knot density is bounded as

$$\begin{aligned}
\mathbb{E}\left[\lambda^{\boldsymbol{\gamma}}_{\text{Maxout}}\right] &\leq \frac{1}{|S|}\mathbb{E}\left[\sum_{1\leq p<q\leq K}(\boldsymbol{\gamma}(S)\cap H_{p,q}\neq\emptyset)\right] \\
&= \frac{1}{|S|}\sum_{1\leq p<q\leq K}\mathbb{E}\left[(\boldsymbol{\gamma}(S)\cap H_{p,q}\neq\emptyset)\right] \\
&\leq \frac{1}{|S|}\sum_{1\leq p<q\leq K}\sqrt{\mathbb{E}\left[(w_p - w_q)^2\right]}\sup_{t\in\mathbb{R}}\rho_b(t) \\
&= \frac{1}{|S|}\sum_{1\leq p<q\leq K}\sqrt{2}\sigma_w\sup_{t\in\mathbb{R}}\rho_b(t) \\
&= \frac{1}{|S|}\sqrt{2}\binom{K}{2}\sigma_w\sup_{t\in\mathbb{R}}\rho_b(t),
\end{aligned} \tag{C.1}$$

where we have taken advantage of the results derived in the proof of Proposition 5 to bound the probability that a randomly generated hyperplane intersects a segment of length $|S|$, along with the independence of the random variables. We also notate $(A \neq \emptyset)$ to encode the variable that takes the value 0 if $A = \emptyset$ and 1 otherwise. When the random variables are normally distributed, the reasoning is similar to the one in the proof of Proposition 5. □

Proof of Proposition 7. A knot of f along a line $\boldsymbol{\gamma}$ must lie on a hyperplane $H_{p,q} = \{\mathbf{x} : (\boldsymbol{w}_p - \boldsymbol{w}_q)^T\mathbf{x} + (b_p - b_q) = 0\}$, where $1 \leq p < q \leq K$ and where p, q belong to the same sorting group, since elsewhere the GroupSort layer is affine. One can now follow the same steps as those in the proof of Proposition 6 with $n_g\binom{g_s}{2} = n_g g_s(g_s-1)/2 = d(g_s-1)/2$ hyperplanes. □

Appendix D. Proofs of the bounds on the expected knot density of CPWL NNs

Proof of Lemma 4. In what follows, the technical developments originate from the fact that $\boldsymbol{f}_\theta$ is not differentiable everywhere. The function $\boldsymbol{f}_\theta \circ \boldsymbol{\gamma}$ is the composition of two CPWL functions, hence it is CPWL and, therefore, differentiable for almost every $t \in S$. Note, however, that we cannot assert that the Jacobian of $\boldsymbol{f}_\theta$ is well defined at $\boldsymbol{\gamma}(t)$ for almost every $t \in S$. Indeed, whenever $\boldsymbol{\gamma}$ follows the boundary of two projection regions of $\boldsymbol{f}_\theta$, the Jacobian of $\boldsymbol{f}_\theta$ along $\boldsymbol{\gamma}$ becomes ill-posed. This is why the notion of directional derivative is better suited. The characteristic function $\boldsymbol{\varphi}^{\boldsymbol{\gamma}}_{\boldsymbol{f}_\theta}$ of $\boldsymbol{f}_\theta$ along $\boldsymbol{\gamma}$ is piecewise-constant on S: We can partition S into finitely many convex regions where $\boldsymbol{\varphi}^{\boldsymbol{\gamma}}_{\boldsymbol{f}_\theta}$ is constant. Let $P \subset S$ denote one of these regions and let $Q \subset S$ be a linear convex region of $\boldsymbol{\gamma}$. Following the definition of the characteristic function, there exists a projection region Ω of $\boldsymbol{f}_\theta$ such that $\boldsymbol{\gamma}(\text{int}(P \cap Q))$ either lies entirely in the interior of Ω, or entirely on its boundary. In the first case, $\boldsymbol{f}_\theta$ is differentiable in $\boldsymbol{\gamma}(\text{int}(P \cap Q))$ and we have that $(\boldsymbol{f}_\theta \circ \boldsymbol{\gamma})'(t) = J_{\boldsymbol{f}_\theta}(\boldsymbol{\gamma}(t))\boldsymbol{\gamma}'(t) = D_{\boldsymbol{\gamma}'(t)}\{\boldsymbol{f}_\theta\}(\boldsymbol{\gamma}(t))$ for $t \in \text{int}(P \cap Q)$. In the second case, $\boldsymbol{\gamma}$ is differentiable on $\text{int}(P \cap Q)$ as well, but the Jacobian of $\boldsymbol{f}_\theta$ is undefined. Fortunately, the directional derivative of $\boldsymbol{f}_\theta$ is well defined along $\boldsymbol{\gamma}(t)$ since, for any $t \in \text{int}(P \cap Q)$, there exists $\epsilon > 0$ such that $\tau \mapsto \boldsymbol{f}_\theta(\boldsymbol{\gamma}(t) + \tau\boldsymbol{\gamma}'(t))$ is affine on $(-\epsilon, \epsilon)$. All in all, the relation $(\boldsymbol{f}_\theta \circ \boldsymbol{\gamma})'(t) = D_{\boldsymbol{\gamma}'(t)}\{\boldsymbol{f}_\theta\}(\boldsymbol{\gamma}(t))$ is well defined for any $t \in \text{int}(P \cap Q)$ and, more generally, for almost any $t \in S$ because of the properties of P and Q. We can now write that

$$\begin{aligned}
\mathbb{E}\left[\text{Len}(\boldsymbol{f}_\theta \circ \boldsymbol{\gamma})\right] &= \mathbb{E}\left[\int_{t\in S}\left\|(\boldsymbol{f}_\theta \circ \boldsymbol{\gamma})'(t)\right\|_2 \mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_{t\in S}\left\|D_{\boldsymbol{\gamma}'(t)}\{\boldsymbol{f}_\theta\}(\boldsymbol{\gamma}(t))\right\|_2 \mathrm{d}t\right] \\
&= \int_{t\in S}\mathbb{E}\left[\left\|D_{\boldsymbol{\gamma}'(t)}\{\boldsymbol{f}_\theta\}(\boldsymbol{\gamma}(t))\right\|_2\right]\mathrm{d}t \\
&\leq \sup_{\substack{\mathbf{x},\mathbf{u}\in\mathbb{R}^d \\ \|\mathbf{u}\|_2=1}}\mathbb{E}[\|D_{\mathbf{u}}\{\boldsymbol{f}_\theta\}(\mathbf{x})\|_2]\int_{t\in S}\left\|\boldsymbol{\gamma}'(t)\right\|_2 \mathrm{d}t \\
&= \text{Len}(\boldsymbol{\gamma})\sup_{\substack{\mathbf{x},\mathbf{u}\in\mathbb{R}^d \\ \|\mathbf{u}\|_2=1}}\mathbb{E}[\|D_{\mathbf{u}}\{\boldsymbol{f}_\theta\}(\mathbf{x})\|_2],
\end{aligned} \tag{D.1}$$

where we have used Tonelli's theorem to interchange the expectation and the integral. □

Proof of Theorem 3. Let $\boldsymbol{F}_\ell = \boldsymbol{f}_{\theta_\ell} \circ \cdots \circ \boldsymbol{f}_{\theta_1}$. With Proposition 4, we have that

$$\mathbb{E}\left[\text{kt}^{\boldsymbol{\gamma}}_{\boldsymbol{F}_\ell}\right] = \mathbb{E}\left[\text{kt}^{\boldsymbol{\gamma}}_{\boldsymbol{f}_{\theta_\ell}\circ\boldsymbol{F}_{L-1}}\right] \leq \mathbb{E}\left[\text{kt}^{\boldsymbol{F}_{\ell-1}\circ\boldsymbol{\gamma}}_{\boldsymbol{f}_{\theta_\ell}}\right] + \mathbb{E}\left[\text{kt}^{\boldsymbol{\gamma}}_{\boldsymbol{F}_{\ell-1}}\right]. \tag{D.2}$$

We now apply the law of the iterated expectation to obtain that

$$\mathbb{E}\left[\text{kt}^{\boldsymbol{F}_{\ell-1}\circ\boldsymbol{\gamma}}_{\boldsymbol{f}_{\theta_\ell}}\right] = \mathbb{E}_{\theta_1,\ldots,\theta_{\ell-1}}\left[\mathbb{E}_{\theta_\ell}\left[\text{kt}^{\boldsymbol{F}_{\ell-1}\circ\boldsymbol{\gamma}}_{\boldsymbol{f}_{\theta_\ell}}|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{\ell-1}\right]\right]$$

$$\leq \mathbb{E}_{\theta_1,\dots,\theta_{\ell-1}}\left[d\,\lambda_0 \mathrm{Len}(\boldsymbol{F}_{\ell-1}\circ\boldsymbol{\gamma})\right], \tag{D.3}$$

where the inequality follows from the first assumption of the theorem, the application of Proposition 3, and requires the independence of the random variables. We can now apply Lemma 4 recursively to $\boldsymbol{F}_\ell$ and invoke the second assumption of the theorem to infer that

$$\mathbb{E}\left[\mathrm{kt}^{\boldsymbol{F}_{\ell-1}\circ\boldsymbol{\gamma}}_{f_{\theta_\ell}}\right] \leq d\,\lambda_0 \mathrm{Len}(\boldsymbol{\gamma}) D_0^{\ell-1}. \tag{D.4}$$

All in all, we just proved that

$$\mathbb{E}\left[\mathrm{kt}^{\boldsymbol{\gamma}}_{\boldsymbol{F}_\ell}\right] \leq \mathbb{E}\left[\mathrm{kt}^{\boldsymbol{\gamma}}_{\boldsymbol{F}_{\ell-1}}\right] + d\,\lambda_0 \mathrm{Len}(\boldsymbol{\gamma}) D_0^{\ell-1}, \tag{D.5}$$

which reads in term of linear densities as

$$\mathbb{E}\left[\lambda^{\boldsymbol{\gamma}}_{\boldsymbol{F}_\ell}\right] \leq \mathbb{E}\left[\lambda^{\boldsymbol{\gamma}}_{\boldsymbol{F}_{\ell-1}}\right] + d\,\lambda_0 D_0^{\ell-1}. \tag{D.6}$$

This recurrence relation directly yields the announced bound. □

Proof of Corollary 3. The proof is similar to the proof of Theorem 2 except that, with the different second assumption, the quantity $\mathbb{E}_{\theta_1,\dots,\theta_{\ell-1}}[\mathrm{Len}(\boldsymbol{F}_{\ell-1}\circ\boldsymbol{\gamma})]$ can be bounded by $D_0\mathrm{Len}(\boldsymbol{\gamma})$ (Lemma 4). In the end the recurrence relation (D.6) is changed into

$$\mathbb{E}\left[\lambda^{\boldsymbol{\gamma}}_{\boldsymbol{F}_\ell}\right] \leq \mathbb{E}\left[\lambda^{\boldsymbol{\gamma}}_{\boldsymbol{F}_{\ell-1}}\right] + D_0\lambda_0\, W. \quad \square \tag{D.7}$$

References

[1] Y. Bengio, Learning deep architectures for AI, Foundations and Trends® in Machine Learning (1) (2009) 1–127, http://dx.doi.org/10.1561/2200000006.
[2] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.
[3] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Vol. 15, 2011, pp. 315–323.
[4] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proceedings of the 30th International Conference on Machine Learning, Vol. 30, Citeseer, 2013, p. 3.
[5] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, http://dx.doi.org/10.1038/nature14539.
[6] G.F. Montúfar, R. Pascanu, K. Cho, Y. Bengio, On the number of linear regions of deep neural networks, Adv. Neural Inf. Process. Syst. 27 (2014).
[7] R. Arora, A. Basu, P. Mianjy, A. Mukherjee, Understanding deep neural networks with rectified linear units, in: 6th International Conference on Learning Representations, ICLR 2018, 2018, arXiv:1611.01491.
[8] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[9] W. Shang, K. Sohn, D. Almeida, H. Lee, Understanding and improving convolutional neural networks via concatenated rectified linear units, in: 33rd International Conference on Machine Learning, 2016, pp. 2217—-2225, arXiv:1603.05201.
[10] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, in: 30th International Conference on Machine Learning, International Machine Learning Society (IMLS), 2013, pp. 2356–2364, arXiv:1302.4389.
[11] F. Agostinelli, M. Hoffman, P. Sadowski, P. Baldi, Learning activation functions to improve deep neural networks, 2015, arXiv:1412.6830 [cs, stat].
[12] M. Unser, A representer theorem for deep neural networks, J. Mach. Learn. Res. 20 (110) (2019) 1–30, arXiv:1802.09210.
[13] C. Anil, J. Lucas, R. Grosse, Sorting out Lipschitz function approximation, in: International Conference on Machine Learning, PMLR, 2019, pp. 291–301.
[14] S. Singla, S. Singla, S. Feizi, Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100, 2021, http://dx.doi.org/10.48550/ARXIV.2108.04062, URL https://arxiv.org/abs/2108.04062.
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.
[16] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, PMLR, 2015, pp. 448–456.
[17] R. Eldan, O. Shamir, The power of depth for feedforward neural networks, in: 29th Annual Conference on Learning Theory, Vol. 49, PMLR, Columbia University, New York, New York, USA, 2016, pp. 907–940, URL https://proceedings.mlr.press/v49/eldan16.html.
[18] H. Mhaskar, T. Poggio, Deep vs. Shallow networks : an approximation theory perspective, Anal. Appl. 14 (06) (2016) 829—-848, arXiv:1608.03287.
[19] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, Q. Liao, Why and when can deep—but not shallow – Networks avoid the curse of dimensionality: A review, Int. J. Autom. Comput. 14 (5) (2017) 503—-519, arXiv:1611.00740.
[20] A. Goujon, J. Campos, M. Unser, Stable parameterization of continuous and piecewise-linear functions, Appl. Comput. Harmon. Anal. (2023) 101581, http://dx.doi.org/10.1016/j.acha.2023.101581.
[21] R. Balestriero, R. Baraniuk, A spline theory of deep learning, in: Proceedings of the 35th International Conference on Machine Learning, Vol. 80, 2018, pp. 374–383.
[22] R. Balestriero, R. Baraniuk, Mad Max: Affine spline insights into deep learning, Proc. IEEE 109 (5) (2021) 704–727, http://dx.doi.org/10.1109/JPROC.2020.3042100.
[23] J.M. Tarela, E. Alonso, M.V. Martínez, A representation method for PWL functions oriented to parallel processing, Math. Comput. Modelling (1990) http://dx.doi.org/10.1016/0895-7177(90)90090-A.
[24] J.M. Tarela, M.V. Martínez, Region configurations for realizability of lattice piecewise-linear models, Math. Comput. Modelling (1999) http://dx.doi.org/10.1016/S0895-7177(99)00195-8.
[25] B. Hanin, D. Rolnick, Deep ReLU networks have surprisingly few activation patterns, Adv. Neural Inf. Process. Syst. 32 (NeurIPS) (2019).
[26] R. Balestriero, R. Cosentino, B. Aazhang, R. Baraniuk, The geometry of deep networks: Power diagram subdivision, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.
[27] T. Serra, C. Tjandraatmadja, S. Ramalingam, Bounding and Counting Linear Regions of deep neural networks, in: 35th International Conference on Machine Learning, ICML 2018, Vol. 10, 2018, pp. 7243–7261, arXiv:1711.02114.

[28] P. Hinz, S. de Geer, A framework for the construction of upper bounds on the number of Affine Linear Regions of ReLU reed-forward neural networks, IEEE Trans. Inform. Theory 65 (11) (2019) 7304–7324.
[29] H. Xiong, L. Huang, M. Yu, L. Liu, F. Zhu, L. Shao, On the number of linear regions of convolutional neural networks, in: H.D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 10514–10523, URL https://proceedings.mlr.press/v119/xiong20a.html.
[30] G. Montúfar, Y. Ren, L. Zhang, Sharp bounds for the number of regions of maxout networks and vertices of Minkowski sums, SIAM J. Appl. Algebra Geom. 6 (4) (2022) 618–649, http://dx.doi.org/10.1137/21M1413699.
[31] B. Hanin, D. Rolnick, Complexity of linear regions in deep networks, in: 36th International Conference on Machine Learning, 2019, pp. 4585–4600, arXiv:1901.09021.
[32] H. Tseran, G.F. Montúfar, On the expected complexity of maxout networks, Adv. Neural Inf. Process. Syst. 34 (2021) 28995–29008.
[33] U. Tanielian, G. Biau, Approximating Lipschitz continuous functions with GroupSort neural networks, in: A. Banerjee, K. Fukumizu (Eds.), Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 130, PMLR, 2021, pp. 442–450, URL https://proceedings.mlr.press/v130/tanielian21a.html.
[34] Z. Zhu, Y. Zhou, Y. Dong, Z. Zhong, PWLU: Learning specialized activation functions with the piecewise linear unit, IEEE Trans. Pattern Anal. Mach. Intell. (2023) 1–19, http://dx.doi.org/10.1109/TPAMI.2023.3286109.
[35] E. León, G.M. Ziegler, Spaces of convex n-partitions, Bolyai Soc. Math. Stud. 27 (2018) 279–306, http://dx.doi.org/10.1007/978-3-662-57413-3_11, arXiv:1511.02904 URL CPWL.
[36] S. Wang, X. Sun, Generalization of hinging hyperplanes, IEEE Trans. Inform. Theory 51 (12) (2005) 4425—-4431, http://dx.doi.org/10.1109/TIT.2005.859246.
[37] T. Zaslavsky, Facing Up To Arrangements: Face-Count Formulas for Partitions of Space by Hyperplanes, Vol. 154, American Mathematical Society, 1975.
[38] D. Bulavka, A. Goodarzi, M. Tancer, Optimal bounds for the colorful fractional helly theorem, 2020, http://dx.doi.org/10.48550/ARXIV.2010.15765, URL https://arxiv.org/abs/2010.15765.
[39] G. Montúfar, Notes on the number of linear regions of deep neural networks, in: 2017, 12th International Conference on Sampling Theory and Applications (SampTA), 2017.
[40] P. Bohra, J. Campos, H. Gupta, S. Aziznejad, M. Unser, Learning activation functions in deep (spline) neural networks, IEEE Open J. Signal Process. 1 (2020) 295–309, http://dx.doi.org/10.1109/OJSP.2020.3039379.
[41] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, J. Sohl-Dickstein, On the expressive power of deep neural networks, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 2847–2854, URL https://proceedings.mlr.press/v70/raghu17a.html.
[42] B. Hanin, R.S. Jeong, D. Rolnick, Deep ReLU networks preserve expected length, in: International Conference on Learning Representations, 2022, URL https://openreview.net/forum?id=ci7LBzDn2Q.
[43] P. Pauli, A. Koch, J. Berberich, P. Kohler, F. Allgöwer, Training robust neural networks using Lipschitz bounds, IEEE Control Syst. Lett. 6 (2022) 121–126, http://dx.doi.org/10.1109/LCSYS.2021.3050444.
[44] P. Bohra, D. Perdios, A. Goujon, S. Emery, M. Unser, Learning Lipschitz-controlled activation functions in neural networks for plug-and-play image reconstruction methods, in: NeurIPS 2021 Workshop on Deep Learning and Inverse Problems, 2021, URL https://openreview.net/forum?id=efCsbTzQTbH.
[45] S. Neumayer, A. Goujon, P. Bohra, M. Unser, Approximation of Lipschitz functions using deep spline neural networks, SIAM J. Math. Data Sci. 5 (2) (2023) 306–322.