Polynomial Distribution of Feedforward Neural Network Output

Lokman Abbas-Turki, Brian Alexandrine, Qinghua Li

▶ **To cite this version:**

Lokman Abbas-Turki, Brian Alexandrine, Qinghua Li. Polynomial Distribution of Feedforward Neural Network Output. 2023. hal-04217029

HAL Id: hal-04217029
https://hal.science/hal-04217029

Preprint submitted on 25 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Polynomial Distribution of Feedforward Neural Network Output

Lokman A. Abbas-Turki*, Brian Alexandrine† and Qinghua Li‡

September 25, 2023

Abstract

This paper studies the output distribution of Leaky ReLU neural networks. The input is a vector of independent uniformly distributed random variables on $[0,1]$. The cumulative distribution function of the output is piecewise polynomial. We prove this characterization and give an implementable algorithm to compute the exact expression. We also discuss how to extend our result to piecewise linear neural networks with other input distributions.

Keywords: neural network, probabilistic distribution, volumes of polytopes

MSC codes: 68T01, 60E05, 60D05, 52A38, 52B11

1 Introduction

This paper is motivated by approximating the probabilistic model of a given dataset Y. Standard references in distribution learning solve this problem by direct approximations of the distribution of Y. In generative models, this approximation takes the form of training a neural network to generate other realizations of Y like with generative adversarial networks (cf. the recent survey in Gui et al. (2023)) and with diffusion models (we refer for example to Croitoru et al. (2023)). Other contributions focus on the approximation of the distribution itself with a neural network using maximum likelihood (cf. Dinh et al. (2017)). This is the case for latent variables models (cf. Gresele et al. (2020) and the references therein) where the density of Y is obtained through the expansion of the likelihood expressed in terms of the density of a latent variable applied to the output of a neural net, that has Y as its input, plus the determinant of a Jacobian term.

Different from the direct approximation methods, we propose computing the analytical distribution of $G_L(X)$, which is a neural network approximation of a random variable Y. The random vector X has a known distribution. The feedforward neural network $G_L(\cdot)$ with L-layers is a minimizer of some loss function measuring the distance, the L^2-distance for example, between Y and its approximator. To be more specific, let $\mathscr{N}_L$ denotes the collection of all L-layer feedforward neural networks with proper input domain, then

$$G_L(\cdot) \in \arg\min \left\{ E\left[|\phi(X) - Y|^2\right] \middle| \, \phi \in \mathscr{N}_L \right\}. \tag{1.1}$$

*LPSM (UMR 8001), Sorbonne Université (lokmane.abbas_turki@sorbonne-universite.fr). The research of L. A. Abbas-Turki has benefited from the support of the Chair *Capital Markets Tomorrow: Modeling and Computational Issues* under the aegis of the Institut Europlace de Finance, a joint initiative of Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Université Paris Cité and Crédit Agricole CIB.

†Fotonower, France (brian.alexandrine@gmail.com)

‡School of Statistics, Shandong Technology and Business University, China (202013904@sdtbu.edu.cn) Qinghua Li acknowledges the Doctoral Research Starting Grant No. BS202139 from her affiliated university.

When G_L is a solution to Equation (1.1), the random variable $G_L(X)$ is an approximation of $E[Y|X]$. When Y is measurable with respect to X, $G_L(X)$ approximates Y and thus their distributions should be close. When Y is not measurable with respect to X, computing the distribution of $G_L(X) \approx E[Y|X]$ is also different from the conditional latent variables methods that approximate the distribution of Y conditional to X.

Neural networks are indeed a reasonable class of models to approximate functions. Their infinite approximation power is state as the Universal Approximation Theorem, for shallow networks in Hornik (1991) and for deep networks in Kidger and Lyons (2020). Because of their modelling efficiency, neural networks are advantageous in approximating functions particularly on high-dimensional input spaces. Theoretical proofs in multiple situations suggest that neural networks are hopeful in breaking the curse of dimensionality in terms of sample size required, for example the approximation of several structures by shallow networks with non-decreasing homogeneous activation functions in Bach (2017), and the approximation of Lipschitz continuous functions by networks with ReLU activation functions in Poggio and Liao (2018). In their setting, Poggio and Liao (2018) further suggests that deep networks are even more efficient than shallow networks.

Despite their approximating power and modeling efficiency, there is in general no simple explicit expression of an inductively defined neural network. One exception is neural networks with piecewise affine activation functions, where $G_L(\cdot)$ is affine on polyhedra. This category of neural nets is already studied in multiple contributions like in Bunel et al. (2018); Tao et al. (2022). We choose Leaky ReLU activation functions for their explicit expressions. The input is uniformly distributed on a unit hypercube. The results can be extended to piecewise affine activation functions with notational changes only. In addition, this distribution can approximate the output distributions of neural networks with other commonly seen activation functions and input distributions, since an activation function can be approximated by piecewise affine activation functions and an input distribution can be approximated by histogram distributions. Prior to our paper, the output distribution of a neural network is not yet widely studied. The only existing work that we have found is Pan et al. (2022), which provides the output distributions for neural networks with respectively Sigmoid and smooth ReLU activation functions and Gaussian input, by applying the density formula of transformed random variables.

Our distributional computation is realized via theories and methods in polytopes. In the setting of piecewise linear neural networks with uniform input, the output cumulative distribution function $\mathbf{F}(y)$ is identified with the total volume of finitely many polytopes $\mathscr{P}_i \cap H(y)^-$. Each polytope $\mathscr{P}_i \cap H(y)^-$ is the part of a constant polytope $\mathscr{P}_i$ below a parallelly moving hyperplane $H(y)$. There exist several methods and algorithms to compute the exact volume of a constant polytope (cf. Büeler et al. (2000)). Among those methods, the signed decomposition methods can accommodate the cutting hyperplane $H(y)$. Among the signed decomposition methods, the volume induction formula in Lasserre (1983) is chosen for its numerical robustness. We propose an algorithm to compute the volume of the cut polytope $\mathscr{P}_i \cap H(y)^-$ with y being a variable. When expressing the facets of a polytope into polytopes of one dimension lower, the proposed algorithm takes a different approach from the variable substitution method in Lasserre (1983). In addition to showing the piecewise polynomial property of the volume of the cut polytope with respect to y, our approach enables a partition of the real line into finitely many intervals by the "levels", which are points determined by the vertices of the constant polytope and the constant coefficients of the cutting hyperplane. The volume becomes polynomial on each of these intervals defined using the set of partitioning levels. The latter set is a key element toward an implementable algorithm that computes the expression of $\mathbf{F}(y)$. We present a simplified version of this algorithm and illustrate it with a numerical example. We give also hints on how the implemented algorithm can be further optimized.

Section 2 sets the considered piecewise linear architecture and states the main result on the cumulative distribution function as Theorem 2.1. Section 3 presents an analytical formula for the volume of the cut polytope, and characterizes the volume as a continuous, piecewise polynomial

function of y, which allows to establish Theorem 2.1. Section 4 designs a detailed algorithm to numerically implement the distribution computation. This algorithm is then illustrated on a simple numerical example in Section 5. Appendix A lists the polytope definitions and properties needed in volume computation. Appendix B provides an analytical expression of the cumulative distribution function in terms of the parameters of the neural network.

2 Output distribution

Each neuron in the hidden layer processes information received from the previous layer and passes it on to the next layer; this is achieved via the composition of an affine transformation and an activation function. Affine transformations are specified by weight matrices and bias vectors. For the l-th hidden layer, $l = 1, \cdots, L-1$, the $d_l \times d_{l-1}$ weight matrix and the d_l-dimensional bias vector are respectively given by

$$W_l := (w_{lij})_{i=1,\cdots,d_l, j=1,\cdots,d_{l-1}}, \quad b_l := (b_{l1}, \cdots, b_{ld_l})^T. \tag{2.1}$$

Since $d_L = 1$, the output layer has the weight given by a d_{L-1}-dimensional row vector

$$W_L := \left(w_{L11}, \cdots, w_{L1d_{L-1}}\right) \tag{2.2}$$

and the bias b_L is a real number.

We choose an activation function of the Leaky ReLU type, having the expression

$$a(z) := a_+ \max\{z, 0\} + a_- \min\{z, 0\}, \text{ for } z \in \mathbb{R}, \tag{2.3}$$

where the parameters a_+ and a_- are two positive real numbers. Although Leaky ReLU has $a_+ = 1$ and $a_- = 0.01$, we use the general expression (2.3) for its symmetry and its genericity. For a d-dimensional real vector $u = (u_1, \cdots, u_d)^T$, the mapping $\mathbf{a}(u)$ is defined as

$$\mathbf{a}(u) := (a(u_1), \cdots, a(u_d))^T. \tag{2.4}$$

The neural network is a function

$$G_L : \mathbb{R}^{d_0} \to \mathbb{R}; z_0 \mapsto G_L(z_0), \tag{2.5}$$

defined through the iteration

$$\begin{cases} G_0(z_0) := z_0; \\ G_l(z_0) := \mathbf{a}(W_l G_{l-1}(z_0) + b_l), \, l = 1, \cdots, L-1; \\ G_L(z_0) := W_L G_{L-1}(z_0) + b_L. \end{cases} \tag{2.6}$$

Our computation of the output distribution of the neural network will be restricted to input variables in a polytope domain D_0 satisfying Assumption 2.1. Related definitions and properties about polytopes are available in Appendix A.

Assumption 2.1 *The domain D_0 is a d_0-polytope and is defined as*

$$D_0 := \{x \in \mathbb{R}^{d_0} | C_D x \leq \beta_D\}, \tag{2.7}$$

for an integer $m_0 > d_0$, a constant matrix $C_D \in \mathbb{R}^{m_0 \times d_0}$ and a constant vector $\beta_D \in \mathbb{R}^{m_0}$.

Lemma 2.1 *There exist d_0-polytopes $\mathscr{P}_1, \cdots, \mathscr{P}_n$ with disjoint interiors, such that*

$$G_L(z_0) = \sum_{i=1}^{n} \mathbb{1}_{\{z_0 \in \mathscr{P}_i\}} \left(C_i z_0 - \beta_i\right), \, z_0 \in D_0 = \cup_{i=1}^{n} \mathscr{P}_i, \tag{2.8}$$

for some integer $n > 0$, d_0-dimensional real column vectors $\{C_i\}_{i=1}^n$ and real numbers $\{\beta_i\}_{i=1}^n$.

Proof. It is well known in the literature, for example Bunel et al. (2018), that the output is piecewise affine on a family $\{\mathscr{Q}_i\}_{1\leq i\leq n}$ of polyhedra. It is then sufficient to set $\mathscr{P}_i = D_0 \cap \mathscr{Q}_i$. □

Remark 2.1 *Lemma 2.1 can be extended to the following situations.*

(i) *The activation function $a(\cdot)$ is piecewise affine.*

(ii) *The domain D_0 is a union of finitely many d_0-polytopes with disjoint interiors.*

Theorem 2.1 *If Z_0 is a uniformly distributed random variable on the polytope domain D_0, then the cumulative distribution function $\mathbf{F}(y)$ of $G_L(Z_0)$ is continuous with respect to y, and is polynomial of degree at most d_0 for y between any two adjacent values in the set $\mathscr{O}$ defined in Notation B.1.*

Proof. These properties of $\mathbf{F}(y)$ can be concluded from Property B.1 and theorem 3.2. □

Remark 2.2 *The proof of Theorem 2.1 essentially relies on two facts: piecewise affine expression of the neural network on polytopes and continuous, piecewise polynomial expression of the volume of a cut polytope. Theorem 2.1 can be extended to situations where the first fact remains.*

(i) *By Remark 2.1* (i)*, Theorem 2.1 can be extended to neural networks with piecewise affine activation functions. Activation functions of other types can be approximated by piecewise affine functions.*

(ii) *Even if it is not uniform, the distribution of Z_0 can be approximated by a histogram distribution, which is a sum of uniform distributions on a set $\{D_j\}_{1\leq j\leq k}$ of d_0-polytopes. By Remark 2.1* (ii)*, Theorem 2.1 can be extended to this histogram approximation. The use of histograms to study neural networks is not new and can be found for example in Păsăreanu et al. (2020).*

Remark 2.3 *In Theorem 2.1, announcing the piecewise polynomial property of $\mathbf{F}(y)$ with the set $\mathscr{O}$ is crucial for the numerical implementation. Between any two adjacent levels in $\mathscr{O}$, the polynomial expression of $\mathbf{F}(y)$ does not involve any indicator function, which would otherwise be almost impossible to manage. Moreover, an embarrassingly parallel implementation of the computation of $\mathbf{F}(y)$ is possible with respect to different levels in $\mathscr{O}$.*

3 Volumes of polytopes

The proof of Theorem 2.1 identifies the cumulative distribution function with volumes of polytopes cut by hyperplanes, as in Property B.1. This section is dedicated to expressions and properties of volumes of polytopes. Section 3.1 derives a procedure and formulae to compute the volume of a d-polytope in terms of a weighted sum of edge lengths. These results are developed from the Lasserre (1983) contributions. Specifically for the part of a constant polytope below a parameterized hyperplane, Section 3.2 further characterizes its volume as a function of y. Theorem 3.2 required in the proof of Theorem 2.1 is established. Basic polytope definitions and properties cited in this section are provided in Appendix A.

This section focuses on polytopes of dimension $d \geq 2$. When $d = 1$, a polytope is a line segment and a hyperplane degenerates to a point, for which Theorem 3.2 is trivially true.

3.1 Volume computation procedure and formulae

This subsection presents Theorem 3.1, the main volume formula needed to prove Theorem 3.2. An isometric transformation is designed in Lemma 3.2 in order to successively apply the Lasserre's formula in Lemma 3.3. This procedure combining the edge length expression in Lemma 3.1 and the volume expression (3.23) can be numerically computed.

Notation 3.1 (i) *The symbol $||\cdot||$ is the Euclidean norm of a matrix or vector; for a positive integer i, a a matrix or vector with the subscript i indicates its i-th row or element, and the subscript $-i$ indicates the matrix or vector having its i-th row or element removed; the symbol $det(\cdot)$ is the determinant of a square matrix.*

(ii) *For an integer $r \in \{1, \cdots, d\}$, an $(m-d+r+1) \times r$ constant real matrix $C^{(r)}$ and an $(m-d+r+1)$-column vector $\beta^{(r)}(y)$ parameterized by the real number y, the free placeholder notation for a polytope of dimension at most r is defined as*

$$\mathscr{P}^{(r)}(y) := \left\{ x \in \mathbb{R}^r \,\middle|\, C^{(r)}x \leq \beta^{(r)}(y) \right\}. \tag{3.1}$$

(iii) *For $i \in \{1, \cdots, m-d+r+1\}$, the i-th hyperplane defining $\mathscr{P}^{(r)}(y)$ is*

$$H_i^{(r)}(y) := \left\{ x \in \mathbb{R}^r \,\middle|\, C_i^{(r)}x = \beta_i^{(r)}(y) \right\}. \tag{3.2}$$

The set $F_i^{(r)}(y) := \mathscr{P}^{(r)}(y) \cap H_i^{(r)}(y)$ is a face of $\mathscr{P}^{(r)}(y)$ of dimension at most $r-1$ and is a polytope by Property A.1 (ii).

(iv) *For $r \in \{1, \cdots, d-1\}$ and distinct integers $1 \leq i_1, \cdots, i_{d-r} \leq m+1$, the intersection of hyperplanes*

$$H_{i_1,\cdots,i_{d-r}}(y) := \bigcap_{i \in \{i_1,\cdots,i_{d-r}\}} H_i^{(d)}(y), \tag{3.3}$$

and the intersection of faces

$$F_{i_1,\cdots,i_{d-r}}(y) := \bigcap_{i \in \{i_1,\cdots,i_{d-r}\}} F_i^{(d)}(y) = \mathscr{P}^{(d)}(y) \cap H_{i_1,\cdots,i_{d-r}}(y). \tag{3.4}$$

If the row vectors $\left\{ C_i^{(d)} \,\middle|\, i = i_1, \cdots, i_{d-r} \right\}$ are linearly independent, then by Property A.1 (iv), *$H_{i_1,\cdots,i_{d-r}}(y)$ is of dimension r, and $F_{i_1,\cdots,i_{d-r}}(y)$ is a face of $\mathscr{P}^{(d)}(y)$ of dimension at most r.*

Property 3.1 (ii) and Property 3.2 (ii) will need the following assumption in addition to those in Notation 3.1 (ii).

Assumption 3.1 *In Notation 3.1* (ii), *the elements of the vector $\beta^{(r)}(y)$ are affine with respect to y.*

Definition 3.1 *The (d-dimensional) volume $V_d(D)$ of a domain $D \subset \mathbb{R}^d$ is defined as*

$$V_d(D) := \mu_d(D) = \int_{\mathbb{R}^d} \mathbb{1}_{\{(x_1,\cdots,x_d) \in D\}} dx_1 \cdots dx_d, \tag{3.5}$$

where μ_d is the Lebesgue measure on $\mathbb{R}^d$. Especially, the one-dimensional volume is the length.

Lemma 3.1 *The length of $\mathscr{P}^{(1)}(y)$ defined in Equation* (3.1) *with $r=1$ is given by*

$$V_1(\mathscr{P}^{(1)}(y)) = B_0(y) \left(B_+(y) - B_-(y) \right)^+, \tag{3.6}$$

where

$$\begin{aligned} B_+(y) &:= \min \left\{ \beta_j^{(1)}(y)/C_j^{(1)} \,\middle|\, j \in \{1, \cdots, m-d+2\} \text{ and } C_j^{(1)} > 0 \right\}; \\ B_-(y) &:= \max \left\{ \beta_j^{(1)}(y)/C_j^{(1)} \,\middle|\, j \in \{1, \cdots, m-d+2\} \text{ and } C_j^{(1)} < 0 \right\}; \\ B_0(y) &:= \prod_{j=1}^{m-d+2} \left(1 - \mathbb{1}_{\{C_j^{(1)}=0 \text{ and } \beta_j^{(1)}(y)<0\}} \right). \end{aligned} \tag{3.7}$$

Proof. In Equation (3.1) with $r = 1$, the inequalities with $C_j^{(1)} > 0$ (respectively $C_j^{(1)} < 0$) correspond to the upper (respectively lower) bound constraints; when $C_j^{(1)} = 0$, the j-th inequality is never satisfied if $\beta_j^{(1)}(y) < 0$ and is always satisfied if otherwise. Since $\mathscr{P}^{(1)}(y)$ is at most a line segment, its upper and lower bounds have to be finite when $B_0(y) = 1$. □

Lemma 3.2 *For $r \in \{2, \cdots, d\}$ and for $i \in \{1, \cdots, m - d + r + 1\}$ such that $\left|\left|C_i^{(r)}\right|\right| \neq 0$, there exists a constant $r \times r$ orthogonal real matrix $R^{(r,i)}$ with determinant equal to 1, an $(m - d + r) \times (r - 1)$ matrix $C^{(r-1,i)}$ and an $(m - d + r)$-column vector $\beta^{(r-1,i)}(y)$, such that*

$$\mathscr{P}_i^{(r-1)}(y) := \left\{ x \in \mathbb{R}^{r-1} \,\middle|\, C^{(r-1,i)} x \leq \beta^{(r-1,i)}(y) \right\} \tag{3.8}$$

and $F_i^{(r)}(y)$ are related by the rotation and translation

$$\left\{ R^{(r,i)} \left(\begin{pmatrix} x \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{\beta_i^{(r)}(y)}{\left|\left|C_i^{(r)}\right|\right|} \end{pmatrix} \right) \middle| \, x \in \mathscr{P}_i^{(r-1)}(y) \right\} = F_i^{(r)}(y). \tag{3.9}$$

For $j \in \{1, \cdots, m - d + r\}$, the same rotation and translation relates the hyperplanes

$$H_j^{(r-1,i)}(y) := \left\{ x \in \mathbb{R}^{r-1} \,\middle|\, C_j^{(r-1,i)} x = \beta_j^{(r-1,i)}(y) \right\} \tag{3.10}$$

and

$$H_{i,j}^{(r)}(y) := \begin{cases} H_i^{(r)}(y) \cap H_j^{(r)}(y), \text{ if } j < i; \\ H_i^{(r)}(y) \cap H_{j+1}^{(r)}(y), \text{ if } j \geq i. \end{cases} \tag{3.11}$$

Proof. Let $R_1^{(r,i)}, \cdots, R_{r-1}^{(r,i)}$ be $r - 1$ orthonormal column vectors in the hyperplane

$$\left\{ x \in \mathbb{R}^r \,\middle|\, C_i^{(r)} x = 0 \right\} \tag{3.12}$$

of dimension $r - 1$, obtained using a Gram-Schmidt procedure. Then the vector

$$R_r^{(r,i)} := \frac{\left(C_i^{(r)}\right)^T}{\left|\left|C_i^{(r)}\right|\right|} \det \left(R_1^{(r,i)}, \cdots, R_{r-1}^{(r,i)}, \frac{\left(C_i^{(r)}\right)^T}{\left|\left|C_i^{(r)}\right|\right|} \right) \tag{3.13}$$

is a unit normal vector of the hyperplane Equation (3.12), and the orthogonal matrix

$$R^{(r,i)} := \left(R_1^{(r,i)}, \cdots, R_r^{(r,i)} \right) \tag{3.14}$$

has a determinant equal to 1. Let the matrix $C^{(r-1,i)}$ and the vector $\beta^{(r-1,i)}(y)$ be given by

$$C^{(r-1,i)} := C_{-i}^{(r)} \left(R_1^{(r,i)}, \cdots, R_{r-1}^{(r,i)} \right) \text{ and } \beta^{(r-1,i)}(y) := \beta_{-i}^{(r)}(y) - \frac{C_{-i}^{(r)} R_r^{(r,i)} \beta_i^{(r)}(y)}{\left|\left|C_i^{(r)}\right|\right|}. \tag{3.15}$$

It is then straightforward to verify the relations before and after the rotation and translation. □

Remark 3.1 *For a given row vector $C_i^{(r)}$, the rotation matrix $R^{(r,i)}$ defined in Equation (3.14) is not unique, because the choice of the orthonormal vectors $R_1^{(r,i)}, \cdots, R_{r-1}^{(r,i)}$ is not unique. In this paper, we choose a unique way of implementing the Gram-Schmidt orthonormalization and stick to it to get a unique $R^{(r,i)}$.*

Property 3.1 *In Lemma 3.2,*

(i) *there is an isometry between* $\mathscr{P}_i^{(r-1)}(y)$ *and* $F_i^{(r)}(y)$*, and the same isometry applies between* $H_j^{(r-1,i)}(y)$ *and* $H_{i,j}^{(r)}(y)$*;*

(ii) *the matrix* $C^{(r-1,i)}$ *is constant; if Assumption 3.1 is satisfied, then the elements of the vector* $\beta^{(r-1,i)}(y)$ *are affine with respect to* y*.*

(iii) *the* i*-th row of* $C^{(r)}$ *is linearly independent of all the other rows, if and only if* $C^{(r-1,i)}$ *does not have all-zero rows.*

Proof. (i) The composition of rotation and translation is an isometry.

(ii) According to Notation 3.1 (ii) the matrix $C^{(r)}$ is constant, and by Assumption 3.1 the vector $\beta^{(r)}(y)$ is affine. This property can then be verified from Equation (3.15).

(iii) This property comes from the constructions of $C^{(r-1,i)}$ and $R_1^{(r,i)}, \cdots, R_{r-1}^{(r,i)}$. □

Lemma 3.3 *For* $r \in \{2, 3, \cdots, d\}$*, if the matrix* $\left(C^{(r)}, \beta^{(r)}(y)\right)$ *does not contain any two positively colinear row vectors, then the volume* $V_r(\mathscr{P}^{(r)}(y))$ *is given by*

$$V_r(\mathscr{P}^{(r)}(y)) = \frac{1}{r} \sum_{i=1}^{m-d+r+1} \mathbb{1}_{\left\{\left|\left|C_i^{(r)}\right|\right| \neq 0\right\}} \frac{\beta_i^{(r)}(y)}{||C_i^{(r)}||} V_{r-1}(\mathscr{P}_i^{(r-1)}(y)). \tag{3.16}$$

Otherwise, Equation (3.16) *is modified by keeping in the summation the index of any one of the two-by-two positively colinear row vectors of the matrix* $\left(C^{(r)}, \beta^{(r)}(y)\right)$*.*

Proof. Two positively colinear row vectors of $\left(C^{(r)}, \beta^{(r)}(y)\right)$ correspond to identical inequalities that define $\mathscr{P}^{(r)}(y)$, so it suffices to keep any one of them.

If there is no positive colinearity, we first apply the Lasserre's formula

$$V_r(\mathscr{P}^{(r)}(y)) = \frac{1}{r} \sum_{i=1}^{m-d+r+1} \frac{\beta_i^{(r)}(y)}{||C_i^{(r)}||} V_{r-1}(F_i^{(r)}(y)) \tag{3.17}$$

on $\mathscr{P}^{(r)}(y)$, assuming that $||C_i^{(r)}|| \neq 0$ for any index i. In the case when $\mathscr{P}_i^{(r)}(y)$ is an r-polytope, its volume $V_r(\mathscr{P}_i^{(r)}(y)) \neq 0$. Hence either $V_{r-1}(F_i^{(r)}(y)) \neq 0$ for all indices i and we can apply Proposition 3.3 and Theorem 3.1 in Lasserre (1983) to obtain Equation (3.17), or $V_{r-1}(F_i^{(r)}(y)) = 0$ for some index i and we use Proposition 4.2 in Lasserre (1983). In the degenerate case when the dimension of $\mathscr{P}_i^{(r)}(y)$ is lower than r, its volume $V_r(\mathscr{P}_i^{(r)}(y)) = 0$. By Proposition 4.1 in Lasserre (1983), $V_{r-1}(F_i^{(r)}(y)) = 0$ for all indices i. The equality Equation (3.17) still holds.

If $||C_i^{(r)}|| = 0$ for some index i, then the i-th inequality that defines $\mathscr{P}^{(r)}(y)$ becomes $0 \leq \beta_i^{(r)}(y)$, which is either true for all $x \in \mathbb{R}^r$ or not true for any $x \in \mathbb{R}^r$. If always true, then this equality is redundant for $\mathscr{P}^{(r)}(y)$; if never true, then $F_i^{(r)}(y)$ is the empty set. In either case, the i-th summand does not need to participate in the Lasserre's formula (3.17).

Equation Equation (3.16) is true because of the isometry between $\mathscr{P}_i^{(r-1)}(y)$ and $F_i^{(r)}(y)$ by Property 3.1 (i). □

Notation 3.2 (i) *Starting with* $r = d$ *and an arbitrary* $\mathscr{P}^{(d)}(y)$ *as in Notation 3.1* (ii)*, for* $i_1 \in \{1, \cdots, m+1\}$ *such that* $\left|\left|C_{i_1}^{(d)}\right|\right| \neq 0$*, let* $\mathscr{P}_{i_1}^{(d-1)}(y)$ *be the polytope obtained according to Lemma 3.2 with the expression* (3.8)*.*

(ii) *For* r *being an integer ranging from* $d-1$ *all the way down to 2, and for* $i_1 \in \{1, \cdots, m+1\}$*,* $i_2 \in \{1, \cdots, m\}$*, ...,* $i_{d-r} \in \{1, \cdots, m-d+r+2\}$*, let* $\mathscr{P}_{i_1, \cdots, i_{d-r}}^{(r)}(y)$ *be the polytope*

$$\mathscr{P}_{i_1, \cdots, i_{d-r}}^{(r)}(y) := \left\{ x \in \mathbb{R}^r \,\middle|\, C^{(r, i_1, \cdots, i_{d-r})} x \leq \beta^{(r, i_1, \cdots, i_{d-r})}(y) \right\}. \tag{3.18}$$

(iii) *For the choice of* $i_1, \cdots, i_{d-r}$ *in* (ii)*, and for* $i_{d-r+1} \in \{1, \cdots, m-d+r+2\}$ *such that* $\left|\left|C_{i_{d-r+1}}^{(r,i_1,\cdots,i_{d-r})}\right|\right| \neq 0$*, the hyperplane*

$$H_{i_{d-r+1}}^{(r,i_1,\cdots,i_{d-r})}(y) := \left\{ x \in \mathbb{R}^r \,\middle|\, C_{i_{d-r+1}}^{(r,i_1,\cdots,i_{d-r})} x = \beta_{i_{d-r+1}}^{(r,i_1,\cdots,i_{d-r})}(y) \right\}. \tag{3.19}$$

The set $F_{i_1,\cdots,i_{d-r+1}}^{(r)}(y)$ *defined as*

$$F_{i_1,\cdots,i_{d-r+1}}^{(r)}(y) := \mathscr{P}_{i_1,\cdots,i_{d-r}}^{(r)}(y) \cap H_{i_{d-r+1}}^{(r,i_1,\cdots,i_{d-r})}(y) \tag{3.20}$$

is a face of $\mathscr{P}_{i_1,\cdots,i_{d-r}}^{(r)}(y)$ *of dimension at most* $r-1$*. Let* $\mathscr{P}_{i_1,\cdots,i_{d-r+1}}^{(r-1)}(y)$ *be the polytope obtained from rotating and translating* $F_{i_1,\cdots,i_{d-r+1}}^{(r)}(y)$ *according to Lemma 3.2. The inequality expression of* $\mathscr{P}_{i_1,\cdots,i_{d-r+1}}^{(r-1)}(y)$ *is denoted as in Equation* (3.18) *with* r *replaced by* $r-1$*.*

(iv) *Let* $\mathcal{I}^{(d-1)}(y)$ *be a maximal collection of indices* $(i_1, i_2, \cdots, i_{d-1})$ *in*

$$\{1, \cdots, m+1\} \times \{1, \cdots, m\} \times \cdots \times \{1, \cdots, m-d+3\} \tag{3.21}$$

such that $||C^{(r,i_1,\cdots,i_{d-r})}|| \neq 0$ *and that, for every* $r \in \{2, \cdots, d\}$*, the matrix*

$$\left(C^{(r,i_1,\cdots,i_{d-r})}, \beta^{(r,i_1,\cdots,i_{d-r})}(y)\right) \tag{3.22}$$

does not contain any two positively colinear row vectors. In case there are, the indices corresponding to the first row are kept in $\mathcal{I}^{(d-1)}(y)$*.*

Property 3.2 *Let* r *be an integer in* $\{1, \cdots, d-1\}$*.*

(i) *There exist integers* $1 \leq i'_1 < \cdots < i'_{d-r} \leq m+1$ *irrelevant of* y*, such that there is an isometry between* $\mathscr{P}_{i_1,\cdots,i_{d-r}}^{(r)}(y)$ *and* $F_{i'_1,\cdots,i'_{d-r}}(y)$*. Furthermore, there exists a integer* $i'_{d-r+1} \in \{1, \cdots, m+1\} \setminus \{i'_1, \cdots, i'_{d-r}\}$ *irrelevant of* y*, such that the same isometry applies between* $H_{i_{d-r+1}}^{(r,i_1,\cdots,i_{d-r})}(y)$ *and* $H_{i'_1,\cdots,i'_{d-r+1}}(y)$*. The rows of* $C^{(d)}$ *indexed by* $i'_1, \cdots, i'_{d-r+1}$ *are linearly independent.*

(ii) *The matrices* $C^{(r,i_1,\cdots,i_{d-r})}$ *are constant; if* $\beta^{(d)}(y)$ *satisfies Assumption 3.1, then the elements of the vector* $\beta^{(r,i_1,\cdots,i_{d-r})}(y)$ *are affine with respect to* y*.*

Proof. (i) The composition of isometries, discussed in Notation 3.2 (iii) and Property 3.1 (i), is an isometry. The rotation and translation are performed if and only if $\left|\left|C_{i_{d-r+1}}^{(r,i_1,\cdots,i_{d-r})}\right|\right| \neq 0$. The indexed rows are linearly independent by applying Property 3.1 (iii) at every step.

(ii) We recall from Notation 3.1 (ii) that $C^{(d)}$ is a constant matrix and $\beta^{(d)}(y)$ is a vector affine with respect to y. Inductively applying Property 3.1 (ii) for $r = d-1, \cdots, 1$, the coefficients $C^{(r,i_1,\cdots,i_{d-r})}$ and $\beta^{(r,i_1,\cdots,i_{d-r})}(y)$ are respectively a constant matrix and a vector affine with respect to y. □

Theorem 3.1 *The volume of* $\mathscr{P}^{(d)}(y)$ *is a weighted sum of edge lengths, expressed as*

$$V_d(\mathscr{P}^{(d)}(y)) = \frac{1}{d!} \sum_{\mathcal{I}^{(d-1)}(y)} \frac{\beta_{i_1}^{(d)}(y)}{||C_{i_1}^{(d)}||} \frac{\beta_{i_2}^{(d-1,i_1)}(y)}{||C_{i_2}^{(d-1,i_1)}||} \cdots \frac{\beta_{i_{d-1}}^{(2,i_1,\cdots,i_{d-2})}(y)}{||C_{i_{d-1}}^{(2,i_1,\cdots,i_{d-2})}||} V_1(\mathscr{P}_{i_1,\cdots,i_{d-1}}^{(1)}(y)). \tag{3.23}$$

Proof. We demonstrate Equation (3.23) in the case of no two-by-two positive colinearity. Otherwise, it is a matter of keeping only one of the positively colinear rows as in Lemma 3.3.

Starting with $r = d$, Lemma 3.3 gives

$$V_d(\mathscr{P}^{(d)}(y)) = \frac{1}{d} \sum_{i_1=1}^{m+1} \mathbb{1}_{\left\{\left|\left|C_{i_1}^{(d)}\right|\right| \neq 0\right\}} \frac{\beta_{i_1}^{(d)}(y)}{||C_{i_1}^{(d)}||} V_{d-1}(\mathscr{P}_{i_1}^{(d-1)}(y)). \tag{3.24}$$

For $r = d-1, \cdots, 2$, $i_1 \in \{1, \cdots, m+1\}$, $i_2 \in \{1, \cdots, m\}$, ..., $i_{d-r} \in \{1, \cdots, m-d+r+2\}$, let $\mathscr{P}^{(r)}_{i_1,\cdots,i_{d-r}}(y)$ be the polytope in Equation (3.18). Applying Lemma 3.3 to each $\mathscr{P}^{(r)}_{i_1,\cdots,i_{d-r}}(y)$ gives

$$V_r(\mathscr{P}^{(r)}_{i_1,\cdots,i_{d-r}}(y)) = \frac{1}{r}\sum_{i_{d-r+1}=1}^{m-d+r+1} \mathbb{1}_{\left\{\left|\left|C^{(r,i_1,\cdots,i_{d-r})}_{i_{d-r+1}}\right|\right|\neq 0\right\}} \frac{\beta^{(r,i_1,\cdots,i_{d-r})}_{i_{d-r+1}}(y)}{||C^{(d-r,i_1,\cdots,i_{d-r})}_{i_{d-r+1}}||} V_{r-1}(\mathscr{P}^{(r-1)}_{i_1,\cdots,i_{d-r+1}}(y)). \tag{3.25}$$

Equation Equation (3.24) may be expanded by expressing every r-dimensional volume in terms of $(r-1)$-dimensional volumes as in Equation (3.25), for $r = d-1, \cdots, 2$. Eventually, the d-dimensional volume $V_d(\mathscr{P}^{(d)}(y))$ becomes a weighted sum of edge lengths as in Equation (3.23). □

3.2 Piecewise polynomial volume

This subsection focuses on the volume of $\mathcal{P}(y)$ defined in Notation 3.3 (iv), with the help of the concept of "levels" and their properties in Appendix A.2. Theorem 3.2 characterizes the volume as continuous and piecewise polynomial with respect to y. The volume is a polynomial when y is between two adjacent levels, because by Lemma 3.4 (i) the edge lengths are affine for y within this interval.

Notation 3.3 (i) *For a constant matrix $C \in \mathbb{R}^{m\times d}$ and a constant vector $\beta \in \mathbb{R}^m$,*

$$\mathcal{P} := \{x \in \mathbb{R}^d | Cx \leq \beta\} \tag{3.26}$$

is assumed to be a polytope of dimension d.

(ii) *For $i \in \{1, \cdots, m+1\}$, we define $H_i := \left\{ x \in \mathbb{R}^d \middle| C_i x = \beta_i \right\}$ and $\mathcal{F}_i := \mathcal{P} \cap H_i$ a face of $\mathcal{P}$. For $r \in \{1, \cdots, d-1\}$ and distinct integers $1 \leq i_1, \cdots, i_{d-r} \leq m+1$, the notations*

$$H_{i_1,\cdots,i_{d-r}} := \bigcap_{i\in\{i_1,\cdots,i_{d-r}\}} H_i \text{ and } \mathcal{F}_{i_1,\cdots,i_{d-r}} := \bigcap_{i\in\{i_1,\cdots,i_{d-r}\}} \mathcal{F}_i = \mathcal{P} \cap H_{i_1,\cdots,i_{d-r}} \tag{3.27}$$

respectively represent an intersection of defining hyperplanes and an intersection of faces of $\mathcal{P}$. If the row vectors $\{C_i | \, i = i_1, \cdots, i_{d-r}\}$ are linearly independent, then by Property A.1 (iv)*, $H_{i_1,\cdots,i_{d-r}}$ is of dimension r, and $\mathcal{F}_{i_1,\cdots,i_{d-r}}$ is a face of $\mathcal{P}$ of dimension at most r.*

(iii) *For a constant vector $C_0 \in \mathbb{R}^d$ and a constant real number β_0, the hyperplane $H(y)$ parameterized by the real number y is defined as*

$$H(y) := \left\{ x \in \mathbb{R}^d \middle| C_0 x = \beta_0 + y \right\}. \tag{3.28}$$

The family of hyperplanes $\mathscr{H} := \{H(y) | y \in \mathbb{R}\}$ are parallel to each other.

(iv) *The set*

$$\mathcal{P}(y) := \mathcal{P} \cap H(y)^- = \left\{ x \in \mathbb{R}^d \middle| \begin{pmatrix} C \\ C_0 \end{pmatrix} x \leq \begin{pmatrix} \beta \\ \beta_0 + y \end{pmatrix} \right\} \tag{3.29}$$

is the part of $\mathcal{P}$ below the hyperplane $H(y)$, which is the empty set if $H(y)$ is below $\mathcal{P}$, is identical to $\mathcal{P}$ if $H(y)$ is above $\mathcal{P}$, and by Property A.2 (iv) *is a d-polytope if $H(y)$ cuts $\mathcal{P}$.*

The polytope $\mathcal{P}(y)$ in Notation 3.3 (iv) is an instance of the free placeholder polytope $\mathscr{P}^{(d)}(y)$ in Notation 3.1 (ii) with coefficients

$$C^{(d)} = \begin{pmatrix} C \\ C_0 \end{pmatrix} \text{ and } \beta^{(d)}(y) = \begin{pmatrix} \beta \\ \beta_0 + y \end{pmatrix}. \tag{3.30}$$

It also satisfies Assumption 3.1 for $r = d$. Hence all the procedures and results about $\mathscr{P}^{(d)}(y)$ in Section 3.1 apply to $\mathcal{P}(y)$.

Lemma 3.4 *For y strictly between two adjacent levels in $\mathscr{L}(\mathcal{P}, C_0, \beta_0)$,*

(i) *the length $V_1(\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-1}}(y))$ is affine with respect to y, for the set $\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-1}}(y)$ obtained according to Notation 3.2;*

(ii) *the set of indices $(i_1,\cdots,i_{d-1})$ in $\mathcal{I}^{(d-1)}(y)$ as defined in Notation 3.2* (iv) *such that $V_1(\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-1}}(y)) > 0$ does not depend on the value of y.*

Proof. In this proof, let the value of y be strictly between two adjacent levels.

(i) By Property 3.2 (i), there is an isometry between $\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-1}}(y)$ and $F_{i'_1,\cdots,i'_{d-1}}(y)$, for some integers $1 \le i'_1 < \cdots < i'_{d-1} \le m+1$ irrelevant of y. By Property A.2 (vii, viii), the dimension of $F_{i'_1,\cdots,i'_{d-1}}(y)$ and thus that of $\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-1}}(y)$ is also irrelevant of y. There is the representation

$$F_{i'_1,\cdots,i'_{d-1}}(y) = \begin{cases} \mathcal{F}_{i'_1,\cdots,i'_{d-1}} \cap H(y)^-, \text{ if } i'_{d-1} < m+1; \\ \mathcal{F}_{i'_1,\cdots,i'_{d-2}} \cap H(y), \text{ if } i'_{d-1} = m+1. \end{cases} \tag{3.31}$$

The affine property of the edge length follows from Property A.3.

(ii) For some $r \in \{2,\cdots,d-1\}$, let the two sets of indices $(i_1,\cdots,i_{d-r},j_1,\cdots)$ and $(i_1,\cdots,i_{d-r},j_2,\cdots)$ be such that the one-dimensional polytopes

$$\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-r},j_1,\cdots}(y) \text{ and } \mathscr{P}^{(1)}_{i_1,\cdots,i_{d-r},j_2,\cdots}(y)$$

have positive lengths. Then $\mathscr{P}^{(r)}_{i_1,\cdots,i_{d-r},j_1}(y)$ and $\mathscr{P}^{(r)}_{i_1,\cdots,i_{d-r},j_2}(y)$ are non-empty, because the one-dimensional polytopes are respectively isometric to one of their edges. By Property 3.2 (i), there exist integers $1 \le i'_1 < \cdots < i'_{d-r} \le m+1$ and $j'_1, j'_2 \in \{1,\cdots,m+1\} \setminus \{i'_1,\cdots,i'_{d-r}\}$, all irrelevant of y, such that $H^{(r,i_1,\cdots,i_{d-r})}_{j_1}(y)$ and $H^{(r,i_1,\cdots,i_{d-r})}_{j_2}(y)$ are respectively isometric to $H_{i'_1,\cdots,i'_{d-r},j'_1}(y)$ and $H_{i'_1,\cdots,i'_{d-r},j'_2}(y)$. We assume the existence of some $y_0 \in \mathbb{R}$, such that the j_1-th and the j_2-th rows of the matrix Equation (3.22) are positively colinear for $y = y_0$, which is equivalent to $H_{i'_1,\cdots,i'_{d-r},j'_1}(y_0)$ and $H_{i'_1,\cdots,i'_{d-r},j'_2}(y_0)$ being identical. The positive colinearity of the two rows cannot possibly change with the value of y, except in the following two cases where the representations of $H_{i'_1,\cdots,i'_{d-r},j'_1}(y)$ and $H_{i'_1,\cdots,i'_{d-r},j'_2}(y)$ involve $H(y)$.

(ii.a) If $i'_{d-r} < m+1$ and, without loss of generality, $j'_1 < j'_2 = m+1$, then

$$\begin{aligned} H_{i'_1,\cdots,i'_{d-r},j'_1}(y) =& H_{i'_1,\cdots,i'_{d-r}} \cap H_{j'_1}; \\ H_{i'_1,\cdots,i'_{d-r},j'_2}(y) =& H_{i'_1,\cdots,i'_{d-r}} \cap H(y). \end{aligned} \tag{3.32}$$

The non-empty face $\mathcal{F}_{i'_1,\cdots,i'_{d-r},j'_1} = H_{i'_1,\cdots,i'_{d-r}} \cap H_{j'_1} \cap \mathcal{P}$ of $\mathcal{P}$ is isometric to $\mathscr{P}^{(r)}_{i_1,\cdots,i_{d-r},j_1}(y)$ and is contained in $H(y_0)$. It follows that $H(y_0)$ has to be at one of the levels.

(ii.b) If $i'_{d-r} = m+1$ and $j'_1, j'_2 < m+1$, then

$$\begin{aligned} H_{i'_1,\cdots,i'_{d-r},j'_1}(y) =& H_{i'_1,\cdots,i'_{d-r-1}} \cap H_{j'_1} \cap H(y); \\ H_{i'_1,\cdots,i'_{d-r},j'_2}(y) =& H_{i'_1,\cdots,i'_{d-r-1}} \cap H_{j'_2} \cap H(y). \end{aligned} \tag{3.33}$$

In order that $H_{i'_1,\cdots,i'_{d-r},j'_1}(y_0)$ and $H_{i'_1,\cdots,i'_{d-r},j'_2}(y_0)$ are identical, $H_{j'_1}$ and $H_{j'_2}$ cannot be parallel. For $y = y_0$, it holds that

$$H_{i'_1,\cdots,i'_{d-r-1}} \cap H_{j'_1} \cap H_{j'_2} \cap H(y_0) = H_{i'_1,\cdots,i'_{d-r-1}} \cap H_{j'_2} \cap H(y_0). \tag{3.34}$$

If $H_{j'_1}$ and $H_{j'_2}$ were not identical, then by Property A.1 (iv), equation Equation (3.34) would be of dimension $r-2$ on its left and dimension $r-1$ on its right, which is impossible. □

Theorem 3.2 *The volume $V_d(\mathcal{P}(y))$ as a function of y is continuous and is piecewise polynomial of degree at most d. This function is a polynomial of degree at most d, when y is restricted between any two adjacent values in the augmented levels $\mathscr{L}^\infty(\mathcal{P}, C_0, \beta_0)$ defined in Definition A.6.*

Proof. By the expression (3.29), the coordinates of elements of $\mathcal{P}(y)$ are affine functions of y. This fact plus the continuity of volumes (cf. Eggleston (1958)) ensures the continuity of $V_d(\mathcal{P}(y))$.

By Property A.2 (iii), when $y < p_0$, $\mathcal{P}(y)$ is the empty set with constant volume zero; when $y > p_K$, $\mathcal{P}(y)$ is identical to $\mathcal{P}$ with constant volume $V_d(\mathcal{P})$.

Let y be strictly between two adjacent levels. A formula for $V_d(\mathcal{P}(y))$ is derived in Theorem 3.1. By Equation (3.30), the coefficients $C^{(d)}$ and $\beta^{(d)}(y)$ are respectively a constant matrix and a vector affine with respect to y. By Property 3.2 (ii), so are the coefficients $C^{(r,i_1,\cdots,i_{d-r})}$ and $\beta^{(r,i_1,\cdots,i_{d-r})}(y)$, for $r = d-1,\cdots,2$. Hence in Equation (3.23) the coefficient of each $V_1(\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-1}}(y))$ is a polynomial with respect to y of degree at most $d-1$. By Lemma 3.4, the edge lengths $V_1(\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-1}}(y))$ are affine with respect to y and the summation in Equation (3.23) is over a set of indices irrelevant of y. This concludes the proof. □

Remark 3.2 *In particular for the polytope $\mathcal{P}(y)$ in Notation 3.3* (iv)*, an added value of our approach in this section is to explicitly partition the real line into finitely many constant intervals by the notion of "levels", such that the volume is polynomial when y is within each interval. This result is important also in terms of numerical implementation, now that the levels can be determined "a priori" (cf. Sections 4 and 5, Parts 1, 2 and 3) before volume computation (cf. Sections 4 and 5, Part 4).*

4 An implementable algorithm

This section designs an implementable algorithm to compute the cumulative distribution function $\mathbf{F}(y)$ of the neural network output $G_L(Z_0)$ introduced in Section 2, via volumes of polytopes as in Property B.1. The algorithm consists of four parts that successively computes the polytopes, the vertices, the levels and the volumes. Table 1 lists symbols to appear in the algorithm.

Symbol	Interpretation
$[e_1, e_2, \cdots]$	a list of elements $e_1, e_2, \cdots$
$[L, e]$	a list L with an element e appended
$rowbind(A, B, \cdots)$ $(colbind(A, B, \cdots))$	row (column) concatenation of matrices or vectors $A, B, \cdots$
$nrow(M)$ $(ncol(M))$	number of rows (columns) of a matrix M
$M[i]$	the i-th row (element) of a matrix (vector or list) M
$M[\ , i]$	the i-th column of a matrix M
$M[-i]$ $(M[\ , -i])$	a matrix M with its i-th row (column) deleted
$\mathbf{x}_d$	the vector of length d whose components are identically x
$\mathbf{x}_{d\times d'}$	the $d \times d'$ matrix whose components are identically x
$\leftarrow$	the assignment operator
$q := (q[0], q[1] \cdots, q[d])$	coefficient vector of the polynomial $q(y) := q[0]+\cdots+q[d_0]y^d$

Table 1: Symbols in the algorithm.

Part 1. A partitioning polytope $\mathcal{P}$

Algorithm 4.1 implements Lemma B.1. For any combination $(s_1, \cdots, s_{L-1})$ of values plus or minus one, this part takes in the weight matrices, bias vectors and activation function of the neural network defined in Section 2, and outputs the coefficients that define the polytope

$\mathscr{P}(s_1, \cdots, s_{L-1})$ in Equation (B.2) and the hyperplane $H(y; s_1, \cdots, s_{L-1})$ in Equation (B.21). The output polytope and hyperplane are instances of $\mathcal{P}$ and $H(y)$ defined in Notation 3.3 (i, iii), with $d = d_0$ and $m = m_0 + d_1 + \cdots + d_{L-1}$. Their coefficients $C(s_1, \cdots, s_{L-1})$, $\beta(s_1, \cdots, s_{L-1})$, $C_0(s_1, \cdots, s_{L-1})$ and $\beta_0(s_1, \cdots, s_{L-1})$ correspond to, and thus will be denoted for short as, C, β, C_0 and β_0 in Notation 3.3 (i, iii).

Algorithm 4.1 A partitioning polytope

Input: $s_1 \in \{-1, 1\}^{d_1}, \cdots, s_{L-1} \in \{-1, 1\}^{d_{L-1}}$, $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and $b_l \in \mathbb{R}^{d_l}$, for $l = 1, ..., L$, $C_D \in \mathbb{R}^{m_0 \times d_0}$, $\beta_D \in \mathbb{R}^{m_0}$.
Intermediate: $C_l \in \mathbb{R}^{d_l \times d_0}$ and $\beta_l \in \mathbb{R}^{d_l}$, for $l = 1, \cdots, L-1$
Output: m, $C \in \mathbb{R}^{m \times d_0}$, $\beta \in \mathbb{R}^m$, $C_0 \in \mathbb{R}^{d_0}$, $\beta_0 \in \mathbb{R}$
  $C_1 \leftarrow -\operatorname{diag}(s_1)W_1$, $\beta_1 \leftarrow \operatorname{diag}(s_1)b_1$ {c.f. (B.4)}
  $C \leftarrow rowbind(C_D, C_1)$, $\beta \leftarrow rowbind(\beta_D, \beta_1)$, $m \leftarrow 0$
  for $l = 2, \cdots, L-1$ **do** {c.f. (B.5)}
    $C_l \leftarrow \operatorname{diag}(s_l)W_l \operatorname{diag}(\mathbf{a}(s_{l-1}))C_{l-1}$, $\beta_l \leftarrow \operatorname{diag}(s_l)(W_l \operatorname{diag}(\mathbf{a}(s_{l-1}))\beta_{l-1} + b_l)$
    $C \leftarrow rowbind(C, C_l)$, $\beta \leftarrow rowbind(\beta, \beta_l)$, $m \leftarrow m + d_l$
  $C_0 \leftarrow -W_L \operatorname{diag}(\mathbf{a}(s_{L-1}))C_{L-1}$, $\beta_0 \leftarrow -(W_L \operatorname{diag}(\mathbf{a}(s_{L-1}))\beta_{L-1} + b_L)$ {c.f. (B.6)}
  $m \leftarrow m + m_0$
Return m, C, β, C_0, β_0 {c.f. (B.7)}

Part 2. The vertices of $\mathcal{P}$

Algorithm 4.2 collects all the vertices of a polytope $\mathcal{P}$ in the form of Notation 3.3 (i). Property A.1 (iii, iv) suggest that, a point $v \in \mathbb{R}^{d_0}$ is a vertex of $\mathcal{P}$, if and only if it locates in $\mathcal{P}$ and is the intersection of d_0 defining hyperplanes of $\mathcal{P}$ with linearly independent normal vectors. Let $\mathcal{I}$ be the set of all possible combinations of d_0 elements in $\{1, \cdots, m\}$. Looping over all the combinations $\mathcal{J} \in \mathcal{I}$ for solutions to $\{x \in \mathbb{R}^{d_0} | C_j x = \beta_j, \text{ for all } j \in \mathcal{J}\}$ will give all the vertices, denoted as $\mathscr{V}(\mathcal{P})$ in Definition A.5.

Algorithm 4.2 Vertices list

Input: $C \in \mathbb{R}^{m \times d_0}$, $\beta \in \mathbb{R}^m$
Output: $\mathscr{V}(\mathcal{P})$
  $\mathscr{V}(\mathcal{P}) \leftarrow [\,]$
  for $\mathcal{J} \in \mathcal{I}$ **do**
    if $\{C_j | j \in \mathcal{J}\}$ are linearly independent **then** {the solution is unique}
      $v \leftarrow \{x \in \mathbb{R}^{d_0} | C_j x = \beta_j, \text{ for all } j \in \mathcal{J}\}$
    if $(Cv \le \beta) = \mathbf{1}_{d_0}$ and $v \notin \mathscr{V}(\mathcal{P})$ **then** {vertex of $\mathcal{P}$ not yet collected}
      $\mathscr{V}(\mathcal{P}) \leftarrow [\mathscr{V}(\mathcal{P}), v]$
Return $\mathscr{V}(\mathcal{P})$

Part 3. The levels of $\mathcal{P}$

Algorithm 4.3 computes the levels set $\mathscr{L}(\mathcal{P}, C_0, \beta_0)$ according to Equation (A.3), given the output vertices set $\mathscr{V}(\mathcal{P})$ from Algorithm 4.2.

Algorithm 4.3 Levels list

Input: $\mathscr{V}(\mathcal{P})$ the list of vertices, $C_0 \in \mathbb{R}^{d_0}$ and $\beta_0 \in \mathbb{R}$
Output: $\mathscr{L}(\mathcal{P}, C_0, \beta_0) = \{p_0, \cdots, p_K\}$ with $K+1 \leq$ cardinality of $\mathscr{V}(\mathcal{P})$
  $\mathscr{L}(\mathcal{P}, C_0, \beta_0) \leftarrow$ empty list
  for $v \in \mathscr{V}(\mathcal{P})$ **do**
    $p \leftarrow C_0 v - \beta_0$
    if $p \notin \mathscr{L}(\mathcal{P}, C_0, \beta_0)$ **then** $\mathscr{L}(\mathcal{P}, C_0, \beta_0) \leftarrow [\mathscr{L}(\mathcal{P}, C_0, \beta_0), p]$
Return sorted $\mathscr{L}(\mathcal{P}, C_0, \beta_0)$ in ascending order

Part 4. Volume computation

This part computes the volume $V_{d_0}(\mathcal{P}(y))$ when y varies strictly between any two adjacent levels p_k and p_{k+1} from Algorithm 4.3, for every partitioning polytope $\mathcal{P}$ and cutting hyperplane $H(y)$ from Algorithm 4.1.

We would like to remark that essentially only the coefficients of the edge length (Algorithm 4.6) may change when y crosses levels. It is possible to factor out many other actions in this part, especially the rotation and translation (Algorithm 4.7), to perform them once for all levels. This would be limited, however, by the increased memory size needed to compute the volumes associated with different levels at the same time. This optimization can be implemented given sufficient memory space available. Besides, a parallel implementation for different levels is also possible.

Algorithm 4.4 returns the coefficient vector of the polynomial expression of the volume $V_{d_0}(\mathcal{P}(y))$, by first setting the data structure and then calling the major function *Vol* to be defined in Algorithm 4.5.

The lists $\boldsymbol{C}$, $\boldsymbol{\beta}^a$, $\boldsymbol{\beta}^b$ and V in Algorithm 4.4 provide memory spaces to store the quantities needed to apply Equation (3.23). Algorithms 4.4 to 4.6 share these memories with read and write access. For $d = d_0$ and a generic r-polytope $\mathscr{P}^{(r)}(y)$ defined in Notation 3.1 (ii), the matrix $C^{(r)}$ is stored in $\boldsymbol{C}[r-1]$; the affine vector $\beta^{(r)}(y)$ is stored as $\boldsymbol{\beta}^a[r-1] + \boldsymbol{\beta}^b[r-1]y$. The vector $V[r-1]$ stores the coefficients of a polynomial of degree at most r, which represents the volume $V_r(\mathscr{P}^{(r)}(y)) = \frac{1}{r!}(V[r-1][0] + V[r-1][1]y + \cdots + V[r-1][r]y^r)$.

Algorithm 4.4 The volume of $\mathcal{P}(y)$ for $y \in (p_k, p_{k+1})$

Input: $C \in \mathbb{R}^{m \times d_0}, \beta \in \mathbb{R}^m$, $C_0 \in \mathbb{R}^{d_0}, \beta_0 \in \mathbb{R}, p_k \in \mathbb{R}$ and $p_{k+1} \in \mathbb{R}$
Output: The polynomial coefficients of $V_{d_0}(\mathcal{P}(y))$ for $y \in (p_k, p_{k+1})$
  $\boldsymbol{C} \leftarrow [\mathbf{0}_{m(1)\times 1}, \cdots, \mathbf{0}_{m(d_0-1)\times(d_0-1)}, rowbind(C, C_0)]$ {with $m(r) := m - d_0 + r + 1$}
  $\boldsymbol{\beta}^a \leftarrow [\mathbf{0}_{m(1)}, \cdots, \mathbf{0}_{m(d_0-1)}, rowbind(\beta, \beta_0)]$
  $\boldsymbol{\beta}^b \leftarrow [\mathbf{0}_{m(1)}, \cdots, \mathbf{0}_{m(d_0-1)}, rowbind(\mathbf{0}_{m(d_0-1)}, 1)]$
  $V \leftarrow [\mathbf{0}_2, \cdots, \mathbf{0}_{d_0+1}]$
  $Vol(V, d_0, \boldsymbol{C}, \boldsymbol{\beta}^a, \boldsymbol{\beta}^b, p_k, p_{k+1})$
Return $\frac{1}{d_0!} V[d_0 - 1]$

With its coefficients stored in $\boldsymbol{C}[r-1]$, $\boldsymbol{\beta}^a[r-1]$ and $\boldsymbol{\beta}^b[r-1]$, Algorithm 4.5 defines a recursive function *Vol* to updated the vector $V[r-1]$ with the coefficients of $r!V_r(\mathscr{P}^{(r)}(y))$. Then, the action will depend on the dimension r. If $r = 1$, the function *Vol* gets the coefficients of the affine length of $\mathscr{P}^{(1)}(y)$, by calling the function *1DimVol* to be defined in Algorithm 4.6. If $r > 1$, the function *Vol* implements Equation (3.16) by calling, for each inequality that defines $\mathscr{P}^{(r)}(y)$, the function *Decomposition* to be defined in Algorithm 4.7, *Vol* itself in dimension $r-1$ and the function *Addition* to be defined in Algorithm 4.8.

In the case of $r > 1$, Algorithm 4.5 computes one single summand in Equation (3.16) at a time and adds it to the r-dimensional volume. This design largely reduces the memory spaces needed during the computation. Information about only one polytope in each dimension is saved.

Algorithm 4.5 Implementation of Lemma 3.3

Function: $\mathit{Vol}(V, r, \boldsymbol{C}, \boldsymbol{\beta}^a, \boldsymbol{\beta}^b, p_k, p_{k+1})$
  $V[r-1] \leftarrow \mathbf{0}_{r+1}$
  if $r = 1$ **then** $\mathit{1DimVol}(V[0], \boldsymbol{C}[0], \boldsymbol{\beta}^a[0], \boldsymbol{\beta}^b[0], p_k, p_{k+1})$ {edge length}
  else {dimension higher than one}
    for $i = 0, \cdots, nrow(\boldsymbol{C}[r-1]) - 1$ **do** {c.f. (3.16)}
      if $||\boldsymbol{C}[r-1][i]|| > 0$ **then**
        $\mathit{Decomposition}(i, r, \boldsymbol{C}, \boldsymbol{\beta}^a, \boldsymbol{\beta}^b)$
        $\mathit{Vol}(V, r-1, \boldsymbol{C}, \boldsymbol{\beta}^a, \boldsymbol{\beta}^b, p_k, p_{k+1})$
        $\mathit{Addition}(V, r, [\boldsymbol{\beta}^a[r-1], \boldsymbol{\beta}^b[r-1]]/||\boldsymbol{C}[r-1]||)$

Algorithm 4.6 defines the function *1DimVol* to implement Lemma 3.1. Lemma 3.4 (i) suggests that the length of each $\mathscr{P}^{(1)}_{i_1,\cdots,i_{d-1}}(y)$ is affine for $y \in (p_k, p_{k+1})$. Given that $\boldsymbol{C}[0]$, $\boldsymbol{\beta}^a[0]$ and $\boldsymbol{\beta}^b[0]$ store the coefficients of the generic polytope $\mathscr{P}^{(1)}(y)$ as in Equation (3.1), this function updates a two-dimensional vector $l = (l[0], l[1])$, sharing the memory of $V[0]$, with the coefficients of the affine length. It suffices to calculate the lengths L_0 and L_1 at two different values y_1 and y_2 in (p_k, p_{k+1}). The vector l is then determined by the unique solution $(l[0], l[1])$ to the equations $l[0]+l[1]y_1 = L_0$ and $l[0]+l[1]y_2 = L_1$. In this function, **case**$(C[i] = 0)$ corresponds to $B_0(y[0]) = 0$ in Lemma 3.1, where by Lemma 3.4 (ii) the length is zero for all y; otherwise, once the for loop is finished, there is $\mathrm{MGB} = (B_+(y[0]), B_+(y[1]))$ and $\mathrm{MLB} = (B_-(y[0]), B_-(y[1]))$, hence $L_0 = (\mathrm{MGB}[0] - \mathrm{MLB}[0])^+$ and $L_1 = (\mathrm{MGB}[1] - \mathrm{MLB}[1])^+$. The last two commands in this function assigns the unique solution to $l[0]$ and $l[1]$.

Algorithm 4.6 Implementation of Lemma 3.1 and Lemma 3.4

Function: $\mathit{1DimVol}(l, C, \beta^a, \beta^b, p_k, p_{k+1})$
  $\mathrm{I}_{\mathrm{MGB}} \leftarrow \mathrm{I}_{\mathrm{MLB}} \leftarrow -\mathbf{1}_2$, $\mathrm{MGB} \leftarrow \mathrm{MLB} \leftarrow \mathbf{0}_2$
  $y \in \mathbb{R}^2$ containing two different values in (p_k, p_{k+1})
  for $i = 0, \cdots, \mathrm{length}(C) - 1$ **do**
    for $j = 0, 1$ **do switch**
      case$(C[i] > 0)$
        $\mathrm{MGB}[j] \leftarrow (\mathrm{I}_{\mathrm{MGB}}[j] < 0)\frac{\beta^a[i]+\beta^b[i]y[j]}{C[i]} + (\mathrm{I}_{\mathrm{MGB}}[j] \geq 0)\min(\mathrm{MGB}[j], \frac{\beta^a[i]+\beta^b[i]y[j]}{C[i]})$
        $\mathrm{I}_{\mathrm{MGB}}[j] \leftarrow (\mathrm{MGB}[j] == \frac{\beta^a[i]+\beta^b[i]y[j]}{C[i]})i + (\mathrm{MGB}[j] < \frac{\beta^a[i]+\beta^b[i]y[j]}{C[i]})\mathrm{I}_{\mathrm{MGB}}[j]$
      case$(C[j] < 0)$
        $\mathrm{MLB}[j] \leftarrow (\mathrm{I}_{\mathrm{MLB}}[j] < 0)\frac{\beta^a[i]+\beta^b[i]y[j]}{C[i]} + (\mathrm{I}_{\mathrm{MLB}}[j] \geq 0)\max(\mathrm{MLB}[j], \frac{\beta^a[i]+\beta^b[i]y[j]}{C[i]})$
        $\mathrm{I}_{\mathrm{MLB}}[j] \leftarrow (\mathrm{MLB}[j] == \frac{\beta^a[i]+\beta^b[i]y[j]}{C[i]})i + (\mathrm{MLB}[j] > \frac{\beta^a[i]+\beta^b[i]y[j]}{C[i]})\mathrm{I}_{\mathrm{MLB}}[j]$
      case$(C[i] = 0)$ **if** $\beta^a[i] + \beta^b[i]y[0] < 0$ **then** Exit()
  $l[0] \leftarrow (y[1]\max\{\mathrm{MGB}[0] - \mathrm{MLB}[0], 0\} - y[0]\max\{\mathrm{MGB}[1] - \mathrm{MLB}[1], 0\})/(y[1] - y[0])$
  $l[1] \leftarrow (\max\{\mathrm{MGB}[1] - \mathrm{MLB}[1], 0\} - \max\{\mathrm{MGB}[0] - \mathrm{MLB}[0], 0\})/(y[1] - y[0])$

Algorithm 4.7 implements Lemma 3.2 with the function *Decomposition*. First of all, only the first one of the two-by-two positively colinear rows of $(C^{(r)}, \beta^{(r)}(y))$ (corresponding to identical halfspaces) counts in the Lasserre's formula (3.16); by setting $\boldsymbol{C}[r-2]$ to all zeros, the latter ones do not count. By Lemma 3.4 (ii), it suffices to check positive colinearity for any one value of y in (p_k, p_{k+1}). For the generic r-polytope $\mathscr{P}^{(r)}(y)$ as input to Algorithm 4.5 and the given index i, $\boldsymbol{C}[r-2]$, $\boldsymbol{\beta}^a[r-2]$ and $\boldsymbol{\beta}^b[r-2]$ are updated with the constant matrix $C^{(r-1,i)}$ and the affine vector $\beta^{(r-1,i)}(y)$ from Equation (3.15) (c.f. Property 3.1 (ii)). This prepares Algorithm 4.5 for the calling of *Vol* in dimension $r-1$.

Algorithm 4.7 Implementation of Lemma 3.2

Function: *Decomposition*$(i, r, \boldsymbol{C}, \boldsymbol{\beta}^a, \boldsymbol{\beta}^b)$
  $M \leftarrow colbind(\boldsymbol{C}[r-1], \boldsymbol{\beta}^a[r-1] + \boldsymbol{\beta}^b[r-1](p_k + p_{k+1})/2)$
  for $j = 0, \cdots, i-1$ **do** {remove 2-by-2 positive colinearity}
    if $M[i] = cM[j]$ for some $c > 0$ **then**
      $\boldsymbol{C}[r-2] \leftarrow \mathbf{0}_{nrow(\boldsymbol{C}[r-2]) \times ncol(\boldsymbol{C}[r-2])}$, Exit()
  $R \leftarrow$ matrix $R^{(r,i)}$ from the Proof of Lemma 3.2 {one fixed routine of computation}
  $\boldsymbol{C}[r-2] \leftarrow \boldsymbol{C}[r-1][-i]R[\ , -r]$ {c.f. (3.15)}
  $\boldsymbol{\beta}^a[r-2] \leftarrow \boldsymbol{\beta}^a[r-1][-i] - \boldsymbol{C}[r-1][-i]R[\ , r]\boldsymbol{\beta}^a[r-1][i]/||\boldsymbol{C}[r-1][i]||$
  $\boldsymbol{\beta}^b[r-2] \leftarrow \boldsymbol{\beta}^b[r-1][-i] - \boldsymbol{C}[r-1][-i]R[\ , r]\boldsymbol{\beta}^b[r-1][i]/||\boldsymbol{C}[r-1][i]||$

Algorithm 4.8 adds the current $V_{r-1}(\mathscr{P}_i^{(r-1)}(y))$ weighted by $\frac{\beta_i^{(r)}(y)}{||C_i^{(r)}||}$ to the r-dimensional volume stored as $V[r-1]$. The operations are performed in terms of coefficients of polynomials.

Algorithm 4.8 Adding one summand in the Lasserre's formula (3.16)

Function: *Addition*(V, r, q)
  $V[r-1][0] \leftarrow V[r-1][0] + V[r-2][0]q[0]$, $V[r-1][r] \leftarrow V[r-1][r] + V[r-2][r-1]q[1]$
  for $i = 1, \cdots, r-1$ **do**
    $V[r-1][i] \leftarrow V[r-1][i] + V[r-2][i]q[0] + V[r-2][i-1]q[1]$

5 Numerical example

This section illustrates the algorithm in Section 4 with a simple numerical example. We take $D_0 = [0,1]^2$, Leaky ReLU activation function $a(x) = \frac{1}{100}x\mathbf{1}_{x<0} + x\mathbf{1}_{x\geq 0}$ and a two-layer neural network with weight matrices and bias vectors

$$W_1 = \begin{pmatrix} -1 & 1 \\ -1 & 2 \\ -1 & -1 \end{pmatrix}, \ b_1 = \begin{pmatrix} 0 \\ -\frac{1}{5} \\ 1 \end{pmatrix}, \ W_2 = (1, -1, 1) \text{ and } b_2 = 0.$$

Part 1. A partitioning polytope $\mathcal{P}$

For the combination $s_1 = (1, 1, 1)$, this part finds the expression of the polytope $\mathcal{P} = \mathscr{P}(s_1)$ and Part 2-3 of the algorithm will be illustrated with this polytope. Algorithm 4.1 outputs $m = 7$,

$C_0 = -W_2 \operatorname{diag}(\mathbf{a}(s_1))C_1 = (-1, -2)$, $\beta_0 = -(W_2 \operatorname{diag}(\mathbf{a}(s_1))\beta_1 + b_2) = -\frac{6}{5}$, as well as

$$C = \begin{pmatrix} \mathbf{I}_2 \\ -\mathbf{I}_2 \\ C_1 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_2 \\ -\mathbf{I}_2 \\ -\operatorname{diag}(s_1)W_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \\ 1 & -1 \\ 1 & -2 \\ 1 & 1 \end{pmatrix} \text{ and } \beta = \begin{pmatrix} \mathbf{1}_2 \\ \mathbf{0}_2 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_2 \\ \mathbf{0}_2 \\ \operatorname{diag}(s_1)b_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ -\frac{1}{5} \\ 1 \end{pmatrix}.$$

The first four rows (elements) of C and β set the domain $[0, 1]^2$ and the last three rows of them are constraints from the neural network parameters. The i-th row (element) of C and β defines a hyperplane H_i. Figure 1 (left) shows the polytope $\mathcal{P}$, the hyperplanes $H_1, \cdots, H_7$ and $H(y)$ with $y = 0$.

Part 2. The vertices of $\mathcal{P}$

According to Algorithm 4.2, a point is a vertex of $\mathcal{P}$, if and only if it is the intersection between any two hyperplanes out of $H_1, \cdots, H_7$ and this point is in $\mathcal{P}$. Figure 1 (middle) shows the vertices from intersections between H_5 and another hyperplane. Figure 1 (right) shows all the vertices $\mathscr{V}(\mathcal{P}) = \{(0, 0.1), (0, 1), (0.2, 0.2), (0.5, 0.5)\}$.

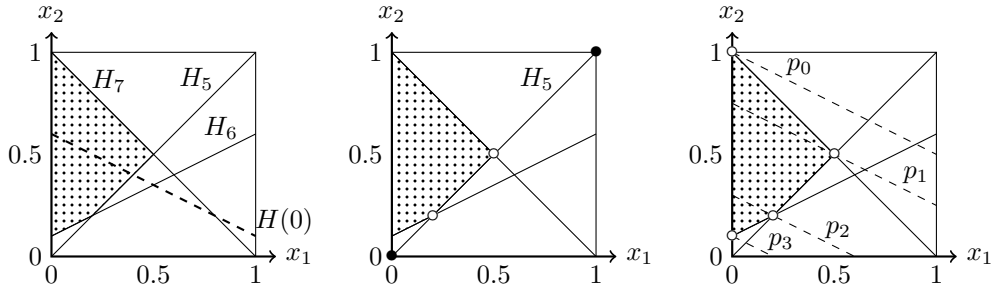

Figure 1: **Left:** The polytope $\mathcal{P}$ (dotted area), with $H_1 := \{x \in \mathbb{R}^2 | x_1 = 1\}$, $H_2 := \{x \in \mathbb{R}^2 | x_2 = 1\}$, $H_3 := \{x \in \mathbb{R}^2 | x_1 = 0\}$, $H_4 := \{x \in \mathbb{R}^2 | x_2 = 0\}$, $H_5 := \{x \in \mathbb{R}^2 | x_1 - x_2 = 0\}$, $H_6 := \{x \in \mathbb{R}^2 | x_1 - 2x_2 = -\frac{1}{5}\}$, $H_7 := \{x \in \mathbb{R}^2 | x_1 + x_2 = 1\}$ and $H(y) := \{x \in \mathbb{R}^2 | -x_1 - 2x_2 = -\frac{6}{5} + y\}$. **Middle:** Intersections (circles) $H_5 \cap H_i$ for $i = 1, 2, 3, 4, 6, 7$. The empty circles are vertices of $\mathcal{P}$ and the solid circles are not. **Right:** Hyperplanes $H(y)$ (dashed lines) that intersect vertices of $\mathcal{P}$ (empty circles).

Part 3. The levels of $\mathcal{P}$

Algorithm 4.3 computes the levels as the sorted values of y such that $H(y)$ intersects at least one vertex. Figure 1 (right) shows all such hyperplanes $H(y)$. Correspondingly, all the levels are $\mathscr{L}(\mathcal{P}, C_0, \beta_0) = \{p_0, p_1, p_2, p_3\} = \{-0.8, -0.3, 0.6, 1\}$.

Part 4. Volume computation

By calling the function *Vol* with $r = 2$, Algorithm 4.4 computes the volume of $\mathcal{P}(y)$ with coefficients $C^{(2)} = \begin{pmatrix} C \\ C_0 \end{pmatrix}$ and $\beta^{(2)}(y) = \begin{pmatrix} \beta \\ \beta_0 + y \end{pmatrix}$, for y between any two adjacent levels. There

is no positive colinearity between any two rows of $(C^{(2)}, \beta^{(2)}(y))$. For $i = 1, \cdots, 8$, this function calls first *Decomposition* for the rotation and translation that produces $\mathscr{P}_i^{(1)}(y)$ and then *Vol* itself with $r = 1$ for the length of $\mathscr{P}_i^{(1)}(y)$. The length is updated by *1DimVol*. The for loop in *Vol* sums up all the eight lengths weighted by $\beta_i^{(2)}(y)/||C_i^{(2)}||$ and updates $V[1]$ with coefficients of this sum. Algorithm 4.4 returns $V[1]/2$, whose components are the coefficients of $V_2(\mathcal{P}(y))$.

Taking $i = 5$ for example, Algorithm 4.7 gives $R^{(2,5)} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$,

$$C^{(1,5)} = C_{-5}^{(2)} R_1^{(2,5)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ 2 \\ -3 \end{pmatrix} \text{ and } \beta^{(1,5)}(y) = \beta_{-5}^{(2)}(y) - \frac{C_{-5}^{(2)} R_2^{(2,5)} \beta_5^{(2)}(y)}{||C_5^{(2)}||} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ -\frac{1}{5} \\ 1 \\ -\frac{6}{5} + y \end{pmatrix}.$$

Besides those apparently redundant ones, the inequalities that define $\mathscr{P}_5^{(1)}(y)$ are $x \le \sqrt{2}/2$, $x \ge \sqrt{2}/5$ and $x \ge (2/5 - y/3)\sqrt{2}$. For $y[0] = (p_1 + p_2)/4$ and $y[1] = 3y[0]$, Algorithm 4.6 computes to get MGB $= (\sqrt{2}/2, \sqrt{2}/2)$, MLB $= (3\sqrt{2}/8, 13\sqrt{2}/40)$, and update $V[0]$ with $(\sqrt{2}/10, \sqrt{2}/3)$. Applying Algorithm 4.6 between every pair of adjacent levels, the edge length is computed as

$$V_1(\mathscr{P}_5^{(1)}(y)) = \begin{cases} 0, \, y < p_1; \\ \dfrac{\sqrt{2}}{10} + \dfrac{\sqrt{2}}{3} y, \, p_1 < y < p_2; \\ \dfrac{3\sqrt{2}}{10}, \, y > p_2. \end{cases}$$

The weighting coefficient of $V_1(\mathscr{P}_5^{(1)}(y))$ is $\beta_5^{(2)}(y)/||C_5^{(2)}|| = 0$.

This algorithm works equally well even in the most complicated case. For example for $i = 8$, $\mathcal{P}_8^{(1)}(y)$ is defined by the inequalities

$$x \le \frac{19}{10\sqrt{5}} + \frac{1}{2\sqrt{5}} y, \, x \ge -\frac{13}{5\sqrt{5}} - \frac{2\sqrt{5}}{5} y, \, x \ge -\frac{3}{5\sqrt{5}} + \frac{\sqrt{5}}{10} y, \, x \le \frac{12}{5\sqrt{5}} - \frac{2\sqrt{5}}{5} y,$$
$$x \le \frac{2}{5\sqrt{5}} - \frac{\sqrt{5}}{15} y, \, x \le \frac{13}{20\sqrt{5}} - \frac{3\sqrt{5}}{20} y \text{ and } x \le \frac{7}{5\sqrt{5}} + \frac{3\sqrt{5}}{5} y,$$

which all involve y. Its length turns out to be

$$V_1(\mathscr{P}_8^{(1)}(y)) = \begin{cases} 0, \, y < p_0 \text{ or } y > p_3; \\ \dfrac{2\sqrt{5}}{5} + \dfrac{\sqrt{5}}{2} y, \, p_0 < y < p_1; \\ \dfrac{\sqrt{5}}{5} - \dfrac{\sqrt{5}}{6} y, \, p_1 < y < p_2; \\ \dfrac{\sqrt{5}}{4} - \dfrac{\sqrt{5}}{4} y, \, p_2 < y < p_3. \end{cases}$$

The weighting coefficient of $V_1(\mathscr{P}_8^{(1)}(y))$ is $\beta_8^{(2)}(y)/||C_8^{(2)}|| = (-6 + 5y)\sqrt{5}/25$.

Applying Algorithm 4.4 between every pair of adjacent levels, together with the continuity

property in Theorem 3.2 at every level, we obtain

$$V_2(\mathcal{P}(y)) = \begin{cases} 0, \ y \le p_0; \\ \frac{1}{4}y^2 + \frac{4}{10}y + \frac{4}{25}, \ p_0 < y \le p_1; \\ -\frac{1}{12}y^2 + \frac{1}{5}y + \frac{13}{100}, \ p_1 < y \le p_2; \\ -\frac{1}{8}y^2 + \frac{1}{4}y + \frac{23}{200}, \ p_2 < y \le p_3; \\ \frac{24}{100}, \ y > p_3. \end{cases}$$

A Polytope preliminaries

Appendix A.1 provides definitions and results needed in Sections 2 and 3. Appendix A.1 lists those about a generic polytope $\mathscr{P}$ from Grünbaum (2003), Ziegler (1995) and Bronsted (1983). Their locations in the books are given in brackets. Appendix A.2 considers the part of a constant polytope cut by a family of parallelly moving parameterized hyperplanes. This is the kind of polytopes that Theorem 2.1 and Theorem 3.2 refer to. Polytopes in this appendix are defined in the Euclidean space $\mathbb{R}^d$, for some positive integer d.

A.1 A generic polytope

Definition A.1 [Section 1.1 on page 2 of Grünbaum (2003)] *A hyperplane H in $\mathbb{R}^d$ is the set $\{x \in \mathbb{R}^d \mid C_0 x = \beta_0\}$, for some row vector $C_0 \in \mathbb{R}^d$ which is not all-zero and some scalar $\beta_0 \in \mathbb{R}$. The sets H^+ and H^- defined by the inequalities*

$$H^+ := \{x \in \mathbb{R}^d \mid C_0 x \ge \beta_0\} \text{ and } H^- := \{x \in \mathbb{R}^d \mid C_0 x \le \beta_0\} \tag{A.1}$$

are called the closed halfspaces respectively above and below H.

Definition A.2 [Definition 0.2 and Theorem 1.1 in Ziegler (1995)] *A polytope has two equivalent definitions as either* (i) *the convex hull of a finite set of points in $\mathbb{R}^d$, or* (ii) *a bounded set obtained as the intersection of a finitely many closed halfspaces in $\mathbb{R}^d$.*

The dimension of a set is the largest number of linearly independent vectors in this set. [Section 1 on page 5 of Bronsted (1983)] *A "d-polytope" is the abbreviation for a "polytope of dimension d".*

This paper takes the inequality definition as in (ii). The polytope $\mathscr{P}$ in this subsection has the expression

$$\mathscr{P} = \bigcap_{1 \le i \le m} H_i^-, \tag{A.2}$$

for some positive integer $m \ge d$ and hyperplanes H_i in $\mathbb{R}^d$, $i = 1, \cdots, m$.

Definition A.3 [Problems and Exercises 2.14 in Ziegler (1995)] *For $j \in \{1, \cdots, m\}$, the halfspace (or inequality) H_j^- is called redundant for $\mathscr{P}$, if $\mathscr{P} = \bigcap_{1 \le i \le m, i \ne j} H_i^-$, and is otherwise called irredundant.*

Definition A.4 *If $\mathscr{P} \cap H \ne \emptyset$, then there are two possible cases:*

(i) [Section 4 on page 25 of Bronsted (1983)] *either* $\max\{C_0 x \mid x \in \mathscr{P}\} = \beta_0$ *or* $\min\{C_0 x \mid x \in \mathscr{P}\} = \beta_0$, *where H is called a supporting hyperplane of $\mathscr{P}$;*

(ii) *there exist x_1, x_2 in $\mathscr{P}$ with $C_0 x_1 > \beta_0$ and $C_0 x_2 < \beta_0$, then we say that H cuts $\mathscr{P}$.*

Definition A.5 [Definition 2.1 in Ziegler (1995)] *A set F is a face of a polytope $\mathscr{P}$, if there exists a hyperplane H such that $\mathscr{P} \subset H^-$ or $\mathscr{P} \subset H^+$, and such that $F = \mathscr{P} \cap H$. A face of dimension r is called an "r-face", for $r \in \{0, 1, \ldots, d\}$. The faces of dimensions 0, 1 and $d-1$ are called vertices, edges and facets. The set of all faces (respectively vertices) of $\mathscr{P}$ is denoted by $\mathscr{F}(\mathscr{P})$ (respectively $\mathscr{V}(\mathscr{P})$).*

Property A.1 *Faces of a polytope $\mathscr{P}$ have the following properties.*

(i) *A polytope $\mathscr{P}$ is the convex hull of its vertices $\mathscr{V}(\mathscr{P})$.* [Theorem 2.15 (3) in Ziegler (1995)]

(ii) *The faces of a polytope are polytopes.* [Proposition 2.3 in Ziegler (1995)]

(iii) *Each r-face of $\mathscr{P}$ is the intersection of $d-r$ facets of $\mathscr{P}$.* [Property 3.1.7 in Grünbaum (2003)]

(iv) *An intersection of faces of $\mathscr{P}$ is a face of $\mathscr{P}$.* [Proposition 2.3 (ii) in Ziegler (1995)] *Especially, for $r \in \{0, \ldots, d\}$ and the collection $\mathcal{I} \subset \{1, \cdots, m\}$ of any $d-r$ indices, if the hyperplanes $\{H_i | i \in \mathcal{I}\}$ have linearly independent normal vectors, then the set $\cap_{i\in\mathcal{I}} H_i$ is non-empty and has dimension r, and the set $\cap_{i\in\mathcal{I}} H_i \cap \mathscr{P}$ is a face of $\mathscr{P}$ of dimension at most r.*

Proof. (iv) The solutions to the system of $d-r$ linear equations given by $\{H_i | i \in \mathcal{I}\}$ are of dimension r, so their intersection with $\mathscr{P}$ is of dimension at most r. □

A.2 Polytope cut by parameterized hyperplanes

This subsection studies the specific polytope $\mathcal{P}(y)$ defined in Notation 3.3 of Section 3, where the constant d-polytope $\mathcal{P}$, the family $\mathscr{H}$ of parallelly moving parameterized hyperplanes $H(y)$ are also defined.

A polytope cut by a hyperplane has been discussed in Bronsted (1983). We shall provide the new definition of "levels" and their properties needed in Sections 2 and 3.

Definition A.6 *The set $\mathscr{L}(\mathcal{P}, C_0, \beta_0)$ of levels of the polytope $\mathcal{P}$ with respect to the family $\mathscr{H}$ of hyperplanes is defined as*

$$\mathscr{L}(\mathcal{P}, C_0, \beta_0) := \{C_0 v - \beta_0 | v \in \mathscr{V}(\mathcal{P})\}, \tag{A.3}$$

which possesses $K+1$ distinct levels $p_0 < p_1 < \cdots < p_K$. The set $\mathscr{L}^\infty(\mathcal{P}, C_0, \beta_0)$ of augmented levels is defined as

$$\mathscr{L}^\infty(\mathcal{P}, C_0, \beta_0) := \mathscr{L}(\mathcal{P}, C_0, \beta_0) \cup \{-\infty, +\infty\}. \tag{A.4}$$

For $k \in \{0, 1, \cdots, K\}$, any $x \in H(p_k)$ is said to be at the k-th level with respect to $(\mathcal{P}, \mathscr{H})$; any $x \in \mathbb{R}^d$ satisfying $C_0 x \geq \beta_0 + p_k$ ($>$, $\leq$ or $<$) is said to be above (strictly above, below or strictly below) the k-th level with respect to $(\mathcal{P}, \mathscr{H})$.

Notation A.1 *For $k \in \{1, \cdots, K\}$, the set $\mathscr{F}_k$ denotes the collection of faces of $\mathcal{P}$ that have at least one vertex above the k-th level and at least one vertex below the $(k-1)$-th level; for $k \in \{0, \cdots, K\}$, the set $\underline{\mathscr{F}}_k$ denotes the collection of faces of $\mathcal{P}$ below the k-th level.*

The well-definedness of $\{p_k\}_{k=0}^K$, $\{\mathscr{F}_k\}_{k=1}^K$, and $\{\underline{\mathscr{F}}_k\}_{k=0}^K$ are verified in the following property.

Property A.2 (i) *The number $(K+1)$ of levels is equal to at least two and at most the cardinal of $\mathscr{V}(\mathcal{P})$.*

(ii) *The value p_0 (respectively, p_K) is the smallest (respectively, the largest) real value of y such that $\mathcal{P} \cap H(y) \neq \emptyset$.*

(iii) *$\mathcal{P}(y)$ is empty when $y < p_0$ and is identical to $\mathcal{P}$ when $y > p_K$.*

(iv) *When $p_0 < y < p_K$, $\mathcal{P}(y)$ is a d-polytope.*

(v) *None of the sets $\mathscr{F}_1, \cdots, \mathscr{F}_K, \underline{\mathscr{F}}_0, \cdots, \underline{\mathscr{F}}_K$ is empty.*

(vi) *For $k \in \{1, \cdots, K\}$, when $p_{k-1} < y < p_k$, the set of faces of $\mathcal{P}$ cut by $H(y)$ is $\mathscr{F}_k$; the set of faces of $\mathcal{P}$ below $H(y)$ is $\underline{\mathscr{F}}_{k-1}$.*

(vii) *Let F be a face of $\mathcal{P}$. For any $k \in \{1, \cdots, K\}$ and any $y \in (p_{k-1}, p_k)$, $F \cap H(y)$ is empty if and only if $F \notin \mathscr{F}_k$; $F \cap H(y)^-$ is empty if $F \notin \mathscr{F}_k \cup \underline{\mathscr{F}}_{k-1}$, is identical to F if $F \in \underline{\mathscr{F}}_{k-1}$, and is neither the empty set nor F if $F \in \mathscr{F}_k$.*

(viii) *For $r \in \{1, \cdots, d-1\}$, $k \in \{1, \cdots, K\}$ and a given $y \in (p_{k-1}, p_k)$, an r-face of $\mathcal{P}(y)$ is either $F \cap H(y)$ for some $(r+1)$-face F in $\mathscr{F}_k$, or $F \cap H(y)^- \subsetneq H(y)$ for some r-face F in $\mathscr{F}_k \cup \underline{\mathscr{F}}_{k-1}$; vice versa. In either case, the choice of the face F is unique.*

Proof. (i) For any $H(y)$ cutting $\mathcal{P}$, there is at least one point in $\mathscr{V}(\mathcal{P})$ below $H(y)$ and one above $H(y)$. The two vertices are at different levels. From the definition of levels, their number is smaller than or equal to the number of vertices.

(ii) Having a part of $\mathcal{P}$ below the value p_0 or above the value p_K leads to a contradiction to Property A.1 (i).

(iii) It follows straightforwardly from (ii).

(iv) This is because of Theorem 11.11 (a) in Bronsted (1983).

(v) $\underline{\mathscr{F}}_k \neq \emptyset$ is trivial since it contains at least the vertex of its level (p_k). Regarding $\mathscr{F}_k \neq \emptyset$: for any $k = 1, \cdots, K$, by the convexity of $\mathcal{P}$, there exists at least an edge connecting one vertex v_i of the i-th level when $i \geq k$ to one vertex v_j of the j-th level when $j < k$. The edge connecting v_k to v_j belongs to $\mathscr{F}_k$.

(vi) By Property A.1 (ii), a face of $\mathcal{P}$ is a polytope. This statement follows from Definition A.4 and the definitions of $\mathscr{F}_k$ and $\underline{\mathscr{F}}_{k-1}$ in Notation A.1.

(vii) This follows from (vi).

(viii) For such a value of y, the hyperplane $H(y)$ cuts the polytope $\mathcal{P}$ and does not contain any non-empty face of $\mathcal{P}$. This statement follows from (vi), Theorem 11.1 (b, d) and Theorem 11.11 (b, c) in Bronsted (1983). □

Property A.3 *Let $E(y)$ be a face of $\mathcal{P}(y)$, represented as either*

(i) *$E(y) = F \cap H(y)^-$ for an edge F of $\mathcal{P}$, or*

(ii) *$E(y) = F \cap H(y)$ for a 2-face F of $\mathcal{P}$, if $d \geq 2$.*

When y varies between two adjacent levels in $\mathscr{L}(\mathcal{P}, C_0, \beta_0)$, the length $V_1(E(y))$ as in Definition 3.1 is affine with respect to y.

Proof. We suppose that $p_{k-1} < y < p_k$ for some $k \in \{1, \cdots, K\}$.

(i) By Property A.2 (vii), the value of $V_1(F \cap H(y)^-)$ is irrelevant of y when $F \notin \mathscr{F}_k$. It remains to show that $V_1(F \cap H(y)^-)$ is affine with respect to y when $F \in \mathscr{F}_k$.

Let v_1 and v_2 be the two vertices of F which are respectively below and above $H(y)$, at the k_1-th and the k_2-th level. Then the edge F has the parametric expression

$$F = \{v_1 + t(v_2 - v_1) | 0 \leq t \leq 1\}. \tag{A.5}$$

Let $t(y)$ be the value of t such that $v_1 + t(y)(v_2 - v_1) = F \cap H(y)$. The value $t(y)$ satisfies

$$C_0(v_1 + t(y)(v_2 - v_1)) = \beta_0 + y \Longleftrightarrow t(y) = \frac{y + \beta_0 - C_0 v_1}{C_0 v_2 - C_0 v_1} = \frac{y - p_{k_1}}{p_{k_2} - p_{k_1}}. \tag{A.6}$$

The edge length $V_1(F \cap H(y)^-) = t(y)||v_2 - v_1||$ is affine with respect to y.

(ii) By Property A.2 (vii, viii), $V_1(F \cap H(y))$ is non-zero if and only if $F \in \mathscr{F}_k$. It suffices to restrict the discussions to the 2-dimensional space spanned by F, as in Figure 2.

Let E_1 and E_2 be the two edges of F that intersect $H(y)$, with respective vertices $\{\nu_{11}, \nu_{12}\}$ and $\{\nu_{21}, \nu_{22}\}$. If the two lines containing E_1 and E_2 are parallel, then $V_1(F \cap H(y))$ is a constant equal to the distance between E_1 and E_2. Otherwise, let P_0 denote the intersection of the two lines containing E_1 and E_2. Let $P_1(y)$ and $P_2(y)$ denote the points where $H(y)$ intersects E_1 and E_2. Without loss of generality, the point P_0 is assumed to be on the same side of $F \cap H(y)$ as ν_{12} and ν_{22}, and the line going through ν_{11} and parallel to $F \cap H(y)$ is assumed to intersect

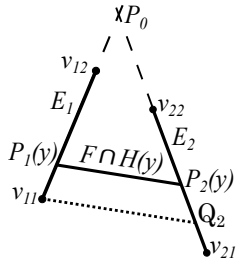

Figure 2: Property A.3 (ii), the case $E(y) = F \cap H(y) \neq \emptyset$.

E_2 at the point Q_2. The triangle $P_0P_1(y)P_2(y)$ is similar to the triangle $P_0\nu_{11}Q_2$. Then their edge lengths have the relation

$$V_1(F \cap H(y)) = \frac{V_1(\nu_{11}Q_2)}{V_1(P_0\nu_{11})} V_1(P_0P_1(y)) = \frac{V_1(\nu_{11}Q_2)}{V_1(P_0\nu_{11})} \left(V_1(P_0\nu_{11}) - V_1(P_1(y)\nu_{11})\right). \tag{A.7}$$

In Equation (A.7), the lengths of the line segments $\nu_{11}Q_2$, $P_0\nu_{11}$ and E_1 are constant numbers determined by the polytope $\mathcal{P}$. By (i), the length of the part of E_1 below $H(y)$ is affine with respect to y, so $V_1(P_1(y)\nu_{11})$ is affine with respect to y, whether it is below or above $H(y)$. It follows that the edge length $V_1(F \cap H(y))$ is affine with respect to y. □

B An expression of the cumulative distribution function

In this appendix, Lemma B.1 explicitly expresses the piecewise decomposition of Lemma 2.1 in terms of the parameters of the neural network; based on this explicit expression, Property B.1 provides a formula of the cumulative distribution function $\mathbf{F}(y)$ as a sum of the volumes of finitely many cut polytopes; Notation B.1 defines the set of partitioning levels.

Lemma B.1 *The neural network G_L has the affine expression*

$$G_L(z_0) = C_0(s_1, \cdots, s_{L-1})z_0 - \beta_0(s_1, \cdots, s_{L-1}), \tag{B.1}$$

when z_0 is in the polytope

$$\mathcal{P}(s_1, \cdots, s_{L-1}) := \left\{ z_0 \in \mathbb{R}^{d_0} \,\middle|\, C(s_1, \cdots, s_{L-1})z_0 \leq \ \beta(s_1, \cdots, s_{L-1}) \right\}, \tag{B.2}$$

for any set of $s_1 \in \{-1, 1\}^{d_1}, \cdots, s_{L-1} \in \{-1, 1\}^{d_{L-1}}$. The collection

$$\mathscr{C} := \left\{ \mathcal{P}(s_1, \cdots, s_{L-1}) \,\middle|\, s_l \in \{-1, 1\}^{d_l},\ l = 1, \cdots, L-1 \right\} \tag{B.3}$$

is a finite partition of the polytope D_0, except for possible overlaps on the boundaries of the polytopes. The coefficients in Equations (B.1) *and* (B.2) *are constant for given values of $(s_1, \cdots, s_{L-1})$, and can be computed explicitly according to the following procedure.*

(1) Initializing with

$$C_1(s_1) := -\operatorname{diag}(s_1)W_1 \text{ and } \beta_1(s_1) := \operatorname{diag}(s_1)b_1. \tag{B.4}$$

(2) If $L \geq 3$, for $l = 2, \cdots, L-1$, defining iteratively

$$\begin{cases} C_l(s_1, \cdots, s_l) := \operatorname{diag}(s_l) W_l \operatorname{diag}(\mathbf{a}(s_{l-1})) C_{l-1}(s_1, \cdots, s_{l-1}); \\ \beta_l(s_1, \cdots, s_l) := \operatorname{diag}(s_l) \left(W_l \operatorname{diag}(\mathbf{a}(s_{l-1})) \beta_{l-1}(s_1, \cdots, s_{l-1}) + b_l \right), \end{cases} \tag{B.5}$$

where $\operatorname{diag}(u)$ *indicates the diagonal matrix with the vector* u *being its diagonal elements. Then the coefficients have the expressions*

$$\begin{cases} C_0(s_1, \cdots, s_{L-1}) = -W_L \operatorname{diag}(\mathbf{a}(s_{L-1})) C_{L-1}(s_1, \cdots, s_{L-1}); \\ \beta_0(s_1, \cdots, s_{L-1}) = -\left(W_L \operatorname{diag}(\mathbf{a}(s_{L-1})) \beta_{L-1}(s_1, \cdots, s_{L-1}) + b_L \right), \end{cases} \tag{B.6}$$

$$C(s_1, \cdots, s_{L-1}) = \begin{pmatrix} C_D \\ C_1(s_1) \\ \vdots \\ C_{L-1}(s_1, \cdots, s_{L-1}) \end{pmatrix} \text{ and } \beta(s_1, \cdots, s_{L-1}) = \begin{pmatrix} \beta_D \\ \beta_1(s_1) \\ \vdots \\ \beta_{L-1}(s_1, \cdots, s_{L-1}) \end{pmatrix}. \tag{B.7}$$

Proof. We recall the inductive definition Equation (2.6) of the neural network as a composition of affine mappings and the Leaky ReLU activation function defined in Equation (2.3).

Let $sign : \mathbb{R} \to \{-1, 1\}$ be the sign function, which takes the value one on the non-negative real line and the value minus one elsewhere. The sign of a d-dimensional real vector $u = (u_1, \cdots, u_d)^T$ is defined as $\mathbf{sign}(u) := (sign(u_1), \cdots, sign(u_d))^T$. The identity

$$\mathbf{a}(u) = \operatorname{diag}(\mathbf{a}(\mathbf{sign}(u))) \operatorname{diag}(\mathbf{sign}(u)))u \tag{B.8}$$

will be needed in the derivation of the piecewise affine expression for $G_L(\cdot)$.

Hidden layer 1: For any $s_1 \in \{-1, 1\}^{d_1}$, when the input z_0 satisfies $C_1(s_1)z_0 \leq \beta_1(s_1)$, or equivalently $\mathbf{sign}(W_1 z_0 + b_1) = s_1$ with possible exceptions only for the zero elements of the argument of the **sign** mapping, the first hidden layer can be expressed as

$$G_1(z_0) = \mathbf{a}(W_1 z_0 + b_1) = \operatorname{diag}(\mathbf{a}(s_1)) \operatorname{diag}(s_1)(W_1 z_0 + b_1) = \bar{W}_1(s_1) z_0 + \bar{b}_1(s_1), \tag{B.9}$$

where

$$\begin{cases} \bar{W}_1(s_1) := \operatorname{diag}(\mathbf{a}(s_1)) \operatorname{diag}(s_1) W_1 = -\operatorname{diag}(\mathbf{a}(s_1)) C_1(s_1); \\ \bar{b}_1(s_1) := \operatorname{diag}(\mathbf{a}(s_1)) \operatorname{diag}(s_1) b_1 = \operatorname{diag}(\mathbf{a}(s_1)) \beta_1(s_1). \end{cases} \tag{B.10}$$

In the case that $L \geq 3$, there are more than one hidden layers, and mathematical induction will be conducted. For $l = 2, \cdots, L-1$, the notations

$$\begin{cases} \widetilde{W}_l(s_1, \cdots, s_{l-1}) := W_l \bar{W}_{l-1}(s_1, \cdots, s_{l-1}); \\ \widetilde{b}_l(s_1, \cdots, s_{l-1}) := W_l \bar{b}_{l-1}(s_1, \cdots, s_{l-1}) + b_l; \\ \bar{W}_l(s_1, \cdots, s_l) := \operatorname{diag}(\mathbf{a}(s_l)) \operatorname{diag}(s_l) \widetilde{W}_l(s_1, \cdots, s_{l-1}); \\ \bar{b}_l(s_1, \cdots, s_l) := \operatorname{diag}(\mathbf{a}(s_l)) \operatorname{diag}(s_l) \widetilde{b}_l(s_1, \cdots, s_{l-1}) \end{cases} \tag{B.11}$$

will simplify expressions in the induction. Matrices $C_l(s_1, \cdots, s_l)$ and vectors $\beta_l(s_1, \cdots, s_l)$ defined in Equation (B.4) and Equation (B.5) satisfy

$$\begin{cases} C_l(s_1, \cdots, s_l) = -\operatorname{diag}(s_l) \widetilde{W}_l(s_1, \cdots, s_{l-1}); \\ \beta_l(s_1, \cdots, s_l) = \operatorname{diag}(s_l) \widetilde{b}_l(s_1, \cdots, s_{l-1}); \\ \bar{W}_l(s_1, \cdots, s_l) = -\operatorname{diag}(\mathbf{a}(s_l)) C_l(s_1, \cdots, s_l); \\ \bar{b}_l(s_1, \cdots, s_l) = \operatorname{diag}(\mathbf{a}(s_l)) \beta_l(s_1, \cdots, s_l). \end{cases} \tag{B.12}$$

Hidden layer $l-1$: Based on the expression of the first hidden layer, for $l = 2, \cdots, L-1$ and $s_1 \in \{-1,1\}^{d_1}, \cdots, s_{l-1} \in \{-1,1\}^{d_{l-1}}$, we may assume that

$$G_{l-1}(z_0) = \bar{W}_{l-1}(s_1, \cdots, s_{l-1})z_0 + \bar{b}_{l-1}(s_1, \cdots, s_{l-1}), \tag{B.13}$$

when z_0 satisfies

$$C_1(s_1)z_0 \le \beta_1(s_1), \cdots, C_{l-1}(s_1, \cdots, s_{l-1})z_0 \le \beta_{l-1}(s_1, \cdots, s_{l-1}). \tag{B.14}$$

Hidden layer l: For $l = 2, \cdots, L-1$ and $s_1 \in \{-1,1\}^{d_1}, \cdots, s_{l-1} \in \{-1,1\}^{d_{l-1}}$, by the assumption on the $(l-1)$-th hidden layer, the l-th hidden layer can be expressed as

$$G_l(z_0) = \mathbf{a}(W_l G_{l-1}(z_0) + b_l) = \mathbf{a}(\widetilde{W}_l(s_1, \cdots, s_{l-1})z_0 + \widetilde{b}_l(s_1, \cdots, s_{l-1})), \tag{B.15}$$

when inequalities Equation (B.14) hold. In addition to inequalities Equation (B.14), for any $s_l \in \{-1,1\}^{d_l}$, when z_0 also satisfies $C_l(s_1, \cdots, s_l)z_0 \le \beta_l(s_1, \cdots, s_l)$, or equivalently $\mathbf{sign}(\widetilde{W}_l(s_1, \cdots, s_{l-1})z_0 + \widetilde{b}_l(s_1, \cdots, s_{l-1})) = s_l$ with possible exceptions only for the zero elements of the argument of the **sign** mapping, the second identity in Equation (B.15) can be further expanded as

$$\begin{aligned} G_l(z_0) =& \operatorname{diag}(\mathbf{a}(s_l)) \operatorname{diag}(s_l)(\widetilde{W}_l(s_1, \cdots, s_{l-1})z_0 + \widetilde{b}_l(s_1, \cdots, s_{l-1})) \\ =& \bar{W}_l(s_1, \cdots, s_l)z_0 + \bar{b}_l(s_1, \cdots, s_l). \end{aligned} \tag{B.16}$$

Hidden layer $L-1$: By induction on the layers, we know that

$$G_{L-1}(z_0) = \bar{W}_{L-1}(s_1, \cdots, s_{L-1})z_0 + \bar{b}_{L-1}(s_1, \cdots, s_{L-1}), \tag{B.17}$$

when z_0 satisfies

$$C_1(s_1)z_0 \le \beta_1(s_1), \cdots, C_{L-1}(s_1, \cdots, s_{L-1})z_0 \le \beta_{L-1}(s_1, \cdots, s_{L-1}). \tag{B.18}$$

The output layer L: The output layer is a linear combination of the $(L-1)$-th hidden layer and can be expressed as

$$G_L(z_0) = W_L G_{L-1}(z_0) + b_L = C_0(s_1, \cdots, s_{L-1})z_0 - \beta_0(s_1, \cdots, s_{L-1}), \tag{B.19}$$

when inequalities Equation (B.18) are satisfied. Given that the input domain is confined to D_0 as in Assumption 2.1, we conclude that the expression (B.1) is valid on each polytope $\mathcal{P}(s_1, \cdots, s_{L-1})$ defined as in Equation (B.2). The construction of the polytopes justifies the fact that the collection $\mathscr{C}$ defined in Equation (B.3) is a partition. □

Property B.1 *With the piecewise affine representation of G_L in Lemma B.1, the cumulative distribution function $\mathbf{F}(y)$ has the expression*

$$\mathbf{F}(y) = \frac{1}{V_{d_0}(D_0)} \sum_{\substack{(s_1, \cdots, s_{L-1}) \ in \\ \{-1,1\}^{d_1} \times \cdots \times \{-1,1\}^{d_{L-1}}}} V_{d_0}(\mathscr{P}(y; s_1, \cdots, s_{L-1})), \tag{B.20}$$

where

$$\begin{aligned} H(y; s_1, \cdots, s_{L-1}) :=& \left\{ z_0 \in \mathbb{R}^{d_0} \,\middle|\, C_0(s_1, \cdots, s_{L-1})z_0 = \beta_0(s_1, \cdots, s_{L-1}) + y \right\}; \\ \mathscr{P}(y; s_1, \cdots, s_{L-1}) :=& \mathscr{P}(s_1, \cdots, s_{L-1}) \cap H^-(y; s_1, \cdots, s_{L-1}). \end{aligned} \tag{B.21}$$

Proof. By Definition 3.1, $\mathbf{F}(y)$ is the integration of the uniform density on the domain $\{G_L(x) \le y\} \cap D_0$, thus equal to the volume of the domain. This volume equals the sum of volumes of $\mathscr{P}(y; s_1, \cdots, s_{L-1})$, which have disjoint interiors and the union $\{G_L(x) \le y\} \cap D_0$. □

Notation B.1 *With the piecewise affine representation of G_L in Lemma B.1 and the definition of levels in Definition A.6, the set $\mathscr{O}$ of partitioning levels is defined as*

$$\mathscr{O} := \bigcup_{\substack{(s_1, \cdots, s_{L-1}) \ in \\ \{-1,1\}^{d_1} \times \cdots \times \{-1,1\}^{d_{L-1}}}} \mathscr{L}^\infty \left(\mathscr{P}(s_1, \cdots, s_{L-1}), C_0(s_1, \cdots, s_{L-1}), \beta_0(s_1, \cdots, s_{L-1}) \right). \tag{B.22}$$

References

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research 18*, 1–53.

Bronsted, A. (1983). *An Introduction to Convex Polytopes*. Graduate Texts in Mathematics 90. Springer–Verlag.

Büeler, B., A. Enge, and K. Fukuda (2000). Exact volume computation for polytopes: A practical study. In G. Kalai and G. M. Ziegler (Eds.), *Polytopes – Combinatorics and Computation*, Basel, pp. 131–154. Birkhäuser.

Bunel, R., I. Turkaslan, P. H. Torr, P. Kohli, and M. P. Kumar (2018). A unified view of piecewise linear neural network verification. *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 4795–4804.

Croitoru, F.-A., V. Hondru, R. T. Ionescu, and M. Shah (2023+). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, early access article*.

Dinh, L., J. Sohl-Dickstein, and S. Bengio (2017). Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Eggleston, H. G. (1958). *Convexity*. Cambridge Tracts in Mathematics and Mathematical Physics. Cambridge University Press.

Gresele, L., G. Fissore, A. Javaloy, B. B. Schölkopf, and A. Hyvarinen (2020). Relative gradient optimization of the Jacobian term in unsupervised deep learning. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 16567–16578. Curran Associates, Inc.

Grünbaum, B. (2003). *Convex Polytopes* (2nd ed.). Graduate Texts in Mathematics 221. Springer–Verlag.

Gui, J., Z. Sun, Y. Wen, D. Tao, and J. Ye (2023). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering 35*(4), 3313–3332.

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

Kidger, P. and T. Lyons (2020). Universal approximation with deep narrow networks. In J. Abernethy and S. Agarwal (Eds.), *Proceedings of Thirty Third Conference on Learning Theory*, Volume 125 of *Proceedings of Machine Learning Research*, pp. 2306–2327. ML Research Press.

Lasserre, J. B. (1983). An analytical expression and an algorithm for the volume of a convex polyhedron in $\mathbb{R}^d$. *Journal of Optimization Theory and Applications 39*(3), 363–377.

Pan, T., W. Pedrycz, J. Cui, J. Yang, and W. Wu (2022). Interpretability of neural networks with probability density functions. *Advanced Theory and Simulations 5*(3), 2100459.

Poggio, T. and Q. Liao (2018). Theory I: Deep networks and the curse of dimensionality. *Bulletin of the Polish Academy of Sciences – Technical Sciences 66*(6), 761–773.

Păsăreanu, C., H. Converse, A. Filieri, and D. Gopinath (2020). On the probabilistic analysis of neural networks. In *SEAMS'20: Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self–Managing Systems*, pp. 5–8. Association for Computing Machinery, New York.

Tao, Q., L. Li, X. Huang, X. Xi, S. Wang, and J. A. K. Suykens (2022). Piecewise linear neural networks and deep learning. *Nature Reviews Methods Primers 2*, 43.

Ziegler, G. M. (1995). *Lectures on Polytopes*. Graduate Texts in Mathematics 152. Springer–Verlag.