

Representing Piecewise-Linear Functions by Functions with Minimal Arity

Christoph Koutschan, Anton Ponomarchuk, and Josef Schicho

Abstract—Any continuous piecewise-linear function $F: \mathbb{R}^n \rightarrow \mathbb{R}$ can be represented as a linear combination of maxima of at most $n + 1$ affine-linear functions. This upper bound is sharp, that is, for every n there exists such a function that is not a linear combination of maxima of n arguments. In this paper, we give a method to determine, for any given such F , the minimal number k such that F can be written as a linear combination of maxima of k arguments.

Index Terms—Continuous piecewise linear function, neural network, machine learning, ReLU activation function.

I. INTRODUCTION

A FUNCTION $F: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous piecewise-linear (CPWL) if it can be written as a composition of affine-linear functions and the max function. It is also true that every CPWL function can be written as a linear combination of maxima of affine-linear functions. For some applications, it is important that the number of arguments of the max summands is small. For instance, [1] suggests the vector space generated by maxima of *two* affine-linear functions as convenient function space for regression, classification, and functional approximation. Even more importantly, the class of CPWL functions is exactly the class of functions that can be computed by neural networks with ReLU activation function, see [2]. By using a result from [3] that any CPWL function F can be written as a linear combination of maxima of at most $n + 1$ affine-linear functions, the authors of [2] show that F can be computed by a ReLU neural network with $\lceil \log_2(n + 1) \rceil + 1$ hidden layers. Moreover, in [4], this upper bound was improved to $\lceil \log_3(n - 1) \rceil + 1$ hidden layers.

In the case of representing F as a linear combination of maxima, the upper bound of arguments for max functions is tight, in the sense that there exist functions which do not allow a representation as a linear combination of maxima with fewer arguments. This result was obtained independently by two different research groups [5], [6]. Both use the witness function $\max(0, x_1, \dots, x_n)$, which cannot be represented as a linear combination of maxima with less than $n + 1$ arguments. However, the function $\max(0, x_1, x_2) + \max(0, -x_1, -x_2)$ can be represented as a linear combination of maxima with two arguments, see Example I.1. The representation is not unique.

C. Koutschan is with the Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria

A. Ponomarchuk is with the Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria

J. Schicho is with the Research Institute for Symbolic Computation, Johannes Kepler University, Hagenberg im Mühlkreis, Austria.

The research reported in this paper has been partly funded by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG in the COMET Module S3AI.

This leads to the question: For a given CPWL function F , what is the minimal number of arguments of maxima that is necessary to represent the function F ? This paper gives an answer to the question above.

A related problem—not treated in this paper—is to represent a CPWL function as a difference of two convex CPWL functions satisfying a certain optimality criterion. This question has been answered in [7] for $n = 2$ and partially by [8] for $n > 2$. Both papers incorporate the duality between the CPWL function F and polyhedral geometry. For analysing ReLU neural networks and their capacity to represent CPWL functions, tropical geometry has been used, for instance in [2], [5], [9]–[12].

This paper is structured as follows. In Section II, we provide basic information from polyhedral geometry [13]. Also, we recall the notion of piecewise-constant functions necessary to formulate and prove Theorem III.5, following [5]. In Section III, we introduce the concept of critical flag length, and we prove that a CPWL function can be represented as a linear combination of max functions with at most $k + 1$ arguments if and only if k is bigger than or equal to the critical flag length. We also give an algorithm that computes the critical flag length and a representation with the minimal number of arguments.

Example I.1. We consider the two functions $F_1(x, y) := \max(0, x, y)$ and $F_2(x, y) := \max(0, -x, -y)$, that are defined on $\mathbb{R}^2$, see Fig. 1. For the sum of the functions F_1 and F_2 the following equality holds:

$$\begin{aligned} \max(0, x, y) + \max(0, -x, -y) = \\ \max(0, x, y, -x, -y, x - y, y - x). \end{aligned} \quad (1)$$

By [5], [6], [14], the functions F_1 and F_2 cannot be represented as a linear combination of max functions with two arguments. In contrast, the function $F_1 + F_2$, see Fig. 2, can be represented as a linear combination of max functions with two arguments:

$$\begin{aligned} (F_1 + F_2)(x, y) = \\ \max(x, y) + \max(-y, x - y) + \max(-x, y - x). \end{aligned} \quad (2)$$

It suffices to show that the right-hand sides of Eqs. (1) and (2) are equivalent. Indeed, by applying basic algebraic transformations, we get:

$$\begin{aligned} \max(x, y) + \max(-y, x - y) + \max(-x, y - x) = \\ \max(-x, y - x) + \max(0, x, 2x - y, x - y) = \\ \max(0, x, y, -x, -y, x - y, y - x). \end{aligned} \quad (3)$$

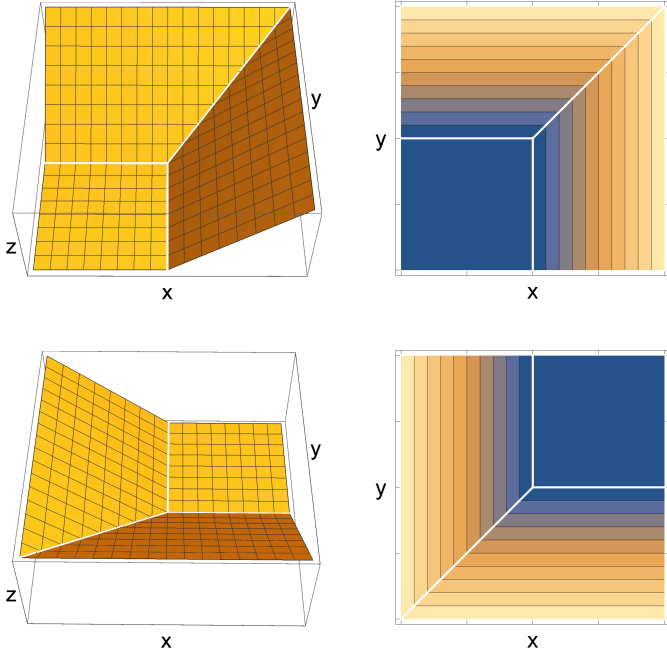


Fig. 1. Two CPWL functions, $F_1(x, y) = \max(0, x, y)$ (first row), and $F_2(x, y) = \max(0, -x, -y)$ (second row). Each row shows the function $\mathbb{R}^2 \rightarrow \mathbb{R}$ as a three-dimensional plot (first column) and as a contour plot (second column).

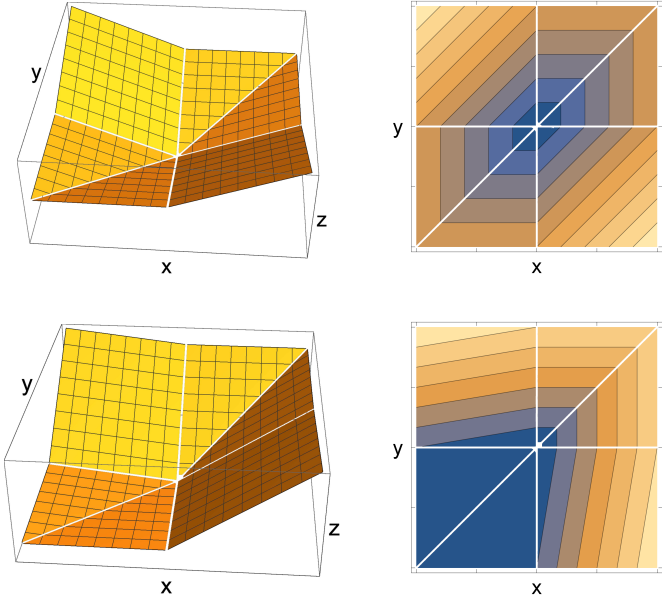


Fig. 2. Two CPWL functions, $F_3(x, y) = \max(0, x, y) + \max(0, -x, -y)$ (first row), and $F_4(x, y) = 6 \max(0, x, y) + \max(0, -x, -y)$ (second row). Each row shows the function $\mathbb{R}^2 \rightarrow \mathbb{R}$ as a three-dimensional plot (first column) and as a contour plot (second column).

Theorem III.5 shows why the function $F_1 + F_2$ can be represented as a linear combination of max with two arguments, and why the functions F_1 and F_2 cannot.

II. PIECEWISE-CONSTANT FUNCTIONS

Before presenting the main theorem of the paper, we need to recall some basic notions and introduce the concept of piecewise-constant functions, which was applied to analyze convex CPWL functions and their underlying polyhedral complexes [5]. All vector spaces and affine spaces in this paper are assumed to be finite-dimensional.

Definition II.1. We say that a property J of points in some non-empty open subset of affine space is defined *almost everywhere* if and only if the set of points for which J does not hold is contained in a finite union of hyperplanes.

Definition II.2. Let Γ be an affine space and G be a vector space. A partial function $f: \Gamma \dashrightarrow G$ is *piecewise-constant* (PC) if it is locally constant and if it is defined for almost every point in Γ .

The set of all points for which a function f is defined is called the *domain of the definition* of f , denoted by $\text{def}(f)$. Note that Definition II.2 implies that the domain of definition of a PC function has finitely many connected components on which the function is constant, and that these components are separated by finitely many hyperplanes. For an affine space Γ , we denote by $\vec{\Gamma}$ the associated vector space of its translations.

Definition II.3. Let $f: \Gamma \dashrightarrow G$ be a PC function defined almost everywhere, then its *lineality space* $L(f)$ is the following set of vectors:

$$L(f) := \{v \in \vec{\Gamma} \mid f(x+v) = f(x) \text{ for all } x \in \Gamma \text{ such that both } f(x) \text{ and } f(x+v) \text{ are defined}\}$$

(note that $f(x)$ and $f(x+v)$ are both defined for almost all x). The dimension of the lineality space (in Lemma II.4 we will show that it is indeed a vector space) is denoted by $\text{lnum}(f) := \dim(L(f))$ and is called the *lineality number* of f .

Lemma II.4. Let $f: \Gamma \dashrightarrow G$ be a PC function, then the set $L(f)$ is a linear subspace of $\vec{\Gamma}$.

Proof. By definition, the vector 0 belongs to $L(f)$. Let $v, w \in L(f)$, then for almost every point $x \in \Gamma$ we have

$$f(x+v+w) = f(x+v) = f(x),$$

i.e., $v+w \in L(f)$, and hence $L(f)$ is closed under addition. Similarly, one can argue that also $-v$ is in $L(f)$. So far, we have established that $kv \in L(f)$ for any $k \in \mathbb{Z}$, and it remains to prove that also $\lambda v \in L(f)$ for any $\lambda \in \mathbb{R}^+$.

Observe that for almost every $x \in \Gamma$, there exists $k \in \mathbb{Z}$ such that $f(x+kv) = f(x+\mu v)$ for all $\mu > k$: we just have to avoid hyperplanes that bound infinite components of $\text{def}(f)$. By choosing $\mu := k + \lambda$ we obtain the following equality:

$$f(x) = f(x+kv) = f(x+\mu v) = f(x+kv+\lambda v) = f(x+\lambda v),$$

from which it follows that $\lambda v \in L(f)$ for any $\lambda \in \mathbb{R}^+$. $\square$

For any point x in an affine space Γ and $\varepsilon > 0$, we denote by $B_\varepsilon(x)$ the open ball with center x and radius ε , i.e., $B_\varepsilon(x) := \{y \in \Gamma \mid \|x - y\| < \varepsilon\}$. An *oriented hyperplane* in Γ is

a hyperplane H together with the two open half-spaces H^+ and H^- . For instance, if $\Gamma = \mathbb{R}^n$, then

$$\begin{aligned} H &= \{x \in \mathbb{R}^n \mid a^T x = b\}, \\ H^+ &= \{x \in \mathbb{R}^n \mid a^T x > b\}, \\ H^- &= \{x \in \mathbb{R}^n \mid a^T x < b\} \end{aligned}$$

for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

Definition II.5. Let Γ be an affine space. An *oriented flag* in Γ is a sequence $\mathcal{H} := (H_1, \dots, H_k)$ of length $k \in \mathbb{N}$ such that H_1 is an oriented hyperplane in Γ , and for any $i \in \{2, \dots, k\}$, H_i is an oriented hyperplane in H_{i-1} .

Definition II.6. Let $f: \Gamma \dashrightarrow G$ be a PC function and let $H \subset \Gamma$ be an oriented hyperplane. Note that for almost every $x \in H$, there exists $\varepsilon > 0$ such that f is defined and constant in the two half balls $B_\varepsilon(x) \cap H^+$ and $B_\varepsilon(x) \cap H^-$. We define the PC function $\Delta_f(H): H \dashrightarrow G$ as the difference of these two constant values, that is, the value in H^+ minus the value in H^- .

Definition II.7. Let $f: \Gamma \dashrightarrow G$ be a PC function, and let $\mathcal{H} = (H_1, \dots, H_k)$ be an oriented flag of length $k \in \mathbb{N}$. Let $f_0 = f$. For $i = 1, \dots, k$, let $f_i: H_i \dashrightarrow G$, $f_i := \Delta_{f_{i-1}}(H_i)$. Then we define $\Delta_f(\mathcal{H}) := f_k$ and call it *delta function*. For the flag $\mathcal{H} = ()$ of length zero, the delta function is defined to be f itself, i.e., $\Delta_f(\mathcal{H}) = f$.

Lemma II.8. Let $f: \Gamma \dashrightarrow G$ be a non-constant PC function and let $H \subset \Gamma$ be an oriented hyperplane. Then for the function f and the delta function $\Delta_f(H)$, the following inequality holds:

$$\text{Inum}(\Delta_f(H)) \geq \text{Inum}(f). \quad (4)$$

Proof. We distinguish two cases.

Case 1: Let $L(f)$ be non-parallel to H . We pick a vector $v \in L(f)$ with $0 < \|v\| < 1$ that is not parallel to H . Then for almost every point $x \in H$ there exists an $\varepsilon > 0$ such that $x + \varepsilon v \in B_\varepsilon(x) \cap H^+$ and $x - \varepsilon v \in B_\varepsilon(x) \cap H^-$ and such that $f(x + \varepsilon v) = f(x - \varepsilon v)$. It follows:

$$\Delta_f(H)(x) = f(x + \varepsilon v) - f(x - \varepsilon v) = 0.$$

Thus $\Delta_f(H)(x) = 0$ for almost all $x \in H$ and therefore $\text{Inum}(\Delta_f(H)) = \dim(\Gamma) - 1$. Since f is not constant, the inequality (4) holds.

Case 2: Let $L(f)$ be parallel to H . Our goal is to show that $L(f) \subseteq L(\Delta_f(H))$, which is equivalent to show that for any $v \in L(f)$ and for almost all $x \in H$, the following equality holds:

$$\Delta_f(H)(x + v) = \Delta_f(H)(x).$$

For almost all $x \in H$ and for any vector w with $0 < \|w\| < 1$ that is not parallel to H and that points towards H^+ , there exists $\varepsilon > 0$ such that $x + \varepsilon w \in B_\varepsilon(x) \cap H^+ \cap \text{def}(f)$, $x + v + \varepsilon w \in B_\varepsilon(x + v) \cap H^+ \cap \text{def}(f)$, $x - \varepsilon w \in B_\varepsilon(x) \cap H^- \cap \text{def}(f)$ and $x + v - \varepsilon w \in B_\varepsilon(x + v) \cap H^- \cap \text{def}(f)$. It implies that the PC function f is defined for these four points,

and using the property that $v \in L(f)$, the following equalities hold:

$$\begin{aligned} \Delta_f(H)(x) &= f(x + \varepsilon w) - f(x - \varepsilon w) \\ &= f(x + v + \varepsilon w) - f(x + v - \varepsilon w) \\ &= \Delta_f(H)(x + v). \end{aligned}$$

Hence, $\Delta_f(H)(x) = \Delta_f(H)(x + v)$, which implies that $v \in L(\Delta_f(H))$ and therefore $L(f) \subseteq L(\Delta_f(H))$. $\square$

Lemma II.9. The Δ symbol is linear in the index, that is, $\Delta_{f+\lambda g}(H) = \Delta_f(H) + \lambda \Delta_g(H)$ for any two PC functions f, g and real number λ .

Proof. The proof is a direct consequence of the properties of PC functions. $\square$

Lemma II.10. Let $f: \Gamma \dashrightarrow G$ be a PC function such that for all oriented hyperplanes H , the delta function $\Delta_f(H)$ is equal to 0. Then f is constant.

Proof. Straightforward. $\square$

Lemma II.11. Let $f: \Gamma \dashrightarrow G$ be a PC function. Then the set $\mathcal{P}$ of all oriented hyperplanes H such that $\Delta_f(H) \neq 0$ is finite.

Proof. We aim to show that every oriented hyperplane $H \in \mathcal{P}$ contains an open subset of points on which the PC function f is not defined. By the definition of a PC function, there are only finitely many such hyperplanes, which will imply that $\mathcal{P}$ contains finitely many oriented hyperplanes.

For any $H \in \mathcal{P}$ there exists $x \in H$ such that $\Delta_f(H)(x) \neq 0$. By the definition of the PC function $\Delta_f(H)$ for the point x , there exists $\varepsilon > 0$ and a ball $B_\varepsilon(x)$ such that for all points $v_1 \in B_\varepsilon(x) \cap H^+$ and $v_2 \in B_\varepsilon(x) \cap H^-$ the function f is defined and $f(v_1) \neq f(v_2)$ holds. Consequently, f is not defined for the set $H \cap B_\varepsilon(x)$. $\square$

By iterating Lemma II.11 we get the following corollary:

Corollary II.12. Let $f: \Gamma \dashrightarrow G$ be a PC function. There are only finitely many flags $\mathcal{H}$ such that the delta function $\Delta_f(\mathcal{H})$ is not 0.

III. MINIMAL DECOMPOSITION

In our study of PC functions from Γ to G , we are especially interested in the case when both $\Gamma = \mathbb{R}^n$ and $G = \mathbb{R}^n$. For a CPWL function $F: \mathbb{R}^n \rightarrow \mathbb{R}$ we define the PC function f as follows:

$$f := \nabla F.$$

We are also interested in the connection between the possible decomposition of the function F as a linear combination of max functions and the properties of the corresponding PC functions.

Definition III.1. Let Γ, G be vector spaces and let V be a linear subspace of Γ . An almost everywhere defined function $f: \Gamma \dashrightarrow G$ is called *V-invariant* if $f(x + v) = f(x)$ for all $v \in V$ and for all $x \in \Gamma$ such that both $f(x)$ and $f(x + v)$ are defined. (Note that if f is a PC function, then this is equivalent to saying that V is a subspace of $L(f)$.)

Lemma III.2. *Let Γ, G be vector spaces and let V be a linear subspace of Γ . Let $F: \Gamma \rightarrow G$ be a continuous and almost everywhere differentiable function. If ∇F is V -invariant then F can be written as $F_1 + F_2$, where F_1 is V -invariant and F_2 is linear.*

Proof. Let W be a complement of V , i.e., $\Gamma = V \oplus W$, the direct sum of V and W . Without loss of generality, we assume that $\Gamma = \mathbb{R}^\ell \times \mathbb{R}^m$, $V = \mathbb{R}^\ell \times 0_m$, and $W = 0_\ell \times \mathbb{R}^m$, where 0_m and 0_ℓ are the zero vectors of length m and ℓ , respectively.

Before we proceed with the proof of the lemma, we will examine the case in which the function F satisfies the additional property that its restriction to W is equal to 0, i.e., $F|_W \equiv 0$. We claim that this stronger assumption implies that the function F is linear. For any point $y \in \mathbb{R}^m$, where $y = (y_1, \dots, y_m)$, we define a *slice function* F_y as follows:

$$F_y: \mathbb{R}^\ell \rightarrow G, \\ (x_1, \dots, x_\ell) \mapsto F(x_1, \dots, x_\ell, y_1, \dots, y_m).$$

Let $(\mathbb{R}^\ell)^*$ be the space of all linear maps from $\mathbb{R}^\ell$ to G . Since F is differentiable almost everywhere, it follows that for almost all y , the slice function F_y is differentiable almost everywhere as well. Because the derivative ∇F is assumed to be V -invariant, it follows that ∇F_y is $\mathbb{R}^\ell$ -invariant. This implies that ∇F_y is constant and hence F_y is an affine-linear function, i.e., $F_y(x) = M_y x + b_y$, where $M_y \in (\mathbb{R}^\ell)^*$ and $b_y \in G$. Note that $F_y(0) = 0$ holds by our additional assumption, which implies that the function F_y is actually a linear function, i.e., $F_y(x) = M_y x$, for almost all y .

We now argue that this last statement is actually true for all $y \in \mathbb{R}^m$. For this purpose, let $M: \mathbb{R}^m \rightarrow (\mathbb{R}^\ell)^*$ be the function that maps each $y \in \mathbb{R}^m$ to the linear functional that is defined by $e_i \mapsto F(e_i, y)$ for every unit vector e_i , $i = 1, \dots, \ell$, and extends to all $x \in \mathbb{R}^\ell$ by linearity. For those y for which $F_y(x) = M_y x$ holds, we have $M_y = M(y)$. Also, the equation $F(x, y) = M(y)x$ is fulfilled almost everywhere, and both sides are continuous, so the equation is fulfilled everywhere. We conclude that F is differentiable at (x, y) if and only if it is differentiable at $(0, y)$.

We claim that the function M is constant. To show this, note that for almost all points (x, y) the function F is differentiable, which implies that the function M is differentiable for almost all $y \in \mathbb{R}^m$. By the assumption that ∇F is V -invariant, we have, for every point (x, y) , the differential ∇F has the same value and the following equality holds: $\nabla F(x, y) = \nabla F(0, y)$. Thereby, for any $i \in \{1, \dots, m\}$ and for almost every $y \in \mathbb{R}^m$, we have

$$\frac{\partial M(y)}{\partial y_i} = \frac{\partial F(e_i, y)}{\partial y_i} = \frac{\partial F(0, y)}{\partial y_i} = \frac{\partial M(y)}{\partial y_i} \cdot 0 = 0.$$

Hence M is constant and F is linear.

The remaining part of the proof is dedicated to the lemma that has initially been stated. Let $F: \Gamma \rightarrow G$ be an almost everywhere differentiable function such that ∇F is V -invariant. As above, we denote by W a linear subspace that is complementary to V . Let $\mu: \Gamma \rightarrow W$ be the projection operator with kernel equal to V . We define the function $F_1: F|_W \circ \mu$ as the composition of the projection operator

μ with the input function F restricted to W . Since μ is linear and maps points from the linear subspace V to 0, it follows that F_1 is V -invariant. We define F_2 as the difference between F and F_1 , i.e., $F_2 := F - F_1$. We have to show that F_2 is linear.

The derivative ∇F_2 is V -invariant, and also the function F_2 restricted to W is equal to 0, i.e., $F_2|_W \equiv 0$. As a result, by the first part of the proof, the function F_2 is linear. $\square$

Lemma III.3. *Let $F := \max(g_1, \dots, g_{k+1})$, where $k \leq n$, be a CPWL function such that $g_i: \mathbb{R}^n \rightarrow \mathbb{R}$ is an affine-linear function for each $i \in \{1, \dots, k+1\}$. Then, the lineality space for the function $f := \nabla F$ has at least dimension $n - k$.*

Proof. Since the following equality holds:

$$F = \max(g_1, \dots, g_{k+1}) \\ = \max(0, g_2 - g_1, \dots, g_{k+1} - g_1) + g_1,$$

we also have the correspondent equality for the derivatives:

$$f = \nabla \max(0, \bar{g}_2, \dots, \bar{g}_{k+1}) + c,$$

where $c := \nabla g_1$, $\bar{g}_i := g_i - g_1$, for all $i \in \{2, \dots, k+1\}$. Thus, the equality holds for the correspondent lineality spaces:

$$\text{lnum}(f) = \text{lnum}(\nabla \max(0, \bar{g}_2, \dots, \bar{g}_{k+1})). \quad (5)$$

Let $A \in \mathbb{R}^{k \times n}$ be the matrix with rows $\nabla \bar{g}_2, \dots, \nabla \bar{g}_{k+1}$. Its kernel $\ker(A)$ is a subset of $\text{L}(\nabla \max(0, \bar{g}_2, \dots, \bar{g}_{k+1}))$. For any $x \in \mathbb{R}^n$ and $y \in \ker(A)$, the following equality holds:

$$\max(0, \bar{g}_2, \dots, \bar{g}_{k+1})(x + y) = \max(0, \bar{g}_2, \dots, \bar{g}_{k+1})(x).$$

Since the matrix A has k rows, it follows that $\dim(\ker(A)) \geq n - k$. $\square$

Definition III.4. For an arbitrary piecewise constant function $f: \Gamma \dashrightarrow G$, we say that an oriented flag $\mathcal{H}$ is *critical* if $\Delta_f(\mathcal{H})$ is non-zero. The set of critical flags is denoted by $\mathcal{F}_f$, and the maximal length of a critical flag is called the *critical flag length* of f , denoted by $\text{cfl}(f)$.

Now we have all the necessary parts to formulate and prove the main theorem of the paper:

Theorem III.5. *Let $F: \mathbb{R}^n \rightarrow \mathbb{R}$ be a CPWL function, let $f := \nabla F$, and let $k \in \mathbb{N}$. The function F can be represented as a linear combination of $\max$ with $k + 1$ arguments if and only if $k \geq \text{cfl}(f)$.*

Proof. By [3], it was shown that the function F can be represented as a linear combination of $\max$ with at most $n + 1$ arguments. Thus, we may assume that $k \leq n$.

Firstly, we show the implication $(\Rightarrow)$. Let F be a CPWL function that can be represented as a linear combination of $\max$ of at most $k + 1$ arguments, i.e., there exists a finite set of indices $I \subset \mathbb{N}$ for which the following equality holds:

$$F = \sum_{i \in I} \sigma_i F_i, \quad (6)$$

where $\sigma_i \in \{-1, 1\}$, $F_i := \max(g_{i,1}, \dots, g_{i,p_i})$, such that $p_i \leq k + 1$ and $g_{i,j}: \mathbb{R}^n \rightarrow \mathbb{R}$ is an affine-linear function for all $j \in \{1, \dots, p_i\}$ and for all $i \in I$. For every $i \in I$, we

denote $f_i := \nabla F_i$. Because the operator ∇ is linear and by Eq. (6), f is decomposed as follows:

$$f = \sum_{i \in I} \sigma_i f_i.$$

Let $\mathcal{H}$ be a critical flag of length k . By Lemma II.9, the function $\Delta_f(\mathcal{H})$ is decomposed as follows:

$$\Delta_f(\mathcal{H}) = \sum_{i \in I} \sigma_i \Delta_{f_i}(\mathcal{H}). \quad (7)$$

For $i \in I$, we have the following chain of inequalities:

$$n - k \leq \text{lnum}(f_i) \leq \text{lnum}(\Delta_{f_i}(\mathcal{H})).$$

The first inequality follows from Lemma III.3, the second from repeated application of Lemma II.8; the assumption that $\mathcal{H}$ is critical permits this repeated application because all intermediate delta functions (except possibly the last one) will be non-constant. Since the function $\Delta_{f_i}(\mathcal{H})$ is defined on an affine subspace of dimension $n - k$, it follows that the function $\Delta_{f_i}(\mathcal{H})$ is constant. Thus, $\Delta_f(\mathcal{H})$ is a linear combination of constant functions and is constant as well. Then, Δ_f is zero for any flag of length bigger than k , and it follows that $k \geq \text{cfl}(f)$.

Secondly, to show the implication $(\Leftarrow)$, we prove the statement: “ F can be written as a linear combination of maxima of $\text{cfl}(\nabla(F)) + 1$ affine-linear functions”, by induction on the number r of critical flags. If this number is zero, then the derivative is constant and the function itself is affine-linear, and the statement is true.

Assume that $r \geq 1$. Let $\mathcal{H}$ be a critical flag of maximal length $\text{cfl}(f)$. We construct a CPWL function $G_1: \mathbb{R}^n \rightarrow \mathbb{R}$ with $g_1 := \nabla G_1$ for which the following properties hold:

- (a) The function G_1 is a linear combination of maxima of at most $\text{cfl}(f) + 1$ affine-linear functions.
- (b) Any flag which is critical for g_1 is also critical for f .
- (c) The PC function $\Delta_{g_1}(\mathcal{H})$ is equal to the PC function $\Delta_f(\mathcal{H})$.

Before proving the existence of a function G_1 with the declared properties, let us assume that such a function exists. We define a CPWL function $F_1: \mathbb{R}^n \rightarrow \mathbb{R}$ as follows:

$$F_1 := F - G_1.$$

The PC function $f_1 := \nabla F_1$ satisfies the following equality:

$$f_1 = f - g_1.$$

Moreover, the set $\mathcal{F}_{f_1}$ is a proper subset of $\mathcal{F}_f$, where “subset” follows from (b) and “proper” follows from (c). Therefore, the cardinality of $\mathcal{F}_{f_1}$ is smaller than the cardinality of $\mathcal{F}_f$. By induction hypothesis, F_1 can be written as a linear combination of maxima with at most $\text{cfl}(f_1) + 1$ affine-linear functions. Since $\text{cfl}(f_1) \leq \text{cfl}(f)$, the induction is finished.

To finish the proof, one is left to show that the function G_1 exists. By $\mathcal{N}$ we define the minimal family of hyperplanes such that the PC function f is defined on its complement. Let $\mathcal{H} := (H_1, \dots, H_{\text{cfl}(f)})$ be an oriented flag in $\mathcal{F}_f$. Let $x_0 \in H_{\text{cfl}(f)}$ be a point for which all the hyperplanes in $\mathcal{N}$ that contain x_0 also contain $H_{\text{cfl}(f)}$ – this is true for almost all points in $H_{\text{cfl}(f)}$. There exists $\varepsilon > 0$ such that the ball $U :=$

$B_\varepsilon(x_0)$ has empty intersection with any hyperplane $H \in \mathcal{N}$ for which $x_0 \notin H$. For the sake of simplicity and without loss of generality, we assume that $x_0 = 0$ and $F(0) = 0$. In two steps, we define the function $G_1: \mathbb{R}^n \rightarrow \mathbb{R}$.

Step 1: For every point $x \in U$, we define $G_1(x) := F(x)$. Because $F(0) = 0$, the function F , restricted to the ball U , is *positively homogeneous*, that is for every $\lambda > 0$, and $x \in U$ such that $\lambda x \in U$ holds, $G_1(\lambda x) = \lambda G_1(x)$.

Step 2: For any point $x \in \mathbb{R}^n$, there exists $\lambda \geq 0$ and $y \in U$, such that $x = \lambda y$. In this case, we define $G_1(x) := \lambda G_1(y)$.

The function G_1 is a CPWL function. It defines a PC function $g_1 := \nabla G_1$. The remainder of the proof shows that the functions G_1 and g_1 satisfy properties (a)–(c).

Property (b): Let $\mathcal{A} = (A_1, \dots, A_p)$, where $p \leq \text{cfl}(f)$, be a flag in $\mathcal{F}_{g_1}$. Then A_p passes through 0. Hence, there is a nonempty open subset $V \subset U$ so that $\Delta_{g_1}(\mathcal{A})$ is defined in $(V \cap A_p)$. But inside U , the two functions f and g_1 coincide. So, $\Delta_{g_1}(\mathcal{A}) = \Delta_f(\mathcal{A})$ inside U and therefore $\mathcal{A} \in \mathcal{F}_f$.

Property (c): Since $\Delta_{g_1}(\mathcal{H})$ and $\Delta_f(\mathcal{H})$ coincide in $U \cap H_{\text{cfl}(f)}$ and are constant, it follows that $\Delta_{g_1}(\mathcal{H}) = \Delta_f(\mathcal{H})$.

Property (a): To prove that the function G_1 can be represented as a linear combination of max with at most $\text{cfl}(f) + 1$ arguments, firstly we are going to show that the lineality space of g_1 contains the vector space $H_{\text{cfl}(f)}$. The proof is done by contradiction. Assume that there exists a vector $v \in H_{\text{cfl}(f)}$ such that $v \notin L(g_1)$, which implies, that there exist two points $x, x + v \in \text{def}(g_1)$ for which the PC function g_1 is defined and $g_1(x) \neq g_1(x + v)$. By the definition of g_1 and the fact that it is defined for x and $x + v$, there exists a scalar $\lambda > 0$ such that $\lambda x, \lambda(x + v) \in \text{def}(g_1) \cap B_\varepsilon(0)$. The PC functions f and g_1 coincide in $B_\varepsilon(x_0)$ which implies that $f(\lambda x)$ and $f(\lambda(x + v))$ are defined and $f(\lambda x) \neq f(\lambda(x + v))$. The points λx and $\lambda(x + v)$ belong to different open sets on which f is defined and takes different constant values. On the line segment between the points λx and $\lambda(x + v)$ there exists a point $y := (1 - \gamma)\lambda x + \gamma\lambda(x + v)$, where $\gamma \in (0, 1)$, such that $y \in B_\varepsilon(x_0)$ and f is not defined in y . So, the point y belongs to a hyperplane H in $\mathcal{N}$. The hyperplane H has non-trivial intersection set with the ball $B_\varepsilon(0)$ (this set contains y). Hence, the hyperplane H also contains the affine subspace $H_{\text{cfl}(f)}$. Since $0 \in H_{\text{cfl}(f)}$, the affine subspaces H and $H_{\text{cfl}(f)}$ are linear subspaces. It implies that for every scalar $\gamma' \in \mathbb{R}$ the point $\gamma'v + y$ belongs to H as well. For $\gamma' = \lambda(1 - \gamma)$ resp. $\gamma' = -\lambda\gamma$ we get that $\lambda(x + v)$ resp. λx belong to H , which contradicts the statement that the points λx and $\lambda(x + v)$ belong to the open sets on which f and g_1 are defined. Our initial assumption is incorrect and $H_{\text{cfl}(f)} \subseteq L(g_1)$.

The only thing left is to show that the function G_1 can be represented as a linear combination of max with at most $\text{cfl}(f) + 1$ arguments. Let $W \subseteq \mathbb{R}^n$ be a linear subspace of dimension $\text{cfl}(f)$, that is complement to $L(g_1)$. Let $\mu_W: \mathbb{R}^n \rightarrow W$ be the projection operator with kernel $H_{\text{cfl}(f)}$. By Lemma III.2 and the fact that $L(g_1)$ contains $H_{\text{cfl}(f)}$, we can construct an $H_{\text{cfl}(f)}$ -invariant function $F_1: \mathbb{R}^n \rightarrow \mathbb{R}$ and a linear function $F_2: \mathbb{R}^n \rightarrow \mathbb{R}$ such that $G_1 = F_1 + F_2$. Moreover, we have $F_1 = G_1|_W \circ \mu_W$. The function $G_1|_W$ is defined on a $\text{cfl}(f)$ -dimensional linear space W and by [3] the function $G_1|_W$ is a linear combination of max with at

most $\text{cfl}(f) + 1$ arguments. It implies that the function F_1 is a composition of the projection operator μ_W and a linear combination of $\max$ with at most $\text{cfl}(f) + 1$ arguments. Such composition is also a linear combination of $\max$ with at most $\text{cfl}(f) + 1$ arguments, and hence this also holds for the function $G_1 = F_1 + F_2$. $\square$

Example III.6. Let us consider two functions $F_3(x, y) := \max(0, x, y) + \max(0, -x, -y)$ and $F_4 := 6 \max(0, x, y) + \max(0, -x, -y)$. These functions are defined on $\mathbb{R}^2$, see Fig. 2. The functions F_3 and F_4 split the input space into the same family of open sets. In Example I.1, we have shown that F_3 can be represented as a linear combination of $\max$ with two arguments. Since F_3 and F_4 tessellate the input space identically, one may think that the function F_4 can be represented as a linear combination of $\max$ with two arguments as well. However, the function F_4 cannot be represented as a linear combination of $\max$ with two arguments.

To show this, we define the two PC functions $f_3 := \nabla F_3$ and $f_4 := \nabla F_4$. There are in total three hyperplanes on which f_3 is not defined. The delta function Δ_{f_3} with respect to those hyperplanes is constant. On any other hyperplane the delta function Δ_{f_3} is 0. By Theorem III.5 the function F_3 can be represented as a linear combination of $\max$ functions with at most two arguments. We have already seen its decomposition in Example I.1. On the other hand, for the PC function f_4 , the delta function $\Delta_{f_4}((H))$, where the hyperplane $H := \{(x, y) \in \mathbb{R}^2 \mid x = y\}$, is non-constant. To see this, let us pick two points $(1, 1)$ and $(-1, -1)$. The values of $\Delta_{f_4}((H))$ in these points are $(-6, 6)$ and $(-1, 1)$ respectively. Since $\Delta_{f_4}((H))$ is non-constant, Theorem III.5 says that the function F_4 cannot be represented as a linear combination of $\max$ with two arguments.

The proof of Theorem III.5 is algorithmic, it allows us to compute a decomposition of a given CPWL function F into a linear combination of maxima of at most $\text{cfl}(\nabla F) + 1$ affine linear functions, see Algorithm Decompose. The algorithm recursively calls itself; it terminates because in any recursive call, either the number of arguments of the function F is decreasing or the number of arguments stays the same and the number of critical flags is decreasing. The algorithm also calls `reduceMax` [6], an algorithm that computes a decomposition of a given CPWL function $F: \mathbb{R}^n \rightarrow \mathbb{R}$ into a linear combination of maxima of at most $n + 1$ affine-linear functions.

Example III.7. We execute the algorithm on the function $F_3: \mathbb{R}^2 \rightarrow \mathbb{R}$, $F_3(x, y) = \max(0, x, y, -x, -y, x - y, y - x)$ (see also Example I.1 and Example III.6). The gradient f_3 is defined everywhere except on the three lines $H_1: x = y$, $H_2: x = 0$, and $H_3: y = 0$. For $i = 1, 2, 3$, we obtain $\Delta_{f_3}(H_i)$ is constant, and so the set of critical flags is $\{(H_1), (H_2), (H_3)\}$. The critical flag length is $k = 1$, and so we expect a linear combination of maxima of at most two affine linear functions.

We pick the flag (H_1) , and the point $x_0 = (1, 1)$. The extension of the local restriction is $G_1(x, y) = \max(x, y)$. This is already a maximum of two affine linear functions—

Algorithm CriticalFlags

Input: PC function f

Output: Set of Critical Flags

```

1:  $\mathcal{F} := \emptyset$ ;
2:  $\mathcal{N}$  is the set of hyperplanes where  $f$  is not defined almost everywhere;
3: For  $H$  in  $\mathcal{N}$  Do
4:    $f_1 := \Delta_f(H)$ ;
5:    $\mathcal{F}_1 := \text{CriticalFlags}(f_1)$ ;
6:   For  $\mathcal{H} \in \mathcal{F}_1$  Do
7:     add  $(H, \mathcal{H})$  to  $\mathcal{F}$ ;
8:   End For
9: End For
10: Return  $\mathcal{F}$ .

```

so we skip the restriction to an orthogonal complement and recursive call. We get the function

$$\begin{aligned} F_1(x, y) &= \max(0, x, y, -x, -y, x - y, y - x) - \max(x, y) \\ &= \max(0, -x, -y, -x - y). \end{aligned}$$

Its critical flags are (H_2) and (H_3) . We pick the flag (H_2) and the point $x_1 = (0, 1)$. The extension of the local restriction is $G_2(x, y) = \max(0, -x)$. Again, this is already a maximum of two affine linear functions. We compute the difference

$$F_2(x, y) = F_1(x, y) - G_2(x, y) = \max(0, -y)$$

and get the final decomposition

$$\begin{aligned} F_3(x, y) &= G_1(x, y) + G_2(x, y) + F_2(x, y) \\ &= \max(x, y) + \max(0, -x) + \max(0, -y). \end{aligned}$$

Theorem III.5 has the implication for different representations as well. By applying Theorem III.5 to the results from [2], we get the following corollary:

Corollary III.8. For $n \geq 3$, a CPWL function $F: \mathbb{R}^n \rightarrow \mathbb{R}$ can be represented as a ReLU-based neural network with depth at most $\lceil \log_3(k) \rceil + 1$, if for any flag $\mathcal{H}$ of length k the delta function $\Delta_{\nabla F}(\mathcal{H})$ is zero.

Proof. The statement holds by applying Theorem III.5 in Theorem 2 from [4]. $\square$

Note that the reversed statement of Corollary III.8 is an open problem. More drastically, we do not even know if there is a CPWL function that cannot be represented by a ReLU neural network with two hidden layers. See [4], [5], [12] for an extended discussion.

ACKNOWLEDGMENT

We are indebted to the two anonymous reviewers for pointing out flaws in a previous version and making many insightful comments and suggestions.

Algorithm Decompose**Input:** CPWL function $F: \mathbb{R}^n \rightarrow \mathbb{R}$ **Output:** List of signed maxima of at most $\text{cfl}(\nabla F) + 1$ affine linear functions that sum up to F

```

1:  $f := \nabla(F)$ ;
2:  $\mathcal{F} := \text{CriticalFlags}(f)$ ;
3: If  $\mathcal{F} \neq \emptyset$  Then
4:   pick  $\mathcal{H} = (H_1, \dots, H_k)$  from  $\mathcal{F}$  of maximal length;
5:   If  $k < n$  Then
6:     pick  $x_0 \in H_k$  such that any hyperplane through  $x_0$  where  $f$  is not defined contains  $H_k$ ;
7:      $G_1 :=$  the extension by positive homogeneity of the restriction of  $F$  to a small neighborhood of  $x_0$ ;
8:      $F_1 := F - G_1$ ;
9:     Choose an orthogonal complement  $W$  of  $H_k$ , and let  $\pi: \mathbb{R}^n \rightarrow W$  be the projection with kernel  $H_k$ ;
10:     $\Sigma := \text{Decompose}(G_1|_W)$ ;
11:    For  $m \in \Sigma$  Do
12:       $m := m \circ \pi$ ;
13:    End For
14:    Return  $\Sigma \cup \text{Decompose}(F_1)$ ;
15:   Else
16:     Compute  $\text{Decompose}(F)$  using Algorithm reduceMax in [6];
17:     Return  $\text{Decompose}(F)$ ;
18:   End If
19: Else (in this case,  $F$  is affine linear)
20:   Return  $\max(F)$ ;
21: End If

```

REFERENCES

- [1] L. Breiman, "Hinging hyperplanes for regression, classification, and function approximation," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 999–1013, 1993.
- [2] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding deep neural networks with rectified linear units," in *ICLR*, 2018.
- [3] S. Wang and X. Sun, "Generalization of hinging hyperplanes," *IEEE Trans. Inf. Theory*, vol. 51, pp. 4425–4431, 12 2005.
- [4] E. Bakaev, F. Brunk, C. Hertrich, J. Stadel, and A. Yehudayoff, "Better neural network expressivity: Subdividing the simplex," arXiv, Tech. Rep. 2505.14338, 2025.
- [5] C. Hertrich, A. Basu, M. Di Summa, and M. Skutella, "Towards lower bounds on the depth of ReLU neural networks," *SIAM J. Discrete Math.*, vol. 37, no. 2, pp. 997–1029, 2023.
- [6] C. Koutschan, B. Moser, A. Ponomarchuk, and J. Schicho, "Representing piecewise linear functions by functions with small arity," *AAECC*, vol. 36, no. 4, pp. 1–16, 2025.
- [7] N. M. Tran and J. Wang, "Minimal representations of tropical rational functions," *Algebr. Stat.*, vol. 15, no. 1, pp. 27–59, 2024.
- [8] M.-C. Brandenburg, M. Grillo, and C. Hertrich, "Decomposition polyhedra of piecewise linear functions," arXiv, Tech. Rep. 2410.04907, 2024.
- [9] G. Montúfar, Y. Ren, and L. Zhang, "Sharp bounds for the number of regions of maxout networks and vertices of Minkowski sums," *SIAGA*, vol. 6, no. 4, pp. 618–649, 2022.
- [10] M.-C. Brandenburg, G. Loho, and G. Montúfar, "The real tropical geometry of neural networks for binary classification," *TMLR*, 2024.
- [11] K.-L. Chen, H. Garudadri, and B. D. Rao, "Improved bounds on neural complexity for representing piecewise linear functions," *NeurIPS*, vol. 35, pp. 7167–7180, 2022.
- [12] C. Haase, C. Hertrich, and G. Loho, "Lower bounds on the depth of integral ReLU neural networks via lattice polytopes," in *ICLR*, 2023.
- [13] A. Brøndsted, *An introduction to convex polytopes*. Springer Science & Business Media, 2012, vol. 90.
- [14] A. Mukherjee and A. Basu, "Lower bounds over Boolean inputs for deep neural networks with ReLU gates," arXiv, Tech. Rep. 1711.03073, 2017.

Christoph Koutschan received the Master degree in Computer Science from Friedrich-Alexander University in Erlangen, Germany, and the Ph.D. degree in Symbolic Computation from the Johannes Kepler University in Linz, Austria. He worked as a researcher at the Research Institute for Symbolic Computation (RISC, Linz, Austria), at Tulane University (New Orleans, USA), and at INRIA (Institut national de recherche en informatique et en automatique, France). Currently he is with the Johann Radon Institute for Computational and Applied Mathematics (RICAM) of the Austrian Academy of Sciences. His research interests are on methods related to the holonomic systems approach, particularly symbolic summation and integration algorithms, and their application to problems from combinatorics and statistical physics.

Anton Ponomarchuk received the B.S. degree in Applied Mathematics from the National University of Kyiv-Mohyla Academy, Kyiv, Ukraine, the M.S. degree in Computer Science from the Ukrainian Catholic University, Lviv, Ukraine, and the Ph.D. degree in Applied Mathematics from Johannes Kepler University, Linz, Austria. From 2020 to 2025, he was a Research Scientist at the Johann Radon Institute for Computational and Applied Mathematics (RICAM). His research lies in applied and computational algebra and geometry, with particular interests in tropical geometry, polyhedral and convex geometry.

Josef Schicho received his Master degree in Mathematics from the University of Vienna, Austria, and the Ph.D. degree in Symbolic Computation from the Johannes Kepler University in Linz, Austria. He has been working at the Johann Radon Institute for Computational and Applied Mathematics and is still partially working there. His main affiliation is the Research Institute of Symbolic Computation at the Johannes Kepler University Linz. He has been working on computational algebraic geometry (rational surfaces), computational kinematics, and algebraic methods in computer vision. Recently, he has become interested in tropical geometry.