Combinatorial Geometric Structures in Parameter Spaces

Contents

Conventions

- A vector $x \in \mathbb{R}^n$ would be regarded as a column vector, $\mathbb{R}^{n\times 1}$.
- For two sets X and Y, we let $\mathrm{Fct}(X, Y)$ denote the set of functions from X to Y, and $\mathrm{Cts}(X, Y)$ the set of continuous functions from X to Y.

1 Introduction

Given a model architecture $\mathcal{A}$, such as transformers, CNN, multilayer perceptron, designed to interpolate a task $\mathcal{T}$, we investigate the encoded information within the parameter space $\mathrm{Par}_{\mathcal{A}}$, playing a similar role as *hypothesis space* in parametric learning theory. We define the mapping:

$$\mathcal{A}_{(-)} : \mathrm{Par}_{\mathcal{A}} \to \mathcal{T} \quad \Theta \mapsto \mathcal{A}_{\Theta} \tag{1}$$

which assigns each parameter Θ, $\mathcal{A}_{\Theta} \in \mathcal{T}$, an object of architecture $\mathcal{A}$, designed for a task, $\mathcal{T}$. For instance $\mathcal{T}$ can be:

- $\mathrm{Fct}(\mathcal{D}, \mathbb{R}^n) := \{f : \mathcal{D} \to \mathbb{R}^n\}$, the set of functions for a domain $\mathcal{D} \subseteq \mathbb{R}^n$. This is typical for the task of classification or pattern retrieval. At this point, we do not *fit* data points. We discuss the case of Hopfield networks in section 3.
- Random variables on measurable space containing a set of data $\mathcal{D}$. This is typical in the context of generative networks, such as diffusion models.

The pair $(\mathcal{A}(-), \mathcal{T})$, partitions the collection of parameters which induces the same object $f \in \mathcal{T}$. We obtain a collection

$$\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}] = \{\mathrm{Par}_f\}_{f \in I(\mathcal{D})} \quad \mathrm{Par}_f = \{\Theta \in \mathrm{Par}_{\mathcal{A}} \,:\, \mathcal{A}_{\Theta} = f\}$$

is the set of regions inducing the same object. Often, $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}]$ is more than a *set*, but is a set with *structure*. [1]

For deep feedforward neural networks, this structure manifests as the face poset of hyperplane arrangement, encapsulating both model expressivity and decision boundaries, see subsection 1.2. The study of face posets is a deep and old question in mathematics, appearing from number theory to geometry. We hope this framework facilitates a coarse-grained unified analysis of various networks. We initiate this exploration with the simplest recurrent networks, specifically associative memory networks or Hopfield networks, in section 3. To extend this to *higher order networks*, which includes transformer networks, and simplicial hopfield network, $\mathrm{Par}_{\mathcal{A}}$ decomposes into semi-algebraic sets, 12, and one may approach with the theory of splines, (Lai et al., 2024).

1.1 Problems we wish to explore

I list a number of problems which we hope to discuss in the paper.

1. Equivalence classes of neural network, subsection 1.3.
2. Network expressivity, subsection 1.2.
3. Dynamics of parameter space, subsection 1.6.
4. Parameter estimation and optimization, subsection 1.5.
5. Data-Architecture-Parameter duality, subsection 1.4.
6. Parameter space for stochastic models.

1.2 Model expressitivity

Consider a typical deep neural net, such as CNN or L-layer feedforward neural network, designed for task $\mathcal{T}$ of classification. One obtains a map

$$\mathcal{A}_{(-)} : \mathrm{Par}_{\mathcal{A}} \to \mathrm{Fct}(\mathbb{R}^n, \mathbb{R}^m) \quad \Theta \mapsto \mathcal{A}_{\Theta}$$

which is *dense*[2] in the subset of continuous functions in $\mathrm{Fct}(\mathbb{R}^n, \mathbb{R}^m)$ when $L \geq 2$, see example 1. This is often referred to as the *universal approximation theorem*, (Cybenko, 1989; Pinkus, 1999; Augustine, 2024), see 1. One may also want study the image of the map:

[1] In various mathematical disciplines, associating combinatorial and algebraic invariants with objects—such as Betti cohomology theories for topological spaces—is a common practice. Our approach aligns with this tradition: for each architecture class $\mathcal{A}$, and task $\mathcal{T}$, we examine $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}]$.

[2] Every in $f \in \mathrm{Fct}(\mathbb{R}^n, \mathbb{R}^n)$, every neighborhood of f, contains a point of the image of $\mathcal{A}(-)$.

Definition 1. Given Equation 1, denote $\mathcal{N}_{\mathcal{A}} := \mathcal{A}_{(\mathrm{Par}_{\mathcal{A}})}$, also called *neuromanifold*.

Neural networks learn by following a gradient flow by minimizing distance over the neural manifold. $\mathcal{N}_{\mathcal{A}}$ is studied in certain cases, see 2, 3.

Example 1. Let $\mathcal{A} := \mathrm{FF}\,[n, \sigma]$ be the class of L-layer feedforward neural network with hyperparameters: *width* $n = (n_i)_{i=1}^{L+1}$ and collection of *activation functions* $\sigma\,(\sigma_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}})$.

For a set of a parameter

$$\Theta := \{A_i, b_i\}_{i=1}^{L} \in \mathbb{R}^{\sum_{i=1}^{L} n_i(n_i+1)}$$

we can associate a function

$$\mathrm{FF}_{\Theta} : f_L \circ \cdots \circ f_1 : \mathbb{R}^{n_1} \to \mathbb{R}^{n_{L+1}}$$

For each $i = 1, \ldots, L$, and $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ is a linear function of the form

$$\sigma_i\,(A_i x_i + b_i)\,,$$

$$\mathrm{Par}_{\mathcal{A}} \to \mathrm{Fct}(\mathbb{R}^{n_1}, \mathbb{R}^{n_{L+1}})$$

To simplify, consider a two layer perception with ReLU actuation everywhere, $n_1 = n_2 = 2$, $n_3 = 1$, has a parameter consisting of quadrauplet $\Theta = (A_1, b_1, A_2, b_2)$, forming a parameter space $\mathrm{Par}_{\mathcal{A}} = \mathbb{R}^{2(3)+2(3)} \simeq \mathbb{R}^{12}$. This defines a map,

$$\mathcal{A}_{(-)} : \mathrm{Par}_{\mathcal{A}} \to \mathrm{Fct}(\mathbb{R}^n, \mathbb{R})$$

has dense image within the subspace of continuous functions, $\mathrm{Cts}(\mathbb{R}^n, \mathbb{R})$.

Example 2. $\mathcal{A} := \mathrm{CNN}[, L, r, k, s, n]$. This is a polynomial convolutional neural network, (Shahverdi et al., 2024). Here, L is the number of layers, $\sigma_r(x) := x^r$ is a polynomial function for $r \in \mathbb{N}$, $(k = (k_i)_{i=1}^{L}, s = (s_i)_{i=1}^{L}, d = (n_i)_{i=1}^{L+1})$ are all tuples of nonnegative integers, and are hyperparameters for the convolutions at each layer. Here $\mathrm{Par}_{\mathcal{A}} = \bigoplus_{i=1}^{L} \mathbb{R}^{k_i}$. For

$$\Theta = (\Theta_i)_{i=1}^{L} \in \bigoplus_{i=1}^{L} \mathbb{R}^{k_i}$$

$$\varphi_{\Theta} := c_L \circ c_{L-1} \circ \cdots \circ c_1 : \mathbb{R}^{n_1} \to \mathbb{R}^{n_{L+1}}$$

where for $i = 1, \cdots, L-1$, is given by

$$c_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$$

$$c_i(x) : \sigma(\Theta_i \star_{s_i} x)$$

and for $i = L$, we have

$$c_L : \mathbb{R}^{n_L} \to \mathbb{R}^{n_{L+1}}$$

$$c_L(x) = \Theta_L \star_{s_L} x$$

The map

$$\mathrm{Par}_{\mathcal{A}} \to \mathrm{Fct}(\mathbb{R}^{n_1}, \mathbb{R}^{n_{L+1}})$$

has image contained in homogeneous polynomial of degree r^L.

$$\mathcal{N}_{\mathcal{A}} \subseteq \mathrm{Sym}_{r^L}(\mathbb{R}^{d_1})^{d_L}$$

The crucial part the work allowed was to pass to the homogenization. For instance $\sigma(\lambda x) = \lambda^r \sigma(x) \sim \sigma(x)$.

1.3 Equivalence of neural networks

Though many functions exhibit distinct quantitative behaviors, are there metrics that give us *qualitative* properties? Examples include whether a neural network structure is *able* to learn certain algorithms, (Zhou et al., 2023); or whether an architecture (and its training regime) encodes inductive biases for the task (*geometric deep learning*), (Bronstein et al., 2021). In our framework, we can ask

- To what extent does the does the expressivity of an architecture on the depend on the data of $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}]$? What can we say about pairs $(\mathcal{A}, \mathcal{T}), (\mathcal{A}', \mathcal{T}')$ for which

  $$\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}] \simeq_{\mathrm{gis,cis}} \Sigma_{\mathcal{T}'}[\mathrm{Par}_{\mathcal{A}'}]$$

  i.e. architecture, task pairs for which they induce isomorphic hyperplane arrangements, see 16, for definitions.

- As explained in many papers, (Montúfar et al., 2014), in the case when $\mathcal{A}$ is multilayer ReLU networks, each function $\mathcal{A}_\Theta$ naturally induces a face poset $\mathrm{Dom}_{\mathcal{A}_\Theta}$ in its domain. For which collections of faces F, F', in $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}]$ induces isomorphic hyperplane arrangement in the domains of $\mathcal{A}_F$ and $\mathcal{A}_{F'}$:

  $$\mathrm{Dom}_{\mathcal{A}_F} \simeq_{\mathrm{gis,cis}} \mathrm{Dom}_{\mathcal{A}_{F'}}?$$

1.4 Task-parameter duality

It is almost by construction that there is a duality between *task*, $\mathcal{T}$, and *parameter* space. Can we make this precise? Given some model architecture $\mathcal{A}$, we have a map

$$\mathcal{A}_{(-)} : \mathrm{Par}_{\mathcal{A}} \to \mathrm{Fct}(\mathbb{R}^n, \mathbb{R}^m) \quad \Theta \mapsto \mathcal{A}_\Theta$$

By currying, we induce

$$\mathcal{D} \to \mathrm{Fct}(\mathrm{Par}, \mathbb{R}^m)$$
$$x \mapsto x^\vee_{\mathcal{A}}$$

where

$$x^\vee_{\mathcal{A}} : \mathrm{Par} \to \mathbb{R}^m, \quad x^\vee_{\mathcal{A}}(\Theta) = \mathcal{A}_\Theta(x)$$

Thus the subspace of weights which realizes a particular f is given by

$$\mathrm{Par}_f = \bigcap_{x \in \mathcal{D}} (x^\vee_{\mathcal{A}})^{-1}(f(x))$$

forming part the face poset of $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}]$

- To what extent does $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}]$ depends on the architecture - perhaps this is more determined by the dataset? This might shed light into representation hypotheses, (Huh et al., 2024).

- Classical complexity measures in statistical learning, such as VC-dimensions, Rademacher dimension, and fat-shattering dimensions, do not explicitly depend on the dataset.

1.5 Parameter Estimation, Optimization

Parameter estimation, or *learning*, involves the process of using data to infer the values of unknown parameters within a model.[3] Many approaches exist, from maximum likelihood estimates, mean-field theory,

[3] In contrast, physicist often has an inverse problem: they begin with fixed model parameters, such as coupling strengths in an Ising spin glass an infer properties of the system.

contrastive divergence, and score matching to a multitude of Monte Carlo and numerical integration-based methods. Examples such as feed forward ReLU network admit factorizations as

$$\begin{array}{ccc} \mathrm{Par}_{\mathcal{A}} & \longrightarrow & \mathcal{N}_{\mathcal{A}} \subset \mathcal{T} \\ & \searrow^{\mathcal{L}} & \downarrow \\ & & \mathbb{R} \end{array}$$

where according to the task at hand, one solves to find

$$\Theta^* := \mathrm{argmin}_{\Theta \in \mathrm{Par}_{\mathcal{A}}} \mathcal{L}(\Theta)$$

It would be interesting to study the dynamics of parameters within the structures $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}]$ - where are the critical points, and how are parameter reflected? See 3 for an example.

One particular estimation method of interest in *minimum probability flow*, (Sohl-Dickstein et al., 2020), which has been used in the case of Hopfield networks, (Hillar et al., 2014). It would be further interesting to use tools from Morse theory.

Example 3. $\mathcal{A} = \mathrm{LCN}[n, k, s]$, linear convolution structure given by layers $n = (n_1, \ldots, n_{L+1}) \in \mathbb{R}^{L+1}$ kernel and stides $k = (k_1, \ldots, k_L)$, $s = (s_1, \ldots, s_L) \in \mathbb{R}^L$, and $\mathcal{T} = \mathrm{Hom}_{\mathrm{Vect}_{\mathbb{R}}}(\mathbb{R}^{n_1}, \mathbb{R}^{n_{L+1}})$, the set of linear maps from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_{L+1}}$, The image of map

$$\mathrm{Par}_{\mathcal{A}} \to \mathcal{T}$$

$$\Theta \mapsto \mathrm{LCN}_{\Theta}$$

is $\mathcal{N}_{\mathcal{A}} = \mathcal{M}_{d,s,k}$ is a closed subspace studied in (Kohn et al., 2022), whose boundaries have been described by *real root multiplicity patterns*. This was in turn used to study the critical points appearing in $\mathrm{Par}_{\mathcal{A}}$.

1.6 Dynamics of geometric structures

In the context of Hopfield network, one can in fact assign a family of simplicial complex $\left\{\Sigma[\mathrm{Par}^k]\right\}_{k=1}^{\infty}$, where each complex is the complex associated to k iterations of Hopfield network. It would be interesting to study how the face posets evolve as k increases - and how do such results compare with infinite width theory. Similar studies include Horoi et al. (2020).

2 Related works

2.1 Hyperplane arrangement of networks

For a fixed network f expressed as a function $\mathbb{R}^n \to \mathbb{R}^m$, the hyperplane arrangement induced in the domain space has been studied intensively. When f is a continuous piecewise linear function, its domain partitions into regions, $R \subseteq \mathbb{R}^n$, where $f|_R$ is a linear function. For ReLU networks (Pascanu et al., 2014), (Montúfar et al., 2014), and for CNN, (Xiong et al., 2020). It has already been suggested in (Montúfar et al., 2014, Sec. 2.5) to study partitions of parameter space, which induces piecewise linear functions whose linear regions are isomorphic. A recent interest is to study such networks through the lens of algebraic geometry, subsection 2.2.

2.2 Tropical geometry of neural networks

The first connection for modern neural networks was established in Zhang et al, (Zhang et al., 2018). However, the use of tropical geometry to bound the complexity of learning models has older roots and applications

from computational biology to restricted Boltzmann Machine. For instance, it was argued, (Charisopoulos and Maragos, 2019) that the number of linear regions for layers with (leaky) ReLU activations is upper bounded by $\min\left\{2^m, \sum_{j=0}^{n}\binom{m}{j}\right\}$.

3 First example of recurrent network

Associative memory networks, or Hopfield networks, illustrate one of the very first examples of the Hebbian learning rule in computations. For an introductory reference, see (Hertz et al., 1994; Hopfield, 1984).

Definition 2. A *classical Hopfield network* (Hopfield, 1984) of parameter $(W, \theta) \in \mathbb{R}^{n\times n+n}$ on n nodes $[n] := \{1, \dots, n\}$ is a Lenz-Ising model equipped with recurrent (discrete time) dynamics on states:

$$x_i^{(t+1)} := \begin{cases} 1 & \text{if } \sum_j W_{ij}x_j^{(t)} > \theta_i \\ 0 & \text{otherwise} \end{cases} = H\left(\sum_j W_{ij}x_j^{(t)} - \theta_i\right) \tag{2}$$

where

- $x_i^{(t)} \in \{0,1\}^n$, with *binary values*, is the ith component of state $x^{(t)} \in \mathbb{R}^n$ at discrete time $t \in \mathbb{Z}_{\geq 0}$. The values 1 and 0 are to be interpreted as the activation of ith neuron.
- $W \in \mathbb{R}^{n\times n}$, is referred as the *synaptic strength.* , $\theta \in \mathbb{R}^n$ is a threshold term.

There are at least two ways to carry the update rule :

Definition 3. Given a classical Hopfield network of parameter (W, θ), the update rule is :

- *asynchronous* if at each time step, one selects *some* unit $i \in [n]$ to apply the the update rule, Equation 2. The method of selection is an additional choice.
- *synchronous*: if at each time step t, all unit $i \in [n]$, is independently updated according to Equation 2.

We will begin our discussion with *synchronous* updates of Hopfield network in subsection 3.1. In this case, one update of Hopfield network is a special case of ReLU functions restricted to the domain of $\{0,1\}^n \subseteq \mathbb{R}^n$, we redefine Hopfield networks in this setting in 4.

3.1 The case of Hopfield network

Definition 4. We let

- $\mathrm{HamFct}(n,m) := \mathrm{Fct}(\{0,1\}^n, \{0,1\}^m)$ the set of functions from $\{0,1\}^n$ to $\{0,1\}^m$.
- $\mathrm{Par}(m,n) := \mathbb{R}^{m\times n} \times \mathbb{R}^m$ be the space of parameters, we will fix m, n for whole discussion and omit when context is clear. For $\Theta = (W, \theta) \in \mathbb{R}^{m\times n} \times \mathbb{R}^m = \mathrm{Par}$ define the map

$$\mathrm{Par} \to \mathrm{HamFct}$$

$$\mathrm{HN}_\Theta(x) := H(Wx + \theta). \tag{3}$$

where $H : \mathbb{R} \to \{0,1\}$ with

$$H(r) = \begin{cases} 1 & r > 0 \\ 0 & r \leq 0 \end{cases}$$

is the *Heaviside function.*

- When $m = n$, we define higher iterations as

$$\mathrm{HN}_\Theta^k = \underbrace{\mathrm{HN}_\Theta \circ \cdots \circ \mathrm{HN}_\Theta}_{k \text{ times}}.$$

- For a fix $f \in \mathrm{HamFct}$ we consider the space of parameters that induces the function. This is the pullback of Equation 3

$$\begin{array}{ccc} \mathrm{Par}_f := \{\Theta \in \mathrm{Par} \,:\, \mathrm{HN}_\Theta = f\} & \longrightarrow & \mathrm{Par} \\ \downarrow & \lrcorner & \downarrow{\scriptstyle \Theta \mapsto \mathrm{HN}_\Theta} \\ * & \xrightarrow{\quad f \quad} & \mathrm{HamFct} \end{array}$$

HN_Θ is the function of one *synchronous* update of a Hopfield network, associated to a parameter Θ. In the context of Equation 1, the task $\mathcal{T}$ that Hopfield network solves, can simply be rephrased as a function in HamFct. In the following we will study $\Sigma[\mathrm{Par}]$, giving an explicit computation of the number of top dimensional faces for 2 nodes synchronous Hopfield network ($n = 2$), 4, concluding the limited expressivity of Hopfield networks with 2-nodes.

Proposition 1. Let $f \in \mathrm{HamFct}$. Par_f is a convex cone, 1.

Proof. For fixed $x \in \{0,1\}^m, y \in \{0,1\}^n$, set $\mathrm{Par}_{x,y} := \{\Theta \in \mathrm{Par} \,:\, f_\Theta(x) = y\}$. Then

$$\mathrm{Par}_f = \bigcap_x \mathrm{Par}_{x,f(x)}$$

As the intersection of convex cones is again convex cones, we thus check that each $\mathrm{Par}_{x,f(x)}$ is a convex cone. This is the same as checking the following conditions

$$(Wx + \theta)_i > 0 \quad (Wx + \theta)_i \leq 0$$

is closed under convex and $\mathbb{R}_{>0}$ scaling combination of the parameters (W, θ), which is clear. □

Par_f in general is *only* a convex cone, and not closed (or open) *polyhedron* as in 1. It supposedly should be described by a theory of half-open polyhedrons (intersection of closed and open polyhedrons). This is common in the study of Erhart's theory. We thus generalize the definition as follows:

Definition 5. $V \subseteq \mathbb{R}^d$ is a *polyhedron* if it is a finite intersection of closed and open polyhedrons. [4]

By definition, Par_f and Par_g are disjoint for $f \neq g$, and that

$$\mathrm{Par} = \bigcup_{f \in \mathrm{HamFct}} \mathrm{Par}_f$$

3.2 Chambers in weight space

Now let $W = (W_1^t, \ldots, W_n^t)$, so that f can be rewritten in components as

$$\left(H\left(W_i^t x + b_i\right)\right)_{i=1}^m$$

We will regard $(W_i^t, b_i) \in \mathbb{R}^n \oplus \mathbb{R}^1 \simeq \mathbb{R}^{n+1}$ as variables. For fixed $x \in \{0,1\}^n$, this induces hyperplane,

$$\mathcal{H}_{i,x} := \left\{(W_i^t, b_i) \in \mathbb{R}^{n+1} \,:\, (W_i^t, b_i)\begin{pmatrix} x \\ 1 \end{pmatrix} = 0\right\} \subset \mathbb{R}^{n+1}$$

[4] There is a "duality" between the definition of closed polyhedron via convex hulls and the intersection of hyperplanes using the Fourier-Motzkin Elimination method. I believe some results of the flavor of (Ziegler, 1994, Theorem 1.3) should hold for this definition of a polyhedron.

For each fix i, these hyperplanes induces a collection of geometric regions called *fans*, 7, whose order relation induces combinatorial objects called *face posets*. In the context of neural networks, there are various equivalent objects to study, (Brandenburg et al., 2024, Sec. 2). Our interest would be the face poset induced by these affine hyperplanes: ranging for $i \in [m]$ we obtain hyperplane arrangement in Par, by taking the preimage of the projection map,

$$\begin{array}{ccc} \pi_i^{-1}(\mathcal{H}_{i,x}) & \longrightarrow & \text{Par} = \mathbb{R}^{n(n+1)} = \bigoplus_{i\in[n]} \mathbb{R}^{n+1} \\ \downarrow & \lrcorner & \downarrow \pi_i \\ \mathcal{H}_{i,x} & \hookrightarrow & \mathbb{R}^{n+1} \end{array} \tag{4}$$

where π_i is projection in to the ith copy of $\mathbb{R}^{n+1}$ in Par - these corresponds to the (W_i^t, b_i) of (W, b).

Proposition 2. The collection $\left\{\pi_i^{-1}(\mathcal{H}_{i,x})\right\}_{i=1}^m$ induces the same face poset as the convex cones, $\{\text{Par}_f\}$.

- There are $(g(n))^m$ chambers in Par, where $g(n)$ is a value that can be explicitly determined.
- The number of chambers corresponds to the number of functions which can be realized under the assignment $\Theta \mapsto \text{HN}_\Theta$.

The function $g(n)$ is upper bounded by Zaslavsky's theorem, 2, and can be explicitly computed.

Example 4. If $m = n = 2$, then we have $g(n) = 14$. So there are 14^2 chambers Par, giving 196 functions. For instance, we cannot have a function which sends:

$$(0,1) \mapsto (1,1), (1,1) \mapsto (0,0)$$

$$(1,0) \mapsto (0,1), (0,0) \mapsto (1,0)$$

More precisely, if $(\text{HN}_\Theta)_i$ induces an "XOR" matrix for $i = 1$ or 2, then such function cannot be learned. This already yields, $2 * 16 + 2 * 16 - 4 = 60$ functions that cannot be learned. It is not surprising at all that this is true, as this is a reformulation of old observations, (Baum, 1988).

We define the parameter space for k iterations. Though the *space* itself is the same, the hyperplane arrangement induced is distinct. Hence, we add the index k.

Definition 6. Let $k \geq 1$, denote

$$\text{Par}_f^k = \{\Theta \in \text{Par} \mid \text{HN}_\Theta^{\circ k} = f\} \subset \text{Par}^k := \mathbb{R}^{n(n+1)}.$$

and $\Sigma[\text{Par}^k]$ the induced face poset from again the hyperplanes induced by the domain space, as Equation 4.

$\Sigma[\text{Par}^k]$ captures the functional expressivity of Hopfield networks after k iterations.

Proposition 3. A two iteration synchronous Hopfield network induces a an inclusion of face poset, $\Sigma\left[\text{Par}^2\right] \subset \Sigma[\text{Par}]$.

Proof. Sketch. Synchornous updates can be decompose components wise. These operations commute. □

Proposition 4. $\Sigma[a\text{Par}] \subseteq \Sigma[\text{Par}]$.

where $\Sigma[a\text{Par}]$ is the face poset induced from asynchronous Hopfield update. In otherwords, we observe that asynchronous update of Hopfield network is no more expressive as Hopfield network, in the case of 2 nodes, contrary to one's expectation of the additional flexibility that asynchronous updates gives.

4 Higher order memory networks

5 Minimizing Energy Flow

Proposition 5. If a function f can be realized, then minimizing the energy flow will find a point in Par_f.

A Polyhedral Theory

Definition 1. Let $V \subseteq \mathbb{R}^n$ be a subset.

- V is a closed (resp. open) polyhedron if it is equal to
$$P(A,b) := \{x \in \mathbb{R}^n \,:\, Ax \leq (\text{ resp. } <)b\}$$
for some $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$. Equivalently, it is the intersection of a finite number of closed (resp. open) half spaces. In the above presentation, it is a *closed (resp. open) polyhedral cone* if $b = 0$. [5]
- V is convex if it for all $x, y \in V$, $[x, y] := \{tx + (1-t)y \,:\, 0 \leq t \leq 1\} \subseteq V$.
- V is a cone if for all $a \in \mathbb{R}_{>0}$, $a \cdot V \subseteq V$.

Via homogenization, we can always turn a polyhedron into a polyhedral cone, illustrated by the following example.

Example 5. Consider
$$P := P(A,b) = \left\{ \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2 \,:\, \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \leq \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} \quad A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}, b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$
This carves out a square region via the inequalities,
$$\left\{ A_i \begin{pmatrix} a \\ b \end{pmatrix} \leq \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$$
where A_i are the rows of matrix A. Then one can define a "Homogenized cone",
$$C(P) = P(A', 0), \text{ where } A' = \begin{pmatrix} 1 & O \\ -z & A \end{pmatrix}$$
and O is the zero matrix, $z = (z_0, z_1, z_3)$ has three coordinates. In this case, $C(P) \subseteq \mathbb{R}^3$ is carved out by the hyperplane
$$\left\{ -z_i x_0 + A_i x \leq 0 \quad : x \in \mathbb{R}^2 \right\}_{i=1}^4$$
where $x = (a, b) \in \mathbb{R}^2$, so that $P = \left\{ x \in \mathbb{R}^2 \,:\, (1, x) \in C(P) \right\}$.

Definition 7. A *fan* in $\mathbb{R}^N$, is a family of polyhedral cones. $\mathcal{F} := \{C_i\}_{i=1}^N$, such that $\left\{\bar{C}_i\right\}$ satisfies the following properties:

1. Every nonempty face of a cone in $\mathcal{F}$ is also a cone in $\mathcal{F}$.
2. The intersection of any two cones in $\mathcal{F}$ is a face of both.

$\mathcal{F}$ is *complete* if $\bigcup C_i = \mathbb{R}^N$.

As the notion of *interior* point in a polytope is not invariant under affine transformation - if a polytope P in $\mathbb{R}^2$ is regarded as a polytope in $\mathbb{R}^d$ for $d > 2$, then $\text{int}(P) = \emptyset$.

Definition 8. Relative interior.

Corollary 1. Let $\mathcal{F}$ be a complete fan, 7

1. $\bigcup_{i=1}^N \text{relint}(C_i) = \mathbb{R}^N$.
2. Let $\mathcal{H} = \{\mathcal{H}_i\}_{i=1}^M$ be a finite set of hyperplanes, then this induces a complete fan, $\mathcal{F}_{\mathcal{H}}$. The relative interior are of the maximal cones are referred as *chambers*.

[5] The general theory of polyhedron can be reduced to the cone case via homogenization.

B Recollections on tropical geometry

Definition 9.
- Let $(\mathbb{R}_{\max}, \oplus, \odot)$ be the tropical semiring, where $\mathbb{R}_{\max} = \mathbb{R} \cup \{-\infty\}$ with the addition and multiplication defined as $a \oplus b := \max(a, b)$ $a \odot b := a + b$.

Definition 10. The basic polynomial functions are of the following form:

- A *tropical Puiseux polynomial* of degree n, is a formal expression of the form $\bigoplus_{j=1}^{m} a_{\alpha_j} t^{\alpha_j}$, where $\alpha_j \in \mathbb{Q}^n_{\geq 0}$.
-

Proposition 6. Tropical polynomials subdivides its domain into closed polyhedral, 1 where it is linear.

Proof. First suppose $f = \bigoplus_{j=1}^{m} a_{\alpha_j} t^{\alpha_j} : \mathbb{R}^n \to \mathbb{R}$, where $\alpha_j \in \mathbb{Z}_{\geq 0}$ so

$$f(x) = \max_{j \in [m]} \left\{ a_{\alpha_j} + \langle \alpha_j, x \rangle \right\}$$

Then the set of points for which f attains a maximum:

$$M_{\alpha_i} := \{ x \in \mathbb{R}^n \, : \, f(x) = a_{\alpha_i} + \langle \alpha_i, x \rangle \}$$

□

For a tropical polynomial, the *zero* locus, is points of singularity.

Definition 11. Let $f \in \mathrm{TropPoly}(\mathbb{R}^n, \mathbb{R})$, then

$$V(f) = \{ x \in \mathbb{R}^n \, : \, p(x) \text{ is singular} \}$$

Example 6. $m = 1$. $f = 0 \odot t^2$, is simply the linear function $x \mapsto 2x$.

Example 7. Every one layer ReLU network with integer coefficients is difference of two tropical polynomials. Indeed, suppose the network one of the form

$$f(x) = \max \{ Ax + b, t \} : \mathbb{R}^n \to \mathbb{R}^m, t \in \mathbb{R}_{\max}, \quad A \in \mathbb{Z}^{m \times n}, b \in \mathbb{R}^m$$

We refer to (Zhang et al., 2018).

It may be useful to consider the dual interpretation of tropical polynomials of their Newton polytopes.

C Semialgebraic Geometry

In this section, we review the definition of *real semi-algebraic geometry*.

Definition 12. $V \subseteq \mathbb{R}^n$ is a *semialgebraic subset* if it is the solution set of a boolean combination of polynomials and inequality in $\mathbb{R}$.

Example 8. A disk of radius 1 center at $(2, 0)$

$$V = \left\{ (x, y) \in \mathbb{R}^2 \, : \, (x-2)^2 + y^2 \leq 1 \right\}$$

The collection of semi-algebraic sets, satisfy nice properties:

Theorem 1. *If $V \subseteq \mathbb{R}^{m+n} \simeq \mathbb{R}^m \oplus \mathbb{R}^n \xrightarrow{\pi} \mathbb{R}^m$ is semialgebraic, then $\pi(V)$ is semialgebraic.*

D Hyperplane Arrangements and their structures

In this section, we let k be an arbitrary field and $V \simeq k^n$ an n dimensional vector space. Our goal is to discuss the natural *structures* arising from hyperplane arrangements. Here is a Hierarchy of objects:

chamber panel half-space

center

Definition 13.
- A *linear hyperplane* is a codimension one subspace[6] of V. An *affine hyperplane* is a additive translation of a linear hyplane.
- A *hyperplane arrangement in* V is a finite set of hyperplanes $\mathcal{A} := \{\mathcal{H}_i\}_{i=1}^N$ for some $N \in \mathbb{N}$.
- The *rank*, $\operatorname{rank}(\mathcal{A})$, of a hyperplane arrangement is the dimension of the subspace spanned by the normal vectors of $\mathcal{H}_i$.

Definition 14. The *center* of hyperplane arrangement $\mathcal{A} = \{\mathcal{H}_i\}_{i=1}^N$ in V is $Z(\mathcal{A}) := \bigcap_{i=1}^N \mathcal{H}_i$ the intersection of all the hyperplanes.

- $\mathcal{A}$ is *central* if $0 \in Z(\mathcal{A})$.
- $\mathcal{A}$ is *essential* if $\operatorname{rank}(\mathcal{A}) = \dim V$

Definition 15. Let $\mathcal{A}$ be a hyperplane arrangement:

- A *face of* $\mathcal{A}$ is
$$F = \bigcup Y$$
where Y is a closed half space induced by a hyperplane in $\mathcal{A}$. We require the intersection to have *at least one half space* for each hyperplane.
- $\Sigma[\mathcal{A}]$ be the set of faces induced from $\mathcal{A}$.
  - For $F, G \in \Sigma[\mathcal{A}]$, meets, $F \wedge G$, always exists (greatest lower bound, or intersection).
  - It is a post under inclusion.
  - There is a grading, given by
$$\operatorname{rank} : \Sigma[\mathcal{A}] \to \mathbb{N}$$
$$F \mapsto \dim(F) - \dim(O)$$
where O is the dimension of the central face.

  Thus, $\Sigma[\mathcal{A}]$ is *graded meet-semilattice*.

There two notions of isomorphism of arrangement, $\mathcal{A}$ and $\mathcal{A}'$.

Definition 16. Let $(\mathcal{A}, V), (\mathcal{A}', V')$ be two arrangements within two vector spaces V and V' over field k. We say they are

- *gisomorphic*, $\mathcal{A} \simeq_{\text{gis}} \mathcal{A}'$, if exists a linear isomorphism of V, V' that induces bijection on hyperplanes.
- *cisomorphic*, $\mathcal{A} \simeq_{\text{cis}} \mathcal{A}'$, if $\Sigma[\mathcal{A}] \simeq \Sigma[\mathcal{A}']$ as posets.

Definition 17.

[6] by definition of subspace, 0 lies in the hyperplane.

Example 9. Consider the hyperplane arrangement in $\mathbb{R}^3$, with x, y, z coordinates, given by

$$\mathcal{A} = \{\mathcal{H}_i := \{x = i \cdot y\}\}_{i=1}^3$$

Then $Z(\mathcal{A}) = \{x = y = 0\}$, is the z-axis. The essentialization of $\mathcal{A}' := \mathcal{A}/Z(\mathcal{A})$, is the hyperplane arrangement in $\mathbb{R}^2$ given by

$$\{\mathcal{H}_i' := \{x = i \cdot y\}\}_{i=1}^3$$

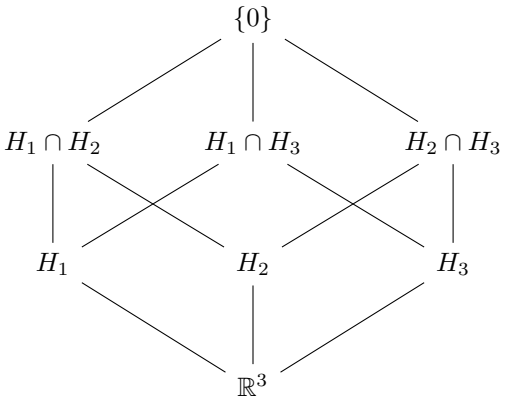
Associated to a hyperplane arrangement. One can obtain various geometric, combinatorial objects.

Definition 18. Let $\mathcal{A} = \{\mathcal{H}_i\}_{i=1}^N$ be a hyperplane arrangement.

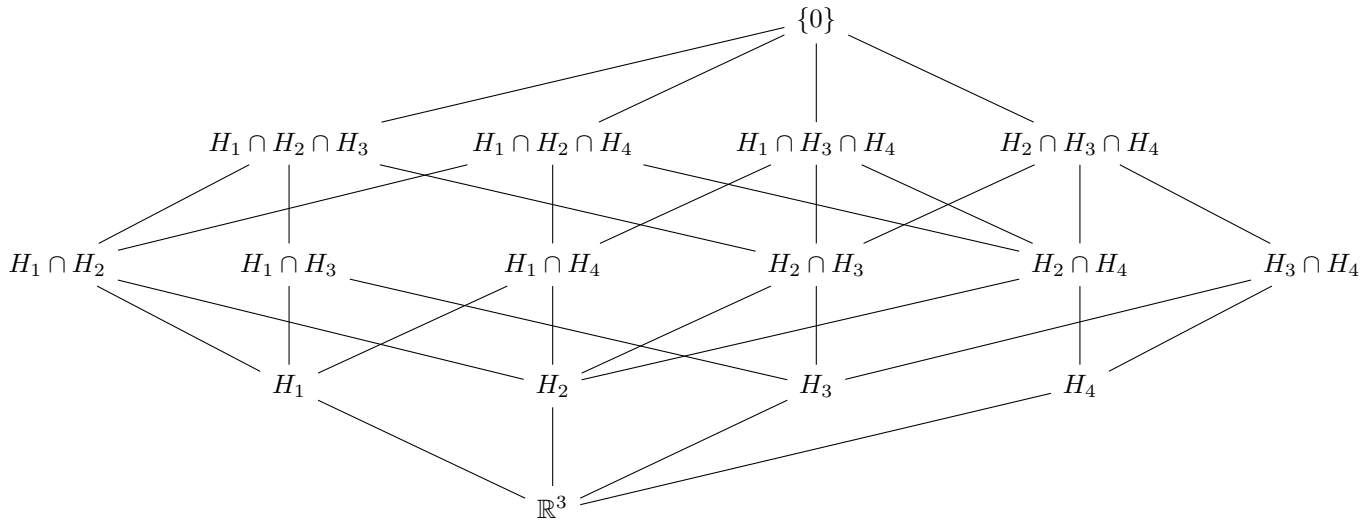
- $L(\mathcal{A}) := \left\{\bigcap_{I\subseteq[n]} \mathcal{H}_i \,:\, \mathcal{H}_i \in \mathcal{A}\right\}$ be the set of intersection posets.
- $\mathrm{Reg}(\mathcal{A}) := \pi_0\left(V \setminus \bigcup_i \mathcal{H}_i\right)$.

Example 10. Consider the hyperplane arrangement $\mathcal{A} := \{H_i := x_i = 0\}_{i=1}^N$ defined in $\mathbb{R}^M$, for $N \leq M$. Then $L(\mathcal{A}) \simeq B_N$, the boolean algebra of all subsets of $[N] = \{1, \ldots, N\}$, ordered by inclusion. Indeed it takes the intersection of subcollection $\{H_i\}_{I\subseteq[N]}$, to the subset I.

Example 11. The *Hasse* diagram for three distinct hyperplanes $\{H_1, H_2, H_3\}$ in $\mathbb{R}^3$ is given as follows.



Hasse diagram for four planes $\{H_1, \ldots, H_4\}$ in general position in $\mathbb{R}^3$:



The general number of regions can be computed using Zaslavsky's formula which we briefly recall.

Definition 19.

Definition 20. Let $\mathcal{A}$ be a hyperplane arrangement, The *characteristic polynomial* is defined by

$$\chi_{\mathcal{A}}(t) := \sum_{x \in L(\mathcal{A})} \mu(x) t^{\dim(x)}$$

As a consequence, we have a result of Thomas Zaslavsky (1975)

Theorem 2. *(Stanley, 2007, Thm 2.5)* $|Reg(\mathcal{A})| = (-1)^n \chi_{\mathcal{A}}(-1)$

References

Midhun T Augustine. A survey on universal approximation theorems, 2024. URL https://arxiv.org/abs/2407.12895.

Eric B Baum. On the capabilities of multilayer perceptrons. *Journal of Complexity*, 4(3):193–215, 1988. ISSN 0885-064X. doi: https://doi.org/10.1016/0885-064X(88)90020-9. URL https://www.sciencedirect.com/science/article/pii/0885064X88900209.

Marie-Charlotte Brandenburg, Georg Loho, and Guido Montúfar. The real tropical geometry of neural networks, 2024. URL https://arxiv.org/abs/2403.11871.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.

Vasileios Charisopoulos and Petros Maragos. A tropical approach to neural networks with piecewise linear activations, 2019. URL https://arxiv.org/abs/1805.08749.

George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989. URL https://api.semanticscholar.org/CorpusID:3958369.

John A. Hertz, Anders Krogh, and Richard G. Palmer. Introduction to the theory of neural computation. In *The advanced book program*, 1994. URL https://api.semanticscholar.org/CorpusID:38623065.

Christopher Hillar, Ram Mehta, and Kilian Koepsell. A hopfield recurrent neural network trained on natural images performs state-of-the-art image compression. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4092–4096, 2014. doi: 10.1109/ICIP.2014.7025831.

John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

Stefan Horoi, Guillaume Lajoie, and Guy Wolf. Internal representation dynamics and geometry in recurrent neural networks, 2020. URL https://arxiv.org/abs/2001.03255.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Kathlén Kohn, Thomas Merkh, Guido Montúfar, and Matthew Trager. Geometry of linear convolutional networks, 2022. URL https://arxiv.org/abs/2108.01538.

Zehua Lai, Lek-Heng Lim, and Yucong Liu. Attention is a smoothed cubic spline, 2024. URL https://arxiv.org/abs/2408.09624.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks, 2014. URL https://arxiv.org/abs/1402.1869.

Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations, 2014. URL https://arxiv.org/abs/1312.6098.

Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999. doi: 10.1017/S0962492900002919.

Vahid Shahverdi, Giovanni Luca Marchetti, and Kathlén Kohn. On the geometry and optimization of polynomial convolutional networks, 2024. URL https://arxiv.org/abs/2410.00722.

Jascha Sohl-Dickstein, Peter Battaglino, and Michael R. DeWeese. A new method for parameter estimation in probabilistic models: Minimum probability flow, 2020. URL https://arxiv.org/abs/2007.09240.

Richard P. Stanley. An introduction to hyperplane arrangements. 2007.

H. Xiong, L. Huang, M. Yu, L. Liu, F. Zhu, and L. Shao. On the number of linear regions of convolutional neural networks, 2020. URL https://arxiv.org/abs/2006.00978.

Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks, 2018. URL https://arxiv.org/abs/1805.07091.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization, 2023. URL https://arxiv.org/abs/2310.16028.

Günter M. Ziegler. Lectures on polytopes. 1994. URL https://api.semanticscholar.org/CorpusID:117286447.