Approche tropicale de l'étude des réseaux de neurones

Grégoire Fournier

Département de mathématiques

March 3, 2025

Université de Lille

Introduction

Ce mémoire vise à mettre en évidence les liens entre les réseaux de neurones et la géométrie tropicale. [3]

Il propose une première une implémentation de ces idées et présente quelques résultats issues de ce rapprochement.

Plan de la présentation

Introduction aux réseaux de neurones

Un réseau ν de profondeur L s'écrit $\mathbb{R}^{n_1} \to \mathbb{R}^{n_2} \to ... \to \mathbb{R}^{n_L}$:

$$\nu = \nu^{(L)} = \sigma^{(L)} \circ \rho^{(L)} \circ \sigma^{(L-1)} \circ ... \circ \sigma^{(1)} \circ \rho^{(1)}$$

Ou ρ sont les transformations affines et σ les activations.
Chaque couple $(\rho^{(l)}, \sigma^{(l)})$ est appelé une couche.
$\rho^{(l)} : \mathbb{R}^{n_{l-1}} \to \mathbb{R}^{n_l}$ est définie par :

$$\rho^{(l)}(\boldsymbol{x}) := A^{(l)}\boldsymbol{x} + \boldsymbol{b}^{(l)}$$

Où $A^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ sont appelées les matrices de poids, $\boldsymbol{b}^{(l)} \in \mathbb{R}^{n_l}$ les biais .

Introduction aux réseaux de neurones

Une des activations les plus populaires est *Relu*, *rectified linear unit* qui correspond à $max(0, .)$.

La première transformation peut donc s'écrire en appliquant le *max* composante par composante $\sigma(\rho(\boldsymbol{x})) = max(A\boldsymbol{x} + b, 0)$.

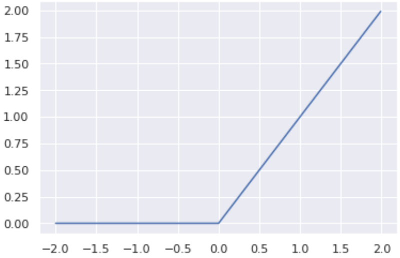

Figure: Activation *Relu*

Introduction aux réseaux de neurones

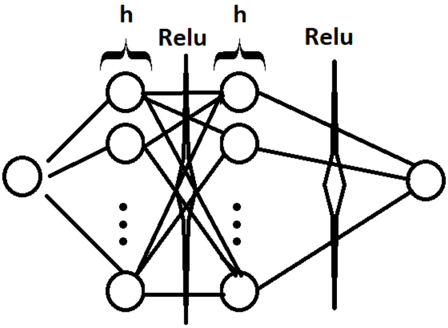

Figure: réseau à activation relu, deux couches cachées de largeur h

Le semi-anneau tropical

Algèbre tropicale sur le semi anneau $(\mathbb{T} := \mathbb{R} \cup \{-\infty\}, \oplus, \odot)$.

On définit deux opérations associatives sur $\mathbb{T}$:

$$x \oplus y := max(x, y) \text{ et } x \odot y := x + y$$

On définit également la division tropicale $\oslash : x, y \to x - y$

La puissance tropicale s'écrit quant à elle:

$$x^{\odot a} := x \odot ... \odot x = a \times x$$

Polynômes et fonctions rationnelles tropicales

En dimension d, pour $\boldsymbol{a} \in \mathbb{N}^d$, un monôme tropical, noté $\boldsymbol{x}^{\boldsymbol{a}}$, est une expression de la forme :

$$\boldsymbol{x}^{\boldsymbol{a}} = c \odot {x_1}^{a_1} \odot \ldots \odot {x_d}^{a_d}$$

$$\mathbb{T}^d \to \mathbb{T}$$

Un polynôme tropical f dans $\mathbb{R}^d$ est la somme tropicale, notée $\sum$ de monômes tropicaux :

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{a} \in S} \boldsymbol{x}^{\boldsymbol{a}} = \sum_{\boldsymbol{a} \in S} c_{\boldsymbol{a}} \odot x_1^{a_1} \odot \ldots \odot x_d^{a_d}$$

pour S un sous ensemble fini de $\mathbb{N}^d$, c des réels, $\forall \boldsymbol{x}$ de $\mathbb{R}^d$.

Une fonction rationnelle tropicale est la différence de deux polynômes tropicaux.

Exemple de polynôme tropical

On peut se servir de l' espace des polynômes tropicaux pour approximer la fonction exponentielle.

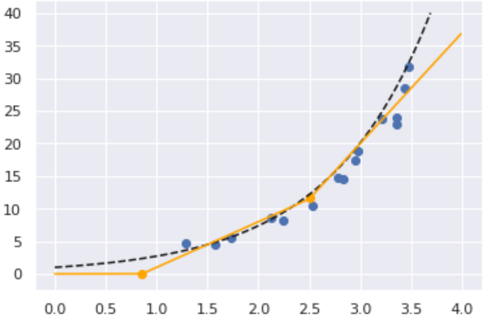

Figure: $f(x) = 0 \oplus (-6 \odot x^7) \oplus (-31 \odot x^{17})$

Équivalence tropicale des réseaux de neurones

On pose 3 conditions concernant les éléments des réseaux de neurones qui permettent d'obtenir une équivalence avec les fonctions tropicales :

a les matrices de poids sont à valeurs entières
b les biais sont à valeurs réelles
c les activations sont de type *Relu*

Équivalence entre réseaux de neurones et fonction rationnelle tropicale

Théorème : Caractérisation tropicale des réseaux de neurones

1. Soit ν un réseau de neurones : $\mathbb{R}^d \to \mathbb{R}$, ν est une fonction rationnelle tropicale si et seulement si ν vérifie (a) - (c)
2. Soit une fonction rationnelle tropicale $\nu = f \oslash g$, ν peut être représentée par un réseau à L couches avec

$$L \leq max(\lceil log_2(r_f) \rceil, \lceil log_2(r_g) \rceil) + 2$$

en notant r_f et r_g le nombre de monomes dans f et g

Preuve de l'équivalence tropicale des réseaux de neurones

Soit $A^{(1)} \in \mathbb{Z}^{n_2 \times n_1}$ et $\boldsymbol{b^{(1)}} \in \mathbb{R}^{n_2}$ les poids de la couche 1:

La première couche peut s'écrire :

$$\nu^{(1)} = \sigma \circ \rho^{(1)}(\boldsymbol{x}) = max(A\boldsymbol{x}+\boldsymbol{b}, \boldsymbol{0}) = max(A_+^{(1)}\boldsymbol{x} - A_-^{(1)}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{0})$$

$$\nu^{(1)} = \sigma \circ \rho^{(1)}(\boldsymbol{x}) = max(A_+^{(1)}\boldsymbol{x} + \boldsymbol{b}, A_-^{(1)}\boldsymbol{x}) - A_-^{(1)}\boldsymbol{x}$$

$$\nu^{(1)} = \sigma \circ \rho^{(1)}(\boldsymbol{x}) = F^{(1)}(\boldsymbol{x}) - G^{(1)}(\boldsymbol{x})$$

Où $F^{(1)}$ et $G^{(1)}$ sont des polynômes tropicaux et $\nu^{(1)}$ une fonction rationnelle tropicale.

Preuve de l'équivalence tropicale des réseaux de neurones

Si à l'étape l le réseau de neurone peut s'écrire:

$$\nu^{(l)} = \sigma \circ \rho^{(l)}(\boldsymbol{x}) = F^{(l)}(\boldsymbol{x}) - G^{(l)}(\boldsymbol{x})$$

Alors:

$$\rho^{(l+1)} \circ \nu^{(l)}(\boldsymbol{x}) = H^{(l+1)}(\boldsymbol{x}) - G^{(l+1)}(\boldsymbol{x}))$$

$$\nu^{(l+1)} = \sigma \circ \rho^{(l+1)} \circ \nu^{(l)}(\boldsymbol{x}) = F^{(l+1)}(\boldsymbol{x}) - G^{(l+1)}(\boldsymbol{x}))$$

avec :

$$H^{(l+1)}(\boldsymbol{x}) = A_+^{(l+1)} F^{(l)}(\boldsymbol{x}) + A_-^{(l+1)} G^{(l)}(\boldsymbol{x}) + \boldsymbol{b^{(l+1)}}$$
$$G^{(l+1)}(\boldsymbol{x}) = A_+^{(l+1)} G^{(l)}(\boldsymbol{x}) + A_-^{(l+1)} F^{(l)}(\boldsymbol{x})$$
$$F^{(l+1)}(\boldsymbol{x}) = max(H^{(l+1)}(\boldsymbol{x}), G^{(l+1)}(\boldsymbol{x}))$$

Preuve de l'équivalence tropicale des réseaux de neurones

On obtient les équations tropicales suivantes pour les coordonnées $f_i^{(l)}, g_i^{(l)}$ et $h_i^{(l)}$ de ces matrices :

$$h_i^{(l+1)} = \bigodot_{j=1}^{n_{l+1}} f_j^{(l)\,a_{i,j}^+} \odot \bigodot_{j=1}^{n_{l+1}} g_j^{(l)\,a_{i,j}^-} \odot b_i$$

$$g_i^{(l+1)} = \bigodot_{j=1}^{n_{l+1}} f_j^{(l)\,a_{i,j}^-} \odot \bigodot_{j=1}^{n_{l+1}} g_j^{(l)\,a_{i,j}^+}$$

$$f_i^{(l+1)} = h_i^{(l+1)} \oplus g_i^{(l+1)}$$

Transformation des hypersurfaces dans un réseau de neurones

Definition (Hypersurface)

Pour un polynôme tropical de la forme:

$$f(\boldsymbol{x}) = c_1\boldsymbol{x}^{\alpha_1} \oplus \ldots \oplus c_d\boldsymbol{x}^{\alpha_d}$$

L'hypersurface est définie par:

$$\mathcal{T}(f) := \left\{\boldsymbol{x} \in \mathbb{R}^d \mid c_i\boldsymbol{x}^{\alpha_i} = c_j\boldsymbol{x}^{\alpha_j} = f(\boldsymbol{x}),\ \alpha_i \neq \alpha_j\right\}$$

L'hypersurface tropicale divise l'espace en régions ou le réseau de neurone se comporte comme une fonction linéaire.

La complexité d'un réseau de neurones, qui correspond aux nombres de régions linéaires que l'on peut représenter.

Transformation des hypersurfaces dans un réseau de neurones

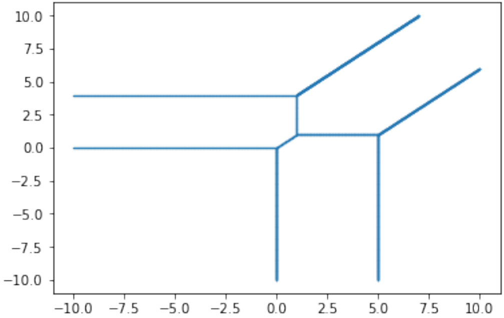

Figure: Hypersurface tropicale de $f(x, y) = 4 + x + y \oplus 1 + 2y \oplus 2x \oplus 5 \oplus 5 + x \oplus 5 + y$

Transformation des hypersurfaces dans un réseau de neurones

Les transformations dans un réseau de neurones sont : combinaison linéaire et passage au max.
La combinaison linéaire pose problème pour l'étude des hypersurfaces, en effet :

Soit f_1, f_2 des polynômes tropicaux de $\mathbb{R}^d \to \mathbb{R}^d$, en notant $\sum$ la somme tropicale, $f_1(\boldsymbol{x}) = \sum_i \boldsymbol{x}^{\boldsymbol{a}_i^1}$, $f_2(\boldsymbol{x}) = \sum_j \boldsymbol{x}^{\boldsymbol{a}_j^2}$, que faire de la somme classique $f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) = \sum_{i,j} \boldsymbol{x}^{\boldsymbol{a}_i^1 + \boldsymbol{a}_j^2}$?

A ma connaissance, il n'existe pas de théorème permettant de lier directement $\mathcal{T}(f_1), \mathcal{T}(f_2)$ à $\mathcal{T}(f_1 + f_2)$.

Transformation des hypersurfaces dans un réseau de neurones

Definition (Polytope de Newton)

Le polytope de Newton d un polynôme tropicale f est l'enveloppe convexe des points représentant ses monomes:

$$\mathcal{N}(f) = conv\left\{(\boldsymbol{c_i}, b_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, ..., d)\right\}$$

$$\mathcal{N}(f) = \left\{\sum_{i=1}^{d} \lambda_i \dot{(}\boldsymbol{c_i}, b_i)| \sum_{i=1}^{d} \lambda_i = 1, 0 \leq \lambda_i\right\}$$

On définie également la subdivision induite par f, $\delta(f)$, en notant UF pour l'enveloppe convexe supérieure:

$$\delta(f) = \left\{\pi(\boldsymbol{p}) \in \mathbb{R}^d \mid \boldsymbol{p} \in UF(\mathcal{N}(f))\right\}$$

Lien entre Hypersurfaces et polytope de Newton

Definition (somme de Minkowski)

La somme de Minkowski de deux ensembles P_1 et P_2 dans $\mathbb{R}^d$ est l'ensemble:

$$P_1 \oplus P_2 := \left\{ \mathbf{x_1} + \mathbf{x_2} \in \mathbb{R}^d \mid \mathbf{x_1} \in P_1, \mathbf{x_2} \in P_2 \right\}$$

Ou $+$ est la somme composante par composante.

Lemma

$f_1, ..., f_d$ des polynômes tropicaux de $\mathbb{R}^d \to \mathbb{R}$, en notant $\bigoplus$ pour la somme de Minkowski et $\sum$ pour la somme tropicale:

$$\mathcal{T}(\sum_{i=1}^{d} f_i) = \bigcup_{i=1}^{d} \mathcal{T}(f_i)$$

$$\mathcal{N}(\sum_{i=1}^{d} f_i) = \bigoplus_{i=1}^{d} \mathcal{N}(f_i)$$

Lien entre Hypersurfaces et polytope de Newton

Quand on a tenté de retranscrire les transformations de couche en couche, on a vu que l'on aboutissait à une impasse.

En revanche l' étude du polytope de Newton (étendu) se révèle plus prometteuse :

Lemma

Soit f un polynôme tropical de $\mathbb{T}[x_1, ..., x_d]$ et a un entier :

$$\mathcal{N}(f^a) = a\mathcal{N}(f)$$

Lemma

Soit f, g deux polynômes tropicaux de $\mathbb{T}[x_1, ..., x_d]$:

$$\mathcal{N}(max(f,g)) = Conv(\mathcal{V}(\mathcal{N}(f)) \cup \mathcal{V}(\mathcal{N}(G))$$

Lien entre Hypersurfaces et polytope de Newton

A l'étape $l \geq 1$:

$$\mathcal{N}(f_i^{(l)}) = Conv[\mathcal{N}(g_i^{(l)}) \cup \mathcal{N}(h_i^{(l)})]$$

Et les polynômes de Newton sont des sommes de Minkowski pondérées :

$$\mathcal{N}(g_{i+1}^{(l)}) = \sum_{j=1}^{n_l} a_{i,j}^{-} \mathcal{N}(f_j^{(l)}) + \sum_{j=1}^{n_l} a_{i,j}^{+} \mathcal{N}(g_j^{(l)})$$

$$\mathcal{N}(h_{i+1}^{(l)}) = \sum_{j=1}^{n_l} a_{i,j}^{+} \mathcal{N}(f_j^{(l)}) + \sum_{j=1}^{n_l} a_{i,j}^{-} \mathcal{N}(g_j^{(l)}) + b_i * e_{d+1}$$

Lien entre Hypersurfaces et polytope de Newton

On définit la subdivision induite par f, $\delta(f)$, en notant UF pour l'enveloppe convexe supérieure:

$$\delta(f) = \left\{ \pi(\boldsymbol{p}) \in \mathbb{R}^d \mid \boldsymbol{p} \in UF(\mathcal{N}(f)) \right\}$$

Théorème Maclagan and Sturmfels [1]

L'hypersurface tropicale $\mathcal{T}(f)$ est le $(d-1)$ squelette du complexe polyèdral dual à la subdivision induite par f, $\delta(f)$.

Exemple

Considérons par exemple un réseau à deux couches, $\nu : \mathbb{R}^2 \to \mathbb{R}$ avec $n_0 = 2$ entrées, $n_1 = 5$ dans la première couche et $n_2 = 1$ en sortie.

En reprenant les notations précédentes:

$$\boldsymbol{y} = \nu^{(1)}(\boldsymbol{x}) = max\left\{\begin{pmatrix} -1 & 1 \\ 3 & 0 \\ 1 & -2 \\ 2 & 1 \\ 3 & -3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \\ 2 \\ 0 \\ -2 \end{pmatrix}, 0\right\}$$

$$\nu^{(2)}(\boldsymbol{y}) = max(-3y_1 + 2y_2 + y_3 - 2y_4 - y_5, 0)$$

Exemple

Pour ce faire il suffit de séparer:

$$\nu^{(1)} = max(A\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{0}) = max(A_+^{(1)}\boldsymbol{x} - A_-^{(1)}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{0})$$

donc

$$g^{(1)}(\boldsymbol{x}) = \begin{pmatrix} x_1 \\ 0 \\ x_2^2 \\ 0 \\ x_2^3 \end{pmatrix} \quad h^{(1)}(\boldsymbol{x}) = \begin{pmatrix} 2 \odot x_2 \\ -1 \odot x_1^3 \\ 2 \odot x_1 \\ x_1^2 x_2 \\ -2 \odot x_1^3 \end{pmatrix}$$

De la même manière on cherche à écrire $f^{(2)}(\boldsymbol{x}) = h^{(2)}(\boldsymbol{x}) \oplus g^{(2)}(\boldsymbol{x})$, en posant $\boldsymbol{z} = g^{(1)}$ et $\boldsymbol{y} = f^{(1)}$.

La couche $\nu^{(2)}$ peut donc s'écrire :

$$g^{(2)}(\boldsymbol{x}) = y_1^3 \odot z_2^2 \odot z_3 \odot y_4^2 \odot y_5$$
$$h^{(2)}(\boldsymbol{x}) = z_1^3 \odot y_2^2 \odot y_3 \odot z_4^2 \odot z_5$$

Exemple : variétés tropicales associées au réseau

On remarque que les $g_j^{(1)}$ et $h_j^{(1)}$ sont des monomes tropicaux, les polytopes de Newton associés, $\mathcal{N}(g_j^{(1)})$ et $\mathcal{N}(h_j^{(1)})$ sont donc des points.

Comme $f^{(1)} = h^{(1)} \oplus g^{(1)}$, $\mathcal{N}(f_j^{(1)})$ est l'enveloppe convexe de deux points donc un segment.

En notant $+$ la somme de Minkowski :

$$\mathcal{N}(g^{(2)}) = 3\mathcal{N}(f_1^{(1)}) + 2\mathcal{N}(g_2^{(1)}) + \mathcal{N}(g_3^{(1)}) + 2\mathcal{N}(f_4^{(1)}) + \dots$$
$$\mathcal{N}(h^{(2)}) = 3\mathcal{N}(g_1^{(1)}) + 2\mathcal{N}(f_2^{(1)}) + \mathcal{N}(f_3^{(1)}) + 2\mathcal{N}(g_4^{(1)}) + \dots$$

Enfin $\mathcal{N}(f^{(2)})$ est l'enveloppe convexe de l'union de $\mathcal{N}(g^{(2)})$ et $\mathcal{N}(h^{(2)})$. On passe à l'hypersurface tropicale $\mathcal{T}(\nu)$.

Exemple : variétés tropicales associées au réseau

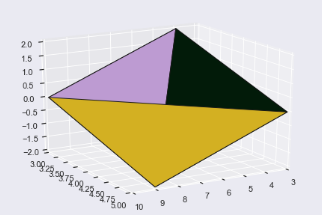

Figure: Le polytope convexe $\mathcal{N}(g^{(2)})$

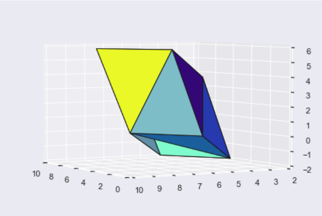

Figure: Le polytope convexe $\mathcal{N}(h^{(2)})$

Exemple : variétés tropicales associées au réseau

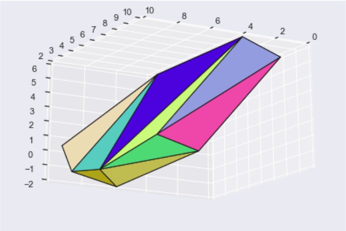

Figure: L'enveloppe convexe du polytope de Newton $\mathcal{N}(f^{(2)})$

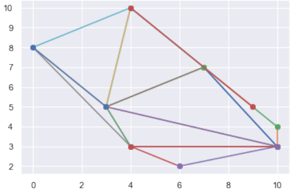

Figure: Projection par π de l'enveloppe supérieure de $\mathcal{N}(f^{(2)})$

Thématiques connexes

- Géométrie computationelle
  - Triangulation de Delaunay
  - Enveloppe convexe de points
  - Somme de Minkowski de polytopes
- Structure des polytopes
  - Théorème de Minkowski Weyl
  - Lemme de Farkas
- Approche combinatoire-polyédrale des réseaux de neurones

Conséquence de l'approche tropicale

L'approche tropicale permet d'appliquer des théorèmes sur les combinatoires afin de mieux comprendre les réseaux de neurones.

L'étude des polytopes nous permet d'obtenir un résultat sur la somme de Minkowski.

Théorème : Gritzmann Sturmfels[2]

Soit $P_1, ...P_k$ des polytopes de $\mathbb{R}^d$, soit m le nombre de cotés de $\{P_1, ...P_k\}$ n'étant pas parallèles.

Alors le nombre de sommets dans la somme de Minkowski de ces polytopes : $\bigoplus_{i=1}^{k} P_i$ est inférieur à $2\sum_{j=0}^{d-1}\binom{m-1}{j}$.

Conséquence de l'approche tropicale

Théorème [3]

Soit $\nu : \mathbb{R}^d \to \mathbb{R}$ un réseau de neurones vérifiant les conditions $(a) - (c)$ précédentes, et si $n_l \geq d \; \forall l$ dans $[1, L]$;

Alors ν a au plus $\prod_{l=1}^{L-1} \sum_{i=0}^{d} \binom{n_l}{i}$ régions linéaires.

Et pour n réel tel que $n \geq n_l \; \forall l$ alors le nombre de région de ν est borne en $O(n^{d(L-1)})$.

Proof.

Si $L = 2$: la preuve est rédigé dans le mémoire.

Pour $L \geq 3$ la démonstration est plus longue et est disponible dans Zhang et Al. [3].

□

Exploration d'une application

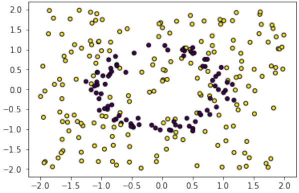

Figure: Exemple de problème de classification en donut 2d

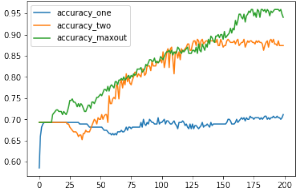

Figure: Précision des réseaux en fonction de leur entraînement

Exploration d'une application

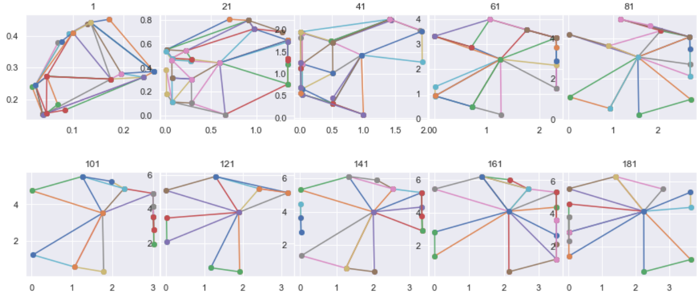

Figure: $\delta(f)$ pour un réseau à 1 couche cachée, de largeur h = 10 à différentes étapes de l'entraînement

Exploration d'une application

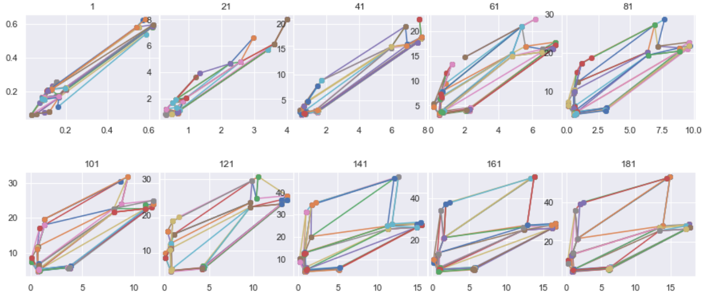

Figure: $\delta(f)$ pour un réseau à 2 couches cachées, de largeurs h = 4 à différentes étapes de l'entraînement

Conclusion et Références I

Des possibilités intéressantes de poursuite de cette approche comprenne l'interprétation de propriétés empiriques des réseaux de neurones concernant les noyaux, la régularisation ou la profondeur des réseaux.

[1] Bernd Sturmfels Diane Maclagan. *Introduction to Tropical Geometry*. 2015.

[2] Peter Gritzmann and Bernd Sturmfels. Minkowski addition of polytopes: Computational complexity and applications to gro¨bner bases. *SIAM J. Discret. Math.*, 6(2):246–269, May 1993.

[3] Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *ICML*, 2018.