Compréhension théorique des comportements non linéaires dans les grands réseaux de neurones

Colloque GRETSI 2019

Zhenyu Liao, Romain Couillet

CentraleSupélec, Université Paris-Saclay
GIPSA-lab, Université Grenoble-Alpes

Lille, 26 août 2019

Plan

1. Introduction

2. Résultats Principaux

3. Conclusion

Motivation: non-linéarité dans d'apprentissage automatique

Apprentissage automatique repose sur transformation **non linéaire**:

- la *fonction noyau* f dans les méthodes à noyaux (e.g., spectral clustering, SVM)
- la *fonction d'activation* σ pour les réseaux de neurones

Dans ce travail:

- lien entre réseaux de neurones (à poids aléatoires) et matrices à noyaux (i.e., σ et f)
- étude de l'effet des non-linéarités (σ et f) et des interactions aux données (nombre, dimension et statistiques)
- conséquence pratique: "**prédire**" la performance d'un réseau de neurones en fonction de la fonction d'activation σ appliquée

Réseau de neurones à une seule couche cachée

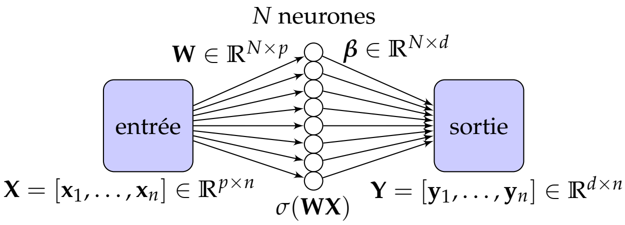

Figure: Illustration d'un réseau à une seule couche cachée.

- $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \Rightarrow$ matrice de "features": $\boldsymbol{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) = [\sigma(\mathbf{W}\mathbf{x}_1), \ldots, \sigma(\mathbf{W}\mathbf{x}_n)]$.
- Seconde couche $\boldsymbol{\beta} \in \mathbb{R}^{N \times d}$ apprise sur $(\mathbf{X}, \mathbf{Y})$:

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \|\mathbf{Y} - \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{\Sigma}\|_F^2 + \lambda \|\boldsymbol{\beta}\|_F^2 = \frac{1}{n} \boldsymbol{\Sigma} \left(\frac{1}{n} \boldsymbol{\Sigma}^{\mathsf{T}} \boldsymbol{\Sigma} + \gamma \mathbf{I}_n \right)^{-1} \mathbf{Y}^{\mathsf{T}}$$

- **Objet clé**: $\frac{1}{n} \boldsymbol{\Sigma}^{\mathsf{T}} \boldsymbol{\Sigma}$: matrice de corrélation dans l'espace "features".

Comprendre la performance d'un réseau de neurones simple

Hypothèse 1: régime asymptotique

- $n, p, N \to \infty$ avec $n \sim p \sim N$;
- $(\mathbf{X}, \mathbf{Y})$ déterministe, $\|\mathbf{X}\| = O(1)$ et $\mathbf{Y}_{ij} = O(1)$.

Hypothèse 2: poids aléatoires

$\mathbf{W}_{ij} \sim \mathcal{N}(0,1)$ i.i.d.

Dans [Louart et al., 2018], pour σ *Lipschitziennes*:

Conclusion 1: impact de la fonction d'activation, lien entre σ et f

Performance du réseau fonction de σ via matrice à noyau

$$\mathbf{\Phi}(\mathbf{X}) \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})], \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$$

ou son entrée (i,j): $\mathbf{\Phi}(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{x}_i^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j)] \equiv f(\mathbf{x}_i, \mathbf{x}_j)$.

[1] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2) :1190–1248, 2018.

Calculer f pour différentes fonctions d'activation σ

Intuition:

- l'objet clé $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} = \frac{1}{n}\sigma(\mathbf{X}^\mathsf{T}\mathbf{W}^\mathsf{T})\sigma(\mathbf{W}\mathbf{X}) = \frac{1}{n}\sum_{i=1}^N \sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)\sigma(\mathbf{w}_i^\mathsf{T}\mathbf{X})$
- $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $i = 1, \ldots, N$, indépendant
- asymptotique $N \to \infty$

$$\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} \Rightarrow \frac{1}{n}\mathbb{E}[\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}] = \frac{1}{n}\sum_{i=1}^N \mathbb{E}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)\sigma(\mathbf{w}_i^\mathsf{T}\mathbf{X})] = \frac{N}{n}\mathbf{\Phi}(\mathbf{X}),$$

avec l'entrée (i,j) $\mathbf{\Phi}(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbb{E}_\mathbf{w}[\sigma(\mathbf{x}_i^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j)] \equiv f(\mathbf{x}_i, \mathbf{x}_j)$.

Trouver f pour différentes σ: intégrale dans $\mathbb{R}^p$

$$f(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbb{E}_\mathbf{w}[\sigma(\mathbf{x}_i^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j)] = (2\pi)^{-\frac{p}{2}} \int_{\mathbb{R}^p} \sigma(\mathbf{x}_i^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j) e^{-\frac{\|\mathbf{w}\|^2}{2}} d\mathbf{w}$$

$\Rightarrow$ solution explicite pour certains σ courants, e.g., $\mathrm{ReLU}(t) \equiv \max(t, 0)$, quadratique $\sigma(t) = a_2 t^2 + a_1 t + a_0$, exponentiel $\sigma(t) = \exp(-t^2/2)$.

Feuille de route

$\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} = \frac{1}{n}\sigma(\mathbf{X}^\mathsf{T}\mathbf{W}^\mathsf{T})\sigma(\mathbf{W}\mathbf{X}) \Rightarrow \mathbf{\Phi}(\mathbf{X}) = \{f(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$, par conséquent $\sigma \Rightarrow f$.

Tableau de f pour différentes σ

$\sigma(t)$	$f(\mathbf{x}_i,\mathbf{x}_j)$
t	$\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$
$\max(t,0)$	$\frac{1}{2\pi}\Vert\mathbf{x}_i\Vert\Vert\mathbf{x}_j\Vert\left(\angle(\mathbf{x}_i,\mathbf{x}_j)\arccos\left(-\angle(\mathbf{x}_i,\mathbf{x}_j)\right)+\sqrt{1-\angle(\mathbf{x}_i,\mathbf{x}_j)^2}\right)$
$\vert t\vert$	$\frac{2}{\pi}\Vert\mathbf{x}_i\Vert\Vert\mathbf{x}_j\Vert\left(\angle(\mathbf{x}_i,\mathbf{x}_j)\arcsin\left(\angle(\mathbf{x}_i,\mathbf{x}_j)\right)+\sqrt{1-\angle(\mathbf{x}_i,\mathbf{x}_j)^2}\right)$
$\mathrm{erf}(t)$	$\frac{2}{\pi}\arcsin\left(\frac{2\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\sqrt{(1+2\Vert\mathbf{x}_i\Vert^2)(1+2\Vert\mathbf{x}_j\Vert^2)}}\right)$
$1_{t>0}$	$\frac{1}{2}-\frac{1}{2\pi}\arccos\left(\angle(\mathbf{x}_i,\mathbf{x}_j)\right)$
$\mathrm{sign}(t)$	$\frac{2}{\pi}\arcsin\left(\angle(\mathbf{x}_i,\mathbf{x}_j)\right)$
$\cos(t)$	$\exp\left(-\frac{1}{2}\left(\Vert\mathbf{x}_i\Vert^2+\Vert\mathbf{x}_j\Vert^2\right)\right)\cosh(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$
$\sin(t)$	$\exp\left(-\frac{1}{2}\left(\Vert\mathbf{x}_i\Vert^2+\Vert\mathbf{x}_j\Vert^2\right)\right)\sinh(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$

Table: Valeurs de $f(\mathbf{x}_i,\mathbf{x}_j)$ pour $\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p)$, avec $\angle(\mathbf{x}_i,\mathbf{x}_j)\equiv\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\Vert\mathbf{x}_i\Vert\Vert\mathbf{x}_j\Vert}$.

Toujours peu explicite avec **non linéarité** et pas facile à comprendre ou interpréter!

Interaction entre f et les statistiques des données

Classification binaire: modèle de mélange gaussien

Les $\mathbf{x}$ sont aléatoires et tirés indépendamment d'un modèle de mélange Gaussien à deux classes $\mathcal{C}_1, \mathcal{C}_2$:

$$\mathcal{C}_1 : \sqrt{p}\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1) \quad \text{versus} \quad \mathcal{C}_2 : \sqrt{p}\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$$

pour $\boldsymbol{\mu}_a \in \mathbb{R}^p$ et $\mathbf{C}_a \in \mathbb{R}^p$, $a = 1, 2$.

Objective: influence des différentes σ pour classifier $\mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$.

Hypothèse 3: "distance" minimale

Pour $p \to \infty$, on demande

- $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = O(1)$;
- $\|\mathbf{C}_a\| = O(1)$, $\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2) = O(\sqrt{p})$ et $\|\mathbf{C}_1 - \mathbf{C}_2\|_F^2 = O(p)$.

Théorème 1: comportement asymptotique de $\mathbf{\Phi}(\mathbf{X})$ [Liao et Couillet, 2018]

Sous les Hypothèses 1–3, pour $\mathbf{\Phi}(\mathbf{X}) = \{f(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$ et (σ, f) listés dans Table 1,

$$\|\mathbf{\Phi} - \tilde{\mathbf{\Phi}}\| \xrightarrow{p.s.} 0, \quad \tilde{\mathbf{\Phi}} = d_1 \cdot \mathbf{M}_1(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) + d_2 \cdot \mathbf{M}_2(\mathbf{C}_1, \mathbf{C}_2) + *$$

quand $n, p \to \infty$.

Conclusion 2: impact des non linéarités dans la classification

- l'influence de la non linéarité (σ et f) ne dépend que **deux** scalaires d_1 et d_2;
- ces deux paramètres "contrôlent" **indépendamment** les moyennes $\boldsymbol{\mu}_a$ et les covariances $\mathbf{C}_a$.

[2] Zhenyu Liao, Romain Couillet, "On the Spectrum of Random Features Maps of High Dimensional Data". *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, 80: 3063–3071, 2018.

Comprendre la fonction d'activation σ

$\sigma(t)$	$f(\mathbf{x}_i,\mathbf{x}_j)$	d_1	d_2
t	$\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$	1	0
$\sin(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{x}_i\|^2+\|\mathbf{x}_j\|^2\right)\right)\sinh(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$	$e^{-\tau}$	0
$\operatorname{erf}(t)$	$\frac{2}{\pi}\arcsin\left(\frac{2\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\sqrt{(1+2\|\mathbf{x}_i\|^2)(1+2\|\mathbf{x}_j\|^2)}}\right)$	$\frac{4}{\pi(2\tau+1)}$	0
$\|t\|$	$\frac{2}{\pi}\|\mathbf{x}_i\|\|\mathbf{x}_j\|\left(\angle(\mathbf{x}_i,\mathbf{x}_j)\arcsin\left(\angle(\mathbf{x}_i,\mathbf{x}_j)\right)+\sqrt{1-\angle(\mathbf{x}_i,\mathbf{x}_j)^2}\right)$	0	$\frac{1}{2\pi\tau}$
$\cos(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{x}_i\|^2+\|\mathbf{x}_j\|^2\right)\right)\cosh(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$	0	$\frac{1}{4}e^{-\tau}$
$\exp(-t^2/2)$	$\frac{1}{\sqrt{(1+\|\mathbf{x}_i\|^2)(1+\|\mathbf{x}_j\|^2)-(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)^2}}$	0	$\frac{1}{4(\tau+1)^3}$
$\max(t,0)$	$\frac{1}{2\pi}\|\mathbf{x}_i\|\|\mathbf{x}_j\|\left(\angle(\mathbf{x}_i,\mathbf{x}_j)\arccos\left(-\angle(\mathbf{x}_i,\mathbf{x}_j)\right)+\sqrt{1-\angle(\mathbf{x}_i,\mathbf{x}_j)^2}\right)$	$\frac{1}{4}$	$\frac{1}{8\pi\tau}$

Table: Valeur de $f(\mathbf{x}_i,\mathbf{x}_j)$ et les coefficients associés d_1, d_2 dans Théorème 1, $\angle(\mathbf{x}_i,\mathbf{x}_j)\equiv\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}$.

Conséquence: différents σ en trois groupes

Avec $\tilde{\mathbf{\Phi}} = d_1\cdot\mathbf{M}_1(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)+d_2\cdot\mathbf{M}_2(\mathbf{C}_1,\mathbf{C}_2)+*$,

1. **moyenne-orienté**: $d_1\neq 0$ et $d_2=0$, contient t, $\sin(t)$ et $\operatorname{erf}(t)$ qui "effacent" l'information dans les covariances $\mathbf{M}_2(\mathbf{C}_1,\mathbf{C}_2)$;
2. **cov-orienté**: $d_1=0$ et $d_2\neq 0$ $|t|$, $\cos(t)$ et $e^{-\frac{t^2}{2}}$ et efface les moyennes $\mathbf{M}_1(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)$;
3. **équilibré**: $d_1,d_2\neq 0$, ici pour $\operatorname{ReLU}(t)\equiv\max(t,0)$.

"Choisir" σ en fonction des statistiques discriminantes des données

Feuille de route

$\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} = \frac{1}{n}\sigma(\mathbf{X}^\mathsf{T}\mathbf{W}^\mathsf{T})\sigma(\mathbf{W}\mathbf{X}) \Rightarrow \mathbf{\Phi}(\mathbf{X}) = \{f(\mathbf{x}_i,\mathbf{x}_j)\}_{i,j=1}^n$, on sait $\sigma \Rightarrow f \Rightarrow (d_1, d_2)$.

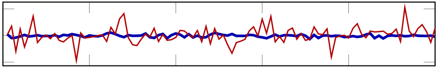

Figure: Base de données MNIST [3] et EEG épileptiques.[4]

	$\Vert\hat{\mu}_1 - \hat{\mu}_2\Vert$	$\Vert\hat{\mathbf{C}}_1 - \hat{\mathbf{C}}_2\Vert$
MNIST (6 vs. 8)	172.4	86.0
EEG (B vs. E)	1.2	182.7

Table: Estimation empirique des statistiques de la base MNIST et EEG.

[3] http://yann.lecun.com/exdb/mnist/

[4] http://www.meb.unibonn.de/epileptologie/science/physik/eegdata.html.

Validation numérique sur la base de données MNIST et EEG

Application: spectral clustering en utilisant $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$.

	$\|\hat{\mu}_1 - \hat{\mu}_2\|$	$\|\hat{\mathbf{C}}_1 - \hat{\mathbf{C}}_2\|$
MNIST (6 vs. 8)	172.4	86.0
EEG (B vs. E)	1.2	182.7

Table: Estimation empirique des statistiques de la base MNIST et EEG.

	$\sigma(t)$	$n = 64$	$n = 128$
moyenne-orienté	t	**88.94**%	87.30%
	$\sin(t)$	87.81%	**87.50**%
	$\mathrm{erf}(t)$	87.28%	86.59%
covariance-orienté	$\|t\|$	60.41%	57.81%
	$\cos(t)$	59.56%	57.72%
	$\exp(-t^2/2)$	60.44%	58.67%
équilibré	$\mathrm{ReLU}(t)$	85.72%	82.27%

Table: Précision de classification sur MNIST.

	$\sigma(t)$	$n = 64$	$n = 128$
moyenne-orienté	t	70.31%	69.58%
	$\sin(t)$	70.34%	68.22%
	$\mathrm{erf}(t)$	70.59%	67.70%
covariance-orienté	$\|t\|$	99.69%	99.50%
	$\cos(t)$	99.38%	99.36%
	$\exp(-t^2/2)$	**99.81**%	**99.77**%
équilibré	$\mathrm{ReLU}(t)$	87.91%	90.97%

Table: Précision de classification sur EEG.

Conclusion

Messages:

- étude de la matrice de corrélation dans l'espace "features (non linéaires)": $\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$, pour comprendre la fonction d'activation σ
- comportement lié à la matrice à noyau $\boldsymbol{\Phi}(\mathbf{X}) = f\{\mathbf{x}_i, \mathbf{x}_j\}_{i,j=1}^n$, $\sigma \Rightarrow f$
- dans la classification des mélanges de gaussiennes, $\boldsymbol{\Phi}(\mathbf{X})$ est (asymptotiquement) accessible, et dépend de f seulement via la paire (d_1, d_2)
- conséquence: on sait comment "choisir" la fonction σ pour différents types de problèmes/données

References:

- Cosme Louart, Zhenyu Liao, and Romain Couillet. "A random matrix approach to neural networks". *The Annals of Applied Probability*, 28(2) :1190–1248, 2018.
- Zhenyu Liao, Romain Couillet, "On the Spectrum of Random Features Maps of High Dimensional Data". *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, 80: 3063–3071, 2018.

Merci!