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

Zhenyu LIAO[1,2], Romain COUILLET[1,2] *

[1]Laboratoire des signaux et systèmes, 3 rue Joliot Curie, 91192 Gif sur Yvette, France

[2]Laboratoire GIPSA-lab, 11 rue des Mathématiques, 38400 St Martin d'Hères, France

zhenyu.liao@l2s.centralesupelec.fr,romain.couillet@gipsa-lab.grenoble-inp.fr

Résumé – Dans cet article, nous faisons une revue des progrès récents pour la compréhension théorique des comportements non linéaires dans les grands réseaux de neurones. En combinant les avancées récentes de la théorie des grandes matrices aléatoires et l'outil de la concentration des mesures, nous apportons un éclairage nouveau sur la compréhension théorique des fonctions d'activation non linéaires ainsi que la dynamique d'apprentissage des réseaux de neurones simples.

Abstract – In this paper we review recent progress in the theoretical understanding of nonlinear behaviors in large neural networks. Combining recent advances in random matrix theory and the concentration of measure phenomenon, we shed new light on the theoretical understanding of nonlinear activation functions as well as the nonlinear learning dynamics of simple neural networks.

1 Introduction

Basé sur l'exploitation de grands jeux de données, les réseaux de neurones grands et profonds sont devenus aujourd'hui des incontournables pour l'apprentissage (classification, modèles prédictifs non linéaires). Ces outils extrêmement puissants atteignent parfois des performances surhumaines sur des tâches diverses. Malgré tous ces résultats exceptionnels, la compréhension théorique de ces grands systèmes complexes progresse à un rythme beaucoup plus modeste.

Un des principaux obstacles à une compréhension approfondie du mécanisme sous-jacent aux réseaux neuronaux modernes est la non linéarité de ces systèmes. Par exemple, dans [2] les auteurs ont analysé le "landscape" d'un réseau *linéaire* à une seule couche cachée et montré que tous les minimaux locaux sont essentiellement globaux. Ce résultat a ensuite été étendu au réseau linéaire *profond* avec un nombre arbitraire de couches [6]. Toutefois, lorsque l'on considère les réseaux de neurones non linéaires, la plupart des résultats théoriques ne sont établis que pour quelques fonctions d'activation spécifiques (en particulier, les fonctions ReLU et quadratique [5, 4]). Une compréhension plus générale des différentes non-linéarités, et de leur interaction aux données (leurs nombre, dimension et statistiques), fait encore défaut.

Un autre phénomène non linéaire intéressant est la dynamique d'apprentissage lorsque les réseaux sont obtenus par le biais de méthodes d'optimisation (notamment par descente de gradient). Alors que les réseaux de neurones modernes sont toujours issus de la méthode de descente de gradient, en raison de la nature non linéaire et non convexe des fonctions de coût du problème d'apprentissage, il est extrêmement difficile d'évaluer à quel point le réseau est "bien entraîné". Il a même été montré dans [13] que, dans le cas du réseau linéaire, la dynamique d'apprentissage au cours de la descente de gradient est en fait fortement non linéaire. Par conséquent, la compréhension de la dynamique *non linéaire* d'apprentissage dans la descente de gradient est d'une importance capitale pour quantifier les performances d'apprentissage.

Par ailleurs, les analyses théoriques des réseaux de neurones se limitent souvent au régime asymptotique où le nombre de neurones N et la dimension des données p sont considérées constantes, tandis que le nombre de données n tend vers l'infini. Cette hypothèse n'est plus valable pour les réseaux de neurones modernes qui sont souvent "sur-paramétrés" ($N > n$), et encore plus erronée à l'époque actuelle du "BigData" où les données sont intrinsèquement de très grandes dimensions. Dans cet article, nous nous plaçons dans le régime plus pertinent où n, p, N sont grands mais commensurables. En se basant sur l'outil de la théorie des grandes matrices aléatoires, nous traitons la question délicate de la non-linéarité intervenant dans ces systèmes, que nous maitrisons au moyen du phénomène de concentration de la mesure [8]. Cette dernière est précisément issue des tailles n, p, N simultanément grandes et des degrés d'indépendance structurels du modèle de données.

2 Réseaux à poids aléatoires

Considérons un réseau simple muni d'une seule couche cachée de N neurones, comme en Figure 1. Les poids du ré-

*Ce travail a été soutenu par le projet GSTATS UGA IDEX Chaire de DataScience et le projet ANR RMT4GRAPH (ANR-14-CE28-0006).

seau sont "appris" à partir de données d'entraînement $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ et de sorties désirées associées $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$. Une fonction d'activation $\sigma(\cdot)$, souvent non linéaire, est appliquée (entrée par entrée) à la sortie de chaque neurone. Par ailleurs, nous considérons le cas où les entrées de $\mathbf{W} \in \mathbb{R}^{N\times p}$ sont échantillonnées indépendamment et aléatoirement à partir d'une certaine distribution, par exemple $\mathbf{W}_{ij} \sim \mathcal{N}(0,1)$. On note $\mathbf{\Sigma} \equiv \sigma(\mathbf{WX}) = [\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_n] \in \mathbb{R}^{N\times n}$ la matrice des "représentations" aléatoires (ce réseau est aussi connu sous le nom de *random feature maps* [12]). Dans ce contexte, les poids de $\boldsymbol{\beta}$ sont choisis de manière à minimiser :

$$L(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{Y} - \boldsymbol{\beta}^\mathsf{T}\mathbf{\Sigma}\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{\beta}\|^2$$

pour un certain $\gamma > 0$. On obtient alors la solution explicite :

$$\boldsymbol{\beta} = \frac{1}{n}\mathbf{\Sigma}\left(\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} + \gamma \mathbf{I}_T\right)^{-1}\mathbf{Y}^\mathsf{T} \tag{1}$$

de sorte que l'erreur d'entraînement $E_{\text{train}} = \frac{1}{n}\|\mathbf{Y} - \boldsymbol{\beta}^\mathsf{T}\mathbf{\Sigma}\|_F^2$ et celle de test $E_{\text{test}} = \frac{1}{\hat{n}}\|\hat{\mathbf{Y}} - \boldsymbol{\beta}^\mathsf{T}\sigma(\mathbf{W}\hat{\mathbf{X}})\|_F^2$ (sur l'ensemble de données de test $\hat{\mathbf{X}} \in \mathbb{R}^{p\times\hat{n}}$ et les sorties correspondantes $\hat{\mathbf{Y}} \in \mathbb{R}^{d\times\hat{n}}$) sont directement accessibles à partir de (1).

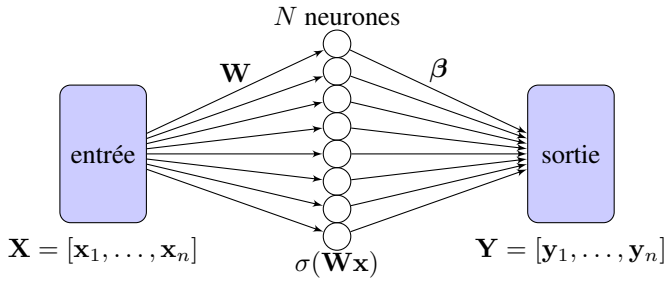

FIGURE 1 – Illustration d'un réseau à une seule couche cachée.

Que ce soit dans le cas présent des réseaux aléatoires ou pour la dynamique de descente de gradient (Section 4), l'apprentissage est fortement lié à la "résolvante"

$$\mathbf{Q}(z) \equiv \left(\frac{1}{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} - z\mathbf{I}_T\right)^{-1} \tag{2}$$

($z \in \mathbb{C} \setminus \mathbb{R}^+$) de la matrice de covariance empirique $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$, qui devient donc naturellement l'objet central à analyser.

Dans un premier temps, nous considérons que les données et les sorties sont déterministes et étudions donc l'influence des dimensions et de la non linéarité du problème. Nous travaillerons (comme dans [3]) sous les hypothèses suivantes.

Hypothèse 1. *Lorsque $n \to \infty$,*

1. $\frac{p}{n} = c_p \to c_1^\infty \in (0,\infty)$, $\frac{N}{n} = c_N \to c_2^\infty \in (0,\infty)$,
2. *la norme d'opérateur $\|\mathbf{X}\| = O(1)$ et $\mathbf{Y}_{ij} = O(1)$.*

La difficulté principale de l'étude de $\mathbf{Q}(z)$ est liée à la non linéarité constitutive de $\mathbf{\Sigma}$. C'est ici que la théorie de la concentration de la mesure prend tout son sens : pour σ Lipschitz (avec constante de Lipschitz indépendante de n), $\mathbf{W} \mapsto \sigma(\mathbf{WX})$ est également Lipschitz et on montre alors que les *fonctionnelles linéaires de* $\mathbf{Q}$ *et* $\mathbf{\Sigma Q}$ (et donc en particulier les erreurs E_{train} et E_{test}) ont un comportement asymptotique déterministe.

Théorème 1 (Performance du réseau)**.** *Sous l'hypothèse 1, pour tout $\varepsilon > 0$ et $\sigma(\cdot)$ Lipschitz,*

$$n^{\varepsilon-\frac{1}{2}}\left(E_{\text{train}} - \bar{E}_{\text{train}}\right) \xrightarrow{p.s.} 0, \quad n^{\varepsilon-\frac{1}{2}}\left(E_{\text{test}} - \bar{E}_{\text{test}}\right) \xrightarrow{p.s.} 0$$

où, pour $\bar{\mathbf{Q}} = \left(c_N\frac{\mathbf{\Phi}}{1+\delta} + \gamma\mathbf{I}_n\right)^{-1}$ et δ l'unique solution positive de l'équation $\delta = \frac{1}{N}\operatorname{tr}(\mathbf{\Phi}\bar{\mathbf{Q}})$, on a défini

$$\bar{E}_{\text{train}} = \frac{\gamma^2}{n}\operatorname{tr}\mathbf{Y}\bar{\mathbf{Q}}\left[\frac{\frac{1}{N}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{\Psi}\bar{\mathbf{Q}})}{1-\frac{1}{N}\operatorname{tr}(\mathbf{\Psi}^2\bar{\mathbf{Q}}^2)}\mathbf{\Psi} + \mathbf{I}_n\right]\bar{\mathbf{Q}}\mathbf{Y}^\mathsf{T}, \tag{3}$$

$$\bar{E}_{\text{test}} = \frac{1}{\hat{n}}\left\|\hat{\mathbf{Y}}^\mathsf{T} - \mathbf{\Psi}_{\mathbf{X}\hat{\mathbf{X}}}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{Y}^\mathsf{T}\right\|_F^2 + \frac{\frac{1}{N}\operatorname{tr}\left(\mathbf{Y}\bar{\mathbf{Q}}\mathbf{\Psi}\bar{\mathbf{Q}}\mathbf{Y}^\mathsf{T}\right)}{1-\frac{1}{N}\operatorname{tr}(\mathbf{\Psi}^2\bar{\mathbf{Q}}^2)}$$
$$\times\left[\frac{1}{\hat{n}}\operatorname{tr}\mathbf{\Psi}_{\hat{\mathbf{X}}\hat{\mathbf{X}}} - \frac{1}{\hat{n}}\operatorname{tr}(\mathbf{I}_n + \gamma\bar{\mathbf{Q}})(\mathbf{\Psi}_{\mathbf{X}\hat{\mathbf{X}}}\mathbf{\Psi}_{\hat{\mathbf{X}}\mathbf{X}}\bar{\mathbf{Q}})\right] \tag{4}$$

avec $\mathbf{\Phi_{AB}} \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{A}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{B})]$ pour toute paire de matrices $(\mathbf{A},\mathbf{B})$ de tailles appropriées, $\mathbf{w} \sim \mathcal{N}(\mathbf{0},\mathbf{I}_p)$ ainsi que

$$\mathbf{\Psi_{AB}} \equiv c_N\frac{\mathbf{\Phi_{AB}}}{1+\delta}, \quad \mathbf{\Phi} = \mathbf{\Phi_{XX}}, \quad \mathbf{\Psi} = \mathbf{\Psi_{XX}}.$$

Les matrices $\mathbf{\Phi_{AB}}$ sont au cœur du Théorème 1. Leur évaluation requiert de calculer, pour tout vecteur $\mathbf{a},\mathbf{b} \in \mathbb{R}^p$, la valeur de $\mathbf{\Phi_{ab}} \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{a})\sigma(\mathbf{w}^\mathsf{T}\mathbf{b})]$. La Table 1 liste un certain nombre de ces valeurs pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0},\mathbf{I}_p)$. Nous renvoyons les lecteurs à [11] pour les détails des preuve du Théorème 1 et calculs de $\mathbf{\Phi_{ab}}$.

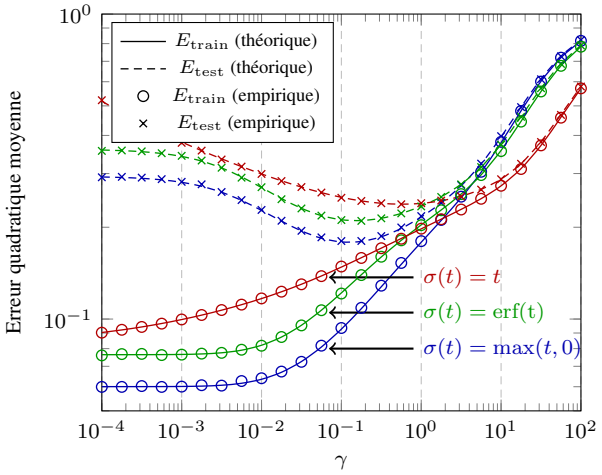

FIGURE 2 – Performance du réseau aléatoire pour σ Lipschitz, en fonction de γ, pour des données MNIST (chiffres sept et neuf), $N = 512$, $n = \hat{n} = 1024$, $p = 784$.

La Figure 2 présente la performance du réseau pour quelque fonctions Lipschitz σ (linéaire, $\operatorname{erf}(t)$ et $\operatorname{ReLU}(t) \equiv \max(t,0)$) en fonction de l'hyper-paramètre γ, sur la base de données

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

TABLE 1 – Valeur de $\mathbf{\Phi_{ab}}$ et les coefficients associés d_1, d_2 dans Théorème 2, $\angle(\mathbf{a},\mathbf{b}) \equiv \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

MNIST [7]. Nous comparons l'erreur quadratique moyenne empirique à l'approximation asymptotique donné par le Théorème 1 et constatons une forte adéquation de nos résultats théoriques, en dépit de petites valeurs de N, n, p.

3 Non-linéarité dans le noyau équivalent

D'après la Section 2, les performances du réseau dépendent de σ à travers le "noyau équivalent" $\mathbf{\Phi}$. Néanmoins, Théorème 1 est peu explicite. Pour mieux étudier l'influence de σ sur l'apprentissage, nous considérons maintenant que les données $\mathbf{X}$ sont aléatoires et tirées indépendamment d'un modèle de mélange Gaussien à K classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$:

$$\mathbf{x}_i = \boldsymbol{\mu}_a/\sqrt{p} + \boldsymbol{\omega}_i, \quad \boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_a/p)$$

pour $\boldsymbol{\mu}_a \in \mathbb{R}^p$ et $\mathbf{C}_a \in \mathbb{R}^{p\times p}$, $a = 1, \ldots, K$, satisfaisant :

Hypothèse 2. *Lorsque $n \to \infty$, pour $a = 1, \ldots, K$,*

1. *le cardinal $n_a = |\mathcal{C}_a|$ est tel que $n_a/n \to c_a \in (0,1)$,*
2. *$\|\boldsymbol{\mu}_a\| = O(1)$,*
3. *$\|\mathbf{C}_a\| = O(1)$ et $\mathrm{tr}(\mathbf{C}_a^\circ) = O(\sqrt{p})$, où $\mathbf{C}^\circ \equiv \sum_{a=1}^K \frac{n_a}{n}\mathbf{C}_a$ et $\mathbf{C}_a^\circ \equiv \mathbf{C}_a - \mathbf{C}^\circ$,*
4. *$\tau \equiv \mathrm{tr}(\mathbf{C}^\circ)/p$ converge dans $(0,\infty)$.*

L'hypothèse 2 garantit que les classes $\mathcal{C}_a$ ne sont ni trop simples ni impossibles à identifier [3]. En exploitant l'hypothèse 2, on obtient en particulier

$$\|\mathbf{x}_i\|^2 = \tau + \underbrace{\frac{1}{p}\mathrm{tr}(\mathbf{C}_a^\circ) + \|\boldsymbol{\omega}_i\|^2 - \mathbb{E}\|\boldsymbol{\omega}_i\|^2}_{O(p^{-1/2})} + \underbrace{\frac{1}{p}\|\boldsymbol{\mu}_a\|^2 + \frac{2}{\sqrt{p}}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i}_{O(p^{-1})}$$

qui permet d'effectuer une expansion de Taylor autour de τ (qui est d'ordre $O(1)$) lorsque $n, p \to \infty$. Ceci donne lieu à approximation des fonctions de $\|\mathbf{x}_i\|$ et $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ apparaissant dans $\mathbf{\Phi}_{\mathbf{x}_i\mathbf{x}_j}$. Sur la base de cette idée, nous obtenons un équivalent asymptotique simplifié $\tilde{\mathbf{\Phi}}$ de $\mathbf{\Phi}$ (i.e., la norme d'opérateur $\|\mathbf{\Phi} - \tilde{\mathbf{\Phi}}\| \overset{p.s.}{\to} 0$ quand $n, p \to \infty$) ne faisant intervenir que des fonctions élémentaires des statistiques.

Théorème 2. *Sous les Hypothèses 1 et 2, notons (la version recentrée) $\mathbf{\Phi}_c \equiv \mathbf{P}\mathbf{\Phi}\mathbf{P}$ avec $\mathbf{P} \equiv \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$. Alors,*

$$\|\mathbf{\Phi}_c - \tilde{\mathbf{\Phi}}_c\| \overset{p.s.}{\to} 0, \quad \tilde{\mathbf{\Phi}}_c = \mathbf{P}\tilde{\mathbf{\Phi}}\mathbf{P}$$

$$\tilde{\mathbf{\Phi}} \equiv d_1\left(\mathbf{\Omega} + \mathbf{M}\frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}}\right)^\mathsf{T}\left(\mathbf{\Omega} + \mathbf{M}\frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}}\right) + d_2\mathbf{U}\mathbf{B}\mathbf{U}^\mathsf{T} + d_0\mathbf{I}_n$$

pour $\mathbf{U} \equiv \left[\frac{\mathbf{J}}{\sqrt{p}}, \boldsymbol{\phi}\right]$, $\mathbf{B} \equiv \begin{bmatrix}\mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S} & \mathbf{t}\\ \mathbf{t}^\mathsf{T} & 1\end{bmatrix}$ *et*

$$\mathbf{\Omega} \equiv \left[\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n\right] \in \mathbb{R}^{p\times n},\ \boldsymbol{\phi} \equiv \left\{\|\boldsymbol{\omega}_i\|^2 - \mathbb{E}\|\boldsymbol{\omega}_i\|^2\right\}_{i=1}^n \in \mathbb{R}^n,$$
$$\mathbf{M} \equiv \left[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\right] \in \mathbb{R}^{p\times K},\ \mathbf{J} \equiv \left[\mathbf{j}_1, \ldots, \mathbf{j}_K\right] \in \mathbb{R}^{n\times K},$$
$$\mathbf{t} \equiv \{\mathrm{tr}\,\mathbf{C}_a^\circ/\sqrt{p}\}_{a=1}^K \in \mathbb{R}^K,\ \mathbf{S} \equiv \{\mathrm{tr}(\mathbf{C}_a\mathbf{C}_b)/p\}_{a,b=1}^K \in \mathbb{R}^{K\times K}$$

où $\mathbf{j}_a$ est le vecteur canonique de la classe $\mathcal{C}_a$ tel que $(\mathbf{j}_a)_i = \delta_{\mathbf{x}_i\in\mathcal{C}_a}$. Les coefficients d_1, d_2 sont donnés dans Table 1 (d_0 n'influence pas les performances et est donc omis).

D'après la Table 1, les fonctions σ se divisent en trois groupes :

1. *moyenne-orienté* : $d_1 \neq 0$ et $d_2 = 0$. C'est le cas de t, $\sin(t)$ et $\mathrm{erf}(t)$ qui "effacent" les covariances $(\mathbf{t}, \mathbf{S})$,
2. *covariance-orienté* : $d_1 = 0$ et $d_2 \neq 0$. Ce groupe contient $\cos(t)$ et $\exp(-t^2/2)$ et efface les moyennes $(\mathbf{M})$,
3. *équilibré* : $d_1, d_2 \neq 0$. Ici pour $\mathrm{ReLU}(t) \equiv \max(t, 0)$.

Ces groupes fournissent des instructions sur le choix de σ en fonction des *statistiques discriminantes* des données. Nous renvoyons les lecteurs à [10] pour plus de détails.

Nous complétons cette section en montrant que nos résultats théoriques dans Théorème 2, obtenus sur des modèles de mélange Gaussien, s'étendent fidèlement à des données réelles. Nous considérons deux jeux de données : MNIST [7] et des séries temporelles d'EEG [1]. Ces données offrent des exemples de moyenne-dominante (MNIST où $\|\mathbf{M}^\mathsf{T}\mathbf{M}\| \approx 2\|\mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S}\|$) et covariance-dominante (EEG où $\|\mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S}\| \approx 100\|\mathbf{M}^\mathsf{T}\mathbf{M}\|$). La Table 2 liste les précisions de classification non-supervisée basée sur $\mathbf{\Phi}_c$ pour différents σ avec $n = 128$. Nous observons un avantage concurrentiel lorsque des "bons" σ sont utilisés.

4 La dynamique de descente de gradient

Dans le réseau de la Figure 1, le poids $\boldsymbol{\beta}$ est explicitement donné par (1). Nous pouvons alternativement imaginer apprendre $\boldsymbol{\beta}$ par la méthode de descente de gradient. Pour simplifier le problème, nous considérons ici que $\mathbf{Y} = \mathbf{y} \in \mathbb{R}^n$ est un vecteur et que $\mathbf{W} = \mathbf{I}_d$ avec $\sigma(t) = t$, et cherchons $\boldsymbol{\beta}$ qui minimise

$$L(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y}^\mathsf{T} - \boldsymbol{\beta}^\mathsf{T}\mathbf{X}\|^2.$$

	$\sigma(t)$	MNIST	EEG
moyenne-orientée	t	87.30%	69.58%
	$\sin(t)$	**87.50%**	68.22%
	$\mathrm{erf}(t)$	86.59%	67.70%
covariance-orientée	$\cos(t)$	57.72%	99.36%
	$\exp(-t^2/2)$	58.67%	**99.77%**
équilibrée	$\mathrm{ReLU}(t)$	82.27%	90.97%

TABLE 2 – Précisions de classification pour $\sigma(\cdot)$ différentes

La descente de gradient effectue des pas de descente de taille $\alpha \ll 1$, de sorte qu'en effectuant une approximation en temps continu, on obtient l'équation différentielle $\frac{d\boldsymbol{\beta}(t)}{dt} = -\alpha \frac{dL(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \frac{\alpha}{n}\mathbf{X}\left(\mathbf{y} - \mathbf{X}^\mathsf{T}\boldsymbol{\beta}(t)\right)$, dont la solution *non-linéaire en* $\mathbf{X}$ est :

$$\boldsymbol{\beta}(t) = e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\boldsymbol{\beta}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\right)(\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}\mathbf{y}. \quad (5)$$

Puisque $e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ partage le même espace propre que $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ et a pour valeurs propres $\exp(-\alpha t\lambda_i)$ ($\lambda_i = \lambda_i(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T})$), pour évaluer la dynamique de l'erreur de classification pour une nouvelle donnée $\hat{\mathbf{x}}$, on doit quantifier $\boldsymbol{\beta}^\mathsf{T}\hat{\mathbf{x}}|\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\mu}_a, \boldsymbol{\beta}^\mathsf{T}\mathbf{C}_a\boldsymbol{\beta})$. Nous sommes donc amenés à analyser $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\mu}_a$ et $\boldsymbol{\beta}^\mathsf{T}\mathbf{C}_a\boldsymbol{\beta}$, qui sont des fonctionnelles de la matrice de covariance empirique $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$. Nous recourons ici à la formule de l'intégrale de Cauchy qui nous permet de dire, par exemple, que

$$\mathbf{a}^\mathsf{T}e^{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\mathbf{b} = -\frac{1}{2\pi i}\oint_\gamma \exp(z)\mathbf{a}^\mathsf{T}\mathbf{Q}(z)\mathbf{b}dz$$

avec γ un contour qui entoure les valeurs propres de $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ et $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ dont on connaît le comportement (établi dans les sections précédentes). En outre, dans certains cas simples, par exemple lorsque $\mathbf{C}_a = \mathbf{I}_p$, les intégrales complexes se réduisent à des intégrales réelles.

Notre analyse permet de montrer que, lorsque $t \to \infty$ et donc $\exp(-\alpha t\lambda_i) \to 0$ ce qui recouvre la solution des moindres carrés, le réseau devient "sur-ajusté" et la performance chute d'un facteur $\sqrt{1 - \min(\frac{p}{n}, \frac{n}{p})}$. Par conséquent, le problème de sur-apprentissage n'est sévère que lorsque $p \approx n$, ce qui justifie l'avantage de sur-paramétrer le réseau. La Figure 3 valide nos résultats théoriques sur la base MNIST. Des détails et discussions supplémentaires sont disponibles dans [9].

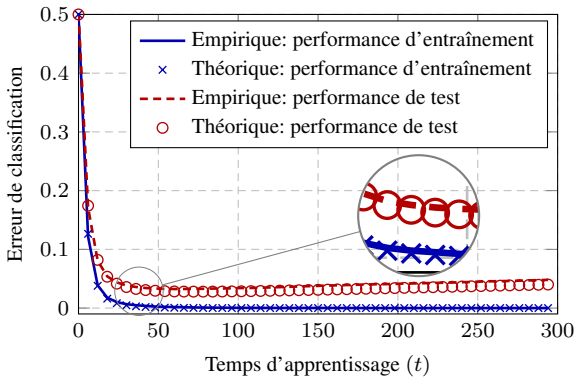

FIGURE 3 – Erreurs de classification (entraînement et test) sur la base MNIST pour $n = p = 784$, $c_1 = c_2 = 1/2$, $\alpha = 0.01$.

Références

[1] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity : Dependence on recording region and brain state. *Physical Review E*, 64(6) :061907, 2001.

[2] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis : Learning from examples without local minima. *Neural networks*, 2(1) :53–58, 1989.

[3] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. Classification asymptotics in the random matrix regime. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1875–1879. IEEE, 2018.

[4] Simon S Du and Jason D Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv :1803.01206*, 2018.

[5] Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn : Don't be afraid of spurious local minima. *arXiv preprint arXiv :1712.00779*, 2017.

[6] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[7] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[8] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

[9] Zhenyu Liao and Romain Couillet. The dynamics of learning : A random matrix approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3072–3081, 2018.

[10] Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3063–3071, 2018.

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

[12] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[13] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv :1312.6120*, 2013.