Une Analyse des Méthodes de Projections Aléatoires par la Théorie des Matrices Aléatoires

Zhenyu Liao, Romain Couillet

Une Analyse des Méthodes de Projections Aléatoires par la Théorie des Matrices Aléatoires

Zhenyu LIAO, Romain COUILLET *

Laboratoire des signaux et systèmes, CentraleSupélec, CNRS, Université Paris-Sud
3, Rue Joliot-Curie, 91192, Gif-sur-Yvette, France
zhenyu.liao@l2s.centralesupelec.fr, romain.couillet@centralesupelec.fr

Résumé – Dans cet article, nous présentons nos résultats théoriques sur les méthodes de projections aléatoires pour les données de très grandes dimensions, exploitées notamment dans les réseaux de neurones aléatoires. En combinant les avancées récentes de la théorie des matrices aléatoires et l'outil de la concentration des mesures, nous étudions en particulier les performances asymptotiques des dites *extreme learning machines*, permettant ainsi une compréhension plus profonde de ce réseau de neurone aléatoire simple.

Abstract – In this paper we present our theoretical findings on *random feature maps* or *random projections* for large dimensional data in large systems, classically found in different types of random neural networks. Combining recent advances in random matrix theory and the concentration of measure phenomenon, we provide a new framework on the theoretical analysis of random feature maps in large dimensional problems. As a concrete example, a study of the asymptotic performance of *extreme learning machines* is provided.

1 Introduction

Les réseaux de neurones grands et profonds sont devenus aujourd'hui des incontournables pour l'apprentissage (à partir de grandes bases de données), notamment de classificateurs ou modèles prédictifs non linéaires. Cependant, l'apprentissage de réseaux de neurones profonds reste aujourd'hui coûteux, peu adaptable et présuppose toujours de grandes bases de données d'apprentissage.

Une alternative possible, contre-intuitive au premier abord, consiste à utiliser des méthodes de projections aléatoires afin d'extraire les caractéristiques (*features* en anglais) des données permettant d'effectuer les tâches d'apprentissage. Dans ce nouvel espace projectif, des méthodes d'apprentissage classiques (régression linéaire ou logistique, machine à vecteurs de support, etc.) peuvent alors être utilisées. Ces méthodes apparaissent sous diverses formes dans la littérature : sous le nom des *extreme learning machines* (ELMs) [3] dans le contexte des réseaux de neurones, de *random feature maps* dans le cas d'approximations aléatoires de noyaux [8], de réseaux de neurones récurrents aléatoires dits *echo state networks* [4], etc.

Ce type de réseaux connait un récent regain d'intérêt ces dernières années, de par ses nombreux avantages : 1) simplicité d'utilisation et adaptabilité à des grands nombres de données, 2) surapprentissage limité sous un nombre faible de données, 3) reparamétrisation facile et adaptabilité aux flux des données (à l'aide, par exemple, de la méthode des moindres carrés récursifs [2]) et 4) une promesse théorique du domaine des projections aléatoires selon laquelle $k\log(k)$ projections aléatoires sont suffisantes pour extraire l'information pertinente de données de très grandes dimensions représentables dans une base de dimension k.

L'étude théorique de ces méthodes de projections aléatoires se limite souvent à l'hypothèse d'un nombre de données T de dimension p constants tandis que le nombre de neurones n tend vers l'infini, comme dans l'exemple populaire des *random fourier features* [8]. Dans cet article, nous nous plaçons dans le régime plus pertinent où n, p, T sont grands mais commensurables. En se basant sur l'outil de la théorie des grandes matrices aléatoires, nous traitons la non-linéarité délicate intervenant dans ces méthodes au moyen du phénomène de concentration de la mesure [6] ainsi qu'en exploitant les degrés d'indépendance structurels du modèle. La contribution principale du travail porte sur l'analyse spectrale de la matrice corrélation des données dans l'espace des projections aléatoires non-linéaires. Cette étude est ensuite appliquée au problème concret de l'analyse des performances asymptotiques des méthodes de réseaux de neurones de type *extreme learning machine* (ELM) [3].

Notations : La norme $\|\cdot\|$ est la norme euclidienne pour des vecteurs et la norme spectrale pour des matrices, tandis que $\|\cdot\|_F$ est la norme de Frobenius. La notation $\xrightarrow{p.s.}$ indique la converge presque sûre.

2 Énoncé du Problème

Les projecteurs aléatoires consistent à construire, à partir des données d'entraînement $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p\times T}$ une

*Ce travail a été soutenu par le projet ANR RMT4GRAPH (ANR-14-CE28-0006). Ce travail était présenté partiellement au 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA.

matrice de "caractéristiques" (*features*) $\boldsymbol{\Sigma} \equiv [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_T] \in \mathbb{R}^{n \times T}$ obtenue en multipliant $\mathbf{X}$ par une matrice de poids aléatoire $\mathbf{W} \in \mathbb{R}^{n \times p}$, souvent telle que les W_{ij} sont i.i.d. $\mathcal{N}(0,1)$, et ensuite en appliquant une fonction d'activation non-linéaire $\sigma : \mathbb{R} \mapsto \mathbb{R}$ élément par élément à la matrice $\mathbf{WX}$, pour obtenir donc la matrice $\boldsymbol{\Sigma} = \sigma(\mathbf{WX}) \in \mathbb{R}^{n \times T}$.

Dans le cas des *random fourier features* ou bien le cas de l'ELM, les méthodes d'apprentissage sont toujours basées sur le spectre de la matrice de covariance empirique dans l'espace des caractéristiques $\frac{1}{T}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$, qui devient donc naturellement l'objet central à analyser. La théorie des grandes matrices aléatoires prédit que le spectre de cette matrice est essentiellement relié à la transformée de Stieltjes de la résolvante associée

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

pour $z \in \mathbb{C} \setminus \mathbb{R}^+$, qui est par conséquent l'objet clé dans la compréhension de ces méthodes.

Afin d'étudier le comportement asymptotique des projections aléatoires dans le régime où $n, p, T \to \infty$, nous travaillerons sous les hypothèses de croissance suivantes.

Hypothèse 1 (Taux de Croissance)**.** *Lorsque $n \to \infty$,*

1. $p/n \to c_0 \in (0, \infty)$, $T/n \to c_T \in (0, \infty)$
2. $\|\mathbf{X}\| = O(1)$
3. *pour évaluer la performance d l'ELM, nous avons besoin en plus de la sortie associée $\mathbf{Y}$ telle que $\mathbf{Y}_{ij} = O(1)$.*

Nous avons par ailleurs besoin d'imposer la condition de régularité suivante de la fonction d'activation σ.

Hypothèse 2 (Fonction d'activation σ)**.** *La fonction σ est Lipschitzienne avec constante de Lipschitz λ_σ indépendante de n.*

La plus grande difficulté de cette analyse porte sur la non-linéarité de la matrice $\boldsymbol{\Sigma}$. Ce problème est résolu en exploitant la caractère Lipschitz de $\sigma(\cdot)$ qui permet alors d'appliquer des inégalités de concentration de la mesure sur l'application $\mathbf{W} \mapsto \sigma(\mathbf{WX})$. Par ce biais, nous parvenons à démontrer que l'espérance $\mathbb{E}[\mathbf{Q}]$ de la matrice aléatoire $\mathbf{Q}$ a un comportement asymptotique dans le régime $n, p, T \to \infty$ qui s'exprime en termes d'objets complètement déterministes.

3 Résultats Principaux

Notre analyse commence par un résultat important lié au phénomène de concentration de la mesure [6]. Les vecteurs aléatoires gaussiens satisfont une propriété dite de *concentration normale* qui peut alors se propager à travers toute application Lipschitzienne. Ceci nous permet en particulier d'assurer que, pour $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_p)$ et $\mathbf{X} \in \mathbb{R}^{p \times T}$ tel que $\|\mathbf{X}\| = 1$, la norme de $\frac{1}{\sqrt{p}}\boldsymbol{\sigma} \equiv \frac{1}{\sqrt{p}}\sigma(\mathbf{X}^\mathsf{T}\mathbf{w}) \in \mathbb{R}^T$ est d'ordre $O(1)$ avec haute probabilité. De plus, conditionnellement au fait que $\frac{1}{\sqrt{p}}\boldsymbol{\sigma}$ est de norme $O(1)$, l'application $\mathbf{w} \mapsto \frac{1}{p}\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\boldsymbol{\sigma}$ est Lipschitzienne. Avec un contrôle précis des résultats de concentration sous conditionnement, nous obtenons le lemme suivant.

Lemme 1 (Concentration de la forme quadratique)**.** *Sous l'hypothèse 1–2, pour $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ et $\boldsymbol{\sigma} \equiv \sigma(\mathbf{X}^\mathsf{T}\mathbf{w}) \in \mathbb{R}^T$. Nous avons pour toute matrice $\mathbf{A} \in \mathbb{R}^{T \times T}$ telle que $\|\mathbf{A}\| \leq 1$ et indépendante de $\boldsymbol{\sigma}$,*

$$P\left(\left| \frac{1}{n}\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\boldsymbol{\sigma} - \frac{1}{n}\operatorname{tr}(\boldsymbol{\Phi}\mathbf{A}) \right| > t \right) \leq Ce^{-cn\min(t,t^2)}$$

pour certains $c, C > 0$ indépendants de n, avec $\boldsymbol{\Phi} \equiv \mathbb{E}[\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}]$.

Le lemme 1 nous permet d'estimer les formes quadratiques $\frac{1}{n}\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\boldsymbol{\sigma}$ par $\frac{1}{n}\operatorname{tr}(\boldsymbol{\Phi}\mathbf{A})$ avec probabilité exponentiellement proche de 1, qui est essentiellement la version non-linéaire du lemme de la trace (e.g., [1, Lemma B.26]) dans la théorie des grandes matrices aléatoires. Ainsi, en utilisant maintenant ce lemme ainsi que le fait que la matrice

$$\boldsymbol{\Sigma}^\mathsf{T} \equiv \sigma(\mathbf{X}^\mathsf{T}\mathbf{W}^\mathsf{T}) = [\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n]$$

avec $\mathbf{W}^\mathsf{T} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$ et $\boldsymbol{\sigma}_i = \sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i) \in \mathbb{R}^T$, a des colonnes indépendantes (mais des lignes fortement dépendantes), nous pouvons utiliser des méthodes dorénavant classiques de la théorie des matrices aléatoires afin d'obtenir le résultat suivant.

Théorème 1 (Équivalent asymptotique de $\mathbb{E}[\mathbf{Q}]$)**.** *Sous l'hypothèse 1–2, pour tout $\varepsilon > 0$, il existe $c > 0$ tel que, pour tout n,*

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \leq cn^{\varepsilon - \frac{1}{2}}.$$

où $\bar{\mathbf{Q}}$ est donnée par

$$\bar{\mathbf{Q}} = \left(\frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\delta} - z\mathbf{I}_T \right)^{-1}$$

avec δ l'unique solution positive de l'équation $\delta = \frac{1}{T}\operatorname{tr}(\boldsymbol{\Phi}\bar{\mathbf{Q}})$.

Trame de la Preuve. Nous introduisons $\alpha \equiv \frac{1}{T}\operatorname{tr}\left(\boldsymbol{\Phi}\mathbb{E}[\mathbf{Q}]\right)$ et $\tilde{\mathbf{Q}} \equiv (\frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\alpha} - z\mathbf{I}_T)^{-1}$. Nous déduisons de l'*identité de la résolvante* $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$ que

$$\mathbb{E}[\mathbf{Q}] - \tilde{\mathbf{Q}} = \mathbb{E}[\mathbf{Q} - \tilde{\mathbf{Q}}] = \mathbb{E}[\mathbf{Q}]\frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\alpha}\tilde{\mathbf{Q}} - \mathbb{E}\left[\mathbf{Q}\frac{1}{T}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}\right]\tilde{\mathbf{Q}}$$

$$= \mathbb{E}[\mathbf{Q}]\frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\alpha}\tilde{\mathbf{Q}} - \frac{1}{T}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{Q}\boldsymbol{\sigma}_i\boldsymbol{\sigma}_i^\mathsf{T}\right]\tilde{\mathbf{Q}}.$$

Avec la *formule de Sherman-Morrison*, pour

$$\mathbf{Q}_{-i} \equiv \left(\frac{1}{T}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} - \frac{1}{T}\boldsymbol{\sigma}_i\boldsymbol{\sigma}_i^\mathsf{T} - z\mathbf{I}_T \right)^{-1}$$

nous obtenons

$$\mathbb{E}[\mathbf{Q}] - \tilde{\mathbf{Q}} = \mathbb{E}[\mathbf{Q}]\frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\alpha}\tilde{\mathbf{Q}} - \frac{1}{T}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathbf{Q}_{-i}\boldsymbol{\sigma}_i\boldsymbol{\sigma}_i^\mathsf{T}}{1 + \frac{1}{T}\boldsymbol{\sigma}_i^\mathsf{T}\mathbf{Q}_{-i}\boldsymbol{\sigma}_i}\right]\tilde{\mathbf{Q}}$$

$$= \mathbb{E}[\mathbf{Q}]\frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\alpha}\tilde{\mathbf{Q}} - \frac{1}{T}\frac{1}{1+\alpha}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{Q}_{-i}\boldsymbol{\sigma}_i\boldsymbol{\sigma}_i^\mathsf{T}\right]\tilde{\mathbf{Q}}$$

$$+ \frac{1}{T}\frac{1}{1+\alpha}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathbf{Q}_{-i}\boldsymbol{\sigma}_i\boldsymbol{\sigma}_i^\mathsf{T}(\frac{1}{T}\boldsymbol{\sigma}_i^\mathsf{T}\mathbf{Q}_{-i}\boldsymbol{\sigma}_i - \alpha)}{1 + \frac{1}{T}\boldsymbol{\sigma}_i^\mathsf{T}\mathbf{Q}_{-i}\boldsymbol{\sigma}_i}\right]\tilde{\mathbf{Q}}.$$

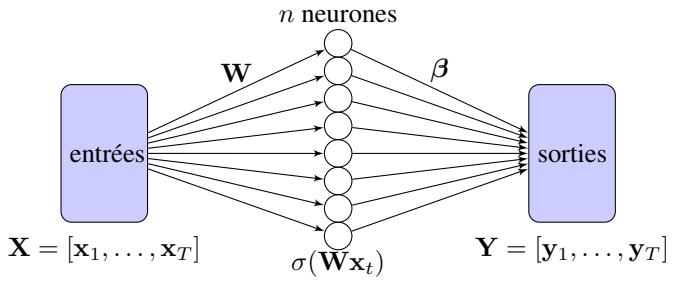

FIGURE 1 – Illustration d l'ELM

Puisque $\mathbf{Q}_{-i}$ est indépendant de $\boldsymbol{\sigma}_i$, nous obtenons alors

$$\mathbb{E}[\mathbf{Q}] - \tilde{\mathbf{Q}} = \mathbb{E}[\mathbf{Q}]\frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\alpha}\tilde{\mathbf{Q}} - \frac{1}{T}\frac{1}{1+\alpha}\sum_{i=1}^{n}\mathbb{E}[\mathbf{Q}_{-i}]\boldsymbol{\Phi}\tilde{\mathbf{Q}} + \frac{1}{T}\frac{1}{1+\alpha}\mathbb{E}\left[\mathbf{Q}\boldsymbol{\Sigma}^\mathsf{T}\mathbf{D}\boldsymbol{\Sigma}\right]\tilde{\mathbf{Q}}$$

avec $\mathbf{D} = \mathrm{diag}(\{\frac{1}{T}\boldsymbol{\sigma}_i^\mathsf{T}\mathbf{Q}_{-i}\boldsymbol{\sigma}_i - \alpha\}_{i=1}^n)$ qui est telle que $\|\mathbf{D}\| \to 0$ quand $n \to \infty$ d'après le lemme 1. De plus, nous montrons que $\|\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}]\| = O(n^{-1})$, ce qui nous dit que

$$\|\mathbb{E}[Q] - \tilde{\mathbf{Q}}\| \to 0,$$

ensuite en observant le fait que

$$\alpha = \frac{1}{T}\mathrm{tr}\left(\boldsymbol{\Phi}\mathbb{E}[\mathbf{Q}]\right) = \frac{1}{T}\mathrm{tr}\left(\boldsymbol{\Phi}(\frac{n}{T}\frac{\boldsymbol{\Phi}}{1+\alpha} - z\mathbf{I}_T)^{-1}\right) + o(1)$$

avec $\delta \equiv \frac{1}{T}\mathrm{tr}(\boldsymbol{\Phi}\bar{\mathbf{Q}})$ nous concluons la preuve. □

4 Exemple Concret de l'ELM

Dans cette section, nous nous concentrons sur l'exemple concret de l'ELM, qui est une simple régression linéaire normalisée de $\boldsymbol{\Sigma}$ par rapport à une sortie désirée $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T] \in \mathbb{R}^{d\times T}$ associée à $\mathbf{X}$ donnée par le régresseur

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

avec γ le facteur de régularisation strictement positif, comme illustré dans la Figure 1.

L'erreur d'entraînement de l'ELM pour les données d'entraînement $\mathbf{X}$ est ainsi donnée par

$$E_{\text{train}} = \frac{1}{T}\left\|\mathbf{Y}^\mathsf{T} - \boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\beta}\right\|_F^2 = \frac{\gamma^2}{T}\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{Y}\mathbf{Q}^2\right). \quad (2)$$

Pendant la *phase de test*, nous considérons l'ensemble de données de test : $\hat{\mathbf{X}} \in \mathbb{R}^{p\times\hat{T}}$ de taille $\hat{T}$, dont les sorties correspondantes sont notées $\hat{\mathbf{Y}} \in \mathbb{R}^{d\times\hat{T}}$. En appliquant le régresseur $\boldsymbol{\beta}$ défini par (1) (qui ne dépend que de $\boldsymbol{\Sigma}$ et $\mathbf{Y}$), nous obtenons l'erreur de test

$$E_{\text{test}} = \frac{1}{T}\left\|\hat{\mathbf{Y}}^\mathsf{T} - \hat{\boldsymbol{\Sigma}}^\mathsf{T}\boldsymbol{\beta}\right\|_F^2 \quad (3)$$

avec $\hat{\boldsymbol{\Sigma}} \equiv \sigma(\mathbf{W}\hat{\mathbf{X}})$. Pour la convenance de l'étude de E_{test} dans la suite, nous étendons la définition de $\boldsymbol{\Phi}$ dans Lemme 1 pour toute paire de matrices $(\mathbf{A}, \mathbf{B})$ à

$$\boldsymbol{\Phi}_{\mathbf{AB}} \equiv \mathbb{E}\left[\sigma(\mathbf{w}^\mathsf{T}\mathbf{A})^\mathsf{T}\sigma(\mathbf{w}^\mathsf{T}\mathbf{B})\right]$$

avec $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ et notons

$$\boldsymbol{\Psi}_{\mathbf{AB}} \equiv \frac{n}{T}\frac{\boldsymbol{\Phi}_{\mathbf{AB}}}{1+\delta}.$$

Pour simplifier les notations, nous notons $\boldsymbol{\Phi} = \boldsymbol{\Phi}_{\mathbf{XX}}$ et $\boldsymbol{\Psi} = \boldsymbol{\Psi}_{\mathbf{XX}}$.

En étendant le lemme 1, nous pouvons démontrer que les erreurs E_{train} et E_{test} convergent (donc vers leur espérance). Concernant $\mathbb{E}[E_{\text{train}}]$, nous avons notamment

$$\mathbb{E}[E_{\text{train}}] = \frac{\gamma^2}{T}\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{Y}\mathbb{E}[\mathbf{Q}^2]\right)$$

qui nous conduit à l'analyse asymptotique de $\mathbb{E}[\mathbf{Q}^2]$ ou plus généralement de $\mathbb{E}[\mathbf{QAQ}]$ pour toute matrice $\mathbf{A}$ déterministe, qui nécessite plus d'efforts et est annoncé sans preuve ici. Nous renvoyons les lecteurs à [7] pour la démonstration complète.

Proposition 1 (Équivalent asymptotique de $\mathbb{E}[\mathbf{QAQ}]$)**.** *Soit $\mathbf{A}$ la matrice $\boldsymbol{\Phi}$ ou une matrice symétrique de norme bornée. Alors, sous l'hypothèse 1–2, avec la matrice $\bar{\mathbf{Q}}$ donnée dans le théorème 1, pour tout $\varepsilon > 0$, il existe une constante c telle que*

$$\left\|\mathbb{E}[\mathbf{QAQ}] - \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} - \bar{\mathbf{Q}}\boldsymbol{\Psi}\bar{\mathbf{Q}}\frac{\frac{1}{n}\mathrm{tr}\left(\boldsymbol{\Psi}\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\right)}{1 - \frac{1}{n}\mathrm{tr}\left(\boldsymbol{\Psi}^2\bar{\mathbf{Q}}^2\right)}\right\| \leq cn^{\varepsilon-\frac{1}{2}}.$$

En prenant $\mathbf{A}$ la matrice d'identité, nous accédons à l'équivalent déterministe de l'erreur d'entraînement E_{train} dans le régime où $n, p, T \to \infty$. La dérivation de l'erreur de test nécessite plus d'efforts et sera annoncée sans démonstration dans le théorème suivant.

Théorème 2 (Performance de l'ELM)**.** *Sous l'hypothèse 1–2, pour tout $\varepsilon > 0$,*

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

où

$$\bar{E}_{\text{train}} = \frac{\gamma^2}{T}\mathrm{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{Y}\bar{\mathbf{Q}}\left[\frac{\frac{1}{n}\mathrm{tr}(\bar{\mathbf{Q}}\boldsymbol{\Psi}\bar{\mathbf{Q}})}{1 - \frac{1}{n}\mathrm{tr}(\boldsymbol{\Psi}^2\bar{\mathbf{Q}}^2)}\boldsymbol{\Psi} + \mathbf{I}_T\right]\bar{\mathbf{Q}}\right) \quad (4)$$

et $\bar{E}_{\text{test}}$ est donnée par l'équation (6).

Notons que $\bar{\mathbf{Q}}$ qui dépend implicitement de la matrice $\boldsymbol{\Phi}$ est au cœur du Théorème 2. L'évaluation de la matrice $\boldsymbol{\Phi}_{\mathbf{AB}}$ pour deux matrices arbitraires $\mathbf{A}, \mathbf{B}$ requiert naturellement l'évaluation de ses entrées et donc revient à calculer, pour deux vecteurs arbitraires $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, la valeur de $\Phi(\mathbf{a}, \mathbf{b}) \equiv \mathbb{E}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{a})\sigma(\mathbf{w}^\mathsf{T}\mathbf{b})]$, avec $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Les résultats pour quelques fonctions d'activation σ couramment utilisées sont fournis dans Table 1. Encore une fois, nous renvoyons les lecteurs à [7] dans lequel tous les détails de calculs sont donnés.

La matrice $\boldsymbol{\Phi}$ joue ici le rôle de la matrices à noyau qui serait obtenue dans la limite $n \to \infty$ seul comme prédit dans [8]. Il est facile de voir en effet ici que pour $T/n, p/n \to 0$, l'ELM devient équivalent à une régression linéaire par noyau. Ces liens entre matrice à noyau et projections aléatoires sont aussi mis en évidence dans le contexte d'autres méthodes d'apprentissage, ce qui confère à cette étude une dimension d'ouverture plus large que seulement restreintes aux ELM.

$$\bar{E}_{\text{test}} = \frac{1}{\hat{T}} \left\| \hat{\mathbf{Y}}^\mathsf{T} - \mathbf{\Psi}_{\mathbf{X}\hat{\mathbf{X}}}^\mathsf{T} \bar{\mathbf{Q}} \mathbf{Y}^\mathsf{T} \right\|_F^2 + \frac{\frac{1}{n} \operatorname{tr}\left(\mathbf{Y}^\mathsf{T} \mathbf{Y} \bar{\mathbf{Q}} \mathbf{\Psi} \bar{\mathbf{Q}} \right)}{1 - \frac{1}{n} \operatorname{tr}(\mathbf{\Psi}^2 \bar{\mathbf{Q}}^2)} \left[\frac{1}{\hat{T}} \operatorname{tr} \mathbf{\Psi}_{\hat{\mathbf{X}}\hat{\mathbf{X}}} - \frac{1}{\hat{T}} \operatorname{tr}(\mathbf{I}_T + \gamma \bar{\mathbf{Q}})(\mathbf{\Psi}_{\mathbf{X}\hat{\mathbf{X}}} \mathbf{\Psi}_{\hat{\mathbf{X}}\mathbf{X}} \bar{\mathbf{Q}}) \right]. \tag{6}$$

$\sigma(t)$	$\Phi_{\mathbf{ab}}$
t	$\mathbf{a}^\mathsf{T}\mathbf{b}$
$\max(t,0)$	$\frac{1}{2\pi}\|\mathbf{a}\|\|\mathbf{b}\| \left(\angle(\mathbf{a},\mathbf{b}) \arccos\left(-\angle(\mathbf{a},\mathbf{b})\right) + \sqrt{1-\angle(\mathbf{a},\mathbf{b})^2} \right)$
$\operatorname{erf}(t)$	$\frac{2}{\pi} \arcsin\left(\frac{2\mathbf{a}^\mathsf{T}\mathbf{b}}{\sqrt{(1+2\|\mathbf{a}\|^2)(1+2\|\mathbf{b}\|^2)}} \right)$
$\cos(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right)\right) \cosh(\mathbf{a}^\mathsf{T}\mathbf{b})$
$\sin(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right)\right) \sinh(\mathbf{a}^\mathsf{T}\mathbf{b})$

TABLE 1 – Valeur de $\Phi_{\mathbf{ab}}$ pour $w \sim \mathcal{N}(0, \mathbf{I}_p)$, $\angle(\mathbf{a},\mathbf{b}) \equiv \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

5 Illustrations Numériques

Nous testons maintenant nos résultats théoriques sur la base de données MNIST [5]. Ici, nous considérons une tâche de classification binaire des images des chiffres manuscrits de taille 28×28, chaque image étant représentée par un vecteur de taille $p = 784$, avec une ELM de $n = 512$ unités de neurones et $\mathbf{W}$ de type Gaussien standard. La base de donnée est composée de $T = \hat{T} = 1\,024$ échantillons : $1\,024$ images du chiffre sept et $1\,024$ du chiffre neuf sont extraits aléatoirement et repartis en 512 échantillons d'entraînement et 512 échantillons de test.

Dans la figure 2 nous évaluons la performance de l'ELM pour certaines fonctions Lipschitziennes σ (linéaire, $\operatorname{erf}(t)$ ainsi que $\operatorname{ReLU}(t) \equiv \max(t,0)$) en fonction de l'hyper-paramètre γ. Nous comparons l'erreur quadratique moyenne empirique à l'approximant asymptotique donné par le théorème 2. Nous pouvons constater une précision impressionnante de nos résultats théoriques dans ce contexte, quand bien même limité à assez peu de données.

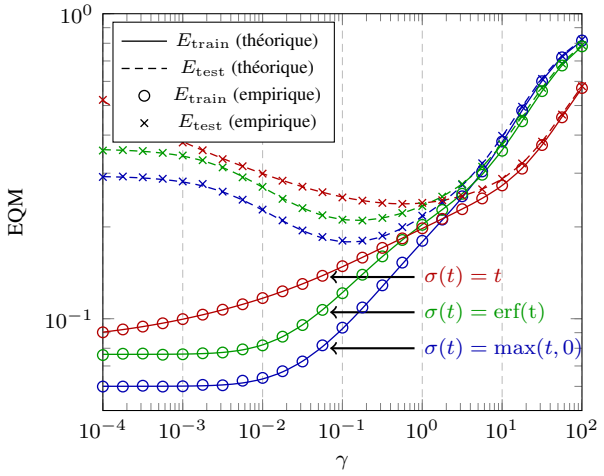

FIGURE 2 – Performance de l'ELM pour σ Lipschitzienne, en fonction de γ, pour 2-classe données de MNIST (sept et neuf), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

6 Conclusion

Ce travail apporte une nouvelle compréhension des méthodes de projections aléatoires à l'aide de l'outil de la théorie des grandes matrices aléatoires. Il ouvre également la voie à l'analyse et l'amélioration d'algorithmes standards dans le contexte plus large de l'apprentissage automatisé des grands systèmes, dont les réseaux de neurones, qui sont au centre des questions modernes du traitement du signal et des données.

Références

[1] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[2] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1996.

[3] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2) :513–529, 2012.

[4] Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany : German National Research Center for Information Technology GMD Technical Report*, 148(34) :13, 2001.

[5] Yann LeCun, Corinna Cortes, and Christopher JC Burges. The MNIST database of handwritten digits, 1998.

[6] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

[7] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *arXiv preprint arXiv :1702.05419*, 2017.

[8] Ali Rahimi, Benjamin Recht, et al. Random Features for Large-Scale Kernel Machines. In *NIPS*, volume 3, page 5, 2007.