Concentration de la Mesure et Apprentissage pour le Traitement des Données de Grande Dimension

Cosme LOUART[1,2], Romain COUILLET[2,3], Mohamed TAMAAZOUSTI[1]

[1]CEA, LIST, LVIC, nano-innov, 91120 Palaiseau

[2]Univ. Grenoble Alpes, CNRS, Grenoble Institute of engenering, GIPSA-lab, 38000 Grenoble

[3]Univ. de Paris-Saclay, CentraleSupélec, L2S, 91190 Gif-sur-Yvette, France

{cosme.louart,romain.couillet}@gipsa-lab.grenoble-inp.fr

Résumé – Nous proposons dans cet article une approche innovante de l'analyse des données réalistes en grande dimension en partant d'hypothèses issues de la théorie de la concentration de la mesure. Ce formalisme autorise des raccourcis efficaces dans l'analyse théorique d'outils au cœur des méthodes de traitement des données et de l'apprentissage en grande dimension. Nous l'illustrons ici à travers l'analyse théorique des performances d'un réseau de neurones aléatoire.

Abstract – In this article an innovative approach to large and realistic dataset analysis is presented. Our study is based on hypotheses issued from the concentration of measure theory. This formalism provides a convenient toolbox to study methods at the heart of data processing and statistical learning in large dimensions. We illustrate this claim through the theoretical performance analysis of a random neural network.

Introduction

L'évaluation statistique des performances de méthodes de traitement des données soulève en premier lieu la question de l'adoption d'un modèle approprié pour les données (images, sons, textes). Pour dépasser l'hypothèse simple mais a priori irréaliste de vecteurs gaussiens, nous étudions ici un modèle de données vérifiant le phénomène de concentration de la mesure (PCM). En deux mots, pour mesurer les variations d'un vecteur aléatoire de grande dimension $X \in \mathbb{R}^p$, on peut évaluer:

- le *diamètre de sa distribution* $\mathcal{D}_d(X) = \mathbb{E}[\|X - \mathbb{E}[X]\|]$
- son *diamètre observable* $\mathcal{D}_o(X) = \sup\{\mathcal{D}_d(f(X)), f : \mathbb{R}^p \to \mathbb{R},\ 1\text{-lipschitzienne}\}$; les fonctionnelles $f(X)$ sont appelées les *observations* de X.

Le PCM a lieu lorsque $\mathcal{D}_d(X) \gg \mathcal{D}_o(X)$. L'exemple historique, donné en Figure 1, est celui d'un vecteur aléatoire X uniformément distribué sur la sphère $\{x \in \mathbb{R}^p, \|x\| = \sqrt{p}\}$; il a été montré en effet que $\mathcal{D}_d(X) = O(\sqrt{p}) \gg \mathcal{D}_o(X) = O(1)$. En pratique, cela signifie que, pour une bonne normalisation, les observations de ces vecteurs aléatoires sont asymptotiquement *prédictibles*: cela induit d'une certaine façon une puissante généralisation de la loi des grands nombres pour des vecteurs à entrées non nécessairement indépendantes.

Cette propriété se retrouve dans un large spectre de distributions. Nous justifierons ici en quoi certains modèles de vecteurs satisfaisant le PCM sont adaptés aux données étudiées en traitement du signal et de l'image en nous appuyant sur l'exemple fondamental des données issues des réseaux de neurones antagonistes génératifs (GAN) [1], récemment développés.

Au delà de la qualité des modèles, nous démontrerons l'efficacité des outils de la théorie que nous avons récemment introduits dans [2, 3] à analyser les performances de techniques avancées en apprentissage. L'atout-clé de la théorie repose ici sur son insensibilité aux transformations non-linéaires fréquentes en apprentissage. En guise d'application, nous estimons les performances asymptotiques d'un réseau de neurones aléatoire sur des données réelles.

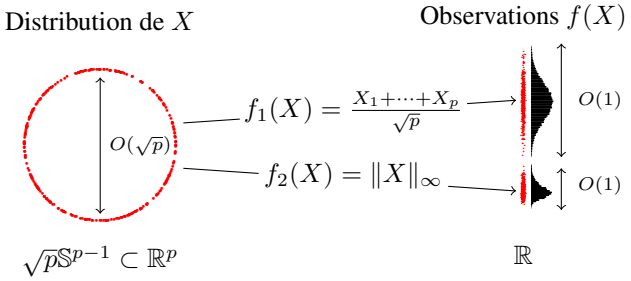

FIG. 1: La distribution uniforme sur la sphère $\sqrt{p}\mathbb{S}^{p-1}$, exemple fondamental du PCM. À gauche, représentation schématique de 500 tirages $X_i \in \mathbb{R}^p$. À droite, concentration des observations des tirages pour des observation linéaire ($f_1(X)$) ou simplement lipschitzienne ($\|X\|_\infty = \sup_{1\leq i\leq p} |X_i|$).

1 Bases de concentration de la mesure

Avant de comprendre les motivations profondes de l'étude du PCM, un certain nombre de notations et de résultats démontrant les facilités calculatrices offertes par la théorie nous sont nécessaires.

Définition 1. *Étant donné un espace vectoriel réel normé* $(E, \|\cdot\|)$ *et une fonction* $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ *(dite fonction de concentration), nous dirons qu'un vecteur aléatoire* $Z \in E$ *est:* α-

***concentré**, ssi pour tout $f : E \to \mathbb{R}$ 1-Lipschitz:*

$$\forall t > 0 \ : \ \mathbb{P}(|f(Z) - \mathbb{E}[f(Z)]| \geq t) \leq \alpha(t).$$

Si, de plus, il existe un vecteur déterministe $\tilde{Z} \in E$ tel que pour toute forme linéaire $u : E \to \mathbb{R}$ de norme opérateur unitaire ($\|u\| \equiv \sup_{\|z\|\leq 1} u(z) = 1$):

$$\forall t > 0 \ : \ \mathbb{P}(|u(Z - \tilde{Z})| \geq t) \leq \alpha(t),$$

*on dira que Z est concentré autour de **l'équivalent déterministe** $\tilde{Z}$ (bien sûr, tout vecteur α-concentré est concentré autour de son espérance).*

Les grands théorèmes du PCM présentés ci-après nous donnent généralement des résultats de concentration dans $(\mathbb{R}^p, \|\cdot\|)$ où $\|\cdot\| : x \mapsto (x_1^2 + \ldots + x_p^2)^{\frac{1}{2}}$ est la norme euclidienne et pour des fonctions de concentration de la forme:

$$\forall t \in \mathbb{R}_+ \ : \ \alpha(t) = \alpha_{\mathcal{N}}(t) \equiv Ce^{-(t/\sigma)^2} \tag{1}$$

où $C \geq 1, \sigma > 0$; on parle alors de *concentration normale* [4].

Pour $X \in E$, $\alpha_{\mathcal{N}}$-concentré et $f : E \to \mathbb{R}$ 1-Lipschitz, on a en particulier la majoration intéressante des moments:

$$\mathbb{E}\left[|f(X) - \mathbb{E}[f(X)]|^r\right] \leq C\left(\frac{r}{2}\right)^{\frac{r}{2}} \sigma^r. \tag{2}$$

Quand les constantes C, σ de concentration normale sont optimisées, C est proche de 1 alors que σ peut être très petit ou très grand; ainsi, du fait de (2), σ est du même ordre que le *diamètre observable* évoqué en introduction.

• **Concentration dans $\mathbb{R}$.** Si X et Y sont deux variables $\alpha_{\mathcal{N}}$-concentrées autour de leurs espérances et $f : \mathbb{R} \to \mathbb{R}$ une fonction λ-Lipschitz, alors $f(X)$, $X + Y$ et XY sont aussi concentrés et on a $\forall t > 0$:

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq Ce^{-(t/\sigma\lambda)^2}, \tag{3}$$

$$\mathbb{P}(|X + Y - \mathbb{E}[X] - \mathbb{E}[Y]| \geq t) \leq 2Ce^{-(t/2\sigma)^2}, \tag{4}$$

$$\mathbb{P}(|XY - \mathbb{E}[X]\mathbb{E}[Y]| \geq t) \tag{5}$$

$$\leq 2C\exp\left(-\frac{t^2}{9\sigma^2|\max(\mathbb{E}[X], \mathbb{E}[Y])|^2}\right) + Ce^{-t/3\sigma^2}.$$

Enfin, si X est $C\alpha_{\mathcal{N}}^{\sigma}$-concentrée autour d'un réel $a \in \mathbb{R}$:

$$\forall r \geq q \ : \ \mathbb{E}[|Z - a|^r] \leq C\left(\frac{r}{2}\right)^{\frac{r}{2}} \sigma^r \tag{6}$$

• **Concentration d'un vecteur aléatoire.** Au delà de la distribution uniforme sur la sphère (Figure 1), l'exemple-phare du PCM est le vecteur gaussien.

Théorème 1.1. *Le vecteur $Z \sim \mathcal{N}(0, I_p)$ est $\alpha_{\mathcal{N}}$-concentré avec C, σ indépendants de p.*

Remark 1.2. *Par définition, le Théorème 1.1 s'étend à toute transformation 1-Lipschitz $F(Z)$ de $Z \sim \mathcal{N}(0, I_p)$.*

On peut ainsi construire des vecteurs concentrés à partir de transformations plus ou moins régulières de vecteurs gaussiens. Cela représente une famille assez large de vecteurs aléatoires (notamment des vecteurs d'entrées à dépendance complexe). Cependant, nous restons incapables de traiter des lois discrètes.

• **Gestion de la norme.** Connaître la concentration d'un vecteur, c'est d'abord pouvoir gérer sa norme. En plus des espaces $(\mathbb{R}^p, \|\cdot\|)$, $p \in \mathbb{N}$ nous pourrons travailler dans:

- $(\mathcal{M}_{p,n}, \|\cdot\|_F)$, où $\forall M \in \mathcal{M}_{p,n}$:
$$\|M\|_F = \sqrt{\mathrm{Tr}(MM^T)},$$
- $(\mathcal{M}_{p,n}, \|\cdot\|)$, où $\forall M \in \mathcal{M}_{p,n}$:
$$\|M\| = \sup_{\|x\|\leq 1} \|Mx\|.$$

Les grands théorèmes de la théorie du PCM, étant généralement énoncés pour des espaces euclidiens, si la norme n'est pas précisée, la concentration a lieu dans $(\mathbb{R}^p, \|\cdot\|)$ ou bien dans $(\mathcal{M}_{p,n}, \|\cdot\|_F)$ (suivant qu'on travaille avec des matrices ou des vecteurs aléatoires).

Nous introduisons dans [3] la notion de *degré de norme* $\eta_{(E,\|\cdot\|)}$ (ou simplement $\eta_{\|\cdot\|}$) que nous ne définirons pas ici mais dont nous donnons quelques exemples en vue de comprendre la Proposition 1.3.

$$\eta_{(\mathbb{R}^p,\|\cdot\|_\infty)} = \log(p) \qquad \eta_{(\mathcal{M}_{p,n},\|\cdot\|)} = n + p$$
$$\eta_{(\mathbb{R}^p,\|\cdot\|)} = p \qquad \eta_{(\mathcal{M}_{p,n},\|\cdot\|_F)} = np.$$

Proposition 1.3. *Pour Z linéairement $\alpha_{\mathcal{N}}$-concentré dans $(E, \|\cdot\|)$ autour de $\tilde{Z}$, on a:*

$$\mathbb{P}(\|Z - \tilde{Z}\| \geq t) \leq Ce^{-\frac{t^2}{\sigma'^2\eta_{\|\cdot\|}^2}} \ \text{ et } \ \mathbb{E}\|Z - \tilde{Z}\| \leq C'\sigma\sqrt{\eta_{\|\cdot\|}}$$

où σ' et C' sont respectivement proportionnels à σ et à C.

• **Concentration des opérations élémentaires.** Pour la concentration de la somme, le cas vectoriel ne se distingue pas du cas d'une somme de variables et on retrouve (4). Pour le produit (matriciel mais aussi produit de Hadamard), on retrouve un résultat analogue à (5), où $|\max(\mathbb{E}[X], \mathbb{E}[Y])|$ est remplacé par $C'\sigma\sqrt{\eta_{\|\cdot\|}}$. Plutôt que de présenter ce résultat général, il convient mieux ici de donner un cas particulier concernant la concentration des formes quadratiques.

Théorème 1.4 (Hanson-Wright)**.** *Étant donné une matrice $A \in \mathcal{M}_p$ telle que $\|A\| \leq 1$ et un vecteur aléatoire $Z \in \mathbb{R}^p$ $\alpha_{\mathcal{N}}$-concentré tel que $\|\mathbb{E}[Z]\| \leq \sigma\sqrt{p}$, on a:*

$$\mathbb{P}\left(\left|\frac{1}{p}Z^TAZ - \frac{1}{p}\mathrm{Tr}(A\Sigma)\right| \geq t\right)$$
$$\leq C'\exp\left(-\frac{pt^2}{\sigma'^4}\right) + C'\exp\left(-\frac{pt}{\sigma'^2}\right)$$

où $\Sigma = \mathbb{E}[ZZ^T]$ (et $C' \propto C$, $\sigma' \propto \sigma$).

2 Vecteurs concentrés et données réelles

Nous discutons ici d'une large famille de vecteurs concentrés, utilisée (vraisemblablement inconsciemment) dans la littérature de l'apprentissage moderne et qui justifie pleinement de modéliser des données réalistes par des vecteurs concentrés. Comme il est a priori impossible de vérifier que les distributions de vraies données vérifient le PCM, nous exploitons le

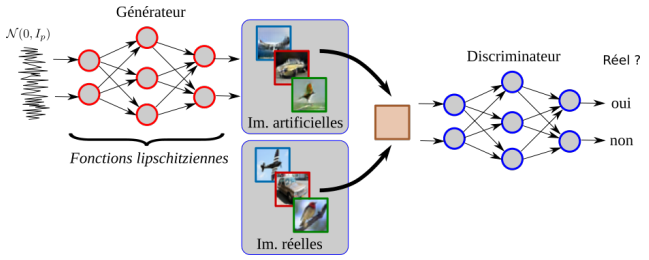

FIG. 2: Réseaux générateur et discriminateur d'un GAN.

fait que des données artificielles *très réalistes* peuvent être issues d'un générateur de vecteurs concentrés.

Pour les images, cette approximation peut être réalisée grâce aux GAN, représentés en Figure 2. Un GAN est un réseau neuronal constitué de deux briques de base (i) un réseau générateur qui tend à construire des images réalistes possible à partir de vecteurs aléatoires $Z \sim \mathcal{N}(0, I_p)$ et (ii) un réseau discriminateur (adverse) qui confronte ces images artificielles à une large banque d'entraînement d'images réelles dont il dispose afin de corriger le réseau générateur. Le réseau générateur étant une succession d'opérateurs linéaires et 1-Lipschitz (rectifieurs de type ReLU en général), du fait de la Remarque 1.2 et du Théorème 1.1, les images produites par les GAN sont des vecteurs concentrés. Cet argument simple justifie la pertinence de l'étude de la concentration de la mesure comme modèle de données, tout du moins en traitement des images [1].

Pour élargir le tableau, nous pouvons ensuite extraire des *représentations* de vecteurs concentrés issus d'un GAN au moyen de réseaux convolutifs (CNN) modernes. Comme illustré en Figure 3, un CNN est un réseau neuronal qui lui aussi opère par successions d'opérateurs Lipschitz et maintient ainsi le PCM de données concentrées. Le traitement a posterio de ces données est donc ici aussi sujet à une analyse par le biais des outils de la PCM, ce qui en l'occurrence couvre bien plus que le cadre restreint du traitement des images.

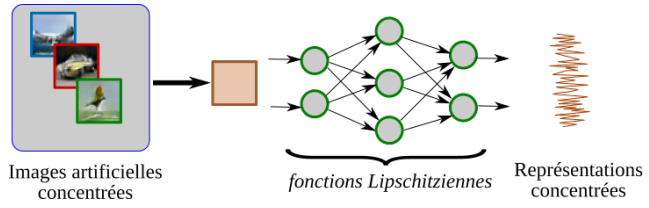

FIG. 3: Concentration de la mesure et CNNs.

3 Application à la prédiction des performances d'un procédé simple de classification

Nous illustrons ici la pertinence de notre modèle avec l'étude de la covariance empirique d'un ensemble de données concentrées qui se retrouve au centre de nombreuses méthodes en traitement du signal et des données. C'est le cas notamment de l'analyse en composantes principales [5], de la régression régularisée, ou même d'algorithmes simples mais efficaces de réseaux neuronaux tels que les dits "extreme learning machines" (ELM) [6] (qui ne sont rien de plus que des projections aléatoires non linéaires). L'étude statistique de la covariance empirique implique le contrôle conjoint de la taille p des données et de leur nombre n que nous allons supposer grands et commensurables.

Hypothèse 1. *$p = O(n)$ et $n = O(p)$.*

On se donne ici n données $x_1, \dots, x_n \in \mathbb{R}^p$ *indépendantes* et identiquement distribuées que l'on range en colonne dans une matrice $X = [x_1, \dots, x_n] \in \mathcal{M}_{p,n}$. La covariance empirique est alors donnée par $S \equiv \frac{1}{n} X X^T$ et la matrice de covariance de la population est notée $\Sigma \equiv \mathbb{E}[S]$ (nous ne soustrayons pas la moyenne $X \mathbb{1}^T \mathbb{1} X^T$, où $\mathbb{1} = (1, \dots, 1)$ car ce terme ne joue aucun rôle dans ce qui suit). Pour rendre possible l'analyse de S, il faut supposer que la matrice des données X vérifie le PCM.

Hypothèse 2. *X est $\alpha_{\mathcal{N}}$-concentré dans $(\mathcal{M}_{p,n}, \|\cdot\|_F)$ pour $C, \sigma = O(n) = O(p)$.*

On ajoute enfin une hypothèse très faible sur le centrage des données pour limiter la norme de Σ (et de S).

Hypothèse 3. $\|\mathbb{E}[X]\|_F = O(\sqrt{np})$.

Assez classiquement l'étude de S, et notamment de ses propriétés spectrales (valeurs propres, vecteurs propres) mais aussi des performances de régresseurs linéaires basés sur X, passe par l'étude de la résolvante $Q = Q(z) = (S + zI_p)^{-1}$ pour $z \in \mathbb{C} \setminus \mathrm{Sp}(S)$ où $\mathrm{Sp}(S)$ désigne le spectre de S [7]. On peut montrer facilement que l'application $X \mapsto Q$ est une application $\frac{2}{\gamma^{3/2}\sqrt{n}}$-Lipschitz, on déduit donc immédiatement la concentration de Q de l'Hypothèse 2.

Proposition 3.1. *Pour toute matrice $A \in \mathcal{M}_p$ telle que $\|A\| \leq 1$, $\forall t > 0$, il existe $\sigma', C' = O(1)$ tels que :*

$$\mathbb{P}(|\operatorname{Tr}(A(Q - \mathbb{E}Q))| \geq t) \leq C' e^{-(t/\sigma'\sqrt{n})^2}.$$

Le dit *équivalent déterministe* $\mathbb{E}[Q]$ (un exemple de $\tilde{Q}$ avec nos notations) n'est pas satisfaisant car nous ne savons pas l'estimer. Un meilleur équivalent déterministe peut être trouvé dans $(\mathcal{M}_{p,n}, \|\cdot\|)$ (et non dans $(\mathcal{M}_{p,n}, \|\cdot\|_F)$ comme précédemment). Pour cela, nous introduisons la notation:

$$\tilde{Q}_\delta = \left(\frac{\Sigma}{1+\delta} + z I_p \right)^{-1}$$

où $\delta > 0$ est une constante à définir. Avec l'identité $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ ($A, B \in \mathcal{M}_p$, inversibles), on a:

$$\begin{aligned} \mathbb{E}[\tilde{Q}_\delta - Q] &= \mathbb{E}\left[Q \left(\frac{1}{n} X X^T - \frac{\Sigma}{1+\delta} \right) \tilde{Q}_\delta \right] \\ &= \sum_{i=1}^n \frac{1}{n} \mathbb{E}\left[Q (x_i x_i^T - \frac{\Sigma}{1+\delta}) \tilde{Q}_\delta \right]. \end{aligned}$$

Posons $X_{-i} = (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n) \in \mathcal{M}_{p,n}$ et $Q_{-i} = (\frac{1}{n} X_{-i} X_{-i}^T + z I_p)^{-1}$ (la résolvante Q privée de la contribution de x_i), on utilise les formules de Schur:

$$Q = Q_{-i} - \frac{1}{n} \frac{Q_{-i} x_i x_i^T Q_{-i}}{1 + \frac{1}{n} x_i^T Q_{-i} x_i} \quad \text{et} \quad Q x_i = \frac{Q_{-i} x_i}{1 + \frac{1}{n} x_i^T Q_{-i} x_i},$$

pour obtenir:

$$\tilde{Q}_\delta - \mathbb{E}Q = \sum_{i=1}^n \frac{1}{n}\mathbb{E}\left[Q_{-i}\left(\frac{x_i x_i^T}{1+\frac{1}{n}x_i^T Q_{-i} x_i} - \frac{\Sigma}{1+\delta}\right)\tilde{Q}_\delta\right] - \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[Q_{-i}x_i x_i^T Q \frac{\Sigma}{1+\delta}\tilde{Q}_1\right]. \quad (7)$$

Pour que $\|\tilde{Q}_\delta - \mathbb{E}Q\|$ soit le plus petit possible, il apparaît judicieux de choisir $\delta = \frac{1}{n}\mathbb{E}[x_i^T Q_{-i} x_i] = \frac{1}{n}\operatorname{Tr}(\Sigma\mathbb{E}[Q_{-i}])$ (on montre en effet qu'avec ce choix, le premier terme de (7) a une norme spectrale d'ordre $O(\frac{1}{\sqrt{n}})$ et le deuxième terme est en $O(\frac{1}{n})$). Mais on en revient au problème de l'estimation de $\mathbb{E}[Q_{-i}]$. Il s'avère alors pertinent de choisir pour δ une solution d'un problème à point fixe en notant qu'un δ idéal vérifierait:

$$\delta \approx \frac{1}{n}x_i^T Q_{-i} x_i \approx \frac{1}{n}\operatorname{Tr}(\Sigma\mathbb{E}[Q_{-i}]) \approx \frac{1}{n}\operatorname{Tr}(\Sigma\tilde{Q}_\delta)$$

Nous avons pu vérifier dans [3] qu'il existe un $\delta > 0$ tel que ces approximations soient bien valides, on a alors le théorème:

Théorème 3.2. *L'équation* $\delta = \frac{1}{n}\operatorname{Tr}(\Sigma\tilde{Q}_\delta)$ *admet une unique solution positive* $\delta > 0$. *Il existe alors deux constantes* $\sigma', C' = O(1)$ *telles que pour tout* $A \in \mathcal{M}_p$ *vérifiant* $\|A\|_1 \equiv \operatorname{Tr}((AA^T)^{\frac{1}{2}}) \leq 1$ ($\|\cdot\|_1$ *est la norme duale de* $\|\cdot\|$*):*

$$\mathbb{P}(|\operatorname{Tr}(A(Q-\tilde{Q}_\delta))| \geq t) \leq C' e^{-(t/\sigma'\sqrt{n})^2}.$$

Ce théorème donne un équivalent déterministe pour Q qui ne dépend que du moment d'ordre 2 (Σ) des données. Ce résultat puissant généralise le cas de matrices X à entrées linéairement dépendantes jusqu'alors étudié dans la littérature et est fondamental lorsqu'on cherche à prédire les performances de nombreuses méthodes d'apprentissage basées sur X.

Prenons ici l'exemple d'une régression linéaire à la sortie d'un projecteur aléatoire non linéaire (dit "random feature map"). On dispose ici d'un vecteur cible (ou d'étiquettes en classification) *déterministe* $Y \in \mathbb{R}^n$ et on cherche un régresseur $\beta \in \mathbb{R}^p$ tel que $Y \approx \beta X$ en minimisant:

$$\frac{1}{n}\|\beta X - Y\|^2 + \gamma\|\beta\|^2 \text{ pour } \gamma > 0 \text{ donné.}$$

La solution fait apparaître la résolvante étudiée plus haut et vaut:

$$\beta = \frac{1}{n}YXQ(\gamma).$$

On attribue alors à toute nouvelle donnée $x \in \mathbb{R}^p$ (indépendante de X) la valeur (ou score ou étiquette)

$$S(x) \equiv \beta x = \frac{1}{n}YXQ(\gamma)x$$

qu'on sait estimer grâce au Théorème 3.2.

Supposons maintenant qu'au lieu d'approcher Y par βX, on l'approche, suivant la même méthode, par $\beta\sigma(WX)$ où $W \in \mathcal{M}_{q,p}$ est une matrice aléatoire (typiquement gaussienne) et σ une application Lipschitz agissant terme à terme sur WX. On se trouve alors dans le cadre d'objets statistiques au cœur de la performances des ELM présentés dans l'introduction. La Figure 4 présente un exemple de performances empiriques et leur approximation théorique par nos outils obtenus pour ces ELMs pour différentes fonctions Lipschitz σ. La figure confirme que notre théorie prédit finement les performances de ce réseau de neurones simple mais non-linéaire.

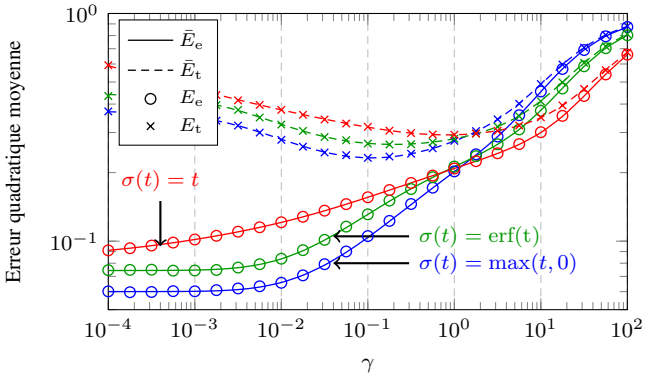

FIG. 4: Performances d'un ELM pour trois applications 1-Lipschitz $\sigma(\cdot)$, $W_{ij} \sim \mathcal{N}(0,1)$, pour deux classes de données MNIST (chiffres sept et neuf) pour $n = 1024$, $p = 784$. $E_e = \frac{1}{n}\|\beta X - Y\|^2$ est l'erreur d'entraînement et $E_t = \frac{1}{n}\|\beta X_t - Y\|^2$ avec X_t une copie indépendante de X est l'erreur test; $\bar{E}_e$ et $\bar{E}_t$ sont nos estimations théoriques.

4 Conclusion

Nous avons montré, par le biais d'un cadre mathématique rigoureux, que le PCM crée un cadre théorique puissant pour l'analyse des performances de méthodes avancées en apprentissage et traitement des données. Les atouts de cet outil sont (i) son aptitude à modéliser finement des données très réalistes et (ii) son aptitude à gérer des algorithmes et méthodes impliquant des opérations non-linéaires potentiellement complexes. Mais le PCM est loin d'avoir révélé toutes ses capacités: nous pensons que c'est le cadre adapté à l'étude d'objets plus complexes, tels que les solutions implicites à des problèmes d'optimisation complexes (rétropropagation du gradient, optimisation convexe parcimonieuse). Le PCM deviendrait ainsi un axe central d'une analyse moderne de l'apprentissage statistique.

References

[1] M. E. A. Seddik, M. Tamaazousti, and R. Couillet, "Kernel random matrices of large concentrated data : the example of gan-generated images," *ICASSP'19*, 2019.

[2] C. Louart, Z. Liao, and R. Couillet, "A random matrix approach to neural networks," *Annals of Applied Probability*, 2017.

[3] C. Louart and R. Couillet, "Concentration of measure and large random matrices with an application to sample covariance matrices," *submitted*, 2019.

[4] M. Ledoux, *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs, Number 89, 2001.

[5] N. E. Karoui, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *The annals of Statistics*, 2008.

[6] G.-B. Huang, Q. Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing vol 70*, pp. 489–501, 2006.

[7] Z. Bai and J. W. Silverstein, *Spectral Analysis of large dimensional Random Matrices*. Springer, 2009.