Une analyse par matrices aléatoires de l'apprentissage en ligne : traiter des grandes données avec des ressources mémoire limitées

Hugo LEBEAU[1], Romain COUILLET[1], Florent CHATELAIN[2]

[1]Université Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France

[2]Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

hugo.lebeau@univ-grenoble-alpes.fr, romain.couillet@univ-grenoble-alpes.fr, florent.chatelain@grenoble-inp.fr

Résumé – Cet article présente une analyse par matrices aléatoires de l'apprentissage en ligne. En supposant que, du fait de limitations mémoire, l'on a accès qu'à un petit nombre de données, la matrice de Gram ne peut être calculée que partiellement. Dans un contexte où les données sont de grande dimension, on étudie sa distribution spectrale limite et ses valeurs et vecteurs propres isolés. Puis, on précise comment ces résultats nous permettent de réaliser un clustering spectral en ligne avec des garanties de performance théoriques. Des simulations confirment nos résultats.

Abstract – This article introduces a random matrix framework for the analysis of online learning. Assuming that, due to memory limitations, only a few data points are available, the Gram matrix can only be partially computed. Under a large dimensional data regime, we study its limiting spectral distribution and its isolated eigenvalues and eigenvectors. We detail how these results can be used to perform efficient online kernel spectral clustering with theoretical performance guarantees. Our findings are empirically confirmed on image classification tasks.

1 Introduction

Alors que la quantité phénoménale de données générées sans relâche ne cesse de croître, le désastre écologique nous pousse à considérer de nouvelles manières, plus sobres, d'utiliser nos ressources de calcul. L'apprentissage en ligne peut ainsi sembler être une approche pertinente pour traiter des données de grande dimension avec des moyens mémoire limités ; que ce soit à cause d'un volume de données trop grand, ou parce que l'on est restreint à l'utilisation d'un ordinateur conventionnel.

De nombreux travaux ont déjà proposé des approches algorithmiques de clustering sur des flux de données [3]. La plupart exploitent les données telles quelles et voient donc leurs performances se dégrader à mesure que la dimension augmente. Il est néanmoins possible de contrer ce fléau de la dimension en projetant les données sur un espace de dimension inférieure.

Au vu des performances séduisantes du *clustering spectral* sur des données de grande dimension, des adaptations en ligne ont été proposées (notamment [6]). Il est en effet bien moins énergivore que la plupart des algorithmes de clustering classiques puisqu'il permet de drastiquement réduire la dimension des données en ne conservant que quelques composantes spectrales dominantes. De plus, cet algorithme jouit d'intéressant résultats d'optimalité. [1]

Récemment, l'analyse par matrices aléatoires de la matrice de Gram $\mathbf{K} = \frac{1}{p}\mathbf{X}^\top\mathbf{X}$ calculée à partir de données $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix} \in \mathbb{R}^{p\times n}$ a permis de montrer que d'énormes gains en coût calculatoire pouvaient être obtenus au prix d'une perte de performance négligeable [1, 5]. À la lumière de ces résultats, de la puissance du clustering spectral sur des données de grande dimension et de la praticité de l'apprentissage en ligne pour traiter de grands flux de données avec peu de ressources mémoire, cet article présente un algorithme d'apprentissage spectral en ligne auquel s'ajoute une analyse de performances rigoureuse grâce à la théorie des matrices aléatoires. Nos résultats sont validés sur des tâches de classification d'images.

2 Modèle d'apprentissage en ligne

2.1 Position du problème

Soit $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix} \in \mathbb{R}^{p\times n}$ un échantillon de n données de dimension p réparties en deux classes $\mathcal{C}^\pm$ de barycentres respectifs $\pm\boldsymbol{\mu} \in \mathbb{R}^p$. [2] On définit aussi le vecteur $\mathbf{j} \in \{\pm 1\}^n$ tel que $\mathbf{j}_i = +1$ si $\mathbf{x}_i \in \mathcal{C}^+$ et $\mathbf{j}_i = -1$ si $\mathbf{x}_i \in \mathcal{C}^-$.

Hypothèse 1. *Les éléments de* $\mathbf{j}$ *sont des réalisations indépendantes d'une loi de mesure* $\pi^-\delta_{-1} + \pi^+\delta_{+1}$ *où* $\pi^-, \pi^+ > 0$ *et* $\pi^- + \pi^+ = 1$. *À savoir : la classe* $\mathcal{C}^\pm$ *est représentée en proportion* $\pi^\pm$ *et la classe de* $\mathbf{x}_i$ *ne dépend pas de celle de* $\{\mathbf{x}_j\}_{j\neq i}$.

1. Il atteint le seuil de transition de phase optimal (c'est-à-dire qu'il fait mieux qu'un choix aléatoire dès que cela est théoriquement possible) et réalise l'erreur de classification optimale sur des modèles de mélange de gaussiennes.

2. On se restreint ici au cas simplifié de données binaires centrées empiriquement. Le cas général à K classes peut être traité par une étude plus approfondie.

Hypothèse 2 (Condition de non-trivialité)**.**

$$0 < \liminf_{p\to+\infty} \|\boldsymbol{\mu}\| < \limsup_{p\to+\infty} \|\boldsymbol{\mu}\| < +\infty.$$

C'est-à-dire que le problème ne devient ni trop facile ($\|\boldsymbol{\mu}\| \to +\infty$) ni trop difficile ($\|\boldsymbol{\mu}\| \to 0$).

Hypothèse 3 (Modèle de bruit additif)**.** $\mathbf{X} = \mathbf{P} + \mathbf{Z}$ *où* $\mathbf{P} = \boldsymbol{\mu}\mathbf{j}^\top$ *est une matrice déterministe (signal) et* $\mathbf{Z}$ *est une matrice aléatoire (bruit) à entrées indépendantes de loi normale centrée réduite* $\mathcal{N}(0,1)$.

Dans un contexte en ligne, seul un petit nombre L de données peut être conservé en mémoire. Le calcul de la matrice de Gram $\mathbf{K} = \frac{1}{p}\mathbf{X}\mathbf{X}^\top$ est alors limité à une bande de rayon L autour de sa diagonale : l'élément $\mathbf{K}_{i,j} = \frac{1}{p}\mathbf{x}_i\mathbf{x}_j^\top$ ne peut être calculé que si $|i-j| < L$. On modélise cela par l'application d'un masque $\mathbf{T} = \left[\mathbf{1}_{|i-j|<L}\right]_{1\leqslant i,j\leqslant n}$:

$$\mathbf{K}_L = \frac{\mathbf{X}^\top\mathbf{X}}{p} \odot \mathbf{T} \quad \text{avec} \quad \mathbf{T} = \begin{bmatrix} 1 & \cdots & 1 & & 0 \\ \vdots & \ddots & & \ddots & \\ 1 & & \ddots & & 1 \\ & \ddots & & \ddots & \vdots \\ 0 & & 1 & \cdots & 1 \end{bmatrix}$$

où $\odot$ représente le produit de Hadamard.

De par l'optimalité du clustering spectral classique, on s'attend à obtenir de bonnes performances avec la matrice $\mathbf{K}_L$ tant que L et n restent du même ordre. Notre objectif est donc de décrire le comportement spectral de $\mathbf{K}_L$ lorsque n, p et L sont grands. Pour cela, on considère le régime où $n, p, L \to +\infty$ avec $p/n \to c \in]0, +\infty[$ et $(2L-1)/n \to \varepsilon \in]0, +\infty[$.

2.2 Approximation circulante

Malgré son apparente simplicité, la matrice de Toeplitz $\mathbf{T}$ s'avérera contraignante dans nos calculs, à cause de l'apparition "d'effets de bord". On lui préférera son approximation circulante $\mathbf{C} = \left[\mathbf{1}_{|i-j|<L} + \mathbf{1}_{|i-j|>n-L}\right]_{1\leqslant i,j\leqslant n}$ qui fait disparaître ces effets indésirables tout en conservant un spectre "proche" de celui de la matrice originale [4].

On note $\{\tau_k\}_{0\leqslant k<n}$ et $\{\psi_k\}_{0\leqslant k<n}$ les valeurs propres de $\mathbf{T}$ et $\mathbf{C}$ respectivement. La décomposition spectrale de l'approximation circulante est bien connue :

$$\psi_k = \nu_L\left(\frac{2k\pi}{n}\right) \quad \text{avec} \quad \nu_L(x) = \frac{\sin((2L-1)\frac{x}{2})}{\sin(\frac{x}{2})}$$

et $\mathbf{C} = \mathbf{F}\boldsymbol{\Psi}\mathbf{F}^*$ où $\mathbf{F}$ est la matrice de Fourier d'ordre n ($\mathbf{F}_{i,j} = \frac{1}{\sqrt{n}}e^{-\frac{2i\pi}{n}(i-1)(j-1)}$) et où $\boldsymbol{\Psi} = \operatorname{diag}(\psi_k)_{0\leqslant k<n}$.

Les résultats des théorèmes 1 et 2 exposés ci-après peuvent être facilement adaptés[3] au cadre qui nous intéresse en remplaçant les ψ_k par τ_k et $\mathbf{F}$ par une base orthonormale $\mathbf{G}$ de vecteurs propres de $\mathbf{T}$ dans les formules présentées.

3. Sans garantie théorique, mais validé par la pratique.

3 Résultats

Le comportement spectral de $\mathbf{K}_L$ en grande dimension peut être analysé via sa résolvante, objet d'étude classique en théorie des matrices aléatoires [2] :

$$\mathbf{Q}(z) = (\mathbf{K}_L - z\mathbf{I}_n)^{-1} \qquad z \in \mathbb{C} \setminus \operatorname{sp}(\mathbf{K}_L)$$

où $\operatorname{sp}(\mathbf{K}_L)$ est l'ensemble des valeurs propres de $\mathbf{K}_L$. En particulier, la transformée de Stieltjes de la distribution spectrale empirique $\mu_n = \frac{1}{n}\sum_{\xi\in\operatorname{sp}(\mathbf{K}_L)}\delta_\xi$ de $\mathbf{K}_L$ est la trace normalisée de sa résolvante : $m_n(z) \equiv \int_{\mathbb{R}} \frac{\mu_n(\mathrm{d}t)}{t-z} = \frac{1}{n}\operatorname{tr}\mathbf{Q}(z)$.

3.1 Distribution spectrale limite

Notre premier théorème présente un équivalent déterministe de la résolvante lorsque le masque $\mathbf{T}$ est approché par $\mathbf{C}$: $\tilde{\mathbf{Q}}(z) = (\tilde{\mathbf{K}}_L - z\mathbf{I}_n)^{-1}$ avec $\tilde{\mathbf{K}}_L = \frac{\mathbf{X}^\top\mathbf{X}}{p} \odot \mathbf{C}$. Précisément, on définit une matrice $\bar{\mathbf{Q}}(z)$ telle que, pour toutes suites de matrices $\mathbf{A}_n \in \mathbb{R}^{n\times n}$ et vecteurs $\mathbf{a}_n, \mathbf{b}_n \in \mathbb{R}^n$ de norme (spectrale et euclidienne respectivement) unitaire, $\frac{1}{n}\operatorname{tr}\mathbf{A}_n(\tilde{\mathbf{Q}}(z) - \bar{\mathbf{Q}}(z)) \to 0$ et $\mathbf{a}_n^\top(\tilde{\mathbf{Q}}(z) - \bar{\mathbf{Q}}(z))\mathbf{b}_n \to 0$ presque sûrement lorsque $n, p, L \to +\infty$. On note cela $\tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\mathbf{Q}}(z)$.

Théorème 1 (Équivalent déterministe de $\tilde{\mathbf{Q}}(z)$)**.** *Sous les hypothèses 1 – 3,* $\tilde{\mathbf{K}}_L$ *admet une distribution spectrale limite* μ *lorsque* $n, p, L \to +\infty$*. Sa transformée de Stieltjes* m *est solution de*

$$1 + zm(z) = \frac{p}{n}\sum_{k=0}^{n-1}\frac{m(z)\frac{\psi_k}{p}}{1+m(z)\frac{\psi_k}{p}} \qquad z \in \mathbb{C}\setminus\operatorname{supp}\mu.$$

De plus, si $\operatorname{dist}(z, \operatorname{supp}\mu) > \frac{2L-1}{p}$,

$$\tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\mathbf{Q}}(z) \equiv m(z)\left[\mathbf{D}_\mathbf{j}\mathbf{F}\right]\left(\mathbf{I}_n + \|\boldsymbol{\mu}\|^2\boldsymbol{\Lambda}(z)\right)^{-1}\left[\mathbf{D}_\mathbf{j}\mathbf{F}\right]^*$$

où $\boldsymbol{\Lambda}(z) = m(z)\frac{\boldsymbol{\Psi}}{p}\left(\mathbf{I}_n + m(z)\frac{\boldsymbol{\Psi}}{p}\right)^{-1}$ *et* $\mathbf{D}_\mathbf{j} = \operatorname{diag}\mathbf{j}$ *est la matrice diagonale induite par le vecteur* $\mathbf{j}$*.*

Une première observation sur ce théorème est que $\bar{\mathbf{Q}}(z)$ est l'inverse d'une perturbation de l'identité qui n'est pas de petit rang, contrairement à ce que l'on a l'habitude de voir dans les modèles dits *spikes* classiques. Néanmoins la plupart des éléments diagonaux de $\boldsymbol{\Lambda}(z)$ seront généralement assez petits pour que seul un petit nombre de valeurs propres s'isole des autres.

La Figure 1 présente des exemples de distributions spectrales empirique et limite de $\tilde{\mathbf{K}}_L$. Deux types de données sont considérés : bruit seul, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ (ligne du haut) et mélange de deux classes, $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ (ligne du bas). Dans ce dernier cas, on observe effectivement l'apparition de *plusieurs* valeurs propres s'isolant du support limite. Il convient de s'y intéresser plus précisément.

3.2 Comportement des *spikes*

Par une analyse de la matrice $\bar{\mathbf{Q}}(z)$, le théorème suivant indique la position asymptotique des valeurs propres isolées ainsi que l'aspect des vecteurs propres associés.

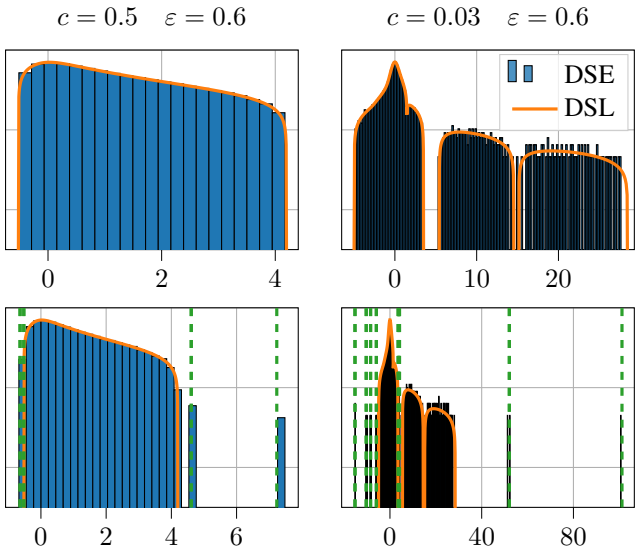

FIGURE 1 – Distribution spectrale empirique (DSE) et distribution spectrale limite (DSL) de $\tilde{\mathbf{K}}_L$. **L'axe des ordonnées est en échelle logarithmique**. **En haut** : bruit seul, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. **En bas** : mélange de deux classes, $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ avec $\|\boldsymbol{\mu}\| = 2$. Les tirets verts indiquent les positions asymptotiques des *spikes* $\bar{\xi}_k$. **Paramètres** : $n = 2\,500$, $L = 750$ et $p = 1\,250$ (**gauche**) ou $p = 75$ (**droite**).

Théorème 2 (Transition de phase, valeurs propres isolées et alignements des vecteurs propres)**.** *Pour chaque entier* $k \in \{0, \dots, n-1\}$*, on définit*

$$\bar{\xi}_k = \left(\|\boldsymbol{\mu}\|^2 + 1\right) \frac{\psi_k}{p} \left(1 + \frac{p}{n} \sum_{l=0}^{n-1} \frac{1}{\left(\|\boldsymbol{\mu}\|^2 + 1\right) \frac{\psi_k}{\psi_l} - 1}\right)$$

et

$$\bar{\zeta}_k = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + 1} \left(1 - \frac{p}{n} \sum_{l=0}^{n-1} \frac{1}{\left[\left(\|\boldsymbol{\mu}\|^2 + 1\right) \frac{\psi_k}{\psi_l} - 1\right]^2}\right).$$

Les trois propositions suivantes sont alors équivalentes.

1. $\psi_k \neq 0$ *et* $\bar{\xi}_k \notin \operatorname{supp} \mu$.
2. $\bar{\zeta}_k > 0$.
3. *Presque sûrement,* $\bar{\xi}_k$ *est la position asymptotique d'une valeur propre isolée de* $\tilde{\mathbf{K}}_L$.

De plus, dans ce cas, la matrice $\mathbf{U}_k = \left[\mathbf{u}_l\right]_{\substack{\psi_k = \psi_l \\ 0 \leqslant l < n}}$ *rassemble tous les vecteurs propres de* $\tilde{\mathbf{K}}_L$ *dont la valeur propre associée converge p.s. vers* $\bar{\xi}_k$ *et*

$$\mathbf{U}_k \mathbf{U}_k^\top \leftrightarrow \bar{\zeta}_k \left[\mathbf{D}_\mathbf{j}\mathbf{F}\right] \mathcal{D}_k \left[\mathbf{D}_\mathbf{j}\mathbf{F}\right]^*$$

où $\mathcal{D}_k = \operatorname{diag}\left(\mathbf{1}_{\psi_k = \psi_l}\right)_{0 \leqslant l < n}$.

Pour bien comprendre ce théorème, il faut voir $\bar{\xi}_k$ comme la position asymptotique *potentielle* d'un *spike* et $\bar{\zeta}_k$ comme une valeur indicatrice : lorsqu'elle est positive, un *spike* existe bien autour de $\bar{\xi}_k$ et alors, plus $\bar{\zeta}_k$ est proche de 1, meilleure est la "qualité" de l'information portée par le vecteur propre associé.

Remarquons que le nombre de valeurs propres isolées peut potentiellement devenir très grand lorsque $\|\boldsymbol{\mu}\|$ croît. Néanmoins, de par la condition de non-trivialité (Hypothèse 2) qui restreint la croissance de $\|\boldsymbol{\mu}\|$, et le fait que seuls quelques ψ_k ont une valeur significative, on n'observe qu'un petit nombre de valeurs propres isolées en pratique. Ces dernières peuvent cependant apparaître des deux côtés du spectre, voire même entre deux composantes continues (cf. Figure 1).

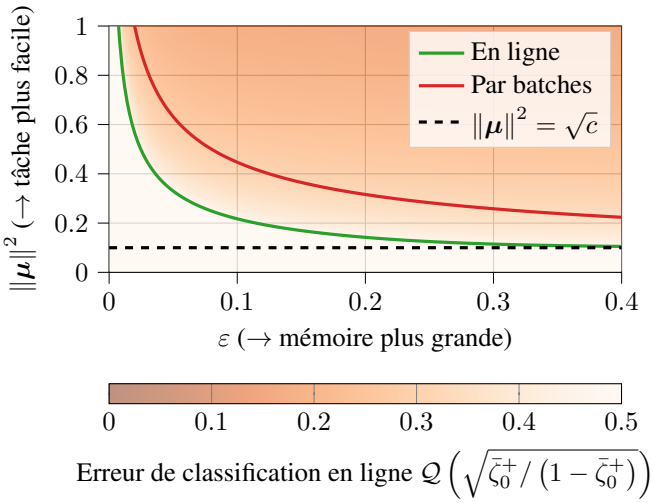

FIGURE 2 – Position de la transition de phase ($\|\boldsymbol{\mu}\|^2$) du vecteur propre dominant de la matrice à noyau en fonction du paramètre $\varepsilon = \frac{2L-1}{n}$ avec $n/p = 100$. La classification selon une méthode donnée n'est possible qu'au-dessus de la courbe correspondante. Les tirets noirs indiquent la transition de phase optimale (lorsque toutes les données sont disponibles). **En vert** : clustering spectral en ligne. **En rouge** : clustering spectral classique avec L données.

4 Clustering spectral en ligne

Les résultats précédents trouvent des applications au clustering en ligne de données en grande dimension.

4.1 Compromis performance-coût

Les résultats du théorème 2 nous permettent de trouver la position de la *transition de phase*, c'est-à-dire la valeur de $\|\boldsymbol{\mu}\|$ en-dessous de laquelle la classification n'est plus réalisable. Elle est donnée par l'équation $\bar{\zeta}_k = 0$.

La Figure 2 présente la position $\|\boldsymbol{\mu}\|^2$ de la transition de phase en fonction de $\varepsilon = \frac{2L-1}{n}$ avec l'erreur de classification (théorique) du clustering spectral en ligne lorsque $n/p = 100$. Celle-ci est comparée avec le clustering *par batches*, c'est-à-dire le clustering spectral classique n'utilisant que les L don-

nées disponibles en mémoire. À mesure que ε croît, la position de la transition de phase atteint rapidement le seuil optimal $\|\boldsymbol{\mu}\|^2 = \sqrt{c}$. De plus, à ε fixé, l'erreur de classification décroît logiquement vers 0 lorsque $\|\boldsymbol{\mu}\|$ croît.

En plus d'avoir de meilleures performances [4] que la classification naïve par batches, notre méthode est aussi capable de classifier les n derniers points à chaque instant, *même si la plupart d'entre eux ne sont plus conservés en mémoire* !

4.2 Principe de l'algorithme

On considère un flux de données de longueur (possiblement infinie) T. À chaque pas de temps, une nouvelle donnée $\mathbf{x}_t$ entre en mémoire tandis que $\mathbf{x}_{t-L}$ est supprimée. La matrice de Gram est ensuite mise à jour :

$$\left[\mathbf{K}_L^{(t)}\right]_{i,j} = \frac{1}{p}\mathbf{x}_{t-n+i}^\top \mathbf{x}_{t-n+j}\mathbf{1}_{|i-j|<L}.$$

L'alignement des vecteurs propres isolés de $\mathbf{K}_L^{(t)}$ avec les vecteurs population (théorème 2) permet alors la classification des n dernières données. [5]

Remarque 1 (Choix de n et L). Tandis que L représente le nombre de données conservées en mémoire, $n \geqslant L$ correspond au nombre de données que l'on classifie à chaque pas de temps. Ces deux paramètres sont fixés par l'utilisateur en prenant en compte les précédentes considérations sur la performance (Figure 2) et les limitations mémoire : l'espace nécessaire est $\mathcal{O}(Lp + Ln)$ pour les données et la matrice de Gram.

4.3 Applications à la classification d'images

Pour illustrer nos résultats, on simule deux flux d'images. Le premier avec $T = 20\,000$ représentations VGG ($p = 4\,096$) d'images de chiens (`collie`) et chats (`tabby`) générées avec un réseau BigGAN. Le second avec $T = 14\,000$ images de Fashion-MNIST ($p = 784$) des classes `coat` et `ankle boot`. Dans les deux cas, on choisit $n = 1000$ et $L = 100$.

La Figure 3 représente l'erreur de classification à $t_0 + \Delta t$ d'une donnée arrivée à t_0, ainsi que l'erreur de classification globale obtenue avec un vote majoritaire (tirets verts) et l'erreur de classification obtenue avec un clustering spectral $T \times T$ *hors ligne* [6] (pointillés noirs). Les performances obtenues avec l'algorithme en ligne sont très proches de celles du clustering spectral complet mais ce premier nécessite beaucoup moins de ressources mémoire ($\mathcal{O}(Lp + Ln)$ contre $\mathcal{O}(Tp + T^2)$).

5 Conclusions et perspectives

Une approche par matrices aléatoires nous a permis de montrer que des performances quasi-optimales pouvaient être obtenues sur des données de grande dimension avec un algorithme de clustering spectral en ligne. La méthode proposée résulte d'une analyse asymptotique précisant les performances atteignables lors de l'apprentissage sur un flux de données. Le cadre en-ligne manque cependant de résultats d'optimalité de la part de la théorie de l'information, qui est une piste à approfondir.

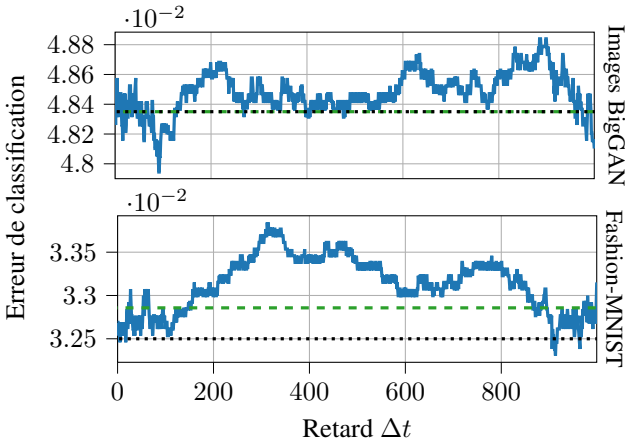

FIGURE 3 – Erreur de classification en fonction du retard Δt sur les images BigGAN (**haut**) et Fashion-MNIST (**bas**). Il s'agit de l'erreur de classification moyenne à $t_0 + \Delta t$ pour une donnée arrivée à t_0. Les tirets verts indiquent l'erreur de classification globale (obtenue par vote majoritaire). Les pointillés noirs représentent l'erreur de classification obtenue par clustering spectral $T \times T$ hors ligne. **Paramètres** : $T = 20\,000$, $n = 1\,000$, $p = 4\,096$, $L = 100$ (images BigGAN) et $T = 14\,000$, $n = 1\,000$, $p = 784$, $L = 100$ (Fashion-MNIST).

Références

[1] R. Couillet, F. Chatelain, and N. Le Bihan. Two-way kernel matrix puncturing : towards resource-efficient PCA and spectral clustering. *arXiv :2102.12293 [cs, stat]*, May 2021. arXiv : 2102.12293.

[2] R. Couillet and Z. Liao. *Random Matrix Methods for Machine Learning : When Theory meets Applications*. 2021.

[3] M. Ghesmoune, M. Lebbah, and H. Azzag. State-of-the-art on clustering data streams. *Big Data Analytics*, 1(1) :13, Dec. 2016.

[4] R. M. Gray. Toeplitz and Circulant Matrices : A Review. *Foundations and Trends® in Communications and Information Theory*, 2(3) :155–239, Jan. 2006. Publisher : Now Publishers, Inc.

[5] Z. Liao, R. Couillet, and M. W. Mahoney. Sparse Quantized Spectral Clustering. *arXiv :2010.01376 [cs, math, stat]*, Oct. 2020. arXiv : 2010.01376.

[6] S. Yoo, H. Huang, and S. P. Kasiviswanathan. Streaming spectral clustering. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 637–648, May 2016.

4. La transition de phase a lieu plus tôt.
5. Dans le cas binaire considéré ici, le vecteur propre dominant suffit.
6. Dont on connaît les résultats d'optimalité.