Projections aléatoires pour l'apprentissage compressif

Antoine Chatalic[1], Nicolas Keriven[2], Rémi Gribonval[1]

[1]Univ Rennes, Inria, CNRS, IRISA, 35000 Rennes

[2]École Normale Supérieure, 45 rue d'Ulm, 75005 Paris

antoine.chatalic@irisa.fr, nicolas.keriven@ens.fr, remi.gribonval@inria.fr

Résumé – L'apprentissage compressif a pour objectif de réduire drastiquement le volume de grandes collections d'entraînement via des sortes de projections aléatoires, tout en préservant l'information nécessaire à l'apprentissage. En s'appuyant sur quelques exemples en apprentissage non-supervisé, nous proposons un tour d'horizon des outils utilisés, des garanties théoriques à la fois en termes de préservation d'information et de respect de la vie privée, pour finir avec quelques problèmes ouverts.

Abstract – Compressive learning is a framework to drastically compress the volume of large training collections while preserving the information needed to learn. Guided by unsupervised learning examples, we survey the involved tools, the existing theoretical guarantees both in terms of information preservation and of privacy preservation, and conclude with some open problems.

1 Introduction

L'abondance de données est bien sûr une bénédiction pour la pertinence statistique en apprentissage automatique. Elle requiert néanmoins des ressources qui en font aussi une malédiction, avec des conséquences écologiques (gourmandise énergétique) mais aussi en termes de déstabilisation de tout un écosystème de recherche et développement, remettant en cause la capacité d'acteurs académiques et de PME à rester au niveau de l'état de l'art.

Peut-on à la fois maîtriser les ressources de calcul et bénéficier de l'abondance de données ? C'est le pari de l'apprentissage compressif, où les données d'entraînement sont compressées drastiquement *avant* apprentissage.

Compression sous forme de croquis. Durant la phase de compression, le principe est de calculer un "croquis" [1] (ou *sketch*) résumant l'ensemble d'une collection d'entraînement sous la forme d'un vecteur de petite dimension construit de manière à préserver l'information nécessaire à la phase d'apprentissage. A partir d'une collection $\mathcal{X}$ constituée de n vecteurs d'entraînement $\mathbf{x}_i \in \mathbb{R}^d$, $1 \leq i \leq n$, le croquis $\hat{\mathbf{z}} \in \mathbb{R}^m$ est construit sous la forme $\hat{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Phi}(\mathbf{x}_i)$ où $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathbb{R}^m$ est une fonction généralement non-linéaire. Le choix de la dimension m (qui gouverne les ressources mémoire nécessaires) et celui de la fonction $\boldsymbol{\Phi}(\mathbf{x}) = (\Phi_j(\mathbf{x}))_{j=1}^m$ résultent d'un compromis entre compression (idéalement, $m \ll nd$) et préservation d'information. Ces choix dépendent de la tâche d'apprentissage, et nous verrons sur des exemples qu'ils peuvent s'appuyer sur des outils issus de l'échantillonnage compressif (*compressive sensing*) [2] et des descripteurs aléatoires (*random features*) [3].

Apprentissage à partir d'un croquis. L'apprentissage classique cherche typiquement des paramètres $\hat{\theta}$ minimisant un risque empirique $\frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}_i, \theta)$ (où $\ell(\mathbf{x}, \theta)$ est une fonction de perte) dont le calcul implique de multiples accès à la collection $\mathcal{X}$. L'apprentissage à partir d'un croquis passe au contraire par la minimisation d'un "succédané" de fonction de risque $f(\theta|\hat{\mathbf{z}})$ qui ne dépend que du seul croquis $\hat{\mathbf{z}}$, sans nouvel accès à la collection $\mathcal{X}$.

Exemples en apprentissage non-supervisé. Nous verrons sur quelques exemples comment la fonction $f(\theta|\hat{\mathbf{z}})$ peut être construite en s'appuyant sur une forme revisitée de la méthode des moments généralisés [4], et comment sa minimisation bénéficie d'algorithmes empiriques inspirés de la reconstruction parcimonieuse pour les problèmes linéaires inverses et de l'échantillonnage compressé [2].

Le croquis, qui dépend non-linéairement des données $\mathbf{x}_i$, est constitué de moments généralisés de la distribution de probabilité P des données. Il dépend donc "linéairement" de P : si $P = \theta P_1 + (1-\theta)P_2$ est un mélange de deux distributions et $\mathbf{z}_\ell = \mathbb{E}_{X\sim P_\ell}\boldsymbol{\Phi}(X)$, alors $\mathbb{E}_{X\sim P}\boldsymbol{\Phi}(X) = \theta\mathbf{z}_1 + (1-\theta)\mathbf{z}_2$. Le croquis $\hat{\mathbf{z}}$ est donc une version bruitée (car calculée sur un tirage empirique de n échantillons) de $\mathbf{z} := \mathcal{A}(P)$, où $\mathcal{A} : P \mapsto \mathcal{A}(P) := \mathbb{E}_{X\sim P}\boldsymbol{\Phi}(X)$ est un opérateur linéaire. En pratique, la fonction $\boldsymbol{\Phi}$ étant choisie de façon aléatoire, le croquis s'apparente à une projection aléatoire de la distribution des données.

Garanties théoriques et défis. Pour certaines tâches, nous décrirons comment le calcul du croquis peut être rendu numériquement efficace [5, 6] et respectueux de la vie privée [7] avec des garanties de préservation de l'information pour l'apprentissage [8]. Nous aborderons enfin les défis posés par le déploiement de l'apprentissage compressif sur une plus large gamme de tâches d'apprentissage.

2 Illustration sur quelques exemples

La littérature sur le *sketching* est vaste avec une branche issue des bases de données (par exemple les tables de hachages) [1] et des méthodes des moments généralisées en statistiques [4]. Le terme *sketch* est également employé pour désigner d'autres procédés que nous n'aborderons pas ici [9].

Analyse en Composante Principale. En Analyse en Composantes Principales (ACP, avec des données centrées pour simplifier), la *matrice de covariance empirique* $\widehat{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$ correspond à un croquis de taille $m = d^2$ calculé via la fonction $\mathbf{\Phi}(\mathbf{x}) = \mathbf{x}\mathbf{x}^\top$. L'ACP consiste à calculer une approximation de rang faible de $\widehat{\mathbf{M}}$:

$$\min_{\mathbf{M},\,\mathrm{rang}(\mathbf{M})\leq k} \|\widehat{\mathbf{M}} - \mathbf{M}\|_F\,,$$

ce qui peut se faire via la décomposition en valeurs singulières de $\widehat{\mathbf{M}}$. Il est possible de montrer que la dimension du croquis peut être réduite davantage à $m = \mathcal{O}(kd)$ [8] en s'appuyant sur de la compression de matrices de rang faible inspirée de l'échantillonnage compressé [2].

Estimation de modèles de mélange. L'estimation de densité consiste à approcher la distribution des données par une distribution appartenant à une famille paramétrique, par exemple un modèle de mélange : étant donné une famille de distributions $\mathcal{P}_\Theta = \{P_\theta, \theta \in \Theta\}$, on considère $\mathcal{P}_k = \{\sum_{\ell=1}^{k} a_\ell P_{\theta_\ell}, a \geq 0, \sum_\ell a_\ell = 1\}$. L'estimation de densité se fait alors classiquement en maximisant la log-vraisemblance des données $\max_{P\in\mathcal{P}_k} \sum_{i=1}^{n} \log P(x_i)$, via l'algorithme Espérance-Maximisation (EM).

Lorsque $\mathcal{A}(\pi_\theta) = \mathbb{E}_{X\sim\pi_\theta}\mathbf{\Phi}(X)$ a une expression explicite, l'estimation compressée d'un mélange peut se faire en cherchant la distribution $P \in \mathcal{P}_k$ dont le croquis est le plus proche du croquis empirique :

$$\min_{a,\theta} \left\| \sum_{\ell=1}^{k} a_\ell \mathcal{A}(P_{\theta_\ell}) - \hat{\mathbf{z}} \right\|_2 . \tag{1}$$

Le choix de la fonction $\mathbf{\Phi}$ est discuté à la section suivante.

Par exemple, pour une tâche de vérification de locuteurs [10], une base de donnée contenant 1000 heures de parole (50 gigaoctets en mémoire) a pu être compressée en un croquis de seulement quelques *kilo-octets* [11] suffisant pour que l'estimation compressée de mélanges de Gaussiennes (GMM) atteigne les performances de vérification de l'algorithme EM, ce dernier accédant itérativement à la base d'entraînement.

L'approche par croquis est également parfois plus flexible que l'approche par maximum de vraisemblance : en effet pour certaines familles de distributions $\mathcal{P}_\Theta$ il n'existe pas d'expression explicite pour la vraisemblance, mais il est tout de même possible de calculer $\mathcal{A}(P_\theta)$. C'est le cas des distributions α-stables (où le paramètre α varie d'une composante du mélange à l'autre), pour certains choix de la fonction $\mathbf{\Phi}$. L'estimation par croquis de ce type de mélanges a par exemple été exploitée pour la séparation de sources audio [12].

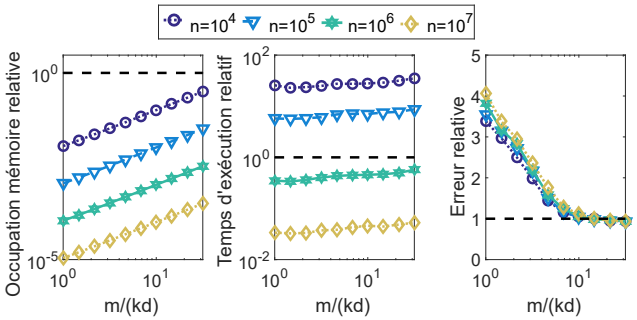

Fig. 1 : Occupation mémoire, temps d'exécution et erreur pour le partitionnement compressif [14] relativement aux k-moyennes classiques, en fonction de la taille du croquis m.

Partitionnement. Le partitionnement non-supervisé de données consiste à grouper entre elles des données similaires ; cela peut se faire par le calcul de centroïdes $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \mathbb{R}^d$, chaque point étant ensuite assigné au groupe correspondant au centroïde le plus proche. Ce calcul est traditionnellement effectué par l'algorithme de Lloyd [13], aussi appelé k-moyennes, qui minimise un certain coût de partitionnement. La complexité de cet algorithme est de $\mathcal{O}(nkdT)$ pour T itérations.

Un algorithme a été proposé [14] pour estimer des centroïdes $\mathbf{c} = \mathbf{c}_1, ..., \mathbf{c}_k$ à partir d'un croquis $\hat{\mathbf{z}}$ (calculé en utilisant une fonction $\mathbf{\Phi}$) via le problème d'optimisation suivant :

$$\min_{\mathbf{c},\, a\geq 0,\, \sum_\ell a_\ell = 1} \left\| \sum_{\ell=1}^{k} a_\ell \mathbf{\Phi}(\mathbf{c}_\ell) - \hat{\mathbf{z}} \right\|_2 . \tag{2}$$

Autrement dit, on cherche les centroïdes (pondérés) dont le croquis est le plus proche de celui des données.

La figure 1 démontre l'utilité de l'approche par croquis : sur de grandes bases de données, elle permet de gagner plusieurs ordres de grandeur en mémoire et temps de calcul, et atteint un coût de partitionnement similaire à celui de l'algorithme des k-moyennes pour une taille de croquis m suffisamment grande (de l'ordre de $m = \mathcal{O}(kd)$).

3 Mise en œuvre pratique

Nous explicitons désormais la fonction $\mathbf{\Phi}$, et discutons des algorithmes permettant de résoudre le problème inverse ainsi que de la complexité du mécanisme.

Choix de la fonction $\mathbf{\Phi}$ de calcul de croquis. Pour le partitionnement compressif et les modèles de mélange compressifs, la fonction $\mathbf{\Phi}$ est choisie en pratique comme une collection d'exponentielles complexes $\mathbf{\Phi}(\mathbf{x}) = [e^{-i\boldsymbol{\omega}_j^\top \mathbf{x}}]_{j=1}^{m}$ [3]. Pour ce choix, $\mathcal{A}(P_\theta)$ est un échantillonnage de la fonction caractéristique de la distribution P_θ, qui a très souvent une expression explicite, notamment dans le cas des distributions α-stables. Par analogie avec l'échantillonnage compressé, les vecteurs de fréquence $\boldsymbol{\omega}_j \in \mathbb{R}^d$ sont tirés aléatoirement selon une distribution judicieusement choisie [11].

Algorithmes pour le problème inverse. En pratique, les problèmes de minimisation non-convexe (1) et (2) peuvent être

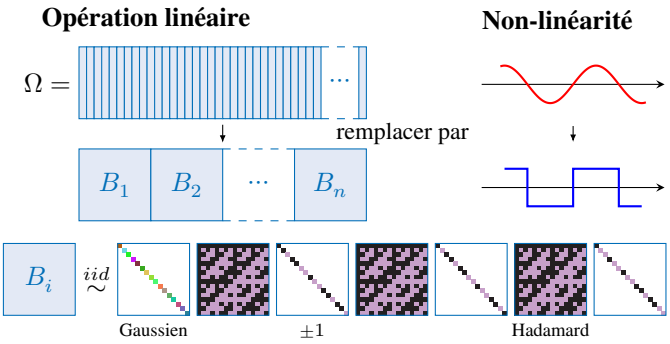

Fig. 2 : Matrices structurées et quantification.

abordés via l'algorithme CL-OMPR [11], inspiré d'*Orthogonal Matching Pursuit with Replacement* (OMPR) [15]. Malgré les excellents résultats empiriques obtenus, l'analyse théorique reste un problème ouvert, même si certains cas particuliers commencent à être abordés [16, 17]. Il est également intéressant de noter que cet algorithme a de fortes similarités avec l'algorithme de Frank-Wolfe (ou gradient conjugué) [18, 19] appliqué à une relaxation convexe du problème (2), plus spécifique au domaine de la super-résolution [20]. Une approche alternative basée sur un algorithme de propagation de croyances [21] formule le problème dans un cadre Bayésien comme l'inférence de la moyenne a posteriori des centroïdes, conditionnellement à l'observation du croquis empirique $\hat{\mathbf{z}}$. Bien que moins générique que CL-OMPR, cette méthode fonctionne avec des croquis plus petits (e.g. $m = 2kd$ plutôt que $m = 10kd$ pour une même erreur).

Réduction de la complexité La fonction $\mathbf{\Phi}$ est typiquement la succession d'une opération linéaire faisant intervenir la matrice des fréquences $\Omega = [\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m]$ et d'une opération non linéaire. Bien qu'il ne requière qu'une seule passe sur les données et soit facilement parallélisable, le calcul du croquis a un coût algorithmique directement lié à l'application de $\mathbf{\Phi}$. La résolution du problème inverse a une complexité limitée de manière inhérente à la taille du sketch, mais qui dépend également fortement du coût de $\mathbf{\Phi}$ étant donnée la nature itérative des algorithmes utilisés en pratique (cf. section 2).

La matrice des fréquences Ω, qui est de taille $d \times m$ et aléatoire, peut être substituée par une alternative [5] impliquant moins de valeurs aléatoires et des matrices structurées orthogonales [22] (cf. bas de Fig. 2) associées à des transformées rapides. La multiplication matrice-vecteur passe ainsi de $\mathcal{O}(md)$ à $\mathcal{O}(m\log(d))$ et le stockage de Ω est réduit de $\mathcal{O}(md)$ à $\mathcal{O}(m)$.

L'utilisation d'unités de calcul optique [23] permettrait de réduire de manière encore plus draconienne le coût de cette opération linéaire.

Dans le cas du calcul de moments de Fourier, l'opération non-linéaire est une exponentielle complexe, qui peut être approchée par un signal triangulaire ou carré sous certaines conditions [6]. Dans ce dernier cas, on obtient un sketch quantifié sur $\{-1, 1\}^{2m}$, plus léger et rapide à calculer. Outre cette quantification binaire, la fonction $\mathbf{\Phi}$ peut être sous-échantillonnée (cf. section 5), ce qui résulte dans le cas extrême en la mesure d'un unique bit d'information par échantillon x_i.

4 Garanties statistiques

L'apprentissage statistique vise à minimiser le risque $\mathcal{R}(P, \theta) = \mathbb{E}_{X\sim P}\ell(X, \theta)$ pour une certaine fonction de perte $\ell(\mathbf{x}, \theta)$ et famille d'hypothèses Θ. En notant $P^\star$ la vraie distribution des données (dans le cas l'apprentissage *supervisé*, il s'agit de la distribution jointe des données et des labels), l'hypothèse idéale est

$$\theta^\star = \arg\min_{\theta\in\Theta} \mathcal{R}(P^\star, \theta).$$

Le risque moyen est classiquement remplacé par le risque empirique, et l'on contrôle l'*excès de risque* qui en résulte.

En apprentissage compressif, notre but est de construire un succédané de risque $f(\theta|\hat{\mathbf{z}})$ de sorte que l'excès de risque associé à $\tilde{\theta} := \arg\min f(\theta|\hat{\mathbf{z}})$ soit contrôlé. Conceptuellement, l'approche retenue est en deux temps : on identifie un "décodeur" Δ produisant une distribution $\tilde{P} = \Delta(\hat{\mathbf{z}})$, tel que la quantité suivante soit contrôlée :

$$\|\tilde{P} - P^\star\| := \sup_\theta |\mathcal{R}(\tilde{P}, \theta) - \mathcal{R}(P^\star, \theta)|$$

et tel qu'il est possible de retrouver dans un second temps l'hypothèse $\tilde{\theta} = \min_\theta \mathbb{E}_{X\sim\tilde{P}}\ell(X, \theta)$ (en pratique, l'estimation de $\tilde{\theta}$ se confondra avec celle de $\tilde{P}$, et l'approche ne passe pas par le problème plus difficile de l'estimation de densité). Dans [8], l'acquisition comprimée sert d'inspiration pour montrer que ce contrôle peut se faire *via* une certaine *Propriété d'Isométrie Restreinte* (RIP).

Théorème 1 ([8]). *Soit Σ un ensemble de distributions de probabilités et $\mathbf{\Phi}$ une fonction de croquis bornée telle que la RIP suivante soit satisfaite :*

$$\forall P, P' \in \Sigma, \|P - P'\| \lesssim \|\mathcal{A}(P) - \mathcal{A}(P')\|_2. \tag{3}$$

Alors, en définissant le décodeur

$$\Delta(\mathbf{z}) = \arg\min_{P\in\Sigma} \|\mathcal{A}(P) - \mathbf{z}\|_2 \tag{4}$$

l'excès de risque de l'apprentissage compressif satisfait

$$\mathcal{R}(P^\star, \tilde{\theta}) - \mathcal{R}(P^\star, \theta^\star) \lesssim d(P^\star, \Sigma) + 1/\sqrt{n} \tag{5}$$

où $d(P^\star, \Sigma) = \min_{P\in\Sigma} d(P^\star, P)$ pour une certaine métrique d. Voir [8] pour détails et constantes multiplicatives.

Tout l'enjeu est de choisir un ensemble Σ qui soit suffisamment grand pour que $d(P^\star, \Sigma)$ soit faible, et un sketch de taille m suffisante pour que la RIP (3) soit satisfaite (avec forte probabilité sur le tirage des fréquences ω_j).

Dans le cadre de l'ACP le théorème ci-dessus mène à des garanties plus explicites lorsque Σ est l'ensemble des distributions de probabilité dont la matrice de covariance est de rang au plus k ; pour le partitionnement (respectivement les modèles de mélange) on considère Σ l'ensemble des mélanges de k Diracs (resp. l'ensemble $\mathcal{P}_k$ des mélanges de k lois, par exemple α-stables). La condition RIP (3) est prouvée pour certains de ces modèles [8] lorsque l'on impose une séparation minimale ε entre les centroïdes (resp. composantes du mélange) et que la fonction $\mathbf{\Phi}$ est une collection d'exponentielles complexes choisies aléatoirement avec une taille de sketch de l'ordre de $m = \mathcal{O}(k^2 d\log(1/\varepsilon))$ aux facteurs logarithmiques près.

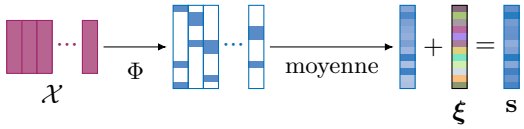

Fig. 3 : Calcul du croquis avec sous-échantillonnage et bruit additif.

5 Croquis pour l'anonymisation

De nombreux jeux de données sont aujourd'hui construits en agrégeant de l'information provenant d'individus différents, chacun fournissant un ou quelques échantillons de la collection totale. La perte d'information causée par le calcul de croquis permet l'apprentissage de comportements moyens tout en garantissant la confidentialité des individus. Là où le contrôle du risque permet de trouver une borne inférieure sur la taille m du croquis – c.-à-d. sur la quantité d'information – nécessaire pour la résolution du problème inverse, la propriété de confidentialité différentielle [24] assure que le résultat d'un algorithme (aléatoire) ne dépend pas de manière trop importante de la présence d'un individu dans le jeu de données. L'ajout de bruit Laplacien sur le croquis permet de garantir cette propriété [7, 25]. Dans ce scénario, les différents propriétaires de données calculent et publient chacun un croquis bruité et sous-échantillonné. L'étude du compromis existant entre niveau de confidentialité et utilité en terme d'apprentissage ultérieur (dans le cas du partitionnement) montre que le calcul des moments peut être effectué en sous-échantillonnant fortement le calcul des $\boldsymbol{\Phi}(x_i)$ (cf. Fig. 3), ce qui en réduit le coût. Dans le cas extrême, une seule mesure $\Phi_j(x_i)$ est calculée par échantillon x_i.

6 Perspectives

Les résultats obtenus pour les modèles de mélanges, le partitionnement et l'ACP invitent naturellement à étendre le mécanisme d'apprentissage compressif à de nouveaux problèmes tels que l'apprentissage de dictionnaire, la factorisation en matrices non-négatives (NMF) ou l'analyse en composantes indépendances (ICA), ainsi qu'à des tâches supervisées comme la régression ou la classification. Le cadre théorique existant devrait aider à caractériser l'adéquation d'une fonction de compression $\boldsymbol{\Phi}$ avec une tâche d'apprentissage, caractérisée par sa fonction de risque.

Les garanties concernant les algorithmes utilisés pour la résolution du problème inverse (cf. section 2) ainsi que les différentes stratégies d'accélération (cf. section 3) restent à ce jour parcellaires et doivent être consolidées.

Références

[1] G. Cormode et al. « Synopses for Massive Data ». In : *FTDB* 4.1 (2011).

[2] S. Foucart et H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, mai 2012.

[3] A. Rahimi et B. Recht. « Random Features for Large Scale Kernel Machines ». In : *NIPS* (2007).

[4] A. R. Hall. *Generalized method of moments*. Oxford university press, 2005.

[5] A. Chatalic et al. « Large-Scale High-Dimensional Clustering with Fast Sketching ». In : *ICASSP*. 2018.

[6] V. Schellekens et L. Jacques. « Quantized Compressive K-Means ». In : *IEEE SPL* 25.8 (août 2018).

[7] V. Schellekens et al. « Differentially Private Compressive k-means ». In : *ICASSP*. 2019.

[8] R. Gribonval et al. « Compressive statistical learning with random feature moments ». In : *preprint arXiv :1706.07180* (2017).

[9] Y. Yang et al. « Randomized sketches for kernels ». In : *Annals of Statistics* 45.3 (2017).

[10] D. A. Reynolds et al. « Speaker Verification Using Adapted Gaussian Mixture Models ». In : *Digital Signal Processing* 10.1-3 (2000).

[11] N. Keriven et al. « Sketching for large-scale learning of mixture models ». In : *Information and Inference* 7.3 (2017).

[12] N. Keriven et al. « Blind Source Separation Using Mixtures of Alpha-Stable Distributions ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2018.

[13] S. P. Lloyd. « Least Squares Quantization in PCM ». In : *IEEE Transactions on Information Theory* 28.2 (1982).

[14] N. Keriven et al. « Compressive K-means ». In : ICASSP. 2017.

[15] P. Jain et al. « Orthogonal matching pursuit with replacement ». In : *NIPS*. 2011.

[16] C. Elvira et al. « A case of exact recovery using OMP with continuous dictionaries ». In : *Curves and Surfaces*. 2018.

[17] Y. Traonmilin et J.-F. Aujol. « The basins of attraction of the global minimizers of the non-convex sparse spikes estimation problem ». In : *Preprint hal-01938239* (2018).

[18] N. Boyd et al. « The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems ». In : (2015).

[19] Q. Denoyelle et al. « The Sliding Frank-Wolfe Algorithm and its Application to Super-Resolution Microscopy ». In : *Arxiv preprint arXiv :1811.06416* (2018).

[20] E. J. Candès et C. Fernandez-Granda. « Towards a mathematical theory of super-resolution ». In : *Communications on Pure and Applied Mathematics* 67.6 (2014).

[21] E. Byrne et al. *Sketched Clustering via Hybrid Approximate Message Passing*. 4 jan. 2019.

[22] F. X. Yu et al. « Orthogonal Random Features ». In : *NIPS*. Sous la dir. de D. D. Lee et al. Curran Associates, Inc., 2016.

[23] A. Saade et al. « Random projections through multiple optical scattering ». In : *ICASSP*. IEEE, 2016.

[24] C. Dwork et al. « Calibrating noise to sensitivity in private data analysis ». In : *Theory of cryptography conference*. 2006.

[25] M. Balog et al. « Differentially Private Database Release via Kernel Mean Embeddings ». In : *ICML*. 3 juil. 2018.