Approximation de Nyström gloutonne par minimisation de la trace

Antoine CHATALIC[1]

[1]CNRS, Univ. Grenoble-Alpes, GIPSA-lab, France

Résumé – Nous proposons un algorithme pour la construction d'approximations de Nyström de noyaux définis positifs, basé sur la minimisation de la trace résiduelle de la matrice à noyau. Cette erreur résiduelle est approchée au moyen de descripteurs de Fourier aléatoires, et minimisée de manière gloutonne. Nous montrons théoriquement et empiriquement que ce choix est particulièrement pertinent lorsque l'approximation induite est utilisée pour résoudre le problème des k-moyennes à noyau.

Abstract – We introduce a greedy algorithm for the construction of kernel Nyström approximations. The targeted objective is the residual trace of the data kernel matrix, which is approximated using random Fourier features and sketching. We motivate theoretically the choice of such a criterion when the induced features are used for kernel k-means clustering, and show empirically that the proposed algorithm indeed outperforms other landmark selection methods classically used in the literature on this task.

1 Introduction

Le formalisme des espaces de Hilbert à noyau reproduisant (RKHS) constitue un cadre puissant pour la définition de modèles non paramétriques, et a été utilisé pour de nombreuses applications en traitement du signal et en apprentissage. La complexité en temps et en espace de ces modèles limite toutefois leur utilisation à grande échelle, et le développement d'approximations efficaces demeure pour cela primordial.

Une stratégie répandue dans ce contexte consiste à approcher les éléments du RKHS en les projetant sur un sous-espace aléatoire de faible dimension. Cette stratégie a été déployée pour diverses applications, et de multiples algorithmes ont été proposés afin de construire le sous-espace en question. Selon l'application considérée, différents critères peuvent être utilisés pour quantifier l'erreur induite par cette approximation, et en conséquent certaines méthodes d'approximation peuvent être plus pertinentes que d'autres.

Cet article propose un algorithme approché pour la construction de sous-espaces d'approximation basé sur la minimisation gloutonne de la trace résiduelle de la matrice à noyau. Nous motivons ce critère en montrant qu'il permet notamment de borner l'excès de risque, pour le problème des k-moyennes à noyau, dû à l'approximation dans un sous-espace. Nous montrons expérimentalement que notre algorithme donne de meilleurs résultats pour ce problème que les autres méthodes de construction de l'état de l'art.

2 Contexte et motivation

Soit $(\mathcal{X}, \Sigma, \rho)$ un espace probabilisé et $\mathcal{H}$ un RKHS séparable de fonctions de $\mathcal{X}$ dans $\mathbb{R}$, de noyau reproduisant κ supposé borné (voir p.ex. [12] pour une introduction aux espaces de Hilbert à noyau reproduisant). Nous noterons par la suite $\phi(x) := \kappa(x, \cdot) \in \mathcal{H}$, supposée mesurable pour tout $x \in \mathcal{X}$, et κ s'écrit $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$ pour tous $x, y \in \mathcal{X}$. La distribution ρ des données est supposée connue seulement via n échantillons $x_i \overset{i.i.d.}{\sim} \rho$, et l'on note $\rho_n := \frac{1}{n} \sum_{i=1}^{n} \delta(x_i)$ la distribution empirique associée.

Soit $\Phi_n : \mathbb{R}^n \to \mathcal{H}$, $\Phi_n w := \sum_{1 \leq i \leq n} w_i \phi(x_i)$, dont l'adjoint $\Phi_n^* : \mathcal{H} \to \mathbb{R}^n$ est l'opérateur d'échantillonnage $\Phi_n^* h = [h(x_i)]_{i=1}^n$. Avec ces notations, la matrice à noyau K de nos n échantillons, c.-à-d. telle que $K_{ij} := \kappa(x_i, x_j)$, s'écrit $K = \Phi_n^* \Phi_n \in \mathbb{R}^{n \times n}$. Lorsque le nombre n d'échantillons est élevé, la plupart des modèles basés sur des RKHS pâtissent de coûts de calcul élevés lié à la manipulation de cette matrice à noyau, dont le coût de calcul est quadratique en le nombre n d'échantillons.

Approximation de Nyström Une technique d'approximation couramment utilisée lorsqu'il est impossible de manipuler la matrice K exacte consiste à travailler dans un sous-espace de dimension finie du RKHS. Pour tout $\mathcal{S} \subseteq [1, \dots, n]$, nous notons $P_{\mathcal{S}}$ le projecteur orthogonal sur le sous-espace $\mathcal{H}_{\mathcal{S}} := \operatorname{span}(\{ \phi(x_i) \mid i \in \mathcal{S} \})$. Pour tout projecteur P, on note par la suite $P^{\perp} := I - P$. L'approximation dite de Nyström [9, 13] de la matrice K est la matrice $\tilde{K} := \Phi_n^* P_{\mathcal{S}} \Phi_n$, de rang au plus $|\mathcal{S}|$. Cette approximation peut être définie pour n'importe quel projecteur, toutefois nous nous limitons volontairement aux espaces de la forme donnée ci-dessus, pour lesquels l'approximation $\tilde{K}$ n'est pas seulement de faible rang, mais par ailleurs efficacement calculable.

Parmi les méthodes existantes dans la littérature pour la construction du support $\mathcal{S}$ définissant le sous-espace, on peut citer le tirage uniforme, le tirage proportionnellement à des scores d'importance [1] ou à leurs approximations [11, 8], la maximisation du déterminant de la matrice $K_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ associée aux points choisis, des méthodes d'échantillonnage itératives [4], différents critères basés sur la covariance [6], ou encore l'utilisation de relaxations continues [7].

Motivation : k-moyennes à noyau La motivation principale pour ce travail est le problème des k-moyennes à noyau, qui consiste à trouver $C = (c_1, \dots, c_k) \in \mathcal{H}^k$ minimisant le risque

$$\mathcal{R}(\rho, C) := \mathbf{E}_{x \sim \rho} \min_{1 \leq i \leq k} \|\phi(x) - c_i\|^2. \tag{1}$$

La minimisation du risque empirique $\mathcal{R}(\rho_n, \cdot)$ via l'algorithme de Lloyd, qui est l'approche la plus courante pour ré-

soudre le problème, nécessite essentiellement de calculer la matrice K. Lorsque n est grand, il est donc courant pour limiter les coûts de calcul d'appliquer l'algorithme de Lloyd non pas sur le plongement des données $(\phi(x_i))_{1\leq i\leq n}$, mais plutôt sur des descripteurs de Nyström $(\phi_{\mathcal{S}}(x_i))_{1\leq i\leq n}$ où l'application $\phi_{\mathcal{S}} : \mathcal{X} \to \mathbb{R}^{|\mathcal{S}|}$ est construite telle que $\langle \phi_{\mathcal{S}}(x), \phi_{\mathcal{S}}(y)\rangle = \langle P_{\mathcal{S}}\phi(x), P_{\mathcal{S}}\phi(y)\rangle$ pour tous $x, y \in \mathcal{X}$.

L'utilisation de descripteurs de Nyström se prête particulièrement à ce problème, car les k moyennes reconstruites en dimension $|\mathcal{S}|$ peuvent directement s'interpréter comme des points de $\mathcal{H}_{\mathcal{S}}$, et peuvent donc s'exprimer comme une combinaison linéaire de $|\mathcal{S}|$ éléments. Lorsque d'autres descripteurs sont utilisés, cette propriété est perdue ; bien qu'il soit toujours possible de reconstruire k éléments de $\mathcal{H}$ en se basant sur les cellules de Voronoi des moyennes calculées dans l'espace d'approximation [2], ces éléments sont des combinaisons linéaires de $O(n)$ points, ce qui peut être préjudiciable selon l'application.

Le lemme suivant permet de borner l'erreur induite par une telle approximation. Par la suite, on note $\|\cdot\|_{\text{HS}}$ la norme de Hilbert-Schmidt, et $\Phi : L^2(\mathcal{X}, \rho) \to \mathcal{H}$ l'opérateur $\Phi f := \int_{\mathcal{X}} f(x)\phi(x)\mathrm{d}\rho(x)$, dont l'adjoint Φ^* est l'opérateur d'inclusion de $\mathcal{H}$ dans $L^2(\mathcal{X}, \rho)$.

Théorème 2.1: *L'excès de risque induit par la recherche d'une solution dans $\mathcal{H}_{\mathcal{S}}$ plutôt que $\mathcal{H}$ est borné comme suit :*

$$\inf_{C\in\mathcal{H}^k_{\mathcal{S}}} \mathcal{R}(\rho, C) - \inf_{C\in\mathcal{H}^k} \mathcal{R}(\rho, C) \leq \|P^{\perp}_{\mathcal{S}}\Phi\|^2_{\text{HS}} =: \mathcal{E}(\mathcal{S}).$$

La démonstration est omise, mais il s'agit d'une généralisation de CALANDRIELLO et al. [3, Th. 1] qui prouve un résultat proche dans le cas d'une distribution empirique.

Minimisation de la trace Le critère $\mathcal{E}(\mathcal{S})$ qui apparaît dans le Théorème 2.1 ressemble au risque pour l'analyse en composantes principales à noyau (non centrée), toutefois nous nous restreignons aux sous-espaces de la forme $\mathrm{span}(\{\, \phi(x_i) \mid i \in \mathcal{S} \,\})$ pour lesquels les projections peuvent être calculées efficacement. En cela, la minimisation de l'équivalent empirique $\mathcal{E}_n(\mathcal{S}) := \|P^{\perp}_{\mathcal{S}}\Phi_n\|^2_{\text{HS}}$ de $\mathcal{E}(\mathcal{S})$ peut être vu comme un problème de sélection de "colonnes" à noyau (en anglais CSSP, pour column subset selection problem). Par ailleurs $\mathcal{E}_n(\mathcal{S}) = \mathrm{tr}(K - \tilde{K})$ correspond à la trace résiduelle de l'approximation de la matrice à noyau, et $\mathcal{E}(\mathcal{S})$ à la trace résiduelle $\mathrm{tr}(P^{\perp}_{\mathcal{S}}C)$ de l'opérateur (non centré) de covariance $C = \Phi\Phi^* = \int \phi(x)\phi(x)^*\mathrm{d}\rho(x)$.

Notons que des garanties sur le critère $\mathcal{E}(\mathcal{S})$ ont été obtenues pour différents algorithmes de construction du support $\mathcal{S}$, et notamment pour des méthodes randomisées qui ne cherchent pas explicitement à minimiser ce critère, voir p.ex. MUSCO et al. [8, Th. 17] pour un résultat de ce type lorsque les indices formant le support $\mathcal{S}$ sont échantillonnés de manière i.i.d. proportionnellement à des scores d'importance.

Dans cet article, nous proposons un algorithme itératif basé sur la minimisation de $\mathcal{E}_n(\mathcal{S})$. L'idée d'optimiser directement un objectif basé sur la trace a récemment été considérée dans un contexte plus large [5]. Toutefois, cet algorithme reste insatisfaisant pour l'approximation de matrices à noyau puisqu'il nécessite, en dépit des stratégies d'approximation déployées, de calculer malgré tout la matrice K de manière exacte.

3 Algorithme

Nous proposons maintenant un algorithme de sélection itératif, basé sur la minimisation gloutonne d'une approximation du critère $\mathcal{E}_n(\mathcal{S}) = \|P^{\perp}_{\mathcal{S}}\Phi_n\|^2_{\text{HS}}$.

Lemme 3.1 (Critère glouton): *Pour tout $\mathcal{S} \subseteq [1, \ldots, n]$:*

$$\operatorname*{arg\,min}_{\substack{s\in[1,\ldots,n]\\ P^{\perp}_{\mathcal{S}}\phi(x_s)\neq 0}} \mathcal{E}_n(\mathcal{S}\cup\{s\}) = \operatorname*{arg\,max}_{\substack{s\in[1,\ldots,n]\\ P^{\perp}_{\mathcal{S}}\phi(x_s)\neq 0}} \frac{\|\Phi^*_n P^{\perp}_{\mathcal{S}}\phi(x_s)\|^2}{\|P^{\perp}_{\mathcal{S}}\phi(x_s)\|^2} =: \mathcal{G}_{\mathcal{S}}(x_s).$$

Ce résultat se dérive facilement en notant que $P_{\mathcal{S}\cup\{s\}} = P_{\mathcal{S}} + u_s u_s^*$ avec $u_s := P^{\perp}_{\mathcal{S}}\phi(x_s)\|P^{\perp}_{\mathcal{S}}\phi(x_s)\|^{-1}$ lorsque $P^{\perp}_{\mathcal{S}}\phi(x_s) \neq 0$, et $u_s = 0$ sinon.

Une première approche pourrait donc consister à directement maximiser $\mathcal{G}_{\mathcal{S}}(x)$, toutefois évaluer $\mathcal{G}_{\mathcal{S}}(x_i)$ pour tous les $i \in [1, \ldots, n]$ nécessite de calculer la matrice à noyau exacte K, et l'optimisation de $\mathcal{G}_{\mathcal{S}}(x)$ sur $\mathcal{X}$ en utilisation une méthode du premier ordre est également coûteux.

Critère approché Nous proposons par la suite deux approximations supplémentaires de l'objectif $\mathcal{G}_{\mathcal{S}}(x)$. Le calcul du numérateur requiert d'évaluer le noyau entre x et le jeu de données entier, ce qui est coûteux en temps. Afin d'éviter cela, nous proposons d'utiliser un premier descripteur $\varphi : \mathcal{X} \to \mathbb{R}^f$ facile à calculer approchant le noyau κ dans le sens où $\langle \varphi(\cdot), \varphi(\cdot)\rangle \approx \kappa(\cdot, \cdot)$. Dans le cas d'un noyau invariant par translation $\phi(x, y) = K(x - y)$ défini sur $\mathcal{X} = \mathbb{R}^d$, l'utilisation de descripteurs de Fourier aléatoires [10] se prête bien à cet usage, on pourra par exemple choisir $\varphi(x) := \sqrt{2}[\cos(\omega_j^T x + b_j)]_{j=1}^f$ où $b_j \overset{i.i.d.}{\sim} \mathcal{U}([0, 2\pi[)$ et ω_j sont tirés de manière i.i.d. proportionnellement à la transformée de fourier du noyau K. Lorsque le coût en mémoire est également limitant, il est possible d'utiliser une matrice aléatoire structurée en remplacement de la matrice $[\omega_1, \ldots, \omega_f]$. L'idée de calculer des descripteur de Fourier aléatoires dans le but de produire des descripteur de Nyström peut sembler ouroborique, mais le descripteur φ est ici rapide à calculer, d'une part car sa construction ne dépend pas des données et est moins coûteuse par nature, mais également car la dimension de ces descripteurs peut être choisie faible en pratique (cf. Section 4). Formellement, on définit $\Psi_n = [\varphi(x_1), \ldots, \varphi(x_n)] \in \mathbb{R}^{f\times n}$. Pour tout $\mathcal{S} \subseteq [1, \ldots, n]$, on note $P_{\varphi,\mathcal{S}}$ le projecteur orthogonal sur le sous-espace $\mathrm{span}(\{\, \varphi(x_i) \mid i \in \mathcal{S} \,\})$. Le critère $\mathcal{G}_{\mathcal{S}}(x)$ peut donc être approché par l'estimateur $\|\Psi_n^T P^{\perp}_{\varphi,\mathcal{S}}\varphi(x)\|^2 / \|P^{\perp}_{\varphi,\mathcal{S}}\varphi(x)\|^2$ à base de descripteurs de Fourier aléatoires.

Approximation de la norme Si l'utilisation de descripteurs aléatoires permet certes de réduire le coût de calcul, le critère ainsi obtenu reste linéaire en le nombre n de points dans le jeu de données. Afin de réduire davantage la complexité (en temps et en espace) de l'algorithme nous proposons simplement d'approcher le critère $\mathcal{G}_{\mathcal{S}}(x)$ par l'estimateur non biaisé

$$\tilde{\mathcal{G}}_{\mathcal{S}}(x) := \frac{\|\Xi^T\Psi_n^T P^{\perp}_{\varphi,\mathcal{S}}\varphi(x)\|^2}{\|P^{\perp}_{\varphi,\mathcal{S}}\varphi(x)\|^2} \tag{2}$$

où Ξ est une matrice aléatoire avec $\Xi_{ij} \overset{i.i.d.}{\sim} \xi^{-1/2}\mathcal{N}(0, 1)$ pour tous $1 \leq i \leq \xi, 1 \leq j \leq n$. L'analyse théorique de

Algorithme 1 : Maximisation gloutonne de (2)

Input : noyau κ, taille l désirée du support, nb. f de descripteurs, taille ξ de l'approximation, données $X \in \mathbb{R}^{d\times n}$

$\Psi_n = \text{RFF}(X, \kappa, f)$ `// matrice` $f \times n$

$\Xi \in \mathbb{R}^{n\times\xi}$ with $\Xi_{ij} \overset{i.i.d.}{\sim} \frac{1}{\sqrt{\xi}}\mathcal{N}(0,1)$

$C = \text{zeros}(l, n)$

$P = [\Psi_n[:,i]^T\Psi_n[:,i]$ for i in $1:n]$

$\text{M} = \Psi_n^T * (\Psi_n * \Xi)$ `//` $n \times \xi$`, coût` $O(fn\xi)$

$\text{N} = \text{zeros}(n, \xi)$ `// cache pour` $C^TC\Xi$

$S = \emptyset$ `// support`

for $k = 1$ **to** l **do**

  $s = (P\ .> 10\epsilon)$ `//` $\epsilon =$ `machine epsilon`

  if $\text{sum}(s) == 0$ **then** break

  $c = \text{rownorms}(\text{M} - N).\hat{}2\ ./\ P$

  $j = \text{rand}(\text{findall}((c\ .\approx \text{maximum}(c[s]))\ .\&\ s))$

  $S = S \cup \{j\}$

  $C[k,s] = (\Psi_n[:,j]^T * \Psi_n[:,s] - C[:,j]^T * C[:,s])$

  $C[k,s] = C[k,s]/\text{sqrt.}(P[j])$

  $N = N + C[k,:]^T * (C[k,:] * \Xi)$ `//` $\Theta(\xi n)$

  $P = P - (C[k,:].\hat{}2)^T$

  $P[j] = -\infty$

end

return S

l'approximation induite est laissée pour des travaux futurs, et permettra de calibrer le choix du paramètre ξ.

Algorithme L'algorithme que nous considérons consiste à itérativement maximiser le critère (2) obtenu en combinant les deux approximations. Bien qu'il soit techniquement possible d'utiliser des techniques d'optimisation du premier ou second ordre afin d'optimiser cet objectif, cette stratégie souffre de problèmes de stabilité numérique. Nous nous contentons donc de maximiser cet objectif sur $\{x_1, \dots, x_n\}$, tout en prenant quelques précautions pour éviter les problèmes de stabilité numérique. La procédure obtenue correspond à l'Algorithme 1, où l'on utilise la notation $\text{RFF}(X, \kappa, f)$ pour la fonction qui calcule $[\varphi(x_1), \dots, \varphi(x_n)] \in \mathbb{R}^{f\times n}$, ainsi que les notations `.op` et `fun.` pour désigner l'application point à point d'un opérateur ou d'une fonction sur un vecteur. La complexité en temps de cet algorithme est $\Theta(nl(f + l + \xi))$, et le coût en mémoire $\Theta(n(l + f + \xi))$.

L'algorithme obtenu se rapproche dans l'esprit de la méthode de Fornace et Lindsey [5], mais évite le calcul explicite de la matrice K, et est par ailleurs bien plus stable ; l'utilisation de [5] sur des descripteurs aléatoires se révèle en effet impossible en pratique en raison de problèmes numériques.

4 Simulations numériques

Nous mesurons la performance de l'algorithme proposé, tout d'abord en termes de trace résiduelle puis pour le problème des k-moyennes à noyau.

Nous considérons par la suite un noyau gaussien $\kappa(x, y) = \exp(-\frac{1}{2\sigma^2}\|x - y\|^2)$ dont le paramètre d'échelle σ est choisi comme la médiane des distances $(\|x_i - x_j\|)_{1\le i\neq j\le n}$, et donc identique pour toutes les méthodes évaluées. Tous les jeux de données sont accessibles sur OpenML.

La Figure 1 (colonne de gauche) présente les résultats obtenus pour la trace résiduelle normalisée $(\text{tr}(K - \tilde{K})/\text{tr}(K))^{1/2}$ en fonction de la taille $|\mathcal{S}|$ du sous-espace d'approximation. Les jeux de données considérés sont "sulfur" ($n = 10081, d = 6$, identifiant OpenML : #44145) et "segment" ($n = 2310, d = 18$, id. OpenML : #36). La figure en haut à droite représente la même mesure d'erreur sur "sulfur", mais en fonction du temps de calcul. On observe de manière similaire dans les deux cas que notre méthode avec $f = \xi = 64$ parvient presque à égaler l'algorithme exact de minimisation de la trace (de coût quadratique), pour un temps de calcul réduit de deux ordres de grandeur. Avec $f = \xi = 32$, les performances de la méthode se dégradent pour $|\mathcal{S}| \geq 50$ environ. L'algorithme proposé obtient toujours de meilleures performances que l'échantillonnage uniforme ou avec des scores d'importance (on utilise dans ce cas des leverage scores "ridge" [1] avec un paramètre de régularisation choisi de sorte à ce que la dimension effective $\text{tr}(C(C + \lambda I)^{-1})$ soit de l'ordre de $|\mathcal{S}|$). La maximisation du déterminant fonctionne en comparaison moins bien, ce qui n'est pas étonnant puisque cette méthode n'est pas adaptative à la distribution des données. Nous comparons également notre méthode à l'approximation gloutonne d'un plongement à noyau moyen $\int \varphi(x)\mathrm{d}\rho_n(x)$ de ρ_n, c'est-à-dire à la minimisation gloutonne de $\|P_{\mathcal{S}}^{\perp}\Psi_n 1\|$, ou encore à l'algorithme "orthogonal least squares" appliqué sur le plongement moyen. Cette approche n'a à notre connaissance pas été utilisée dans ce contexte et est une alternative intéressante en termes de compromis erreur-coûts ; notre Algorithme 1 produit toutefois une meilleure approximation en comparaison.

Par ailleurs, nous rapportons les performances pour le problème des k-moyennes à noyau en bas à droite de la Figure 1. Le jeu de données utilisé est ici "covertype" ($n = 10^4, d = 10, k = 7$, id. OpenML : #1596), dont 70% de données sont utilisées pour l'apprentissage et les 30% restant pour évaluer les modèles entraînés. Le partitionnement est effectué avec l'algorithme de Lloyd avec initialisation k-means++ pour toutes les méthodes, soit sur des descripteurs aléatoires, soit sur des descripteurs de Nyström. Les performances observées pour l'erreur de généralisation sont cohérentes avec nos observations précédentes sur la trace résiduelle, ce qui conforte notre choix d'optimiser un critère basé sur la trace lorsque les descripteurs de Nyström sont destinés à résoudre un problème de partitionnement par k-moyennes à noyau.

Références

[1] Ahmed ALAOUI et Michael W. MAHONEY. "Fast Randomized Kernel Ridge Regression with Statistical Guarantees". In : *Advances in Neural Information Processing Systems*. 2015, p. 775-783.

[2] Gérard BIAU, Luc DEVROYE et Gábor LUGOSI. "On the Performance of Clustering in Hilbert Spaces". In : *IEEE Transactions on Information Theory* 54.2 (2008), p. 781-790.

[3] Daniele CALANDRIELLO et Lorenzo ROSASCO. "Statistical and Computational Trade-Offs in Kernel K-Means". In : *Advances in Neural Information Processing Systems*. T. 31. 2018.

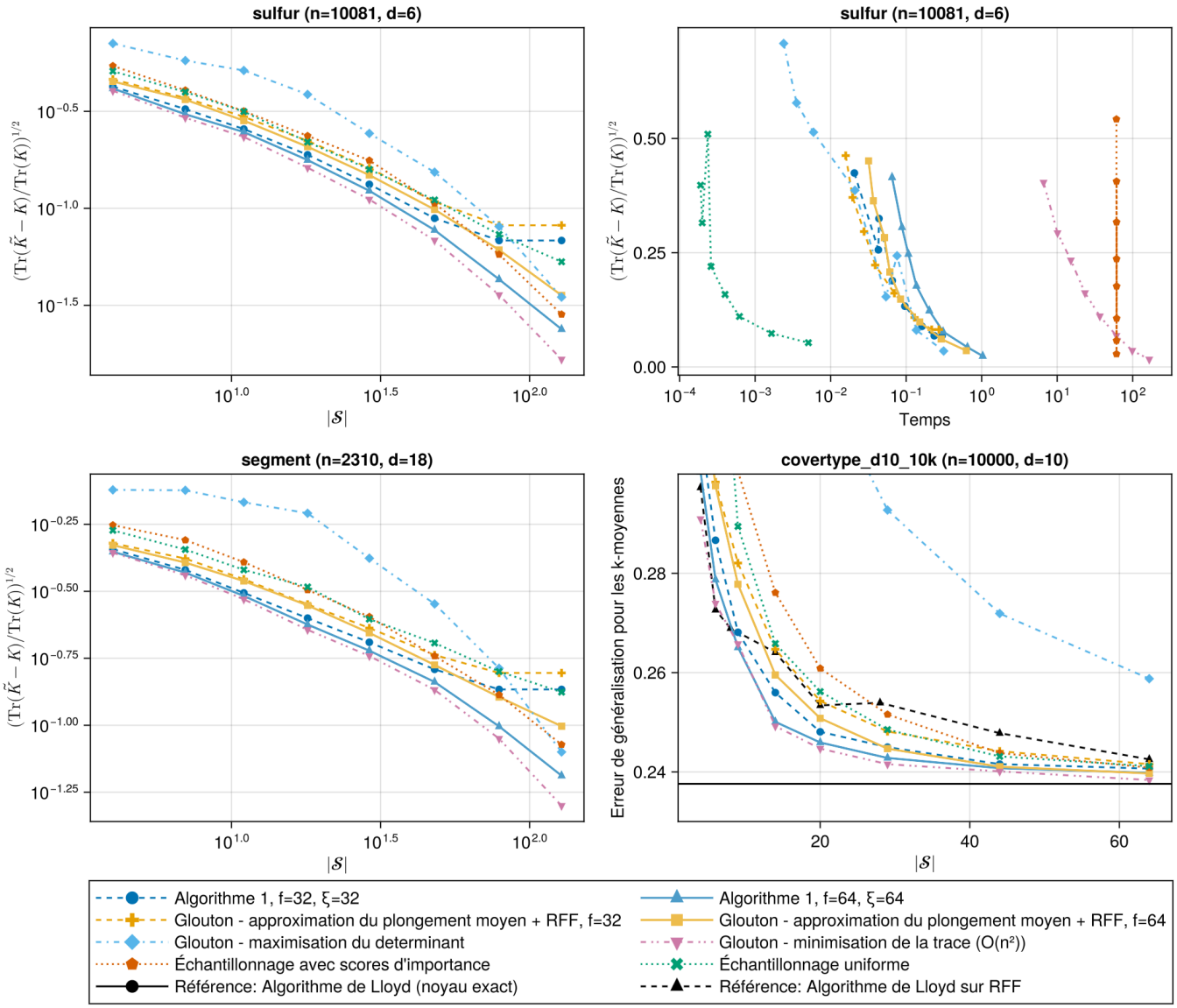

FIGURE 1 : Trace résiduelle (ligne du haut ainsi qu'en bas à gauche) et erreur des k-moyennes à noyau (en bas à droite) de différentes méthodes de construction de l'approximation de Nyström en fonction de la taille $|\mathcal{H}_S|$ du sous-espace ainsi que du temps de calcul (en haut à droite). Chaque point correspond à une médiane sur 10 essais randomisés.

[4] Yifan CHEN, Ethan N. EPPERLY, Joel A. TROPP et Robert J. WEBBER. *Randomly Pivoted Cholesky : Practical Approximation of a Kernel Matrix with Few Entry Evaluations*. 2023. Prépubl.

[5] Mark FORNACE et Michael LINDSEY. *Column and Row Subset Selection Using Nuclear Scores : Algorithms and Theory for Nyström Approximation, CUR Decomposition, and Graph Laplacian Reduction*. 2024. Prépubl.

[6] Bertrand GAUTHIER et Johan A. K. SUYKENS. "Optimal Quadrature-Sparsification for Integral Operator Approximation". In : *SIAM Journal on Scientific Computing* 40.5 (2018), A3636-A3674.

[7] Anant MATHUR, Sarat MOKA et Zdravko BOTEV. *Column Subset Selection and Nyström Approximation via Continuous Optimization*. 2023. Prépubl.

[8] Cameron MUSCO et Christopher MUSCO. "Recursive Sampling for the Nystrom Method". In : *Advances in Neural Information Processing Systems*. 2017, p. 3833-3845.

[9] E. J. NYSTRÖM. "Über Die Praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben". In : *Acta Mathematica* 54 (1930), p. 185-204.

[10] Ali RAHIMI et Benjamin RECHT. "Random Features for Large-Scale Kernel Machines". In : *Advances in Neural Information Processing Systems*. 2008, p. 1177-1184.

[11] Alessandro RUDI, Daniele CALANDRIELLO, Luigi CARRATINO et Lorenzo ROSASCO. "On Fast Leverage Score Sampling and Optimal Learning". In : *Advances in Neural Information Processing Systems*. 2018, p. 5672-5682.

[12] Ingo STEINWART et Andreas CHRISTMANN. *Support Vector Machines*. Springer Science & Business Media, 2008.

[13] Christopher WILLIAMS et Matthias SEEGER. "Using the Nyström Method to Speed up Kernel Machines". In : *Advances in Neural Information Processing Systems*. 2001, p. 682-688.