

Approximation de Nyström gloutonne par minimisation de la trace

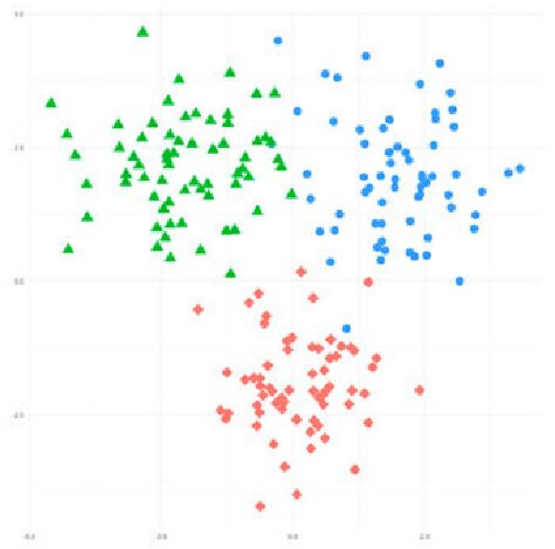
Antoine Chatalic¹

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

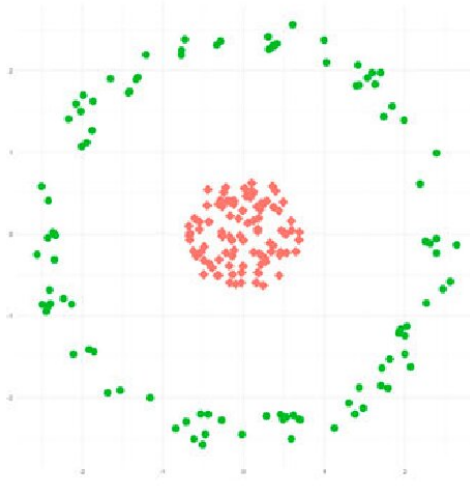
Motivation: k-moyennes à noyau

Données : n échantillons x_i *i.i.d.* ρ à partitionner.

k-moyennes (linéaire)



k-moyennes **à noyau** (c.-à-d. après plongement $x \mapsto \phi(x)$)



Objectif : trouver $C = (c_1, \dots, c_k)$ minimisant

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

Heuristique: algorithme de Lloyd ("k-moyennes")

Nécessite tous les $(\langle \phi(x_i), \phi(x_j) \rangle)_{1 \leq i, j \leq n} \rightarrow O(n^2)$.

Problème

Objectif : passer à l'échelle quand n est grand.

Stratégie : projeter les $(\phi(x_i))_{1 \leq i \leq n}$ sur un **sous-espace aléatoire de faible dimension**.

Problème : comment construire un « bon » sous-espace ?

Contribution : algorithme basé sur la minimisation gloutonne approchée de la trace résiduelle de la matrice à noyau.

Notations et hypothèses

Cadre et hypothèses :

- $\mathcal{H}$ RKHS séparable sur $\mathcal{X} = \mathbb{R}^d$;
- Noyau $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$ pour tous $x, y \in \mathcal{X}$, borné ;

Exemple : noyau Gaussien, $\kappa(x, y) = \exp(-\frac{1}{2\sigma^2}\|x - y\|^2)$.

Définitions :

- Matrice à noyau: $K_{ij} := \kappa(x_i, x_j)$, $K = \Phi_n^* \Phi_n \in \mathbb{R}^{n \times n}$ où $\Phi_n := [\phi(x_1), \dots, \phi(x_n)]$;
- Φ_n = discrétisation de $\Phi f = \int_{\mathcal{X}} f(x) \phi(x) d\rho(x)$.

Approximation de Nyström

Sous-espaces de la forme

$$\mathcal{H}_{\mathcal{S}} := \text{span}(\{\phi(x_i) \mid i \in \mathcal{S}\}), \text{ où } \mathcal{S} \subseteq [1, \dots, n].$$

On note $P_{\mathcal{S}}$ le projecteur orthogonal sur $\mathcal{H}_{\mathcal{S}}$ et $P^{\perp} := I - P$.

Stratégie : effectuer le partitionnement sur des descripteurs de Nyström $P_{\mathcal{S}} \phi(x)$ [1].

Interprétations :

- approximation dans le sous-espace de faible dimension $\mathcal{H}_{\mathcal{S}}$;
- approximation du noyau: $\kappa(x, y) \approx \langle P_{\mathcal{S}} \phi(x), P_{\mathcal{S}} \phi(y) \rangle$;
- approximation de rang faible de la matrice K par la matrice $\tilde{K} := \Phi_n^* P_{\mathcal{S}} \Phi_n$ (de rang au plus $|\mathcal{S}|$, et efficacement calculable).

→ problème de **sélection de points/colonnes**

Approches de sélection existantes

- tirage uniforme ;
- tirage proportionnellement à des scores d'importance [2] et approximations [3] ;
- maximisation du déterminant de la matrice $K_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ associée aux points choisis ;
- échantillonnage itératif [4] ;
- critères basés sur la covariance [5] ;
- relaxations continues [6].

Une borne sur l'erreur d'approximation

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

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

(Démonstration : généralisation de [7, Th. 1].)

Interprétation du critère $\mathcal{E}(\mathcal{S})$:

- $\mathcal{E}(\mathcal{S})$ = trace résiduelle $\text{tr}(P_{\mathcal{S}}^{\perp} C)$ de l'opérateur (non centré) de covariance $C = \Phi \Phi^* = \int \phi(x) \phi(x)^* d\rho(x)$;
- $\mathcal{E}_n(\mathcal{S}) = \|P_{\mathcal{S}}^{\perp} \Phi_n\|_{\text{HS}}^2 = \text{tr}(K - \tilde{K})$ = trace résiduelle de l'approximation de la matrice à noyau ;
- pb. de sélection de "colonnes" à noyau (CSSP).

Algorithme : minimisation gloutonne et approchée de $\mathcal{E}_n(\mathcal{S}) = \|P_{\mathcal{S}}^{\perp} \Phi_n\|_{\text{HS}}^2$.

Critère approché

Lemme : Pour tout $\mathcal{S} \subseteq [1, \dots, n]$:

$$\arg \min_{\substack{s \in [1, \dots, n] \\ P_{\mathcal{S}}^{\perp} \phi(x_s) \neq 0}} \mathcal{E}_n(\mathcal{S} \cup \{s\}) = \arg \max_{\substack{s \in [1, \dots, n] \\ P_{\mathcal{S}}^{\perp} \phi(x_s) \neq 0}} \frac{\|\Phi_n^* P_{\mathcal{S}}^{\perp} \phi(x_s)\|^2}{\|P_{\mathcal{S}}^{\perp} \phi(x_s)\|^2} =: \mathcal{G}_{\mathcal{S}}(x_s).$$

Évaluer $\mathcal{G}_{\mathcal{S}}(x_i)$ pour $i \in [1, \dots, n]$ **nécessite de calculer $\tilde{K}$** .

Approximation 1: réduction de la dimension

Utiliser un premier descripteur $\varphi: \mathcal{X} \rightarrow \mathbb{R}^f$ **facile à calculer**:

$$\varphi(x) := \sqrt{2}[\cos(\omega_j^T x + b_j)]_{j=1}^f, \omega_j \stackrel{i.i.d.}{\sim} \hat{\kappa}, b_j \stackrel{i.i.d.}{\sim} \mathcal{U}([0, 2\pi]).$$

On définit :

- $\Psi_n = [\varphi(x_1), \dots, \varphi(x_n)] \in \mathbb{R}^{f \times n}$;
- $P_{\varphi, \mathcal{S}}$ le projecteur orthogonal sur $\text{span}(\{\varphi(x_i) \mid i \in \mathcal{S}\})$.

Approximation du critère $\mathcal{G}_{\mathcal{S}}(x)$: $\frac{\|\Psi_n^T P_{\varphi, \mathcal{S}}^{\perp} \varphi(x)\|^2}{\|P_{\varphi, \mathcal{S}}^{\perp} \varphi(x)\|^2}$.

Approximation 2: dépendance en n réduite

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

où $\Xi_{ij} \stackrel{i.i.d.}{\sim} \xi^{-1/2} \mathcal{N}(0, 1)$, $1 \leq i \leq \xi$, $1 \leq j \leq n$.

Algorithme

Input : noyau κ , taille l 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 $\Xi_{ij} \stackrel{i.i.d.}{\sim} \frac{1}{\sqrt{\xi}} \mathcal{N}(0, 1)$

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

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

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

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

$S = \emptyset$ // support

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

$s = (P . > 10\varepsilon)$ // $\varepsilon = \text{machine epsilon}$

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

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

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

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

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

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

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

$P = P - (C[k, :] \wedge 2)^T$

$P[j] = -\infty$

end

return S

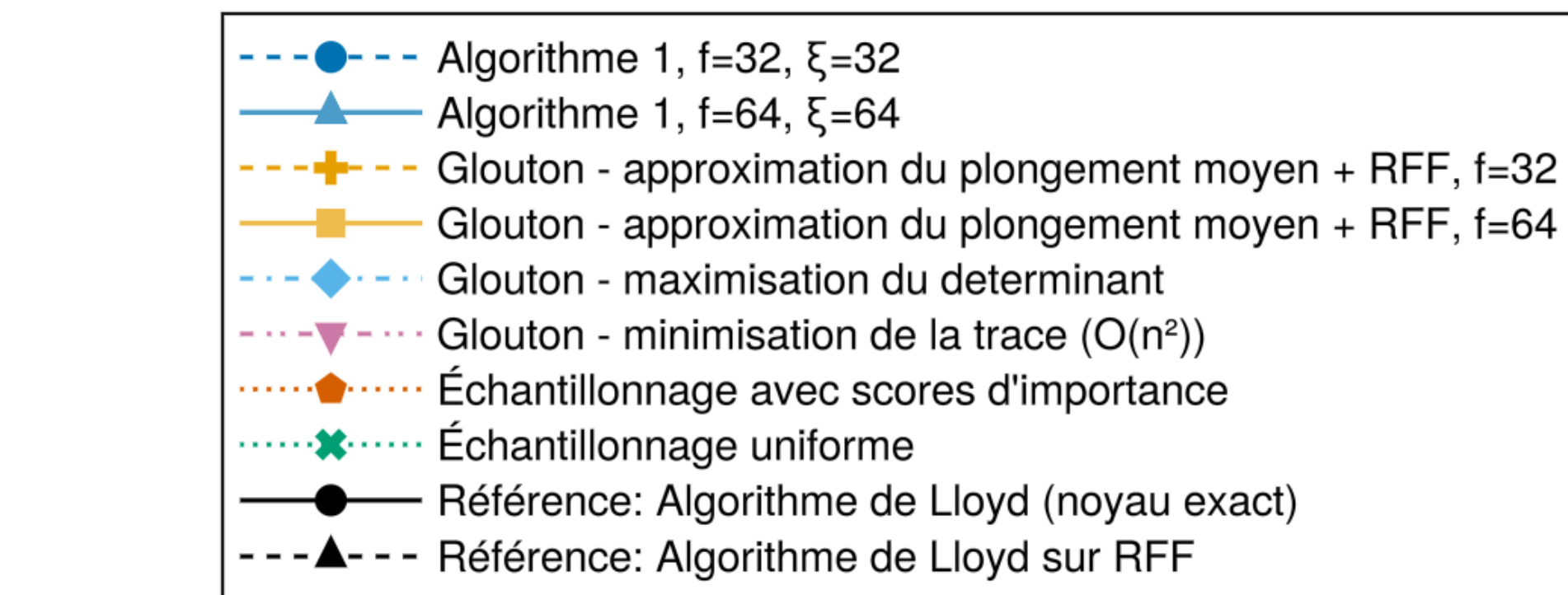
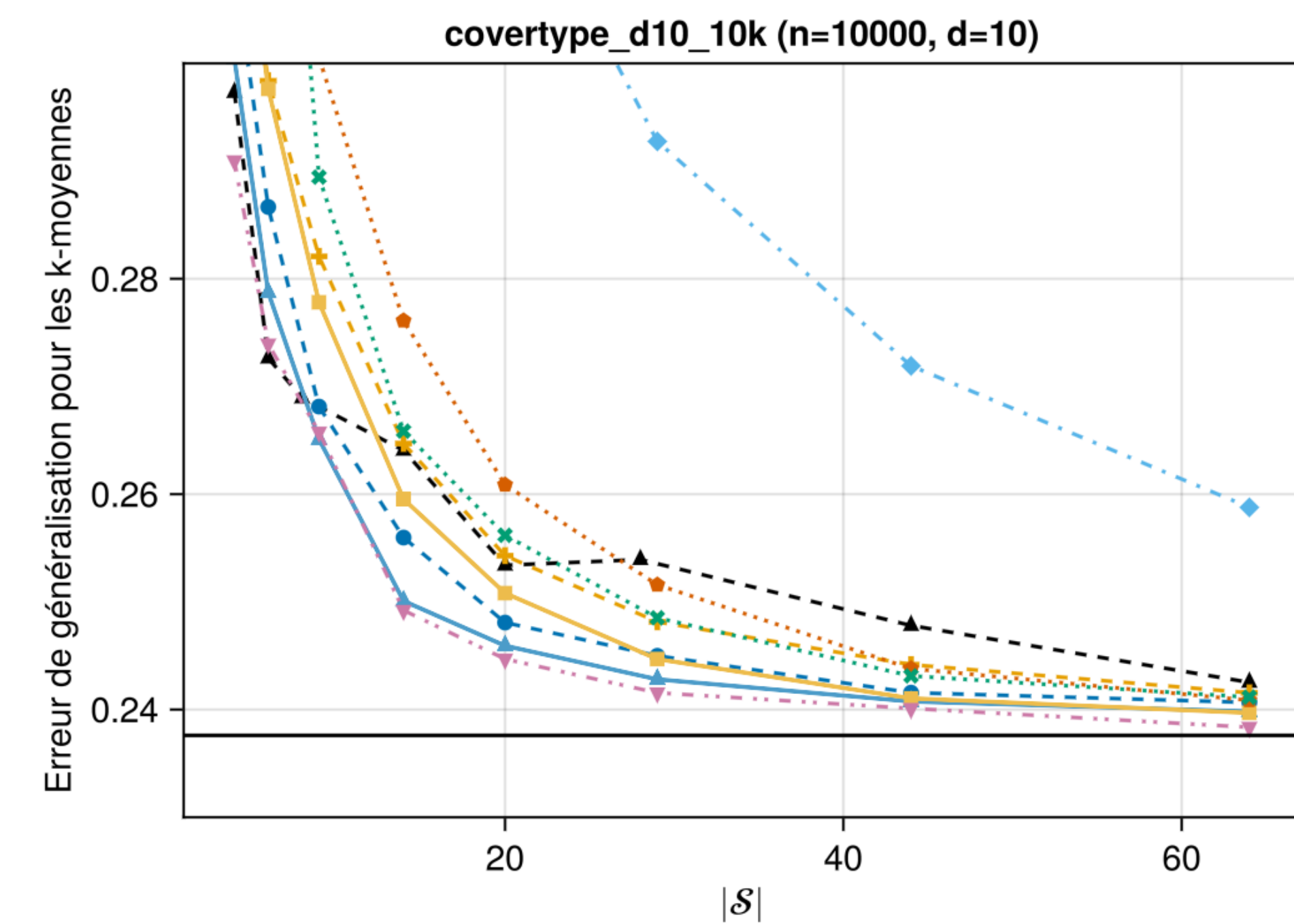
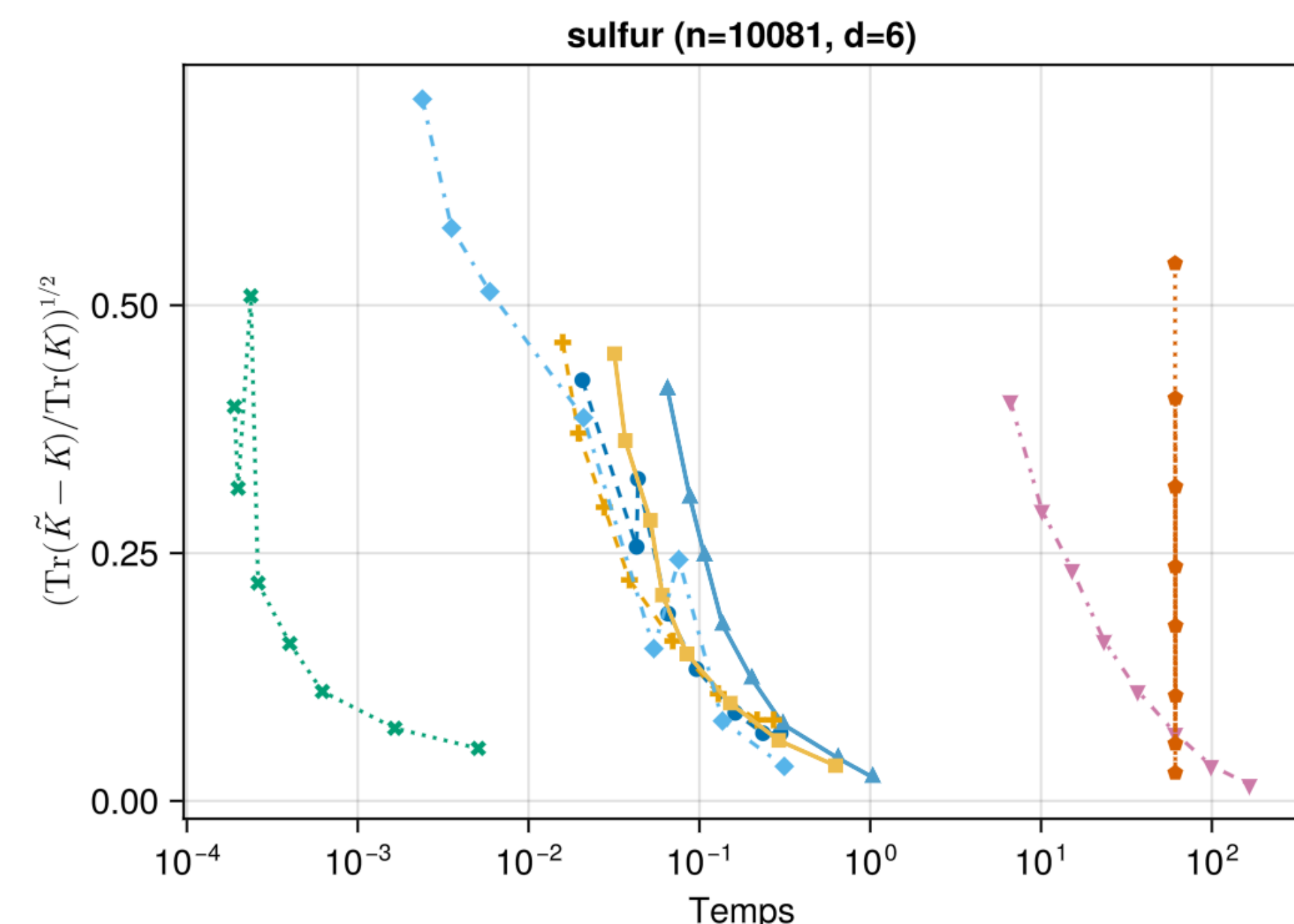
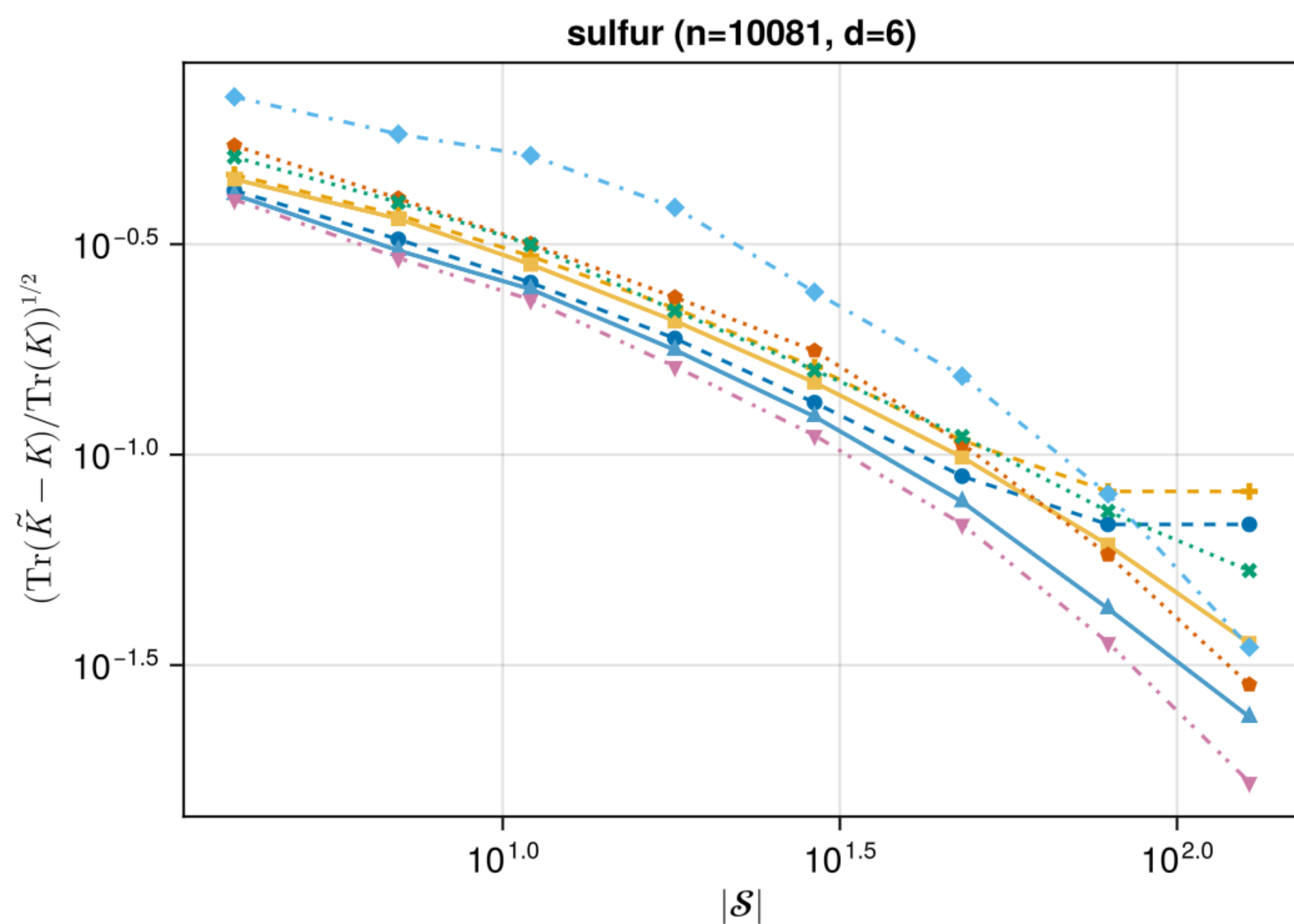
Complexité : temps $\Theta(nl(f + l + \xi))$, mémoire $\Theta(n(l + f + \xi))$.

Simulations numériques

Critères d'évaluation :

- trace résiduelle $(\text{tr}(K - \tilde{K}) / \text{tr}(K))^{1/2}$
- erreur de test pour les k-moyennes à noyau.

Paramètres : Noyau gaussien $\kappa(x, y) = \exp(-\frac{1}{2\sigma^2}\|x - y\|^2)$, σ = médiane des $(\|x_i - x_j\|)_{1 \leq i \neq j \leq n}$, $f = 32, 64$, $\xi = 32, 64$.



Conclusions :

- Avec $f = \xi = 64$, on parvient presque à **égaler l'algorithme exact** de minimisation de la trace (quadratique), pour un temps **réduit de deux ordres de grandeur**.
- Avec $f = \xi = 32$, les performances se dégradent pour $|\mathcal{S}| \geq 50$ environ.
- Notre algorithme obtient de meilleures performances que le tirage uniforme ou avec scores d'importance.
- Résultats pour l'erreur de généralisation cohérents avec les résultats pour la trace résiduelle.
- Bon compromis erreur-coûts avec une approximation gloutonne (OLS) du plongement moyen $\int \varphi(x) d\rho_n(x)$.

Références : [1] Christopher Williams and Matthias Seeger. "Using the Nyström Method to Speed up Kernel Machines." In: *Advances in Neural Information Processing Systems*. 2001, pp. 682–688. [2] Ahmed Alaoui and Michael W. Mahoney. "Fast Randomized Kernel Ridge Regression with Statistical Guarantees." In: *Advances in Neural Information Processing Systems*. 2015, pp. 775–783. [3] Cameron Musco and Christopher Musco. "Recursive Sampling for the Nystrom Method." In: *Advances in Neural Information Processing Systems*. 2017, pp. 3833–3845. [4] Yifan Chen et al. *Randomly Pivoted Cholesky: Practical Approximation of a Kernel Matrix with Few Entry Evaluations*. Dec. 12, 2023. Pre-published. [5] Bertrand Gauthier and Johan A. K. Suykens. "Optimal Quadrature-Sparsification for Integral Operator Approximation." In: *SIAM Journal on Scientific Computing* 40.5 (Jan. 2018), A3636–A3674. [6] Anant Mathur, Sarat Moka, and Zdravko Botev. *Column Subset Selection and Nyström Approximation via Continuous Optimization*. Apr. 19, 2023. Pre-published. [7] Daniele Calandriello and Lorenzo Rosasco. "Statistical and Computational Trade-Offs in Kernel K-Means." In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.