Nyström Kernel Mean Embeddings

Antoine Chatalic[1] Nicolas Schreuder[1] Alessandro Rudi[2] Lorenzo Rosasco[3,1]

[1] DIBRIS and MaLGA, Università di Genova, [2] Inria, École normale supérieure, PSL research university [3] CBMM, MIT, Istituto Italiano di Tecnologia

Introduction

Problem: approximating a kernel mean embedding

$$\mu := \mu(\rho) := \int_{\mathcal{X}} \phi(x)\, \mathrm{d}\rho(x)$$

where $\phi : \mathcal{X} \to \mathcal{H}$ is a feature map associated to a reproducing kernel Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|$.

Main assumption: there exists $K < \infty$ s.t. $\sup_{x \in \mathcal{X}} \|\phi(x)\| \leq K$.

Existing approaches

- **Empirical estimator:** $\hat{\mu} := \mu(\hat{\rho}_n) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$.
  - Rate: $\|\mu - \hat{\mu}\| = O(n^{-1/2})$
  - Time complexity: $O(n)$
  - Space complexity: $O(nd)$
  - Complexity of MMD computation: $O(n^2)$
- **Sampling:** Random features [1], DPPs [2] (no practical algorithm).
- Incoherence-based selection [3] (limited guarantees), Herding [4].
- Estimators based on Stein's effect. [5] Improves constants but not the rate.

Problem statement

Design a new estimator $\hat{\mu}_m$ **computed from m samples** which:

1. can be computed more efficiently than $\hat{\mu}$;
2. preserves the statistical accuracy of $\hat{\mu}$.

Applications

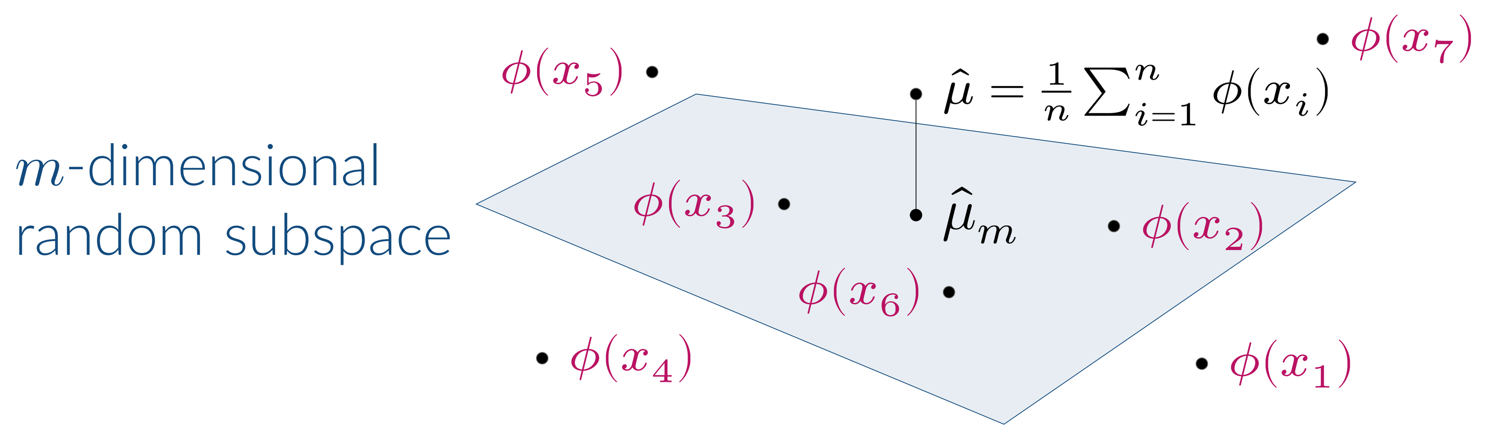
- **Quadratures in RKHS:** The quantity $\left\|\mu - \sum_{j=1}^{m} w_j \phi(x_j)\right\|$ corresponds to the worst-case error (for $f \in \mathcal{H}$) of the approximation

$$\int f(x)\, \mathrm{d}\rho(x) \approx \sum_{j=1}^{m} w_j f(x_j).$$

- Approximate **metrics between distributions**:

$$\mathrm{MMD}(\rho_1, \rho_2) := \|\mu(\rho_1) - \mu(\rho_2)\| \approx \|\hat{\mu}_m(\rho_1) - \hat{\mu}_m(\rho_2)\|.$$

Proposed Method



Idea: project $\hat{\mu}$ on the low-dimensional subspace $\mathcal{H}_m := \mathrm{span}\left\{\phi(\tilde{X}_1), \ldots, \phi(\tilde{X}_m)\right\}$ where the $(\tilde{X}_i)_{1 \leq i \leq m}$ are drawn from the dataset.

$$\hat{\mu}_m := P_m \hat{\mu} = \sum_{1 \leq j \leq m} w_j \phi(\tilde{X}_j)$$

with $m \ll n$ and P_m the projection on $\mathcal{H}_m$.

The weights $(w_j)_{1 \leq j \leq m}$ can be computed in closed form: $w = \frac{1}{n} K_m^{+} K_{mn} 1_n$.

Complexities: time $\Theta(nmd + m^3)$, space $\Theta(md)$.

How small can m be chosen to get the same statistical accuracy as $\hat{\mu}$?

Theoretical Results

We denote:

- $C = \int \phi(x) \otimes \phi(x)\, \mathrm{d}\rho(x)$ the covariance operator.
- $\mathcal{N}(\lambda) := \mathrm{tr}(C(C + \lambda I)^{-1})$ the effective dimension for any $\lambda > 0$.

Theorem: Main result

Assume data points $x_1, \ldots, x_n$ drawn i.i.d. from the probability distribution ρ, and $m \leq n$ sub-samples $\tilde{x}_1, \ldots, \tilde{x}_m$ drawn uniformly with replacement from $\{x_1 \ldots, x_n\}$. Then, it holds with probability $\geq 1 - \delta$ that

$$\|\mu - \hat{\mu}_m\| \leq \frac{c_1}{\sqrt{n}} + \frac{c_2}{m} + \frac{c_3\sqrt{\log(m/\delta)}}{m}\sqrt{\mathcal{N}\left(\frac{12K^2\log(m/\delta)}{m}\right)},$$

provided that $m \geq \max(67, 12K^2\|C\|_{\mathcal{L}(\mathcal{H})}^{-1})\log(m/\delta)$, where c_1, c_2, c_3 are constants of order $K\log(1/\delta)$.

Idea of the decomposition: for any $\lambda > 0$, it holds almost surely

$$\|\mu - \hat{\mu}_m\| \leq \|\mu - \hat{\mu}\| + \|P_m^{\perp}(C + \lambda I)^{1/2}\|_{\mathcal{L}(\mathcal{H})}\|(C + \lambda I)^{-1/2}(\hat{\mu} - \tilde{\mu}_m)\|.$$

Application to the MMD: Similar bound for $\|\hat{\mu}_m(\rho) - \hat{\mu}_m(\nu)\|$ when approximating both $\hat{\mu}_m(\rho)$ and $\hat{\mu}_m(\nu)$ via independent subsamples $\to$ Complexity $O(m^2)$.

Corollary: Rates with Additional Hypotheses

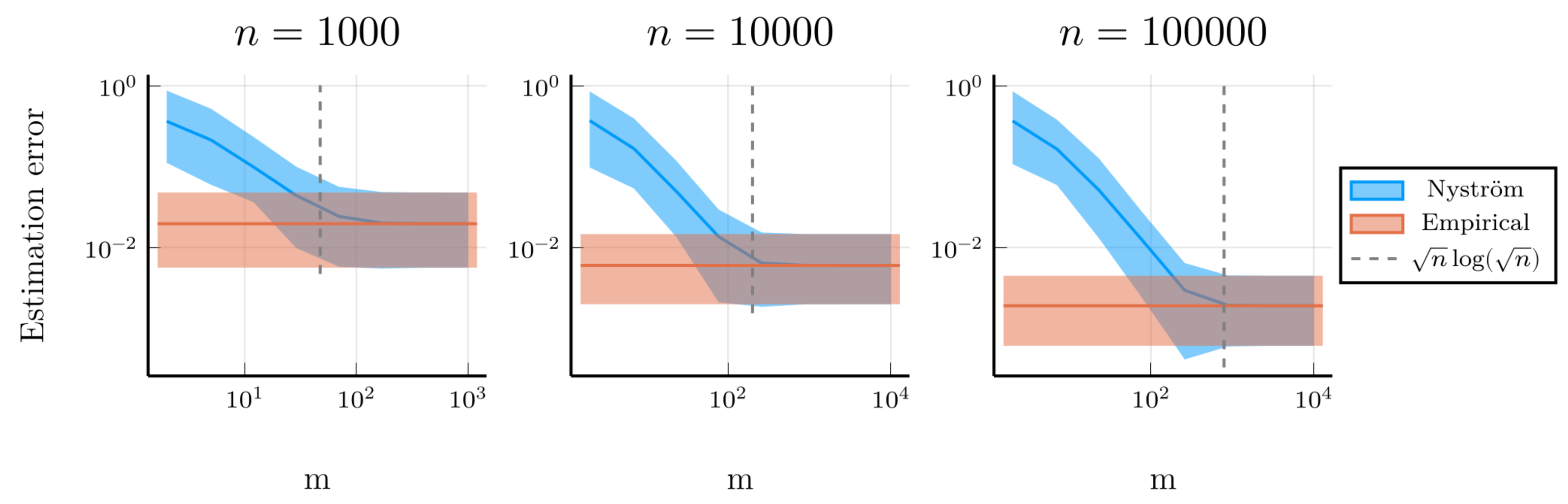
Assume that for some $c > 0$,

- either $\mathcal{N}(\lambda) \leq c\lambda^{-\gamma}$ for some $\gamma \in]0, 1]$ and $m = n^{1/(2-\gamma)}\log(n/\delta)$
- or $\mathcal{N}(\lambda) \leq \log(1 + c/\lambda)/\beta$, for some $\beta > 0$ and $m = \sqrt{n}\log(\sqrt{n}\max(1/\delta, c/(6K^2))$.

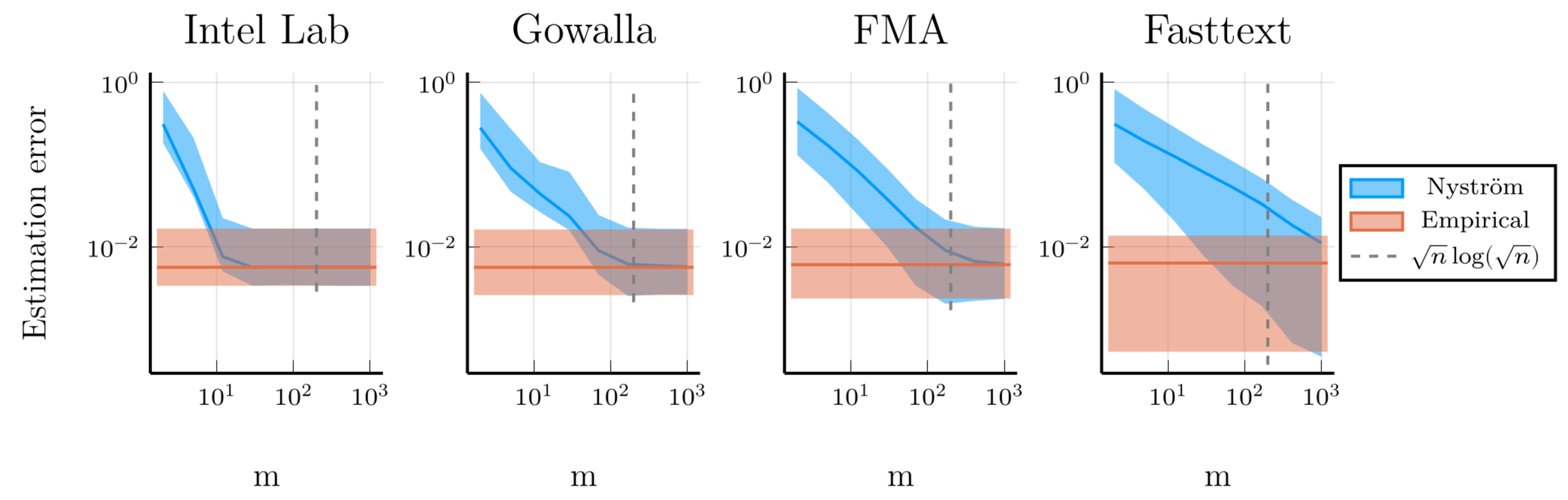
Then we get: $\|\mu - \hat{\mu}_m\| = O\left(\frac{1}{\sqrt{n}}\right)$.

Empirical Results

On synthetic data (gaussian mixture model in dimension $d = 10$):



On four different real datasets:



References

[1] Francis Bach. "On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions." In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 714–751.

[2] Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. "Kernel Quadrature with DPPs." In: Advances in Neural Information Processing Systems. Vol. 32. Dec. 31, 2019, pp. 12927–12937.

[3] Efren Cruz Cortes and Clayton Scott. "Scalable Sparse Approximation of a Sample Mean." In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, May 2014, pp. 5237–5241.

[4] Yutian Chen, Max Welling, and Alex Smola. "Super-Samples from Kernel Herding." In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. UAI'10. Arlington, Virginia, USA: AUAI Press, July 8, 2010, pp. 109–116.

[5] Krikamol Muandet et al. "Kernel Mean Estimation and Stein Effect." In: *Proceedings of the 31st International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, Jan. 27, 2014, pp. 10–18.

antoine.chatalic@dibris.unige.it