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

ICML – July 2022

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 $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ with norm $\|\cdot\|$.

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

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 the unit ball of the RKHS) of the approximation

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

- Approximate **metrics between distributions**:

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

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: $O(n^2)$

Other approaches:

- **Sampling:** Random features [1], DPPs [2] (no practical/efficient algorithms).
- 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 $\widehat{\mu}_m$ **computed from** m **samples** which:

1. can be computed more efficiently than $\widehat{\mu}$;
2. preserves the $O(n^{-1/2})$ statistical accuracy of $\widehat{\mu}$.

Proposed Method

Idea: project $\hat{\mu}$ on the m-dimensional subspace $\mathcal{H}_m := \mathrm{span}\left\{\phi(\tilde{X}_1), ..., \phi(\tilde{X}_m)\right\}$:

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

with:

- $m \ll n$ and P_m the projection on $\mathcal{H}_m$.
- the $(\tilde{X}_i)_{1 \leq i \leq m}$ are drawn from the dataset.

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 $K\log(1/\delta)$.

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

Empirical Results

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