Quick Summary

- Mean embedding: MMD: information theory on kernel-enriched domains.
- Goal: outlier-robust estimation.
- Contribution:
  - Optimal sub-Gaussian deviation bound (minimal 2nd order assumption)
  - Practical algorithms.

Target Quantities

- Mean embedding:
  \[ \mathbb{P} \to \mu = \int f(x) \, d\mathbb{P}(x) \in \mathcal{H}_k. \]

- Maximum mean discrepancy (MMD)
  \[ \text{MMD}(\mathbb{P}, Q) = \|\mu - \mu\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k} \left( \mathbb{E}_f \cdot f(x) - \mathbb{E}_{f} \cdot f(x) \right). \]

Notes:
- Large number of applications; review [1].
- Numerous kernel-ended domains: \( K(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}_k}, \varphi(x) = K(\cdot, x). \)

Goal

- Design outlier-robust estimators.
- Interest: unbounded kernels
  - exponential kernel: \( K(x, y) = e^{\langle \gamma, y \rangle} \)
  - polynomial kernel: \( K(x, y) = \langle \gamma, y \rangle^p \)
  - string, time series or graph kernels.

- Issue with average: A single outlier can ruin it.

Estimator

- Idea (MOM):
  1. Partition: \( x_1, \ldots, x_N/Q, \ldots, x_{N-N/Q}, \ldots, x_N \)
  2. Compute average in each block:
     \[ n_1 = \frac{1}{N} \sum_{i=1}^N x_i, \ldots, n_Q = \frac{1}{N} \sum_{i=1}^N x_i. \]
  3. Estimate \( \mathbb{E}X \): \( \frac{1}{N} \sum_{i=1}^N x_i \).

- On MMD: replace the expectation with MON
  \[ \text{MMD}_0(\mathbb{P}, Q) = \sup_{f} \text{med} \left\{ \frac{1}{N} \sum_{i=1}^N f(x_i) - \frac{1}{N} \sum_{i=1}^N f(y_i) \right\}. \]

Code: https://bitbucket.org/ThiemoMathieu/monk-mmd

Numerical Illustrations

- (a) Gaussian distribution, \( N, P \) (no outlier), RBF kernel.
- (b) Gaussian distribution, \( N, P \) (no outlier), quadratic kernel.
- (c) Pareto distribution, RBF kernel.
- (d) Pareto distribution, quadratic kernel.

Finite-Sample Bound for \( \text{MMD}_0(\mathbb{P}, Q) \) (\( \mu \): Similar)

Assume:
- Contamination: \( \{(x_i, y_i)\}_{i=1}^N, N \leq Q(1/2 - \delta), \delta \in (0, 1/2]. \)
- MMD 2nd-order assumption: \( 3 \text{Tr} \left( \Sigma_{2Q} \right) \leq \text{Tr} \left( \Sigma_{2Q} \right) \).

Then, for any \( \eta \in (0, 1) \) such that \( Q = \frac{256}{\log^2(1/\eta)} \), satisfies \( Q \in (N, (1/2 - \delta), N/2) \), with probability at least \( 1 - \eta \)

\[ \text{MMD}_0(\mathbb{P}, Q) - \text{MMD}_0(\mathbb{P}, Q) \leq 12 \max \left( \sqrt{\frac{\text{Tr} \left( \Sigma_{2Q} \right)}{\eta}}, \sqrt{\frac{N \eta}{\delta}} \right) \]

Discussion

(i) \( N \)-dependence: \( Q \left( \frac{1}{N} \right) \) is optimal for MMD estimation [2].

(ii) \( \Sigma \)-dependence:
- Optimal sub-Gaussian deviation bound for mean estimation under minimal 2nd-order condition even on \( \mathbb{R}^d [3] \) — long-lasting open question.
- They rely on tournament procedure: numerically hard.
- Most practical convex relaxation [4]; \( O(N^3) \).
- After submission: [5]: \( Q(N^3 + d^2) \), \( d \ll \infty \).

(iii) \( \delta \)-dependence:
- Larger \( \delta \) means less outliers, the bound becomes tighter, one needs less blocks.
- Optimal?

(iv) Breakdown point — asymptotic concept:
- median \( \Rightarrow \) Using \( Q \) blocks is resistant to \( Q/2 \) outliers.
- \( Q \) can grow with \( N \), (almost) \( N/2 \).
- Breakdown point can be \( 25\% \).

(v) Unknown \( Q \):
- One choose \( Q \) adaptively by the Lepski method.
- Same guarantee but with increased computational cost.

Acknowledgements

Guillaume Lecué is supported by a grant of the French National Research Agency (ANR), “Investissements d’Avenir” (LabEx Ecodec/ANR-11-LABX-0047).

References