



UCL



GATSBY

Inria

MMD Aggregated Two-Sample Test KSD Aggregated Goodness-of-fit Test

University College London
Centre for Artificial Intelligence
Gatsby Computational Neuroscience Unit
Inria London Programme

Antonin Schrab

a.schrab@ucl.ac.uk
antoninschrab.github.io

Overview

1 MMDAgg: MMD Aggregated Two-Sample Test

- Two-sample problem
- MMD single test
- MMD aggregated test
- Experiments

2 KSDAgg: KSD Aggregated Goodness-of-fit Test

- Goodness-of-fit problem & KSD tests
- Uniform separation rate
- Experiments

MMD Aggregated Two-Sample Test



Antonin
Schrab

†‡§



Ilmun
Kim

*



Mélisande
Albert

★



Béatrice
Laurent

★



Benjamin
Guedj

†§



Arthur
Gretton

‡

† Centre for Artificial Intelligence, UCL

‡ Gatsby Computational Neuroscience Unit, UCL

§ Inria London Programme

* Department of Statistics & Data Science, Yonsei University

★ Institut de Mathématiques, Université de Toulouse

MMDAgg: Two-sample problem

1 MMDAgg: MMD Aggregated Two-Sample Test

- Two-sample problem
- MMD single test
- MMD aggregated test
- Experiments

2 KSDAgg: KSD Aggregated Goodness-of-fit Test

- Goodness-of-fit problem & KSD tests
- Uniform separation rate
- Experiments

Two-sample problem

Two-sample problem:

Given independent samples

- $\mathbb{X}_m := (\mathbb{X}_1, \dots, \mathbb{X}_m)$ where $X_i \stackrel{\text{iid}}{\sim} p$ in \mathbb{R}^d
- $\mathbb{Y}_n := (\mathbb{Y}_1, \dots, \mathbb{Y}_n)$ where $Y_i \stackrel{\text{iid}}{\sim} q$ in \mathbb{R}^d

can we decide whether or not $p \neq q$ holds?

Non-asymptotic regime: no approximation using asymptotic distributions

Balanced sample sizes: $m \leq n$ and $n \leq Cm$

Statistical hypothesis testing:

$$\begin{array}{lll} \mathcal{H}_0: p = q & \text{against} & \mathcal{H}_a: p \neq q \\ \Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1 & \iff & \text{reject } \mathcal{H}_0 \\ \Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0 & \iff & \text{fail to reject } \mathcal{H}_0 \end{array}$$

What is a ‘good’ test Δ ?

By design: probability of type I error is α -controlled

$$\mathbb{P}_{p \times p}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1) \leq \alpha$$

Aim: β -control the probability of type II error

$$\mathbb{P}_{p \times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta \quad (\star)$$

Intuition: Δ cannot β -control (\star) if $\|p - q\|_2$ is too small

Uniform separation rate: What is the smallest value $\delta > 0$ such that Δ β -controls (\star) against all alternative hypotheses satisfying $p - q \in \mathcal{C}$ and $\|p - q\|_2 > \delta$?

$$\rho(\Delta, \mathcal{C}, \beta) := \inf \left\{ \delta > 0 : \sup_{\substack{p, q: \\ p - q \in \mathcal{C} \\ \|p - q\|_2 > \delta}} \mathbb{P}_{p \times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta \right\}$$

What is a ‘good’ test Δ ?

$$\rho(\Delta, \mathcal{C}, \beta) := \inf \left\{ \delta > 0 : \sup_{\substack{p, q: \\ p-q \in \mathcal{C} \\ \|p-q\|_2 > \delta}} \mathbb{P}_{p \times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta \right\}$$

Uniform separation rates we consider are of the form $C(m+n)^{-r}$

Minimax rate: smallest rate a test Δ_α of level α can hope to achieve

$$\underline{\rho}(\mathcal{C}, \alpha, \beta) := \inf_{\Delta_\alpha} \rho(\Delta_\alpha, \mathcal{C}, \beta)$$

Aim: construct a ‘good’ test which is

- **Optimal in the minimax sense:** it achieves the minimax rate (possibly up to some log or log log term)
- **Adaptive:** no dependence on unknown parameters of \mathcal{C}

Question: Which class of functions \mathcal{C} to use?

Sobolev balls

Sobolev balls:

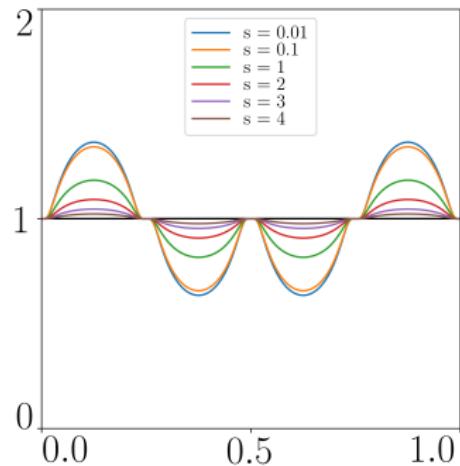
$$\mathcal{S}_d^s(R) := \left\{ f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|\xi\|_2^{2s} |\widehat{f}(\xi)|^2 d\xi \leq (2\pi)^d R^2 \right\}$$

- radius $R > 0$
- dimension $d \in \mathbb{N}$
- smoothness parameter $s > 0$
- Fourier transform $\widehat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-ix^\top \xi} dx$

Minimax rate over Sobolev balls:

$$\underline{\rho}(\mathcal{S}_d^s(R), \alpha, \beta) \asymp (m+n)^{-2s/(4s+d)}$$

Adaptive test: test with no dependence on unknown parameters s and R of $\mathcal{S}_d^s(R)$



MMDAgg: MMD single test

1 MMDAgg: MMD Aggregated Two-Sample Test

- Two-sample problem
- **MMD single test**
- MMD aggregated test
- Experiments

2 KSDAgg: KSD Aggregated Goodness-of-fit Test

- Goodness-of-fit problem & KSD tests
- Uniform separation rate
- Experiments

Kernels

Kernel: positive definite function $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$

Reproducing Kernel Hilbert Space (RKHS): inner product space of real-valued functions \mathcal{H}_k satisfying:

- $k(\cdot, x) \in \mathcal{H}_k$
- $\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$

Translation-invariant kernel: $k(x, y) := \prod_{i=1}^d K_i(x_i - y_i)$

for some functions $K_1, \dots, K_d \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ integrating to 1

Bandwidth: $\lambda = (\lambda_1, \dots, \lambda_d) \in (0, \infty)^d$ giving the kernel

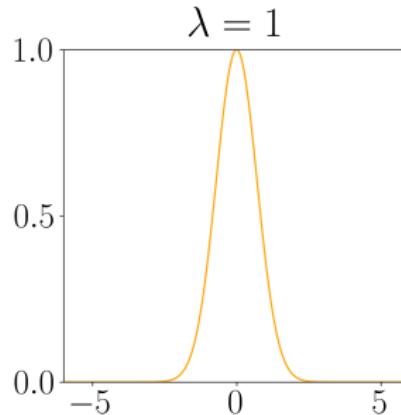
$$k_\lambda(x, y) := \prod_{i=1}^d \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$$

Gaussian kernel

$$k_{\lambda}(x, y) := \prod_{i=1}^d \frac{1}{\lambda_i} K_i \left(\frac{x_i - y_i}{\lambda_i} \right)$$

Gaussian kernel: $K_i(u) = \frac{1}{\sqrt{\pi}} \exp(-u^2), u \in \mathbb{R}, i = 1, \dots, d$

$$k_{\lambda}(x, y) := \frac{1}{\pi^{d/2} \lambda_1 \cdots \lambda_d} \exp \left(- \sum_{i=1}^d \frac{(x_i - y_i)^2}{\lambda_i^2} \right)$$

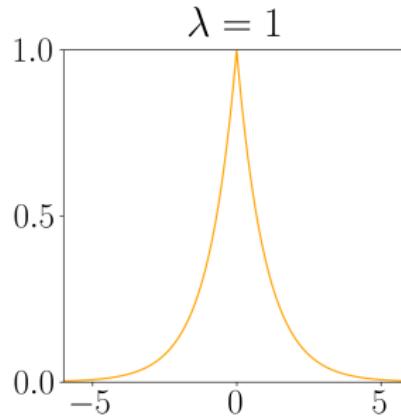


Laplace kernel

$$k_{\lambda}(x, y) := \prod_{i=1}^d \frac{1}{\lambda_i} K_i \left(\frac{x_i - y_i}{\lambda_i} \right)$$

Laplace kernel: $K_i(u) = \frac{1}{2} \exp(-|u|), u \in \mathbb{R}, i = 1, \dots, d$

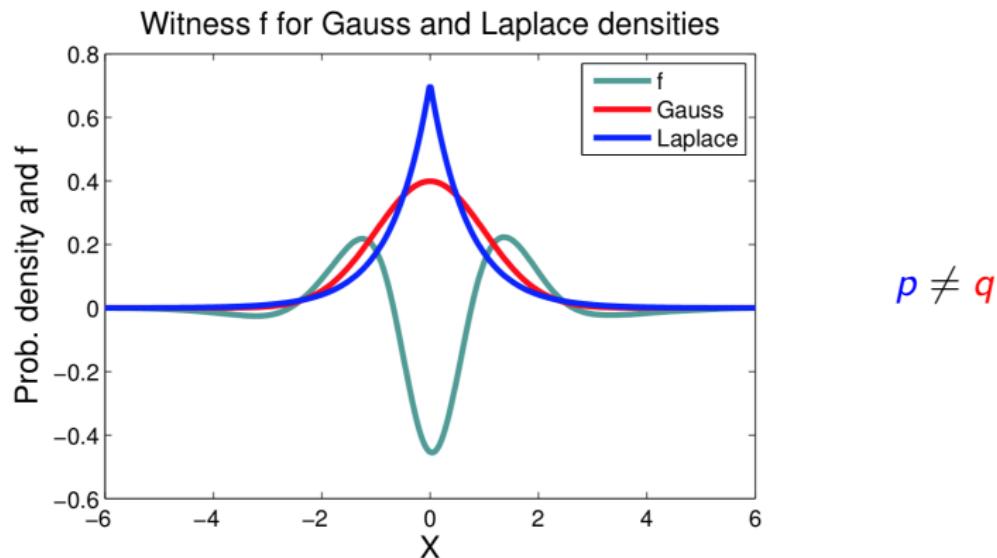
$$k_{\lambda}(x, y) := \frac{1}{2^d \lambda_1 \cdots \lambda_d} \exp \left(- \sum_{i=1}^d \frac{|x_i - y_i|}{\lambda_i} \right)$$



Maximum Mean Discrepancy

Two-sample test based on the MMD:

$$\text{MMD}_{\lambda}(p, q) := \sup_{f \in \mathcal{H}_{\lambda}: \|f\|_{\mathcal{H}_{\lambda}} \leq 1} |\mathbb{E}_{X \sim p}[f(X)] - \mathbb{E}_{Y \sim q}[f(Y)]|$$



Characteristic kernel: property on the kernel which ensures that

$$\text{MMD}_{\lambda}(p, q) = 0 \iff p = q$$

Fukumizu et al. Kernel measures of conditional dependence. *NeurIPS*, 2008.

Bandwidth intuition

- **Small sample sizes:** only global differences are detectable
 - **Small bandwidth:** wrongly detects artificial local differences under \mathcal{H}_0
 - **Large bandwidth:** well-suited to detect global differences under \mathcal{H}_a
- **Large sample sizes:** local differences are detectable
 - **Small bandwidth:** well-suited to detect local differences under \mathcal{H}_a
 - **Large bandwidth:** fails to detect local differences under \mathcal{H}_a

Bandwidths should decrease as the **sample sizes** increase

Bandwidth selection

Choice of kernel bandwidths λ is **crucial** for test power! Common methods:

- **median heuristic:**

$$\lambda_i := \text{median} \left\{ \|z - z'\|_2 : z, z' \in \mathbb{X}_m \cup \mathbb{Y}_n, z \neq z' \right\}, \quad i = 1, \dots, d$$

⇒ no power guarantees

Gretton et al. A kernel two-sample test. *JMLR*, 2012.

- **data splitting:** split the data in two parts

- **1st part:** select bandwidth λ^* maximizing a proxy for asymptotic power

- **2nd part:** perform MMD test with selected bandwidth λ^*

⇒ asymptotic power guarantees (but low power for low samples sizes)

Gretton et al. Optimal kernel choice for large-scale two-sample tests. *NeurIPS*, 2012.

Liu et al. Learning deep kernels for non-parametric two-sample tests. *ICML*, 2020.

Our paper: propose new method which

- aggregates multiple tests with different bandwidths

- avoids using arbitrary heuristics or data splitting

⇒ non-asymptotic power guarantees

MMD single test using permutations

Maximum Mean Discrepancy: $\text{MMD}_{\lambda}^2(\mathbf{p}, \mathbf{q})$ is equal to

$$\mathbb{E}_{X, X' \sim p}[k_{\lambda}(X, X')] - 2\mathbb{E}_{X \sim p, Y \sim q}[k_{\lambda}(X, Y)] + \mathbb{E}_{Y, Y' \sim q}[k_{\lambda}(Y, Y')]$$

Quadratic-time estimator: $\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m, \mathbb{Y}_n)$ defined as

$$\frac{1}{m(m-1)} \sum_{1 \leq i \neq i' \leq m} k_{\lambda}(X_i, X_{i'}) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k_{\lambda}(X_i, Y_j) + \frac{1}{n(n-1)} \sum_{1 \leq j \neq j' \leq n} k_{\lambda}(Y_j, Y_{j'})$$

Permutations: simulate \mathcal{H}_0 by repeating B times: permute with σ the elements of $\mathbb{X}_m \cup \mathbb{Y}_n$ to obtain \mathbb{X}_m^σ and \mathbb{Y}_n^σ and compute $\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m^\sigma, \mathbb{Y}_n^\sigma)$

Quantile: $\hat{q}_{1-\alpha}^{\lambda}$ is the $\lceil (B+1)(1-\alpha) \rceil$ -th largest value of the B permuted test statistics together with the original test statistic

Single test: well-calibrated non-asymptotic level α

$$\Delta_{\alpha}^{\lambda}(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m, \mathbb{Y}_n) > \hat{q}_{1-\alpha}^{\lambda}\right)$$

MMD single test using a wild bootstrap

Assumption: equal sample sizes $m = n$

Different quadratic-time estimator: $\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_n, \mathbb{Y}_n)$ defined as

$$\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} k_{\lambda}(\mathcal{X}_i, \mathcal{X}_j) - k_{\lambda}(\mathcal{X}_i, \mathcal{Y}_j) - k_{\lambda}(\mathcal{Y}_i, \mathcal{X}_j) + k_{\lambda}(\mathcal{Y}_i, \mathcal{Y}_j)$$

Wild bootstrap: simulate \mathcal{H}_0 by repeating B times: generate n Rademacher random variables $\epsilon_1, \dots, \epsilon_n$ in $\{-1, 1\}$ and compute

$$\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \epsilon_i \epsilon_j \left(k_{\lambda}(\mathcal{X}_i, \mathcal{X}_j) - k_{\lambda}(\mathcal{X}_i, \mathcal{Y}_j) - k_{\lambda}(\mathcal{Y}_i, \mathcal{X}_j) + k_{\lambda}(\mathcal{Y}_i, \mathcal{Y}_j) \right)$$

Quantile: $\widehat{q}_{1-\alpha}^{\lambda}$ is the $[(B+1)(1-\alpha)]$ -th largest value of the B bootstrapped test statistics together with the original test statistic

Single test: well-calibrated non-asymptotic level α

$$\Delta_{\alpha}(\mathbb{X}_n, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_n, \mathbb{Y}_n) > \widehat{q}_{1-\alpha}^{\lambda}\right)$$

MMD single test: uniform separation rate

Minimax rate over Sobolev balls: $(m + n)^{-2s/(4s+d)}$

Theorem

For $\alpha \in (0, e^{-1})$, $s > 0$, $R > 0$, $B \in \mathbb{N}$ large enough, using either permutations or a wild bootstrap, the test $\Delta_\alpha^{\lambda^*}$ with

$$\lambda_i^* = (m + n)^{-2/(4s+d)}, i = 1, \dots, d$$

is optimal in the minimax sense over the Sobolev ball $\mathcal{S}_d^s(R)$

$$\rho\left(\Delta_\alpha^{\lambda^*}, \mathcal{S}_d^s(R), \beta\right) \leq C(d, s, R, \alpha, \beta) (m + n)^{-2s/(4s+d)}.$$

⇒ single test $\Delta_\alpha^{\lambda^*}$ is **optimal** but **not adaptive** over Sobolev balls:

- **optimal:** uniform separation rate of $\Delta_\alpha^{\lambda^*}$ achieves the minimax rate
- **not adaptive:** λ^* depends on the unknown smoothness parameter s of $\mathcal{S}_d^s(R)$ (i.e. cannot be implemented)

MMDAgg: MMD aggregated test

1 MMDAgg: MMD Aggregated Two-Sample Test

- Two-sample problem
- MMD single test
- **MMD aggregated test**
- Experiments

2 KSDAgg: KSD Aggregated Goodness-of-fit Test

- Goodness-of-fit problem & KSD tests
- Uniform separation rate
- Experiments

MMDAgg: MMD Aggregated test

Aggregated test MMDAgg: Δ_α^Λ for some finite collection of bandwidths Λ

Reject null hypothesis $\mathcal{H}_0 : p = q$ if

$$\widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-u_\alpha w_\lambda}^\lambda \quad \text{for some } \lambda \in \Lambda$$

How to ensure calibrated non-asymptotic level α for MMDAgg?

Introduce:

- positive weights $(w_\lambda)_{\lambda \in \Lambda}$ satisfying $\sum_{\lambda \in \Lambda} w_\lambda \leq 1$ (chosen by the user)
- correction u_α defined as

$$u_\alpha = \sup \left\{ u > 0 : \mathbb{P}_{p \times p} \left(\max_{\lambda \in \Lambda} \left(\widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}_{1-u w_\lambda}^\lambda \right) > 0 \right) \leq \alpha \right\}$$

The supremum can be estimated using the bisection method.

The probability can be estimated using a Monte-Carlo approximation.

MMDAgg: time complexity

$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1} \left(\widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-u_\alpha w_\lambda}^\lambda \text{ for some } \lambda \in \Lambda \right)$$

$$u_\alpha = \sup \left\{ u > 0 : \mathbb{P}_{p \times p} \left(\max_{\lambda \in \Lambda} \left(\widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}_{1-u w_\lambda}^\lambda \right) > 0 \right) \leq \alpha \right\}$$

For each $\lambda \in \Lambda$:

- compute original MMD test statistic $\widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n)$
- compute B_1 simulated MMD test statistics
 - used to estimate the quantile $\widehat{q}_{1-u w_\lambda}^\lambda$
- compute B_2 simulated MMD test statistics
 - used to estimate the probability $\mathbb{P}_{p \times p}$ in definition of u_α

MMDAgg time complexity:

$$\mathcal{O}(|\Lambda| (B_1 + B_2)(m + n)^2)$$

MMDAgg: uniform separation rate

Minimax rate over Sobolev balls: $(m + n)^{-2s/(4s+d)}$

Theorem

For $\alpha \in (0, e^{-1})$, $s > 0$, $R > 0$, $B_1, B_2, B_3 \in \mathbb{N}$ large enough, using either permutations or a wild bootstrap, the aggregated MMDAgg test $\Delta_\alpha^{\Lambda^*}$ with

$$\Lambda^* := \left\{ 2^{-\ell} \mathbb{1}_d : \ell \in \left\{ 1, \dots, \left\lceil \frac{2}{d} \log_2 \left(\frac{m+n}{\ln(\ln(m+n))} \right) \right\rceil \right\} \right\}$$

and $w_{\Lambda^*} := \frac{6}{\pi^2 \ell^2}$, is (almost) **optimal** and **adaptive** over the Sobolev balls $\{\mathcal{S}_d^s(R) : s > 0, R > 0\}$

$$\rho \left(\Delta_\alpha^{\Lambda^*}, \mathcal{S}_d^s(R), \beta \right) \leq C(d, s, R, \alpha, \beta) \left(\frac{m+n}{\ln(\ln(m+n))} \right)^{-2s/(4s+d)}.$$

⇒ price to pay for **adaptivity** over $\{\mathcal{S}_d^s(R) : s > 0, R > 0\}$ is a $\ln(\ln(m+n))$ term in the **optimal** minimax rate

MMDAgg: summary of theoretical results

Minimax rate over Sobolev balls: $(m + n)^{-2s/(4s+d)}$

Single test: Δ_α^λ for a kernel bandwidth λ

- **optimal:** uniform separation rate of $\Delta_\alpha^{\lambda^*}$ is $(m + n)^{-2s/(4s+d)}$
- **not adaptive:** λ^* depends on unknown parameter s of $\mathcal{S}_d^s(R)$

MMDAgg test: Δ_α^Λ for a collection of kernel bandwidths $\Lambda = \{\lambda^{(j)}\}_{1 \leq j \leq N}$

- **(almost) optimal:** uniform separation rate of $\Delta_\alpha^{\Lambda^*}$ is

$$\left(\frac{m + n}{\ln(\ln(m + n))} \right)^{-2s/(4s+d)}$$

- **adaptive:** Λ^* is independent of unknown parameters s and R of $\mathcal{S}_d^s(R)$

MMDAgg: Experiments

1 MMDAgg: MMD Aggregated Two-Sample Test

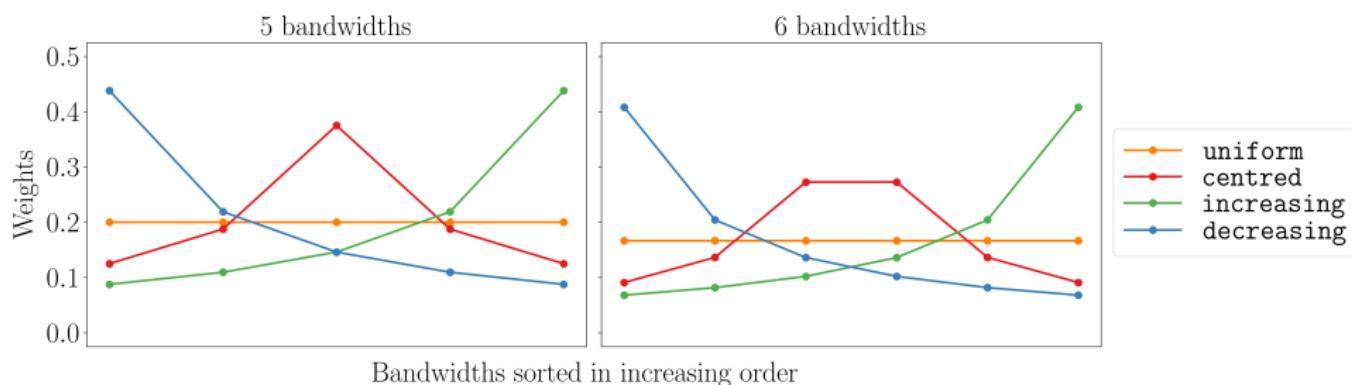
- Two-sample problem
- MMD single test
- MMD aggregated test
- Experiments

2 KSDAgg: KSD Aggregated Goodness-of-fit Test

- Goodness-of-fit problem & KSD tests
- Uniform separation rate
- Experiments

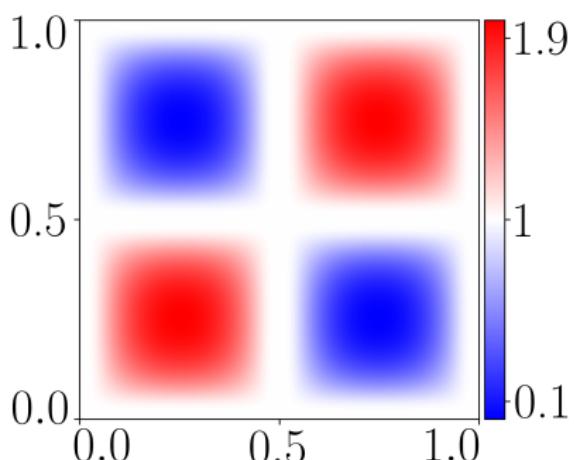
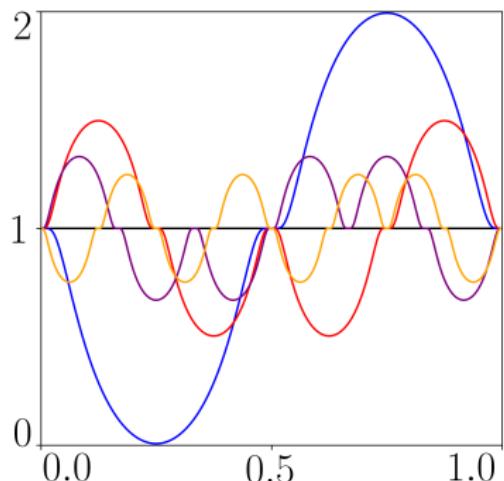
Weighting strategies

- **median bandwidth:** $(\lambda_{med})_i := \text{median}\{|z_i - z'_i| : z, z' \in \mathbb{X}_m \cup \mathbb{Y}_n, z \neq z'\}$
- **finite collection Λ of bandwidths:** for $\ell_-, \ell_+ \in \mathbb{N}$ with $N = 1 + \ell_- + \ell_+$
 $\Lambda(\ell_-, \ell_+) := \left\{ 2^\ell \lambda_{med} \in (0, \infty)^d : \ell \in \{\ell_-, \dots, \ell_+\} \right\} = \left\{ \lambda^{(j)} : j = 1, \dots, N \right\}$
- **uniform weights:** $w_{\lambda^{(j)}} := 1 / N$
- **increasing weights:** $w_{\lambda^{(j)}} := C / (N + 1 - j)$
- **decreasing weights:** $w_{\lambda^{(j)}} := C / j$
- **centred weights:** $w_{\lambda^{(j)}} := C / (|\frac{N+1}{2} - j| + 1)$ or $C / (|\frac{N+1}{2} - j| + \frac{1}{2})$



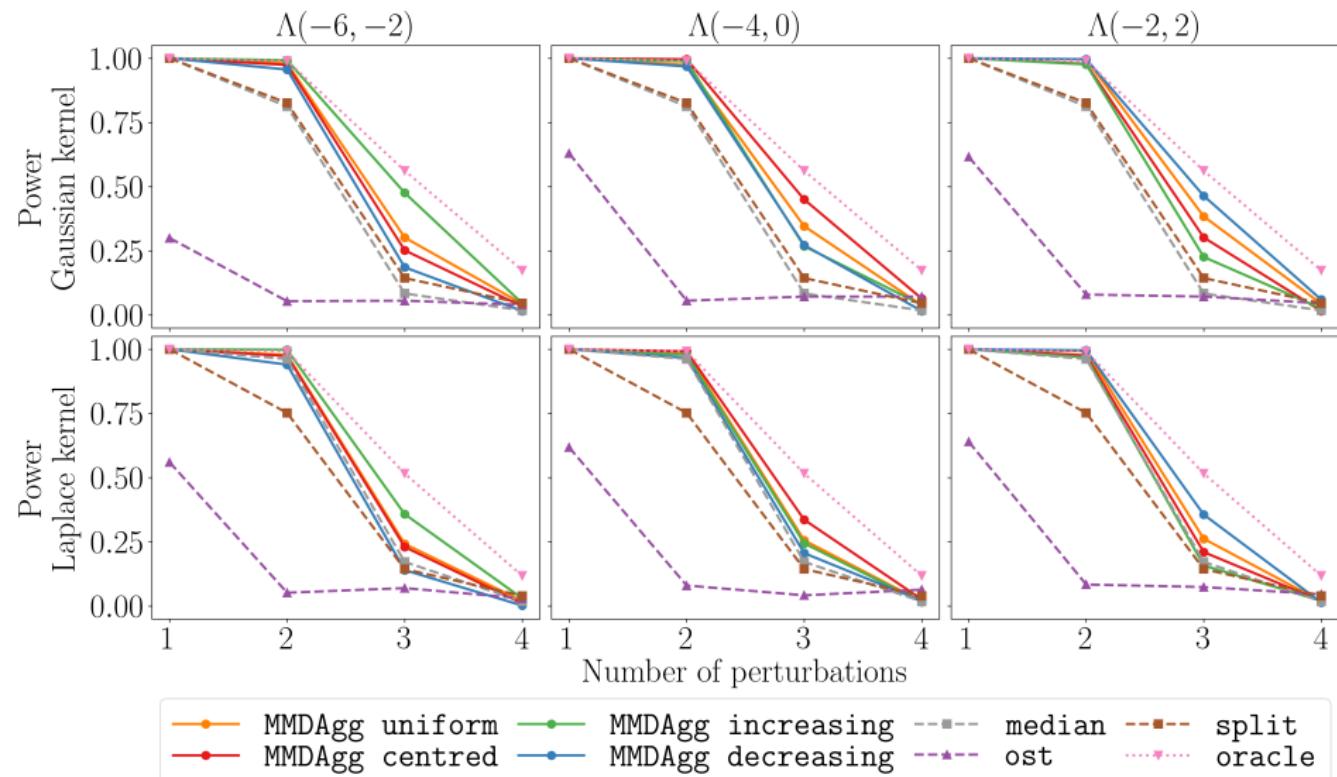
MMDAgg experiment: perturbed uniform distribution

Perturbed uniform densities: belong to Sobolev ball $\mathcal{S}_d^s(R)$



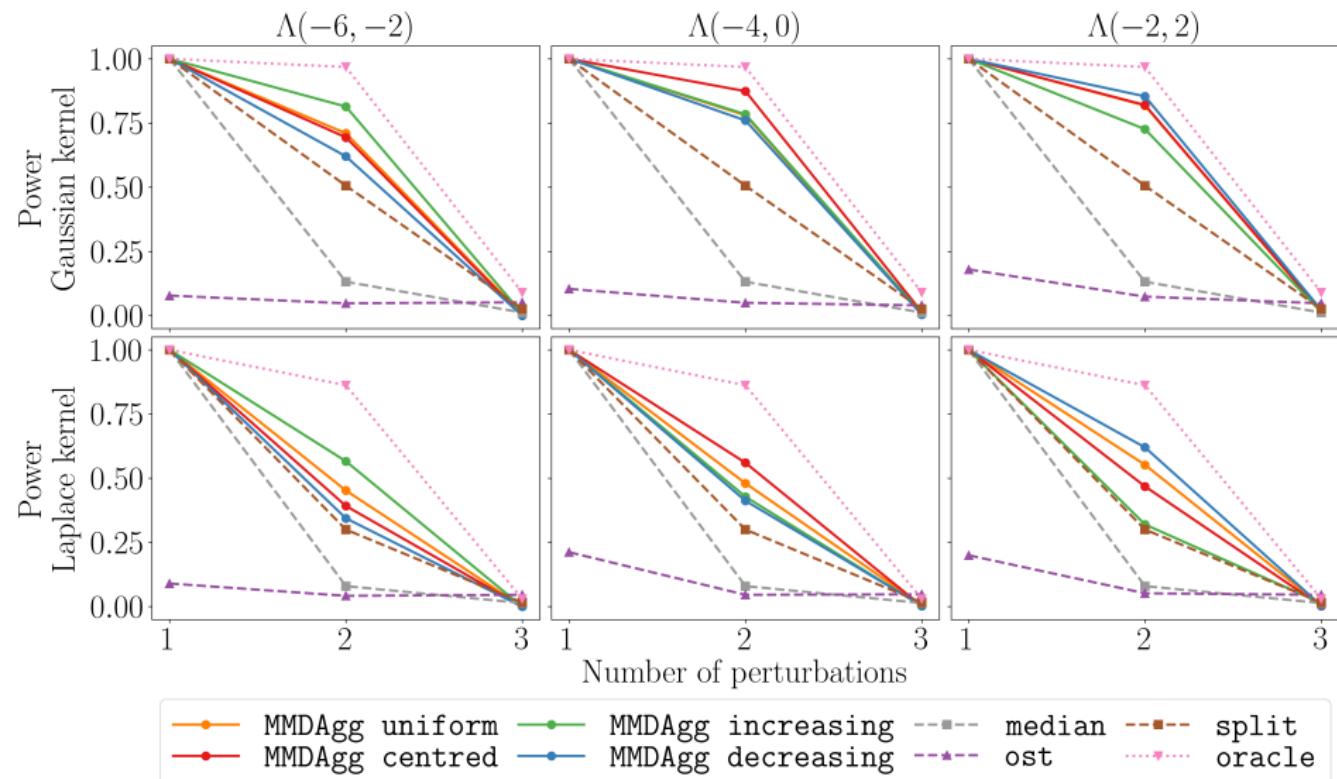
Perturbed uniform densities: used in the proof of lower bound on uniform separation rate over Sobolev balls $(m+n)^{-2s/(4s+d)}$ (**minimax rate**)

MMDAgg experiment: perturbed uniform 1d ($m=n=500$)



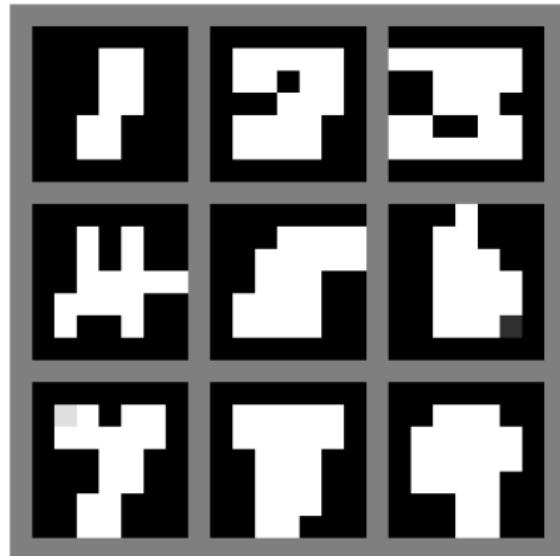
ost: Kübler et al. Learning kernel tests without data splitting. NeurIPS, 2020.

MMDAgg experiment: perturbed uniform 2d ($m=n=2000$)



MMDAgg experiment: MNIST

MNIST digits: images down-sampled to 7×7 resolution



$$\mathcal{P} : 0, 1, 2, 3, 4, 5, 6, 7, 8, 9$$

$$\mathcal{Q}_1 : 1, 3, 5, 7, 9$$

$$\mathcal{Q}_2 : 0, 1, 3, 5, 7, 9$$

$$\mathcal{Q}_3 : 0, 1, 2, 3, 5, 7, 9$$

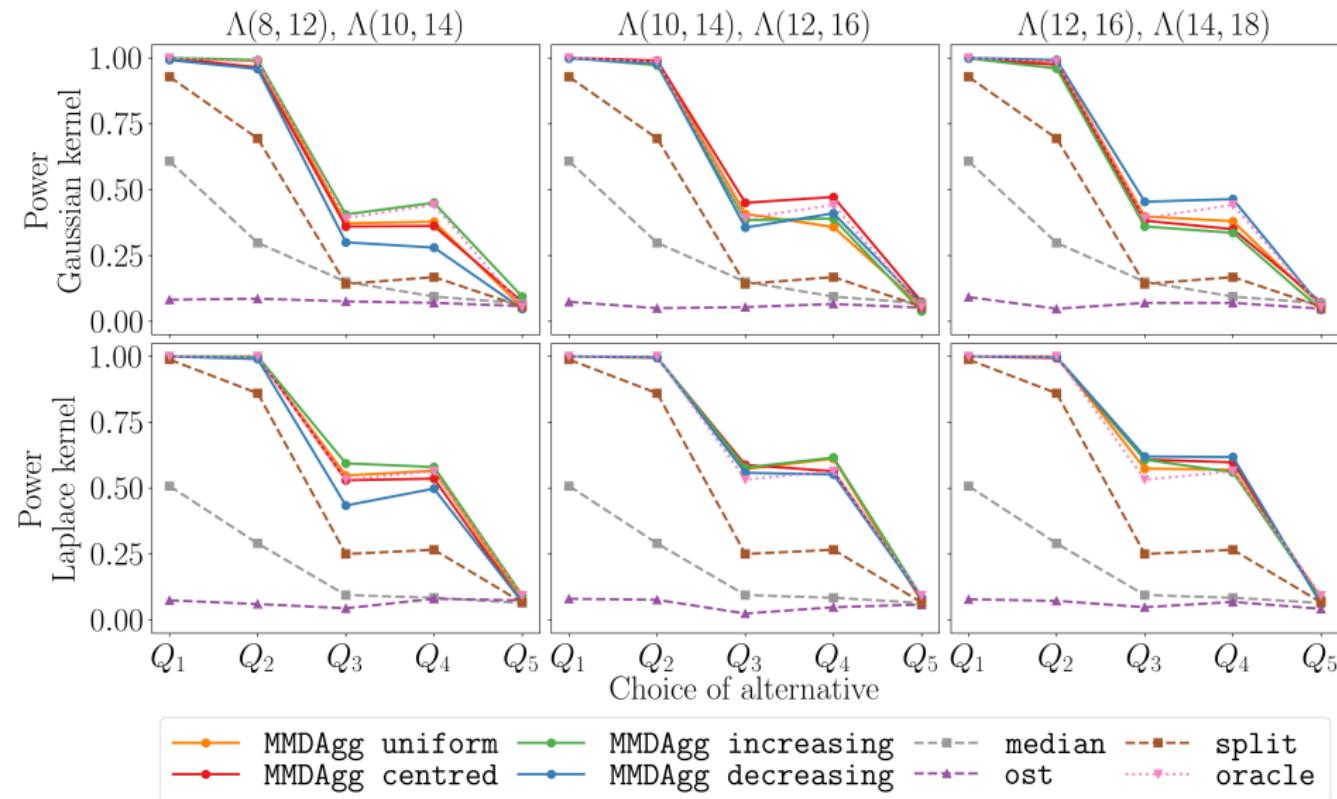
$$\mathcal{Q}_4 : 0, 1, 2, 3, 4, 5, 7, 9$$

$$\mathcal{Q}_5 : 0, 1, 2, 3, 4, 5, 6, 7, 9$$

Images: dimension 49

Problem: inherently lower-dimensional

MMDAgg experiment: MNIST ($m=n=500$)

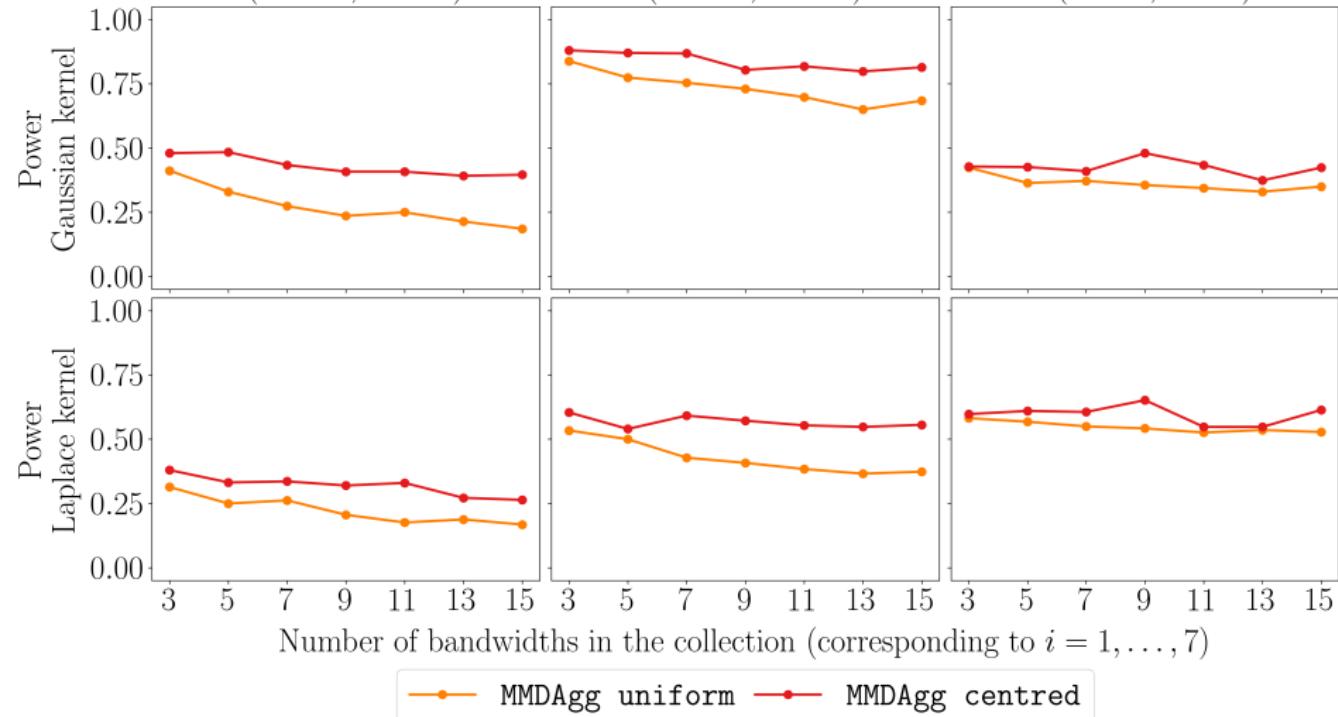


MMDAgg experiment: collections of bandwidths

$d = 1, 3$ perturbations
 $m = n = 500$
 $\Lambda(-2 - i, -2 + i)$

$d = 2, 2$ perturbations
 $m = n = 2000$
 $\Lambda(-2 - i, -2 + i)$

MNIST, Q_3 , $m = n = 500$
 $\Lambda(12 - i, 12 + i)$
 $\Lambda(14 - i, 14 + i)$



MMDAgg: conclusion

MMDAgg test:

- aggregate MMD tests with different kernel bandwidths (or kernels)
- no data splitting

MMDAgg theoretical results:

- optimal in the minimax sense (up to $\log(\log(m + n))$ term)
- adaptive test over Sobolev balls $\{S_d^s(R) : s > 0, R > 0\}$
- quantile estimation: wild bootstrap or permutations
- wide range of kernels

MMDAgg experimental results:

- outperforms state-of-the-art MMD adaptive tests

MMDAgg paper



MMDAgg code



KSD Aggregated Goodness-of-fit Test



Antonin
Schrab

†‡§



Benjamin
Guedj

†§



Arthur
Gretton

‡

† Centre for Artificial Intelligence, UCL

‡ Gatsby Computational Neuroscience Unit, UCL

§ Inria London Programme

KSDAgg: Goodness-of-fit problem & KSD tests

1 MMDAgg: MMD Aggregated Two-Sample Test

- Two-sample problem
- MMD single test
- MMD aggregated test
- Experiments

2 KSDAgg: KSD Aggregated Goodness-of-fit Test

- Goodness-of-fit problem & KSD tests
- Uniform separation rate
- Experiments

Goodness-of-fit problem

Given • a model probability density p on \mathbb{R}^d (or score function $\nabla \log p(z)$)
• samples $\mathbb{Z}_n := (Z_1, \dots, Z_n)$ where $Z_i \stackrel{\text{iid}}{\sim} q$ in \mathbb{R}^d
can we decide whether or not $p \neq q$ holds?

Stein kernel: not translation invariant

$$\begin{aligned} h_{p,\lambda}(x, y) &:= \left(\nabla \log p(x)^\top \nabla \log p(y) \right) k_\lambda(x, y) + \nabla \log p(y)^\top \nabla_1 k_\lambda(x, y) \\ &\quad + \nabla \log p(x)^\top \nabla_2 k_\lambda(x, y) + \sum_{i=1}^d \frac{\partial}{\partial x_i \partial y_i} k_\lambda(x, y) \end{aligned}$$

Stein identity: $\mathbb{E}_p[h_{p,\lambda}(Z, \cdot)] = 0$

Kernel Stein Discrepancy:

$$\text{KSD}_{p,\lambda}^2(q) := \text{MMD}_{h_{p,\lambda}}^2(p, q) = \mathbb{E}_{q,q}[h_{p,\lambda}(Z, Z')]$$

Quadratic-time estimator:

$$\widehat{\text{KSD}}_{p,\lambda}^2(\mathbb{Z}_n) := \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_{p,\lambda}(Z_i, Z_j)$$

Chwialkowski et al. A kernel test of goodness of fit. In ICML, 2016.

Liu et al. A kernelized Stein discrepancy for goodness-of-fit tests. In ICML, 2016.

KSDAgg: KSD Aggregated test

Wild bootstrap: well-calibrated asymptotic level α

$$\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \epsilon_i \epsilon_j h_{p,\lambda}(Z_i, Z_j) \quad \text{for samples } \epsilon_i \sim \text{Unif}\{-1, 1\}$$

Parametric bootstrap: well-calibrated non-asymptotic level α

$$\frac{1}{\tilde{n}(\tilde{n}-1)} \sum_{1 \leq i \neq j \leq \tilde{n}} h_{p,\lambda}(\tilde{Z}_i, \tilde{Z}_j) \quad \text{for samples } \tilde{Z}_i \stackrel{\text{iid}}{\sim} p$$

Quantile $\hat{q}_{1-\alpha}^\lambda$: $[B(1-\alpha)]$ -th largest of the B bootstrapped test statistics

Single test $\Delta_{\alpha,p}^\lambda$: reject null hypothesis $\mathcal{H}_0 : p = q$ if

$$\widehat{\text{KSD}}_{p,\lambda}^2(Z_n) > \hat{q}_{1-\alpha}^\lambda$$

Aggregated test KSDAgg $\Delta_{\alpha,p}^\Lambda$: reject null hypothesis $\mathcal{H}_0 : p = q$ if

$$\widehat{\text{KSD}}_{p,\lambda}^2(Z_n) > \hat{q}_{1-u_\alpha w_\lambda}^\lambda \quad \text{for some } \lambda \in \Lambda$$

- weights $\sum_{\lambda \in \Lambda} w_\lambda \leq 1$
- level correction u_α

KSDAgg: Uniform separation rate

1 MMDAgg: MMD Aggregated Two-Sample Test

- Two-sample problem
- MMD single test
- MMD aggregated test
- Experiments

2 KSDAgg: KSD Aggregated Goodness-of-fit Test

- Goodness-of-fit problem & KSD tests
- **Uniform separation rate**
- Experiments

KSDAgg: uniform separation rate

Integral transform: $(\kappa \diamond f)(y) := \int_{\mathbb{R}^d} \kappa(x, y) f(x) dx$

Kernel assumption: $A_\lambda := \mathbb{E}_{q,q}[h_{p,\lambda}(Z, Z')^2] < \infty$

Theorem

For $\alpha \in (0, e^{-1})$, $\beta \in (0, 1)$, $B_1, B_2, B_3 \in \mathbb{N}$ large enough, using either a wild bootstrap or a parametric bootstrap, the condition

$$\|\mathbf{p} - \mathbf{q}\|_2^2 \geq \min_{\lambda \in \Lambda} \left(\|(\mathbf{p} - \mathbf{q}) - h_{p,\lambda} \diamond (\mathbf{p} - \mathbf{q})\|_2^2 + C \log\left(\frac{1}{\alpha w_\lambda}\right) \frac{\sqrt{A_\lambda}}{\beta n} \right)$$

guarantees β -control over the probability of type II error of KSDAgg

$$\mathbb{P}_{\mathbf{q}}\left(\Delta_{\alpha,p}^\lambda(\mathbb{Z}_n) = 0\right) \leq \beta.$$

Stein kernel is **not** translation invariant:

- ⇒ cannot work in Fourier domain using Plancherel's Theorem
- ⇒ cannot obtain uniform separation rate over Sobolev balls

KSDAgg: Experiments

1 MMDAgg: MMD Aggregated Two-Sample Test

- Two-sample problem
- MMD single test
- MMD aggregated test
- Experiments

2 KSDAgg: KSD Aggregated Goodness-of-fit Test

- Goodness-of-fit problem & KSD tests
- Uniform separation rate
- Experiments

Gaussian-Bernoulli Restricted Boltzmann Machine GBRBM

- Graphical model:**
- binary hidden variable $h \in \{-1, 1\}^{40}$
 - continuous observable variable $z \in \mathbb{R}^{50}$

Joint probability density:

$$p(z, h) = \frac{1}{C} \exp\left(\frac{1}{2} z^\top B h + b^\top z + c^\top h - \frac{1}{2} \|z\|_2^2\right)$$

initialisation: $b \sim \mathcal{N}(0, I_{50})$, $c \sim \mathcal{N}(0, I_{40})$, $B \sim \text{Unif } \{-1, 1\}^{50 \times 40}$

Probability density: intractable

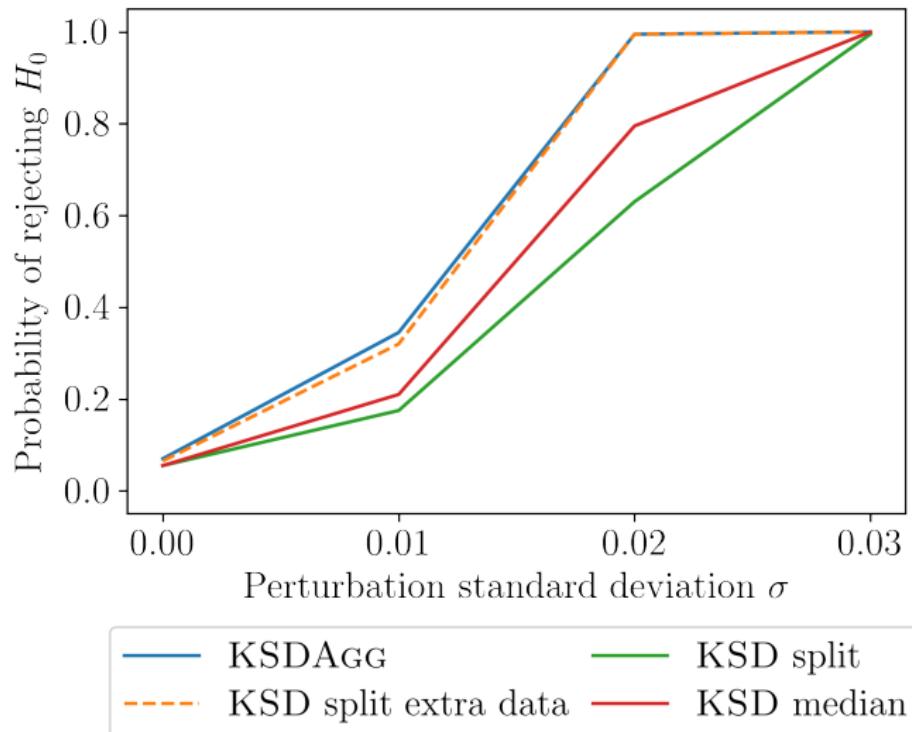
$$\textcolor{blue}{p}(z) = \sum_{h \in \{-1, 1\}^{40}} p(z, h).$$

Score function: closed-form

$$\nabla \log \textcolor{blue}{p}(z) = b - z + B \frac{\exp(2(B^\top z + c)) - 1}{\exp(2(B^\top z + c)) + 1}$$

Experiment: model p and **1000 samples** obtained using a Gibbs sampler for the GBRBM with Gaussian noise $\mathcal{N}(0, \sigma)$ injected into each entry of B

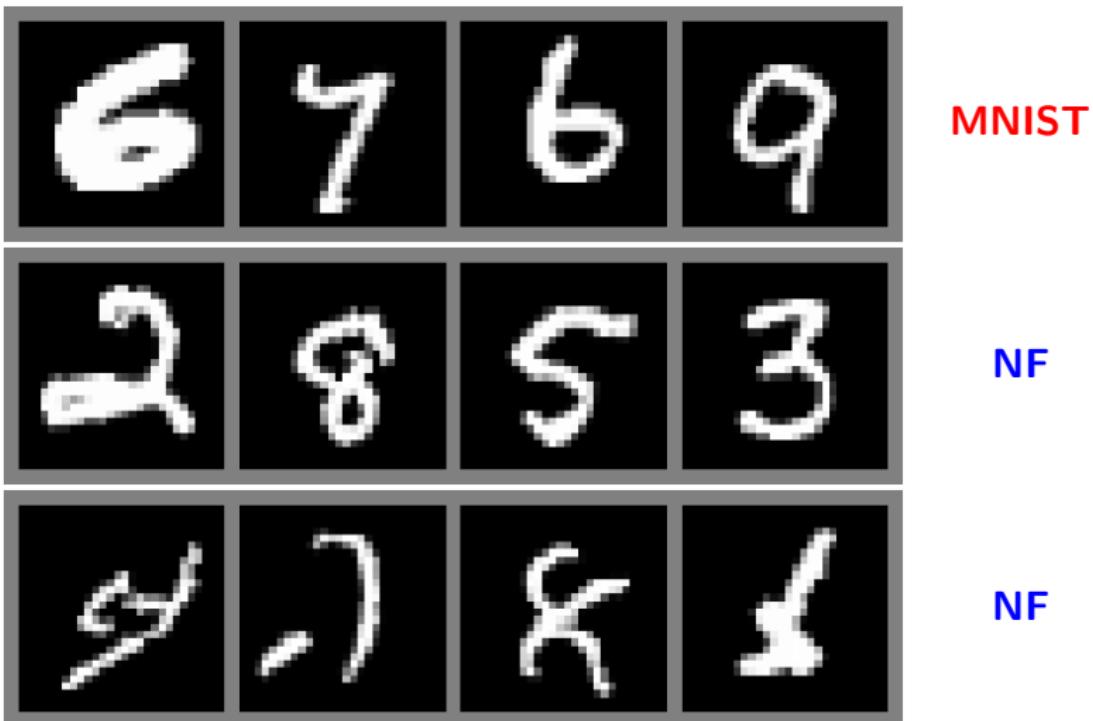
KSDAgg experiment: GBRBM $\Lambda(-20, 0)$



KSDAgg experiment: MNIST Normalizing Flow $\Lambda(-20, 0)$

Model: Normalizing Flow (generative model) with probability density p

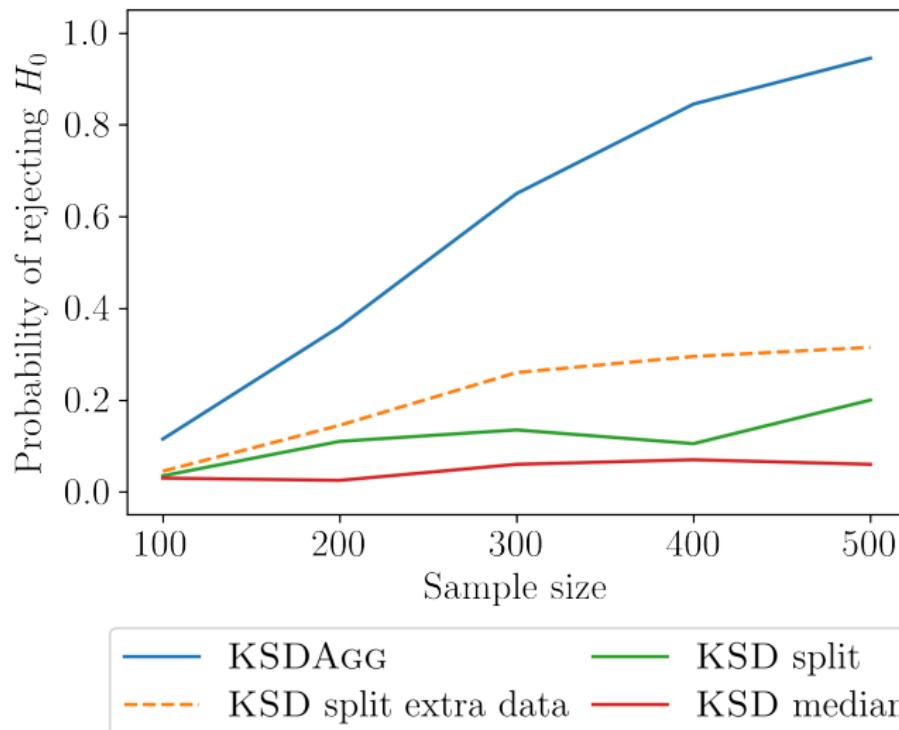
Samples: true MNIST samples in dimension $28^2 = 784$



KSDAgg experiment: MNIST Normalizing Flow $\Lambda(-20, 0)$

Model: Normalizing Flow (generative model) with probability density p

Samples: true MNIST samples in dimension $28^2 = 784$



KSDAgg: conclusion

KSDAgg test:

- aggregate KSD tests with different kernel bandwidths (or kernels)
- no data splitting

KSDAgg theoretical results:

- uniform separation rate upper bound
- quantile estimation: wild bootstrap or parametric bootstrap
- wide range of kernels

KSDAgg experimental results:

- outperforms state-of-the-art KSD adaptive tests

KSDAgg paper



KSDAgg code



What about HSICAgg?

Independence problem:

Given paired samples $((X_1, Y_1), \dots, (X_n, Y_n))$ in $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ with

- joint probability density r
- marginal probability densities p and q

can we decide whether or not $p \otimes q \neq r$ holds?

Hilbert-Schmidt Independence Criterion:

$$\begin{aligned}\text{HSIC}_{k,\ell}(r) &:= \text{MMD}_\kappa(p \otimes q, r) \\ \kappa((X, Y), (X', Y')) &:= k(X, X')\ell(Y, Y')\end{aligned}$$

ADAPTIVE TEST OF INDEPENDENCE BASED ON HSIC MEASURES.

Mélisande Albert^{*,1}, Béatrice Laurent^{†,1}, Amandine Marrel^{‡,2}, and Anouar Meynaoui^{§,1,2}

¹Institut de Mathématiques de Toulouse ; UMR5219, Université de Toulouse ; CNRS, INSA, F-31077
Toulouse, France.

²CEA, DEN, DER, F-13108 Saint-Paul-lez-Durance, France.

References

Aggregated tests:

- M. Fromont, B. Laurent, and P. Reynaud-Bouret. **The two-sample problem for Poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach.** *The Annals of Statistics*, 2013.
- M. Albert, B. Laurent, A. Marrel, and A. Meynaoui. **Adaptive test of independence based on HSIC measures.** *To appear in The Annals of Statistics*, 2019.
- I. Kim, S. Balakrishnan, and L. Wasserman. **Minimax optimality of permutation tests.** *To appear in The Annals of Statistics*, 2020.

References

Tests used for comparison in our experiments:

- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. **A kernel two-sample test.** *In JMLR*, 2012.
- A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. pontil, K. Fukumizu, and B. K. Sriperumbudur. **Optimal kernel choice for large-scale two-sample tests.** *In NeurIPS*, 2012.
- F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. **Learning deep kernels for non-parametric two-sample tests.** *In ICML*, 2020.
- J. M. Kübler, W. Jitkrittum, B. Schölkopf, and K. Muandet. **Learning kernel tests without data splitting.** *In NeurIPS*, 2020.
- Q. Liu, J. Lee, and M. Jordan. **A kernelized Stein discrepancy for goodness-of-fit tests.** *In ICML*, 2016.
- K. Chwialkowski, H. Strathmann, and A. Gretton. **A kernel test of goodness of fit.** *In ICML*, 2016.

Thank you for your attention!

Any questions?

MMDAgg



paper

KSDAgg



paper



code



code