Aggregated Kernel Tests

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

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

Aggregated Kernel Tests

1. MMDAgg: MMD Aggregated Two-Sample Test
2. KSDAgg: KSD Aggregated Goodness-of-fit Test
3. AggInc: Efficient Aggregated Kernel Tests using Incomplete U-statistics

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

Two-sample problem

- samples $\mathbb{X}_m := (X_1, \ldots, X_m)$, $X_i \overset{\text{iid}}{\sim} p$ in $\mathbb{R}^d$
- samples $\mathbb{Y}_n := (Y_1, \ldots, Y_n)$, $Y_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$

$$\mathcal{H}_0 \colon p = q \quad \text{against} \quad \mathcal{H}_a \colon p \neq q$$
$$\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1 \iff \text{reject } \mathcal{H}_0$$
$$\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0 \iff \text{fail to reject } \mathcal{H}_0$$

Type I error: controlled by α by design

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

Type II error: find a condition on $\|p - q\|_2$ to control by β

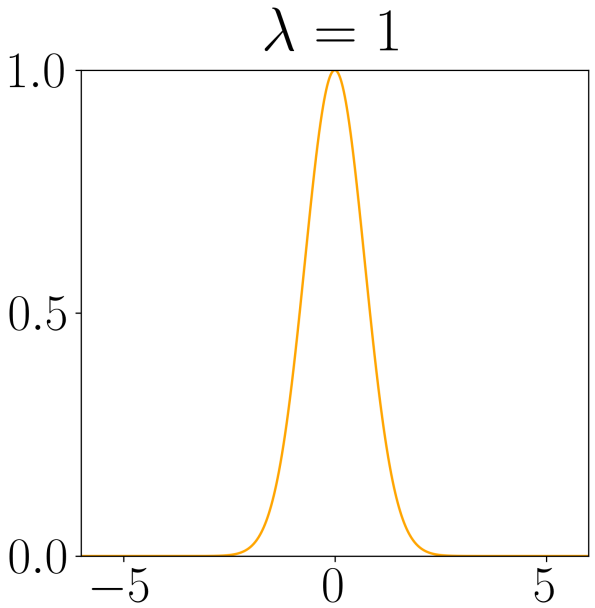
$$\mathbb{P}_{p \times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta$$

Kernels

Kernel: $k_\lambda(x, y) := \prod_{i=1}^{d} \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$ **Bandwidth:** $\lambda \in (0, \infty)^d$

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

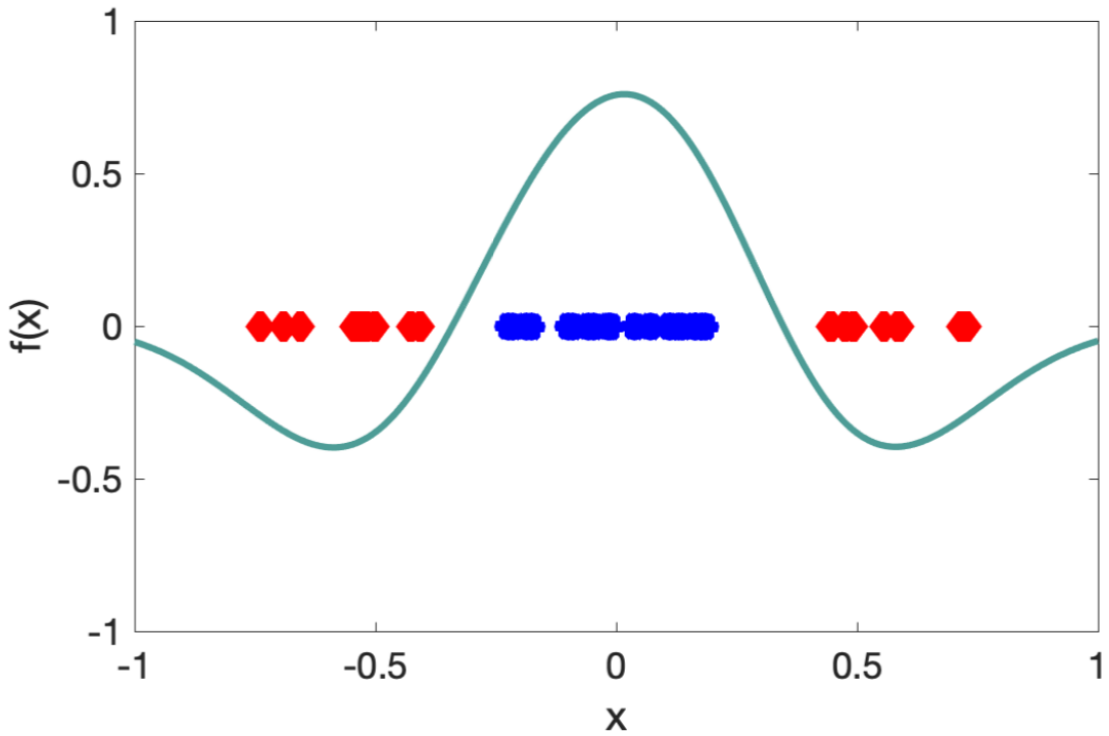
$$k_\lambda(x, y) := \frac{1}{\pi^{d/2} \lambda_1 \cdots \lambda_d} \exp\left(-\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{{\lambda_i}^2}\right)$$



Two-sample test using the Maximum Mean Discrepancy

Kernel: $k_\lambda(x, y) := \prod_{i=1}^{d} \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$ **Bandwidth:** $\lambda \in (0, \infty)^d$

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



$p \neq q$

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 $\mathcal{H}_a$

$\Longrightarrow$ **Bandwidths** should decrease as the **sample sizes** increase

- Choice of **bandwidth** is **crucial** for test power!
- **Bandwidth** selection methods: **median heuristic** & **data splitting**
- **Our method:** aggregate multiple tests with different **bandwidths**

Maximum Mean Discrepancy estimator

$$\begin{aligned}\mathrm{MMD}^2_\lambda(p, q) &:= \mathbb{E}_{p,p}[k_\lambda(X, X')] \\ &\quad - 2\,\mathbb{E}_{p,q}[k_\lambda(X, Y)] \\ &\quad + \mathbb{E}_{q,q}[k_\lambda(Y, Y')]\end{aligned}$$

$$\begin{aligned}\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m, \mathbb{Y}_n) &:= \frac{1}{m(m-1)} \sum_{1 \le i \ne i' \le m} k_\lambda(X_i, X_{i'}) \\ &\quad - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k_\lambda(X_i, Y_j) \\ &\quad + \frac{1}{n(n-1)} \sum_{1 \le j \ne j' \le n} k_\lambda(Y_j, Y_{j'})\end{aligned}$$

MMD test for a fixed bandwidth λ

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

Quantile: $\widehat{q}_{1-\alpha}^\lambda$ is the $\lceil (B+1)(1-\alpha) \rceil$-th largest value of $\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n)$ and B $\mathcal{H}_0$-simulated test statistics

Permutations: $\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m^\sigma, \mathbb{Y}_n^\sigma)$ where $(\mathbb{X}_m^\sigma, \mathbb{Y}_n^\sigma) = \sigma(\mathbb{X}_m \cup \mathbb{Y}_n)$

Wild bootstrap: case $m = n$, $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathrm{Unif}\{-1, 1\}$ (Rademacher)

$$\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \epsilon_i \epsilon_j \left(k_\lambda(X_i, X_j) - k_\lambda(X_i, Y_j) - k_\lambda(Y_i, X_j) + k_\lambda(Y_i, Y_j) \right)$$

Non-asymptotic level: $\mathbb{P}_{p \times p}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1) \ \le \ \alpha$

Time complexity: $\mathcal{O}\left(B\,(m+n)^2\right)$

MMDAgg for a collection of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

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

- time complexity $\mathcal{O}\left(|\Lambda|\, B_1\, (m+n)^2\right)$

MMDAgg: non-asymptotic level α

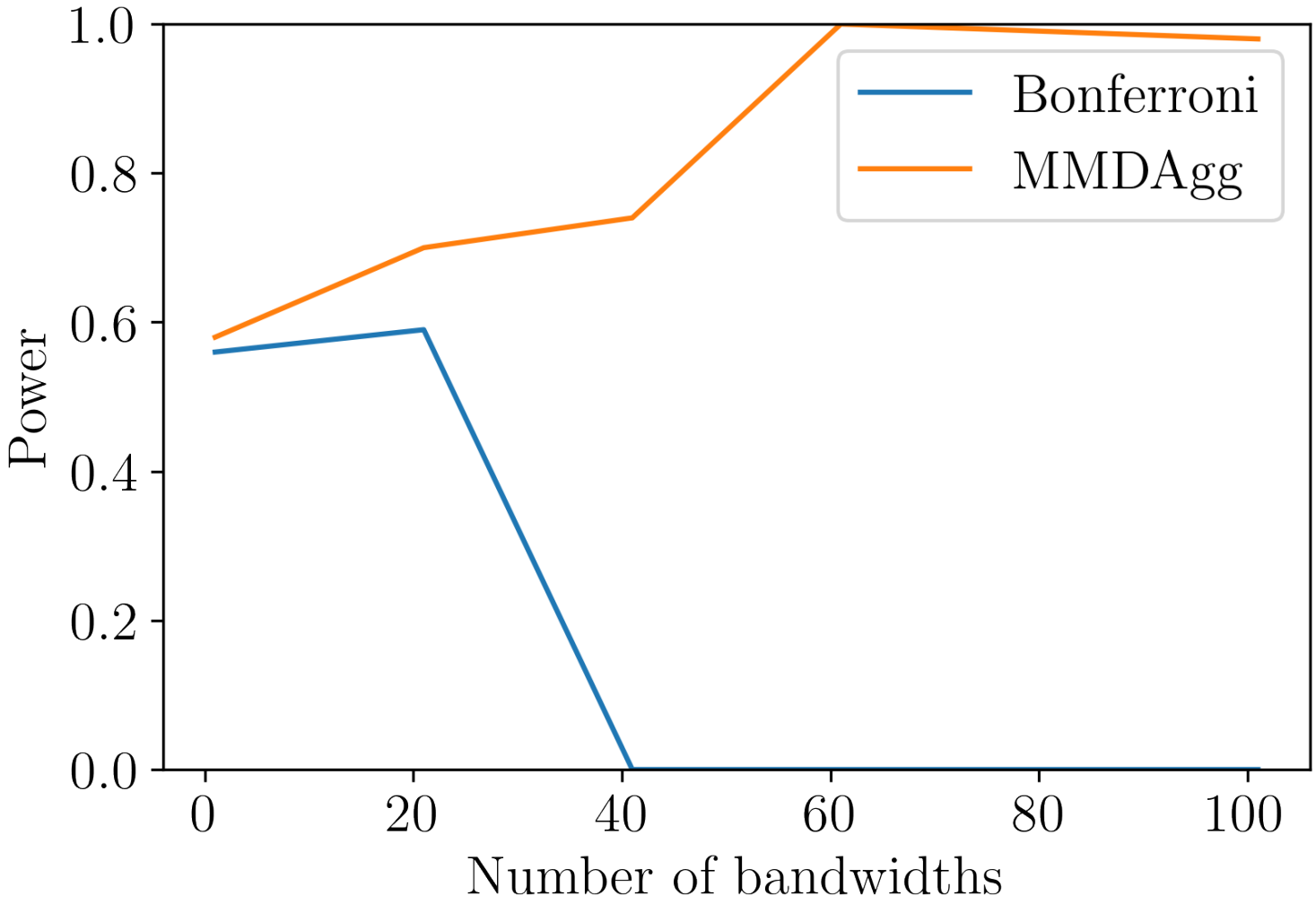
$$\Delta_\alpha^\Lambda(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}^\lambda_{1-u_\alpha w_\lambda} \text{ for some } \lambda \in \Lambda\right)$$

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

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

- more powerful than Bonferroni correction as $u_\alpha \geq \alpha$
- time complexity $\mathcal{O}\left(|\Lambda|\,(B_1+B_2)\,(m+n)^2\right)$

Multiple testing correction comparison



$$\Lambda(\ell_-, \ell_+) := \left\{2^\ell \lambda_{med} : \ell \in \{\ell_-, \ldots, \ell_+\}\right\} \qquad w_\lambda := 1 / |\Lambda|$$

$$\Lambda(-i, i), \quad i \in \{0, 10, 20, 30, 40, 50\}$$

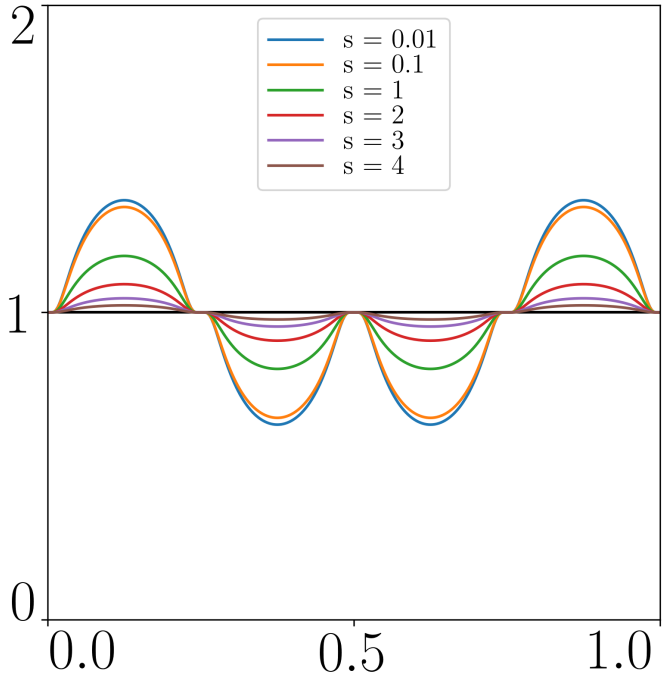
Sobolev balls

Regularity/smoothness assumption: $p - q \in \mathcal{S}_d^s(R)$

Sobolev balls:

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

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



Minimax adaptivity over Sobolev balls

Theorem

$$\Lambda^* := \left\{ 2^{-\ell} \mathbb{1}_d : \ell \in \left\{ 1, \dots, \left\lceil \frac{2}{d} \log_2 \left(\frac{m+n}{\ln(\ln(m+n))} \right) \right\rceil \right\} \right\}, \quad w_\lambda := \frac{6}{\pi^2 \ell^2}$$

Assuming $p - q \in \mathcal{S}_d^s(R)$, *the condition*

$$\|p - q\|_2 \geq C \left(\frac{m+n}{\ln(\ln(m+n))} \right)^{-2s/(4s+d)}$$

guarantees control over the probability of type II error of MMDAgg

$$\mathbb{P}_{p \times q} \left(\Delta_\alpha^{\Lambda^*}(\mathbb{X}_m, \mathbb{Y}_n) = 0 \right) \leq \beta.$$

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

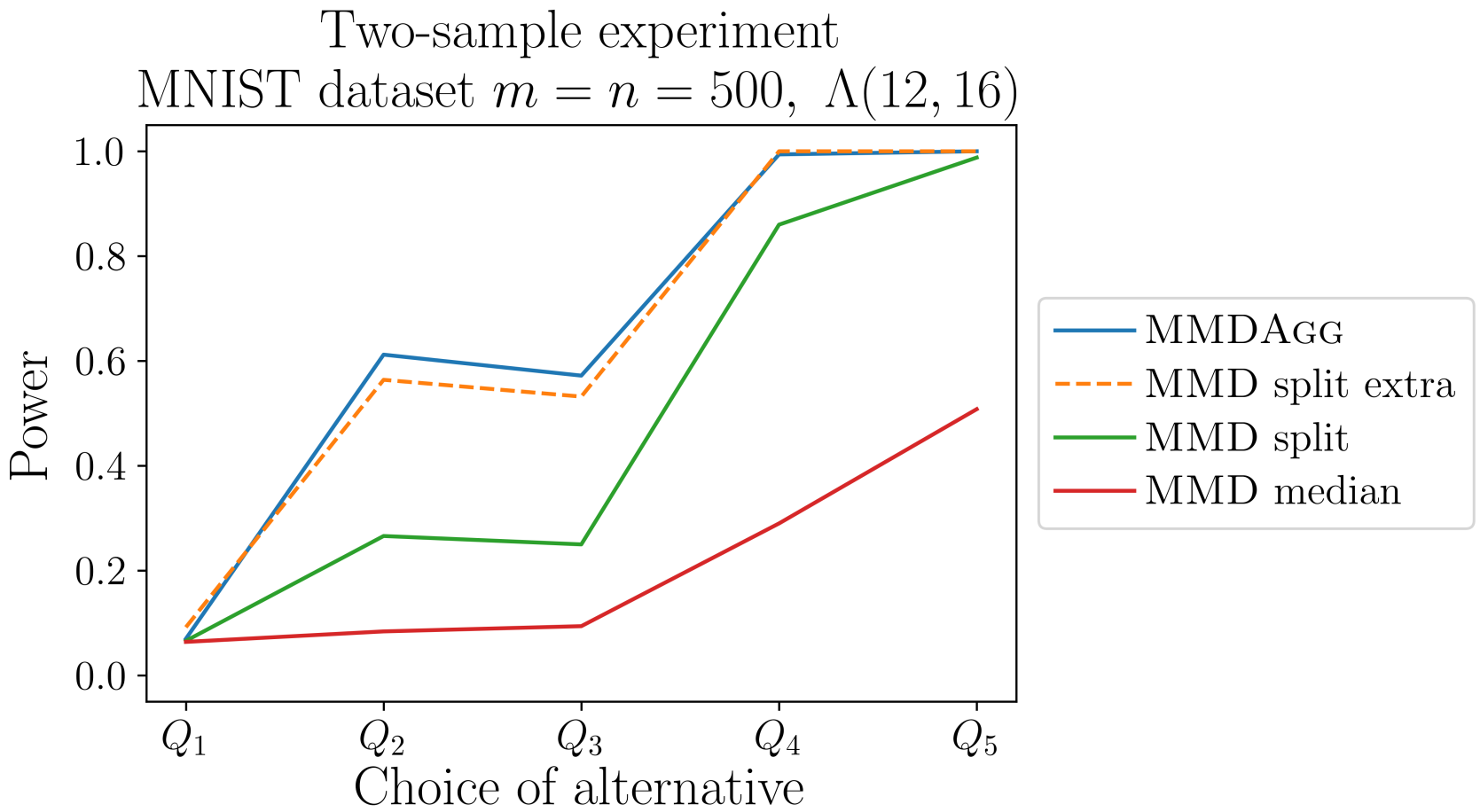
Adaptive over $\{\mathcal{S}_d^s(R) : s > 0, R > 0\}$

MMDAgg Experiment

$$\Lambda(\ell_-, \ell_+) := \left\{2^\ell \lambda_{med} : \ell \in \{\ell_-, \dots, \ell_+\}\right\} \qquad w_\lambda := 1 / |\Lambda|$$

$$\mathcal{P} := \{0, \dots, 9\} \quad \mathcal{Q}_2 := \mathcal{P} \setminus \{8, 6\} \quad \mathcal{Q}_4 := \mathcal{P} \setminus \{8, 6, 4, 2\}$$

$$\mathcal{Q}_1 := \mathcal{P} \setminus \{8\} \quad \mathcal{Q}_3 := \mathcal{P} \setminus \{8, 6, 4\} \quad \mathcal{Q}_5 := \mathcal{P} \setminus \{8, 6, 4, 2, 0\}$$



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

Goodness-of-fit problem & Kernel Stein Discrepancy

- model with probability density p or score function $\nabla \log p(z)$ on $\mathbb{R}^d$
- samples $\mathbb{Z}_n := (Z_1, \dots, Z_n)$, $Z_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$

$$\mathcal{H}_0 \colon p = q \qquad \text{against} \qquad \mathcal{H}_a \colon p \neq q$$

Stein kernel: $h_{p,\lambda}(x, y)$ defined as

$$\left(\nabla \log p(x)^\top \nabla \log p(y)\right) k_\lambda(x, y) + \nabla \log p(y)^\top \nabla_1 k_\lambda(x, y)$$
$$+ \nabla \log p(x)^\top \nabla_2 k_\lambda(x, y) + \sum_{1 \le i \le d} \frac{\partial}{\partial x_i \, \partial y_i} k_\lambda(x, y)$$

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

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

Estimator: $\widehat{\mathrm{KSD}}^2_{p,\lambda}(\mathbb{Z}_n) := \dfrac{1}{n(n-1)} \displaystyle\sum_{1 \le i \neq j \le n} h_{p,\lambda}(Z_i, Z_j)$

KSDAgg: KSD Aggregated test

Wild bootstrap: asymptotic level α

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

Parametric bootstrap: non-asymptotic level α

$$\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_{p,\lambda}(\widetilde{Z}_i, \widetilde{Z}_j) \qquad \text{where} \qquad \widetilde{Z}_i \overset{\text{iid}}{\sim} p$$

KSDAgg:

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

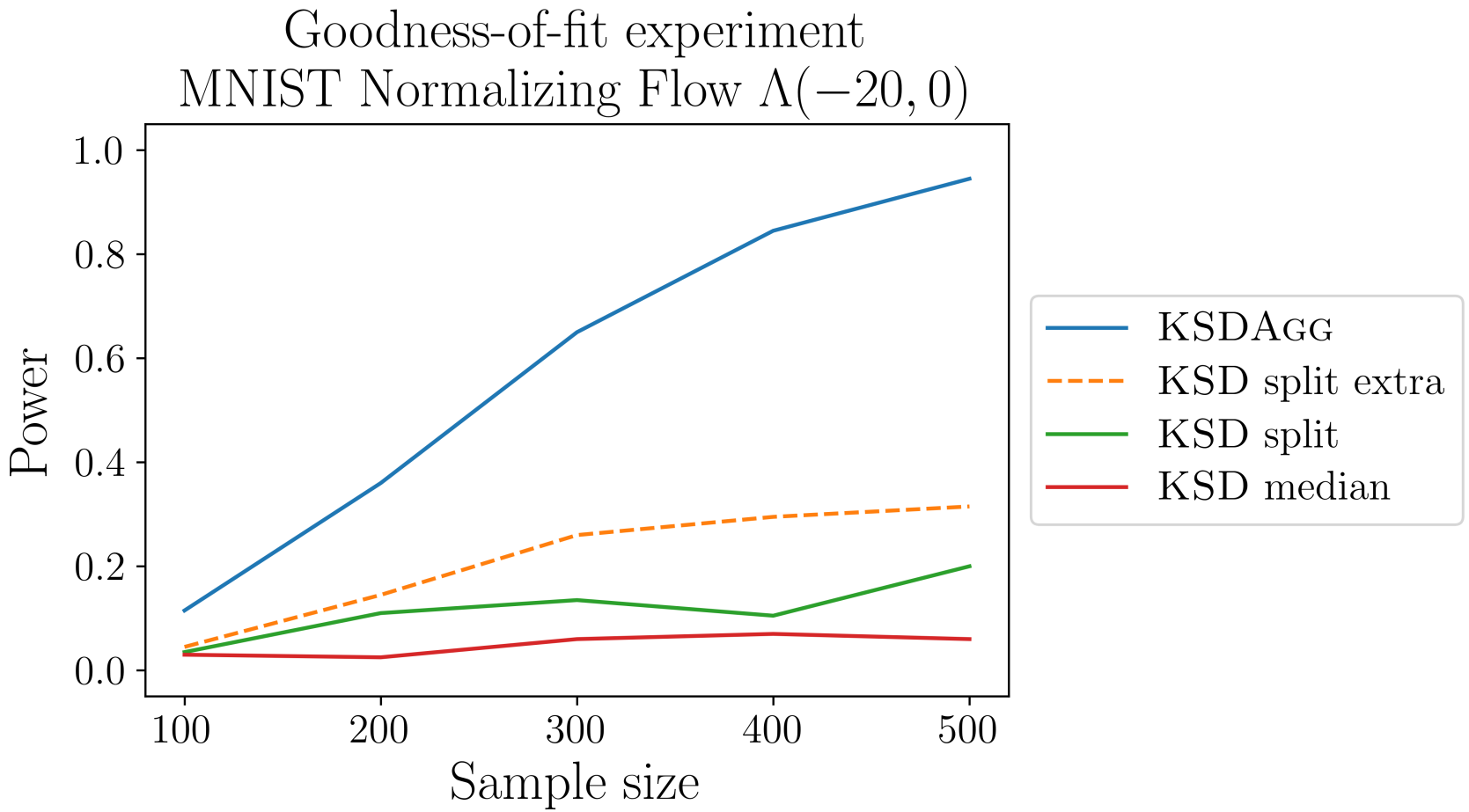
Time complexity: $\mathcal{O}\left(|\Lambda|\,(B_1 + B_2)\,n^2\right)$

KSDAgg Experiment

$$\Lambda(\ell_-, \ell_+) := \left\{2^\ell \lambda_{med} : \ell \in \{\ell_-, \ldots, \ell_+\}\right\} \qquad w_\lambda := 1 / |\Lambda|$$

model: Normalizing Flow density

samples: true MNIST digits



What about HSICAgg?

Independence problem:

Given paired samples $\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$ 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:

$$\mathrm{HSIC}_{k,\ell}(r) := \mathrm{MMD}_{\kappa}(p \otimes q, r)$$
$$\kappa\big((X, Y), (X', Y')\big) := k(X, X')\ell(Y, Y')$$

ADAPTIVE TEST OF INDEPENDENCE BASED ON HSIC MEASURES.

Mélisande Albert[*,1], Béatrice Laurent[†,1], Amandine Marrel[‡,2], and Anouar Meynaoui[§,1,2]

[1]Institut de Mathématiques de Toulouse ; UMR5219, Université de Toulouse ; CNRS, INSA, F-31077 Toulouse, France.
[2]CEA, DEN, DER, F-13108 Saint-Paul-lez-Durance, France.

Efficient Aggregated Kernel Tests using Incomplete U-statistics




Antonin Schrab †‡§

Ilmun Kim *

Benjamin Guedj †§

Arthur Gretton ‡

† Centre for Artificial Intelligence, UCL
‡ Gatsby Computational Neuroscience Unit, UCL
§ Inria London Programme
* Department of Statistics & Data Science, Yonsei University

AggInc

- **Complete U-statistic:**

$$\frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} h(Z_i, Z_j)$$

- **Quadratic-time MMDAgg, KSDAgg & HSICAgg:**

$$\mathcal{O}\left(|\Lambda| \left(B_1 + B_2\right) N^2\right)$$

- **Incomplete U-statistic:**

$$\frac{1}{|\mathcal{D}|} \sum_{(i,j) \in \mathcal{D}} h(Z_i, Z_j) \qquad \mathcal{D} \subseteq \left\{(i,j) : 1 \leq i \neq j \leq N\right\}$$

- **Efficient MMDAggInc, KSDAggInc & HSICAggInc:**

$$\mathcal{O}\left(|\Lambda| \left(B_1 + B_2\right) |\mathcal{D}|\right)$$

- **Linear-time** if $|\mathcal{D}| = cN$

AggInc: uniform separation rates

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

MMDAgg & HSICAgg rates: $\left(\frac{N}{\ln(\ln(N))} \right)^{-2s/(4s+d)}$

MMDAggInc & HSICAggInc rates: $\left(\frac{|\mathcal{D}|/N}{\ln(\ln(|\mathcal{D}|/N))} \right)^{-2s/(4s+d)}$

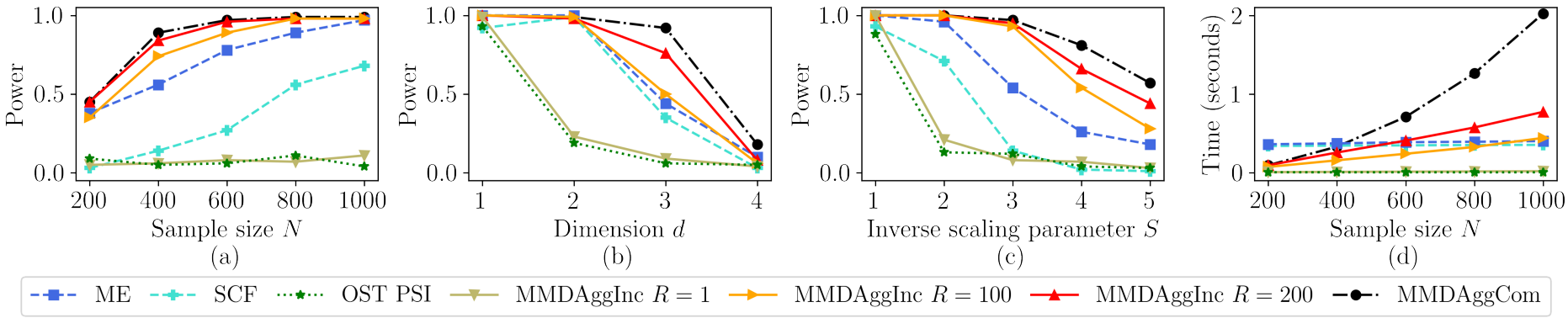
- $|\mathcal{D}| \asymp N^2$: recover the **Agg** rate
- $N \lesssim |\mathcal{D}| \lesssim N^2$: cost $|\mathcal{D}|/N^2$ incurred in the **Agg** rate

Trade-off: computational efficiency / rate of convergence

- $|\mathcal{D}| \lesssim N$: no guarantee that the **AggInc** rate converges to 0

MMDAggInc: experiments

Perturbed uniform densities

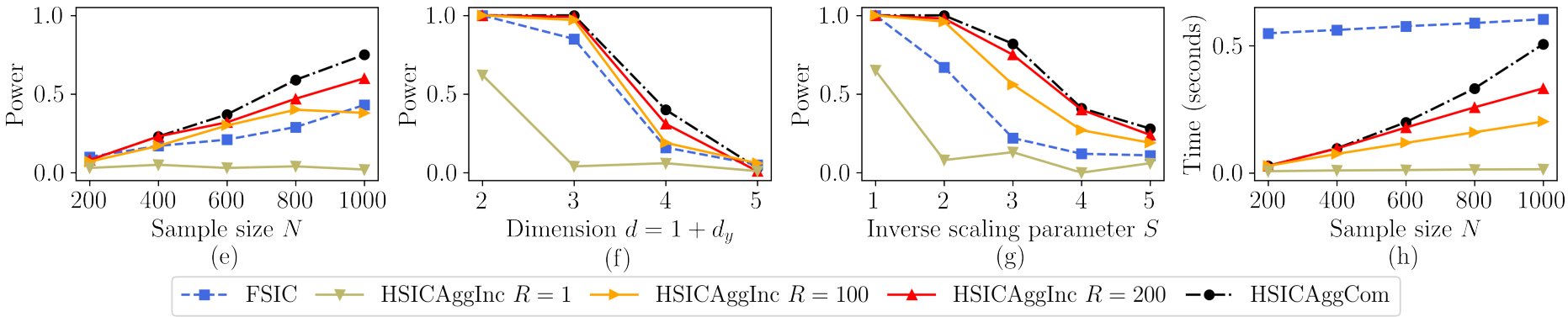


ME (Mean Embeddings) & SCF (Smooth Characteristic Functions): Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. **Interpretable distribution features with maximum testing power**. In NeurIPS 2016.

OST PSI (One-sided Test Post Selection Inference): Kübler, J. M., Jitkrittum, W., Schölkopf, B., and Muandet, K. **Learning kernel tests without data splitting**. NeurIPS 2020.

HSICAggInc: experiments

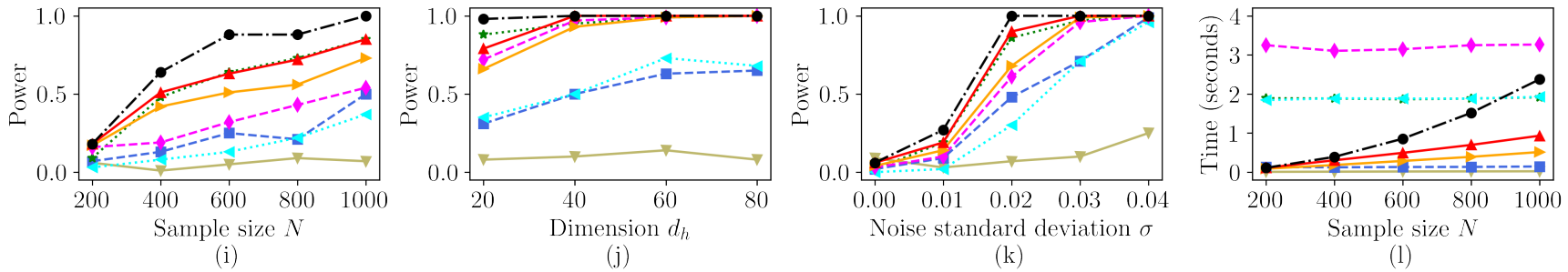
Perturbed uniform densities



FSIC (Finite Set Independence Criterion): Jitkrittum, W., Szabó, Z., and Gretton, A. **An adaptive test of independence with analytic kernel embeddings**. In ICML 2017.

KSDAggInc: experiments

Gaussian-Bernoulli Restricted Boltzmann Machine



FSSD (Finite Set Stein Discrepancy): Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. **A linear-time kernel goodness-of-fit test**. In NeurIPS 2017.

LSD (Learned Stein Discrepancy): Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., and Zemel, R. **Learning the Stein discrepancy for training and evaluating energy-based models without sampling**. In ICML 2020.

L1 IMQ & Cauchy RFF (Random Fourier Features): Huggins, J. and Mackey, L. **Random feature Stein discrepancies**. In NeurIPS 2018.

Thank you for your attention!
Any Questions?

MMDAgg

KSDAgg

AggInc

Code