



UCL



GATSBY

Inria

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

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

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

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

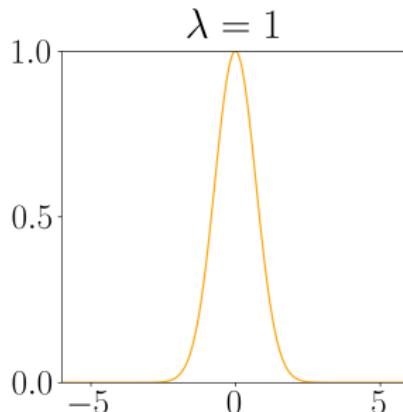
$$\mathbb{P}_{p \times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta$$

Kernels

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

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

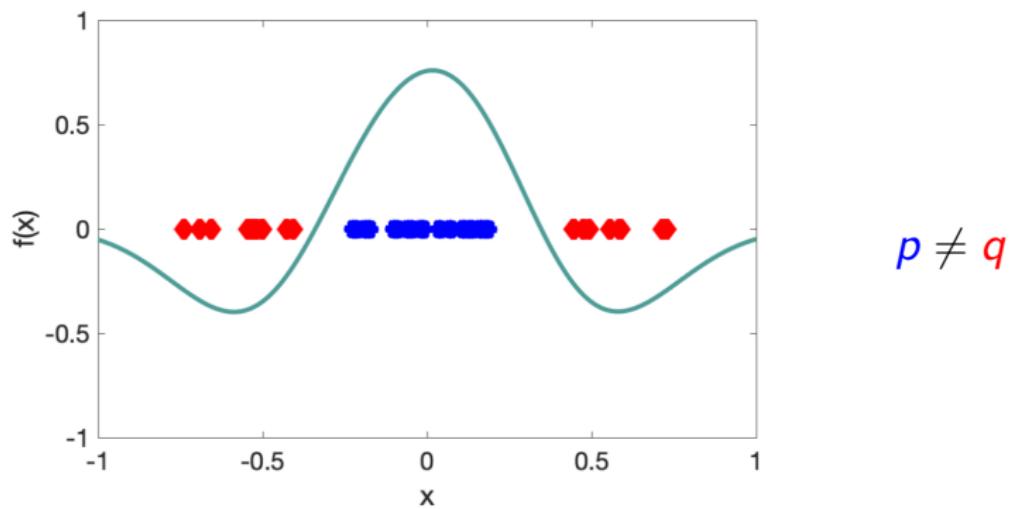
$$k_{\lambda}(\mathbf{x}, \mathbf{y}) := \frac{1}{\pi^{d/2} \lambda_1 \dots \lambda_d} \exp\left(-\sum_{i=1}^d \frac{(\mathbf{x}_i - \mathbf{y}_i)^2}{\lambda_i^2}\right)$$



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:** $\lambda \in (0, \infty)^d$

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



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

- 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

$$\text{MMD}_{\lambda}^2(p, q) := \mathbb{E}_{p,p}[k_{\lambda}(X, X')] - 2 \mathbb{E}_{p,q}[k_{\lambda}(X, Y)] + \mathbb{E}_{q,q}[k_{\lambda}(Y, Y')]$$

$$\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m, \mathbb{Y}_n) := \frac{1}{m(m-1)} \sum_{1 \leq i \neq i' \leq m} k_{\lambda}(X_i, X_{i'}) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k_{\lambda}(X_i, Y_j) + \frac{1}{n(n-1)} \sum_{1 \leq j \neq j' \leq n} k_{\lambda}(Y_j, Y_{j'})$$

MMD test for a fixed bandwidth λ

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

Quantile: $\widehat{q}_{1-\alpha}^\lambda$ is the $[(B+1)(1-\alpha)]$ -th largest value of $\widehat{\text{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n)$ and B permuted test statistics

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

Non-asymptotic level α

Time complexity:

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

MMDAgg for a collection of bandwidths Λ

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

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

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

Non-asymptotic level α

Time complexity:

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

Minimax adaptivity over Sobolev balls

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

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

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

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

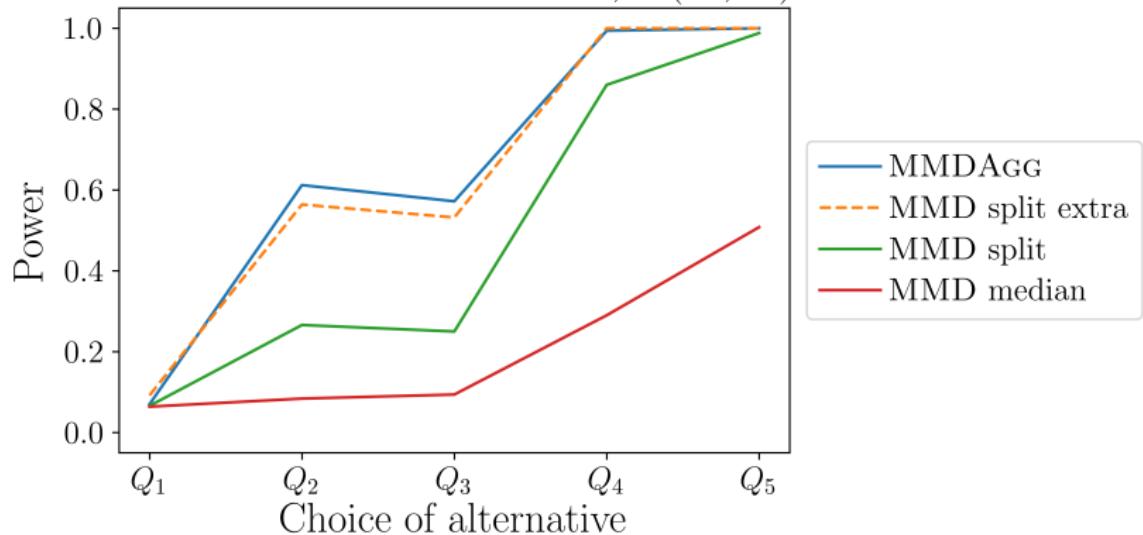
MMDAgg Experiment

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

$$\begin{aligned} \mathcal{P} &:= \{0, \dots, 9\} & \mathcal{Q}_2 &:= \mathcal{P} \setminus \{8, 6\} & \mathcal{Q}_4 &:= \mathcal{P} \setminus \{8, 6, 4, 2\} \\ \mathcal{Q}_1 &:= \mathcal{P} \setminus \{8\} & \mathcal{Q}_3 &:= \mathcal{P} \setminus \{8, 6, 4\} & \mathcal{Q}_5 &:= \mathcal{P} \setminus \{8, 6, 4, 2, 0\} \end{aligned}$$

Two-sample experiment

MNIST dataset $m = n = 500$, $\Lambda(12, 16)$



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 $\mathbb{Z}_n := (Z_1, \dots, Z_n)$, $Z_i \stackrel{\text{iid}}{\sim} q$ in \mathbb{R}^d

$$\mathcal{H}_0: p = q \quad \text{against} \quad \mathcal{H}_a: p \neq q$$

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

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

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

$$\begin{aligned} \text{KSD}_{p,\lambda}^2(q) &:= \text{MMD}_{h_{p,\lambda}}^2(p, q) = \mathbb{E}_{q,q}[h_{p,\lambda}(Z, Z')] \\ \widehat{\text{KSD}}_{p,\lambda}^2(\mathbb{Z}_n) &:= \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_{p,\lambda}(Z_i, Z_j) \end{aligned}$$

KSDAgg Experiment

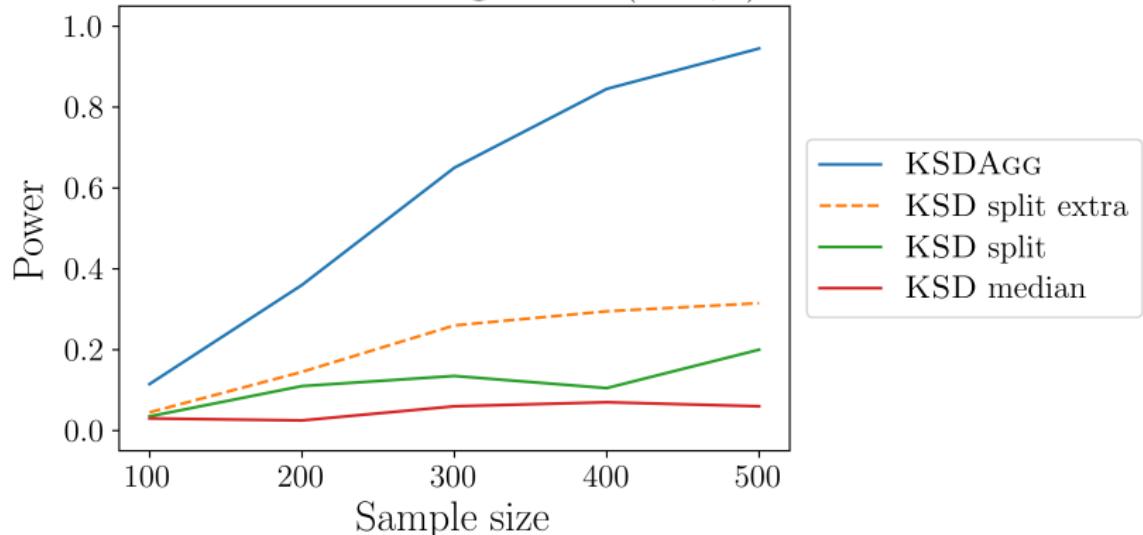
$$\Lambda(\ell_-, \ell_+) := \left\{ 2^\ell \lambda_{med} : \ell \in \{\ell_-, \dots, \ell_+\} \right\} \quad w_\lambda := 1 / |\Lambda|$$

model: Normalizing Flow density

samples: true MNIST digits

Goodness-of-fit experiment

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



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.

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

- Complete U -statistic:

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

- Quadratic-time MMDAgg, KSDAgg & HSICAgg:

$$\mathcal{O}\left(|\Lambda|(\mathcal{B}_1 + \mathcal{B}_2)N^2\right)$$

- Incomplete U -statistic:

$$\sum_{(i,j) \in \mathcal{D}} h(Z_i, Z_j) \quad \mathcal{D} \subseteq \{(i,j) : 1 \leq i \neq j \leq N\}$$

- Linear-time MMDAggInc, KSDAggInc & HSICAggInc:

$$\mathcal{O}\left(|\Lambda|(\mathcal{B}_1 + \mathcal{B}_2)|\mathcal{D}|\right) \quad \mathcal{D} = cN$$

- retain **high power** & **outperform** state-of-the-art linear-time tests

Thank you for your attention!

MMDAgg



[paper](#)

KSDAgg



[paper](#)



[code](#)



[code](#)

AggInc: coming out soon!