

MMD Aggregated Two-Sample Test



Antonin Schrab

University of Cambridge

Department of Computer Science and Technology

Positions:

- **Postdoc:** University of Cambridge, Computer Science & Technology
- **PhD:** University College London, AI Centre & Gatsby Unit
- **MSc ML:** University College London
- **MMath:** University of Oxford, St Hilda's College

Research interests:

- **Topics:** Nonparametric Hypothesis Testing, Kernel Methods
- **Frameworks:** Two-Sample, Independence, Goodness-of-Fit Testing
- **Constraints:** Efficiency, Robustness, Differential Privacy (DP)
- **Applications:** Dataset Shift, Machine Unlearning, DP Audit
- **Theory:** Level, Consistency, Minimax Optimality, Adaptivity
- **Practice:** User-friendly Parameter-free Implementations in JAX

MMD Aggregated Two-Sample Test



Antonin
Schrab

† ‡ §



Ilmun
Kim

*



Mélisande
Albert

★



Béatrice
Laurent

★



Benjamin
Guedj

† §



Arthur
Gretton

‡

† Centre for Artificial Intelligence, UCL

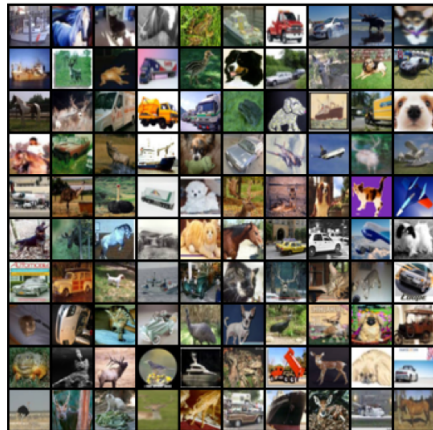
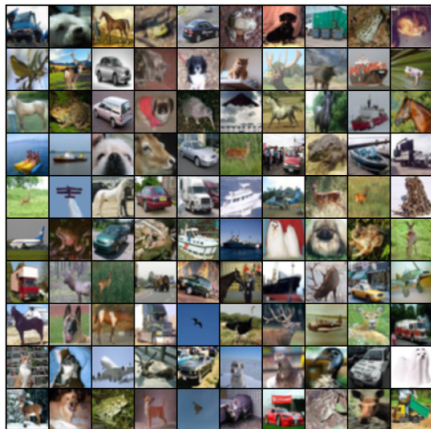
‡ Gatsby Computational Neuroscience Unit, UCL

§ Inria London Programme

* Department of Statistics & Data Science, Yonsei University

★ Institut de Mathématiques, Université de Toulouse

Two-sample motivation



Do these samples come from the same distribution?

Two-sample problem

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

Two-sample problem

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

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

Two-sample problem

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

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

Type I error: controlled by α by design

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

Two-sample problem

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

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

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

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

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

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

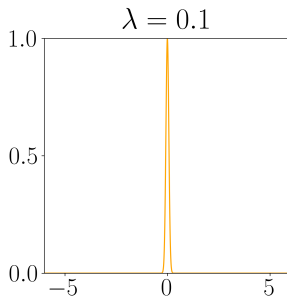
$$k_{\lambda}(x, y) := \exp\left(-\sum_{i=1}^d \frac{(x_i - y_i)^2}{\lambda_i^2}\right)$$

Kernels

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

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

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

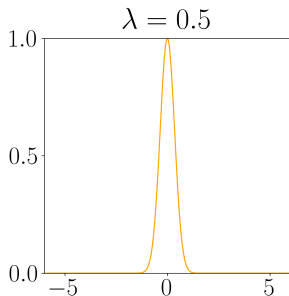


Kernels

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

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

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

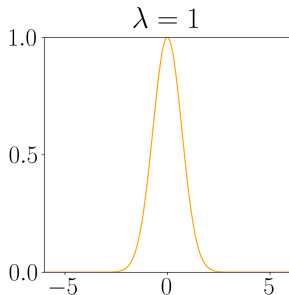


Kernels

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

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

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

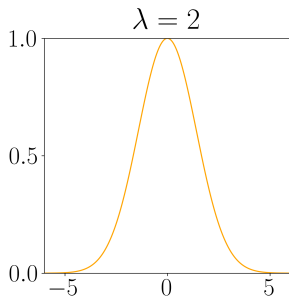


Kernels

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

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

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

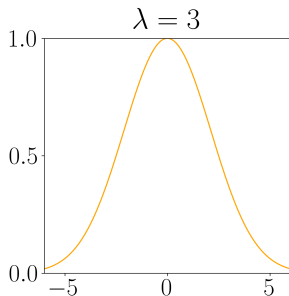


Kernels

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

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

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

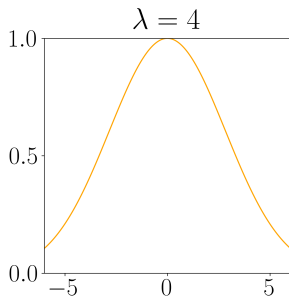


Kernels

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

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

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

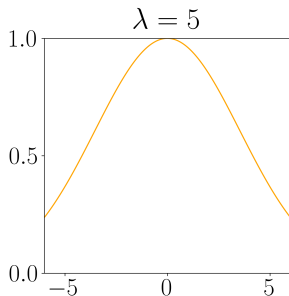


Kernels

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

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

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



Two-sample test using the Maximum Mean Discrepancy

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

Two-sample test using the Maximum Mean Discrepancy

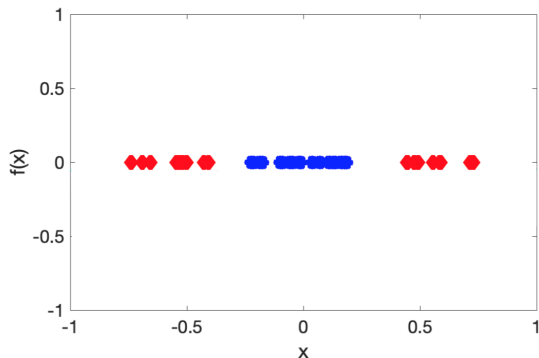
Kernel: $k_{\lambda}(x, y) := \prod_{i=1}^d K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$ **Bandwidth:** $\lambda \in (0, \infty)^d$

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

Two-sample test using the Maximum Mean Discrepancy

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

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

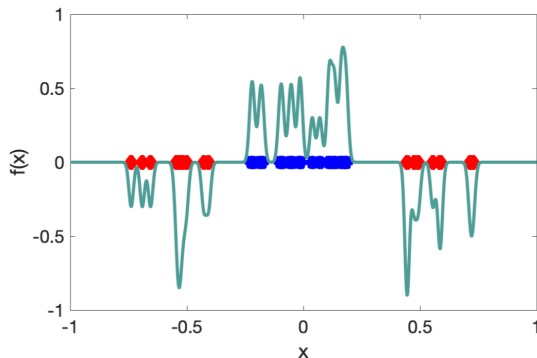


$p \neq q$

Two-sample test using the Maximum Mean Discrepancy

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

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



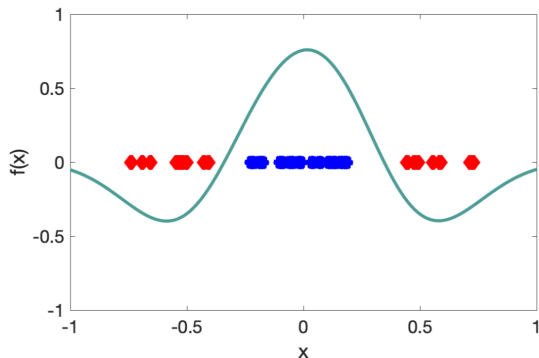
$p \neq q$

bandwidth λ : too small

Two-sample test using the Maximum Mean Discrepancy

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

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



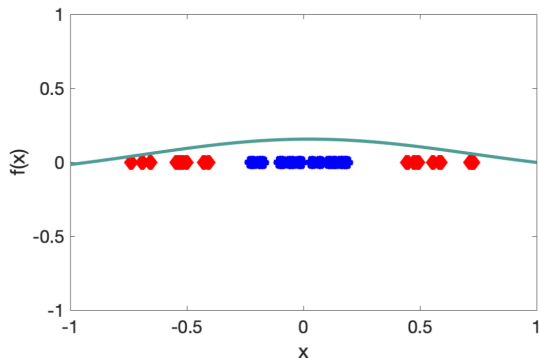
$$p \neq q$$

bandwidth λ : well-calibrated

Two-sample test using the Maximum Mean Discrepancy

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

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



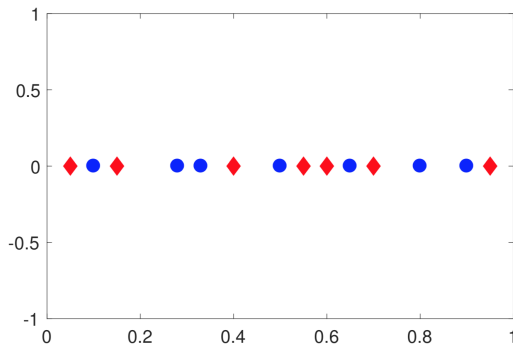
$p \neq q$

bandwidth λ : too large

Two-sample test using the Maximum Mean Discrepancy

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

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

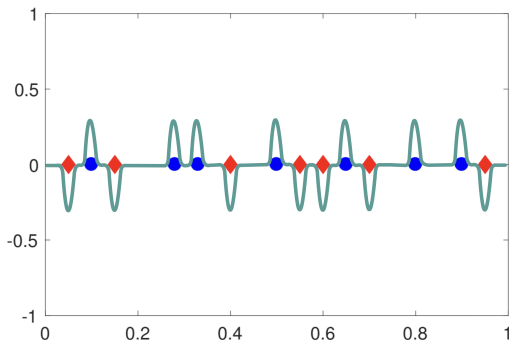


$p = q$

Two-sample test using the Maximum Mean Discrepancy

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

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



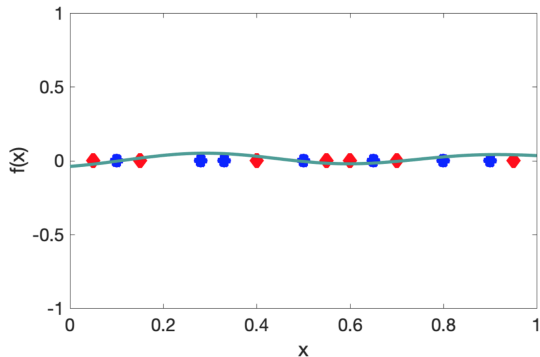
$$p = q$$

bandwidth λ : too small

Two-sample test using the Maximum Mean Discrepancy

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

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



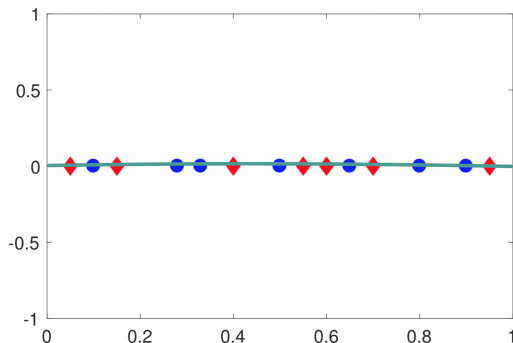
$$p = q$$

bandwidth λ : well-calibrated

Two-sample test using the Maximum Mean Discrepancy

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

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



$$p = q$$

bandwidth λ : too large

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 $\mathcal{H}_a$

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 $\mathcal{H}_a$

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 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 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**

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

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

Maximum Mean Discrepancy estimator

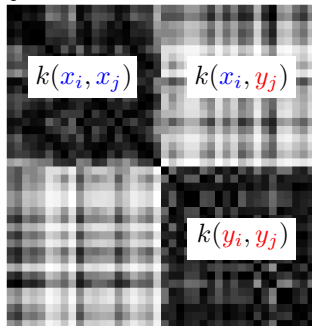
$$\text{MMD}_{\lambda}^2(p, q) := \mathbb{E}_{p,p}[k_{\lambda}(X, X')] - 2\mathbb{E}_{p,q}[k_{\lambda}(X, Y)] + \mathbb{E}_{q,q}[k_{\lambda}(Y, Y')]$$

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

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

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

MMD test for a fixed bandwidth λ

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

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

MMD test for a fixed bandwidth λ

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

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

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

MMD test for a fixed bandwidth λ

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

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

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

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

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

MMD test for a fixed bandwidth λ

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

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

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

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

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

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

MMD test for a fixed bandwidth λ

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

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

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

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

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

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

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

MMDAgg for a collection of bandwidths $\wedge$

MMDAgg for a collection of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

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

MMDAgg for a collection of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

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

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

MMDAgg for a collection of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

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

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

MMDAgg: non-asymptotic level α

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

MMDAgg for a collection of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

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

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

MMDAgg: non-asymptotic level α

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

- correction u_{α} defined as

$$\sup \left\{ u > 0 : \sum_{b=1}^{B_2} 1 \left(\max_{\lambda \in \Lambda} \left(\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m^{\sigma_b}, \mathbb{Y}_n^{\sigma_b}) - \hat{q}_{1-u}^{\lambda} \right) > 0 \right) \leq \alpha \right\}$$

MMDAgg for a collection of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

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

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

MMDAgg: non-asymptotic level α

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

- correction u_{α} defined as

$$\sup \left\{ u > 0 : \sum_{b=1}^{B_2} 1 \left(\max_{\lambda \in \Lambda} \left(\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m^{\sigma_b}, \mathbb{Y}_n^{\sigma_b}) - \hat{q}_{1-u}^{\lambda} \right) > 0 \right) \leq \alpha \right\}$$

- more powerful than Bonferroni correction as $u_{\alpha} \geq \alpha/|\Lambda|$

MMDAgg for a collection of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

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

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

MMDAgg: non-asymptotic level α

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

- correction u_{α} defined as

$$\sup \left\{ u > 0 : \sum_{b=1}^{B_2} 1 \left(\max_{\lambda \in \Lambda} \left(\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m^{\sigma_b}, \mathbb{Y}_n^{\sigma_b}) - \hat{q}_{1-u}^{\lambda} \right) > 0 \right) \leq \alpha \right\}$$

- more powerful than Bonferroni correction as $u_{\alpha} \geq \alpha/|\Lambda|$
- adapts from $\alpha/|\Lambda|$ to α depending on correction required

MMDAgg for a collection of bandwidths Λ

Bonferroni multiple testing: non-asymptotic level α

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

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

MMDAgg: non-asymptotic level α

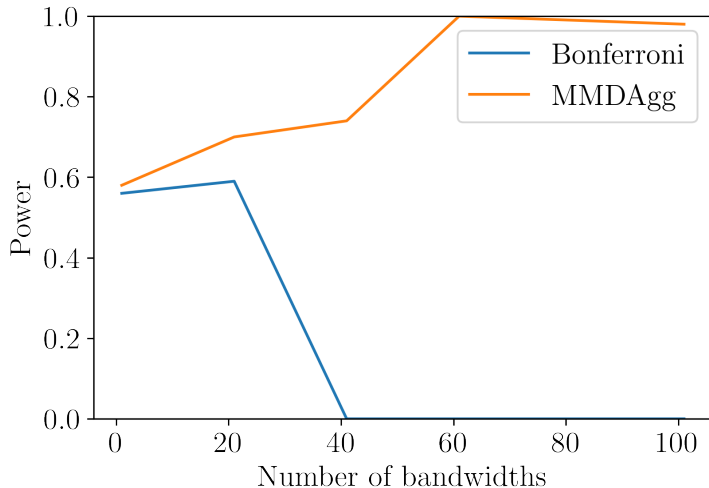
$$\Delta_{\alpha}^{\Lambda}(\mathbb{X}_m, \mathbb{Y}_n) := 1 \left(\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m, \mathbb{Y}_n) > \hat{q}_{1-u_{\alpha}}^{\lambda} \text{ for some } \lambda \in \Lambda \right)$$

- correction u_{α} defined as

$$\sup \left\{ u > 0 : \sum_{b=1}^{B_2} 1 \left(\max_{\lambda \in \Lambda} \left(\widehat{\text{MMD}}_{\lambda}^2(\mathbb{X}_m^{\sigma_b}, \mathbb{Y}_n^{\sigma_b}) - \hat{q}_{1-u}^{\lambda} \right) > 0 \right) \leq \alpha \right\}$$

- more powerful than Bonferroni correction as $u_{\alpha} \geq \alpha/|\Lambda|$
- adapts from $\alpha/|\Lambda|$ to α depending on correction required
- time complexity $\mathcal{O}(|\Lambda| (B_1 + B_2) (m+n)^2)$

Multiple testing correction comparison



$$\Lambda(\ell_-, \ell_+) := \{2^\ell \lambda_{med} : \ell \in \{\ell_-, \dots, \ell_+\}\}$$

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

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

Sobolev balls

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

Sobolev balls:

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

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

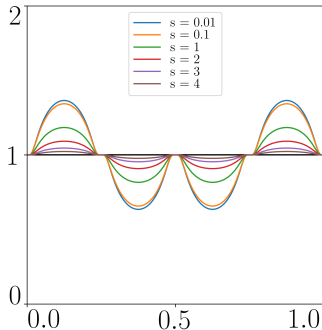
Sobolev balls

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

Sobolev balls:

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

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



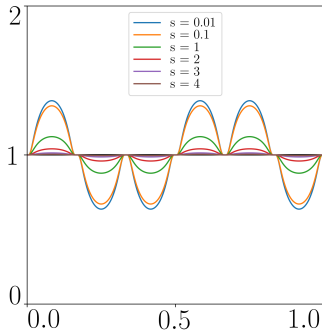
Sobolev balls

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

Sobolev balls:

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

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



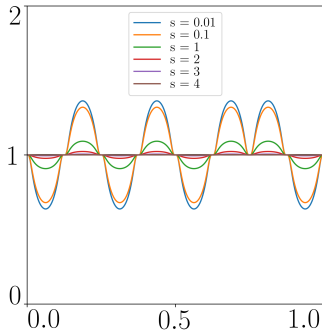
Sobolev balls

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

Sobolev balls:

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

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



Minimax adaptivity over Sobolev balls

Theorem

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

$$\|p - q\|_2 \geq C \left(\frac{m + n}{\ln(\ln(m + n))} \right)^{-2s/(4s+d)} \sqrt{\log\left(\frac{1}{\alpha}\right) \log\left(\frac{1}{\beta}\right)}$$

guarantees control over the probability of type II error of MMDAgg

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

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

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

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

MMDAgg Experiment

$$\Lambda(\ell_-, \ell_+) := \{2^\ell \lambda_{med} : \ell \in \{\ell_-, \dots, \ell_+\}\}$$

$$w_\lambda := 1 / |\Lambda|$$

MMDAgg Experiment

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

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

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

MMDAgg Experiment

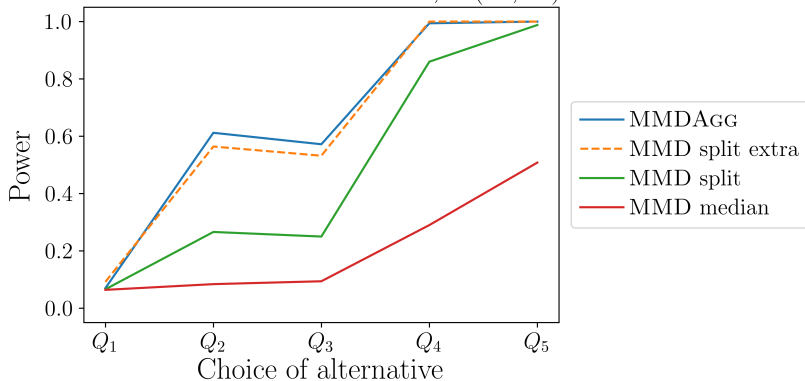
$$\Lambda(\ell_-, \ell_+) := \{2^\ell \lambda_{med} : \ell \in \{\ell_-, \dots, \ell_+\}\} \quad w_\lambda := 1 / |\Lambda|$$

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

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

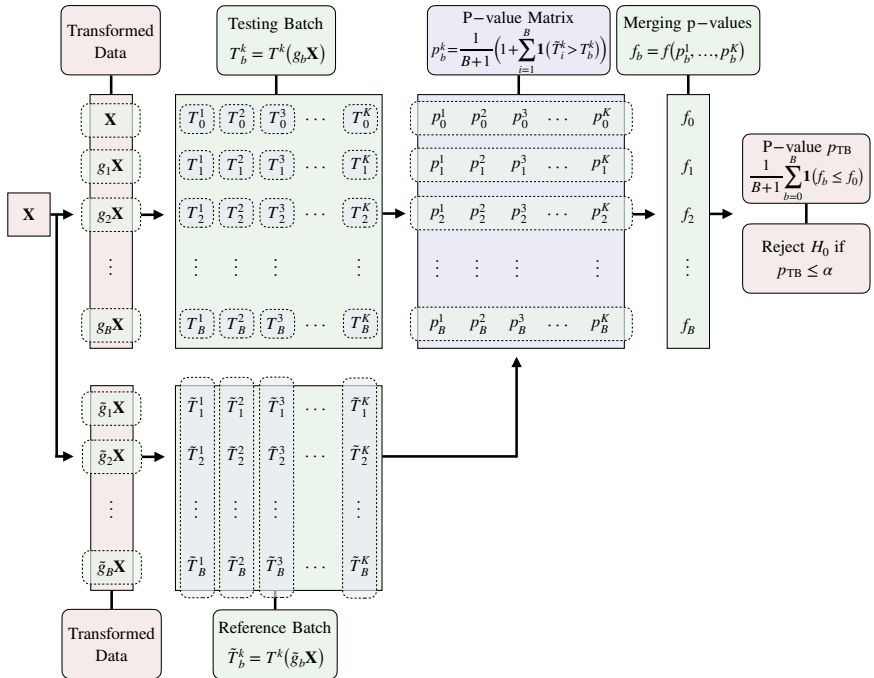
Two-sample experiment

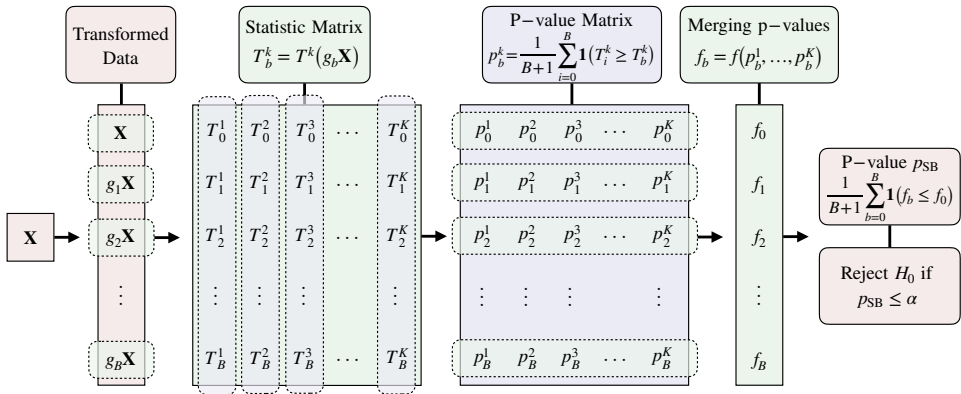
MNIST dataset $m = n = 500$, $\Lambda(12, 16)$

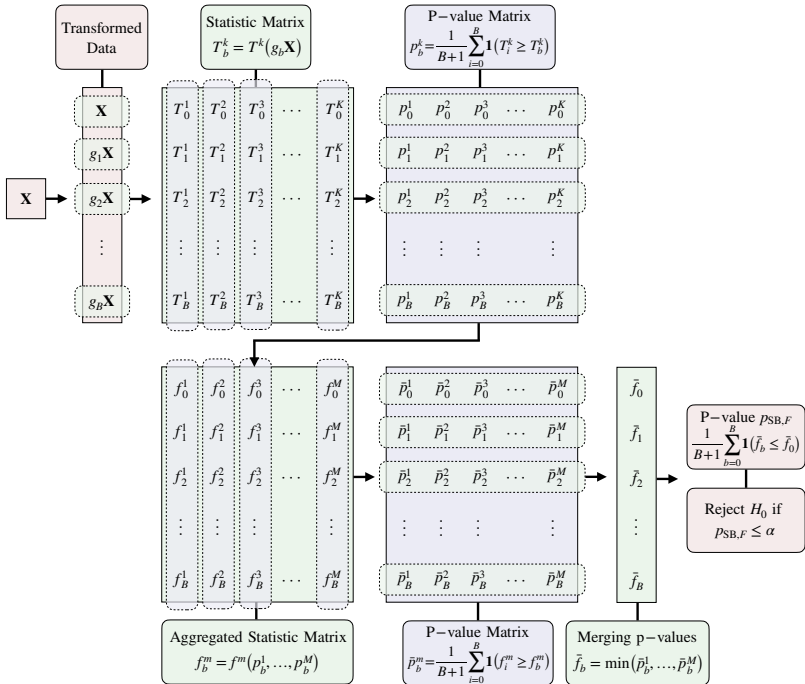


Aggregation of Statistical Evidence under Exchangeability

Antonin Schrab, Rajen Shah, Arthur Gretton, Ilmun Kim







KSD Aggregated Goodness-of-fit Test

Efficient Aggregated Kernel Tests using Incomplete U-statistics

MMD-FUSE: Learning & Combining Kernels for Two-Sample Testing

Differentially Private Permutation Tests

Robust Kernel Hypothesis Testing under Data Corruption

Credal Two-Sample Tests of Epistemic Ignorance

Regularized f-Divergence Kernel Tests

A Practical Introduction to Kernel Discrepancies: MMD, HSIC, KSD

A Unified View of Optimal Kernel Hypothesis Testing

Thank you for your attention!

Any Questions?



SCAN ME

[MMDAagg](#)



SCAN ME

[KSDAagg](#)



SCAN ME

[AggInc](#)



SCAN ME

[Code](#)