MMD-based Aggregated Two-Sample Test

Antonin Schrab[1,2,3], Ilmun Kim[4], Mélisande Albert[5], Béatrice Laurent[5], Benjamin Guedj[2,3], and Arthur Gretton[1]

[1]Gatsby Computational Neuroscience Unit, UCL [2]Centre for Artificial Intelligence, UCL [3]Inria London [4]Statistical Laboratory, University of Cambridge [5]Institut de Mathématiques, Université de Toulouse

UCL

Introduction

Two-sample problem

Given independent samples • $\mathbb{X}_m := (X_1, \dots, X_m)$ where $X_i \overset{\text{iid}}{\sim} p$ in $\mathbb{R}^d$,
• $\mathbb{Y}_n := (Y_1, \dots, Y_n)$ where $Y_j \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$,

can we decide whether or not $p \neq q$ holds?
This corresponds to testing the hypothesis $\mathcal{H}_0 : p = q$ against $\mathcal{H}_a : p \neq q$.

Uniform separation rates & Minimax rate

Given a test Δ, a class of functions $\mathcal{C}$ and some $\beta \in (0,1)$, what is the smallest value $\tilde{\rho} > 0$ such that Δ has power at least $1-\beta$ against all alternative hypotheses satisfying $p-q \in \mathcal{C}$ and $\|p-q\|_2 > \tilde{\rho}$? $(\star)$

$$\rho\left(\Delta, \mathcal{C}, \beta\right) := \inf\left\{\tilde{\rho} > 0 : \sup_{(p,q):(\star)} \mathbb{P}_{p\times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta\right\}$$

Uniform separation rates are rates taking the form $C(m+n)^{-r}$.
The smallest rate achieved by a test of level α is the **minimax rate**

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

For Sobolev balls $\mathcal{S}_d^s(R)$, minimax rate $\underline{\rho}(\mathcal{S}_d^s(R), \alpha, \beta)$ is $(m+n)^{-2s/(4s+d)}$

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

Maximum Mean Discrepancy

The Maximum Mean Discrepancy $\mathrm{MMD}(p,q)$ between p and q is

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

A quadratic-time estimator $\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_m, \mathbb{Y}_n)$ is defined as

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

When $m = n$, another quadratic-time estimator $\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_n, \mathbb{Y}_n)$ is

$$\frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j).$$

We can simulate $\mathcal{H}_0$ using permutations or a wild bootstrap to estimate the $(1-\alpha)$-quantile and construct a non-asymptotic test of level α.

Kernels and choice of bandwidths

For bandwidths $\lambda \in (0,\infty)^d$, we work on $\mathbb{R}^d \times \mathbb{R}^d$ with the kernel

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

for d characteristic kernels $K_i \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfying $\int_{\mathbb{R}} K_i(u)\mathrm{d}u = 1$.
Two common ways to choose the bandwidths:
- median heuristic
- splitting the data

Aim

Construct a non-asymptotic test which is optimal in the minimax sense.

Our contributions

Single test: construction

Consider $\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n)$ and compute B simulated test statistics, let $\widehat{q}_{1-\alpha}^\lambda$ be the $\lceil (B+1)(1-\alpha) \rceil$-th biggest of those $B+1$ values, the single test is

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

Single test: theoretical results

For $\alpha \in (0, e^{-1})$, $\lambda_1 \cdots \lambda_d < 1$ and $B \in \mathbb{N}$ large enough, we have

$$\rho\left(\Delta_\alpha^\lambda, \mathcal{S}_d^s(R), \beta\right)^2 \leq C(d,s,R,\beta)\left(\sum_{i=1}^d \lambda_i^{2s} + \frac{\ln\left(\frac{1}{\alpha}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right).$$

For $\lambda_i^* = (m+n)^{-2/(4s+d)}$, the test $\Delta_\alpha^{\lambda^*}$ is optimal in the minimax sense

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

Aggregated test: construction

Consider a collection Λ of bandwidths and some weights $(w_\lambda)_{\lambda\in\Lambda}$ such that $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$. The aggregated test Δ_α^Λ rejects $\mathcal{H}_0$ if one of the tests $\{\Delta_{u_\alpha w_\lambda}^\lambda\}_{\lambda\in\Lambda}$ rejects $\mathcal{H}_0$ where

$$u_\alpha = \sup\left\{u > 0 : \mathbb{P}_{\mathcal{H}_0}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}_\lambda^2(\mathbb{X}_m, \mathbb{Y}_n) - \widehat{q}_{1-uw_\lambda}^\lambda\right) > 0\right) \leq \alpha\right\}.$$

The probability can be estimated by a Monte-Carlo approximation and the supremum can be estimated using the bisection method.

Aggregated test: theoretical results

For $\alpha \in (0, e^{-1})$ and $B_1, B_2, B_3 \in \mathbb{N}$ all large enough, $\lambda_1 \cdots \lambda_d \leq 1$ for all $\lambda \in \Lambda$ and $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$, we have

$$\rho\left(\Delta_\alpha^\Lambda, \mathcal{S}_d^s(R), \beta\right)^2 \leq C(d,s,R,\beta) \min_{\lambda\in\Lambda}\left(\sum_{i=1}^d \lambda_i^{2s} + \frac{\ln\left(\frac{1}{\alpha w_\lambda}\right)}{(m+n)\sqrt{\lambda_1\cdots\lambda_d}}\right).$$

Consider $\Lambda := \left\{\left(2^{-\ell}, \dots, 2^{-\ell}\right) : \ell \in \left\{1, \dots, \left\lceil \frac{2}{d}\log_2\left(\frac{m+n}{\ln(\ln(m+n))}\right)\right\rceil\right\}\right\}$ and $w_\lambda := \frac{6}{\pi^2 \ell^2}$. Then, Δ_α^Λ is (almost) optimal in the minimax sense

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

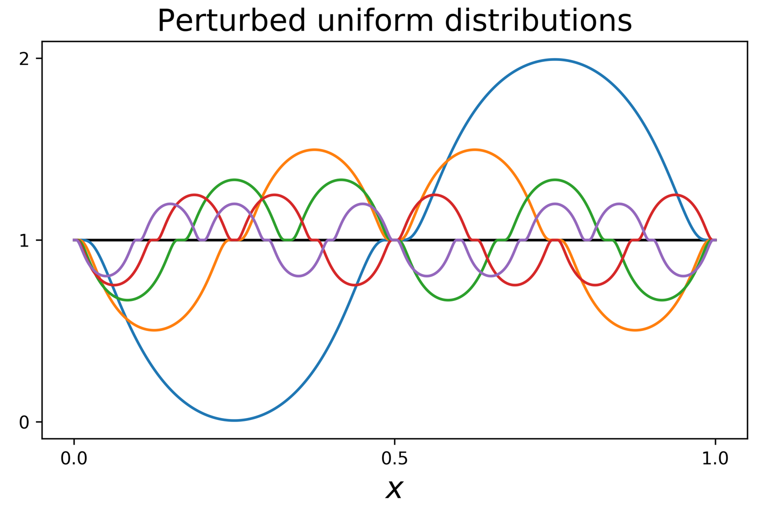
and is adaptive (no dependence on the unknown parameter s of $\mathcal{S}_d^s(R)$).

Summary of key contributions

- (almost) optimal in the minimax sense
- wild bootstrap & permutations
- outperforms state-of-the-art MMD tests
- adaptive test
- no data splitting
- wide range of kernels

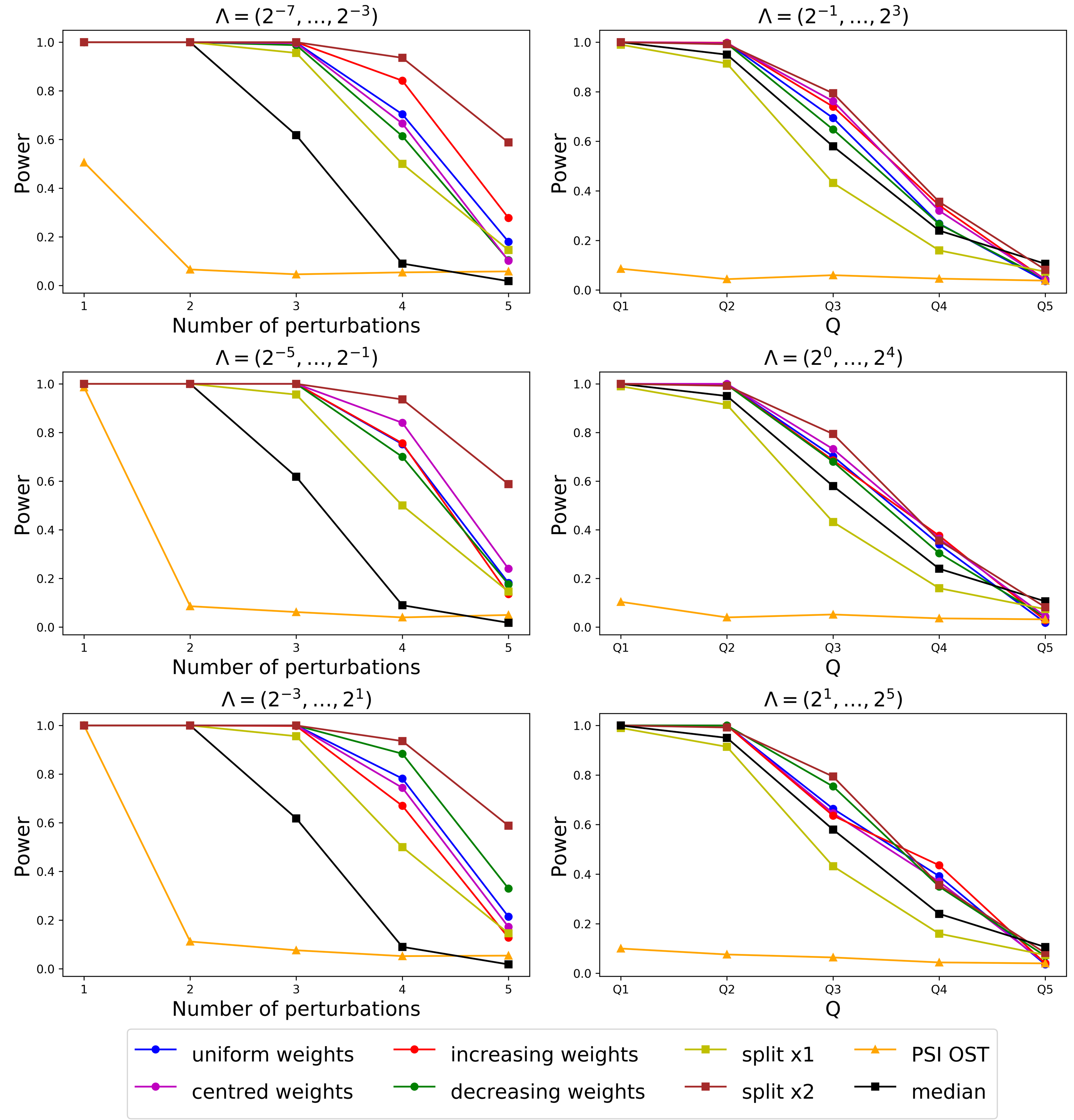
Experiments

Perturbed uniform



MNIST

$P = \{0,1,2,3,4,5,6,7,8,9\}$
$Q1 = \{1,3,5,7,9\}$
$Q2 = \{0,1,3,5,7,9\}$
$Q3 = \{0,1,2,3,5,7,9\}$
$Q4 = \{0,1,2,3,4,5,7,9\}$
$Q5 = \{0,1,2,3,4,5,6,7,9\}$



References

[1] Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on HSIC measures. *arXiv e-prints, pages arXiv–1902, 2019.*

[2] Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. *arXiv preprint arXiv:2003.13208, 2020.*

[3] Tong Li and Ming Yuan. On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302, 2019.*

[4] Jonas M. Kübler and Wittawat Jitkrittum and Bernhard Schölkopf and Krikamol Muandet. Learning Kernel Tests Without Data Splitting. *arXiv preprint arXiv:2006.02286, 2020*