Representing and comparing probabilities

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

UAI, 2017

Comparing two samples

- **Given:** Samples from unknown distributions P and Q.
- **Goal:** do P and Q differ?

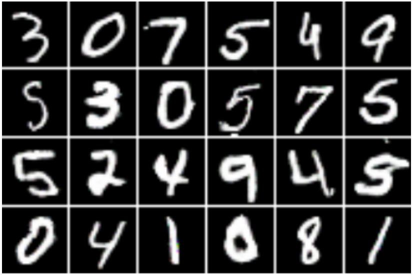
$$\sim P$$

$$\sim Q$$

An example: two-sample tests

- **Have:** Two collections of samples X, Y from unknown distributions P and Q.
- **Goal:** do P and Q differ?

MNIST samples



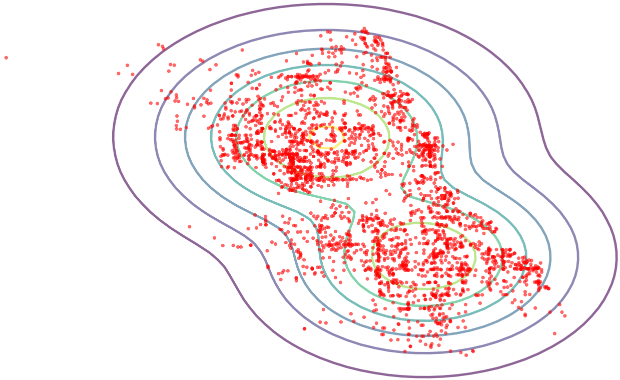
Samples from a GAN

Significant difference in GAN and MNIST?

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, Xi Chen, NIPS 2016.

Testing goodness of fit

- **Given:** **A model** P and samples and Q.
- **Goal:** is P a good fit for Q?



Chicago crime data

Model is Gaussian mixture with **two** components.

Testing independence

- **Given:** Samples from a distribution P_{XY}
- **Goal:** Are X and Y independent?

X	Y
	A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.
	Their noses guide them through life, and they're never happier than when following an interesting scent.
	A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

Text from dogtime.com and petfinder.com

Outline: part 1

Two sample testing

- Test statistic: Maximum Mean Discrepancy (MMD)...
  - ...as a difference in feature means
  - ...as an integral probability metric (not just a technicality!)
- Statistical testing with the MMD
- Troubleshooting GANs with MMD

Outline: part 2

Goodness of fit testing

- The kernel Stein discrepancy

Dependence testing

- Dependence using the MMD
- Depenence using feature covariances
- Statistical testing

Additional topics

Outline: part 2

Goodness of fit testing

- The kernel Stein discrepancy

Dependence testing

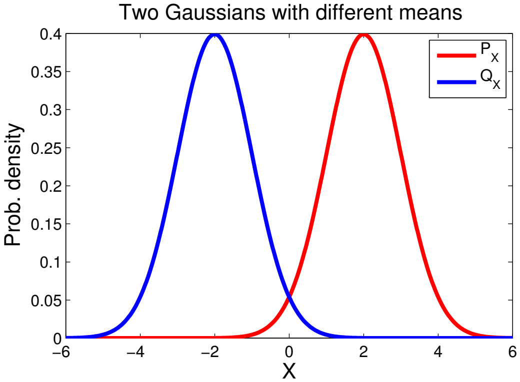
- Dependence using the MMD
- Depenence using feature covariances
- Statistical testing

Additional topics

Maximum Mean Discrepancy

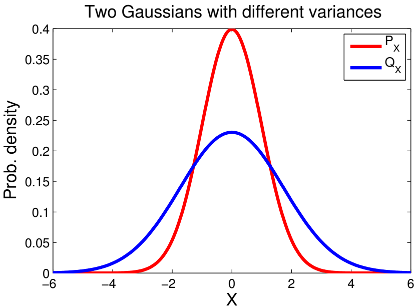
Feature mean difference

- Simple example: 2 Gaussians with different means
- Answer: t-test



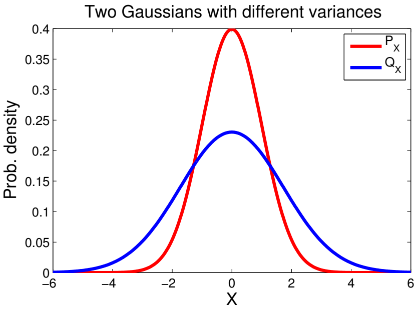
Feature mean difference

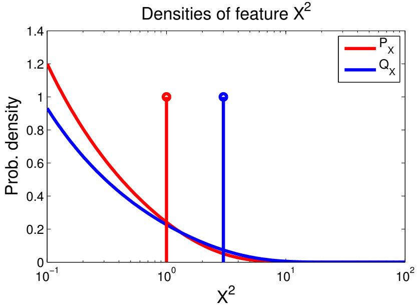
- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

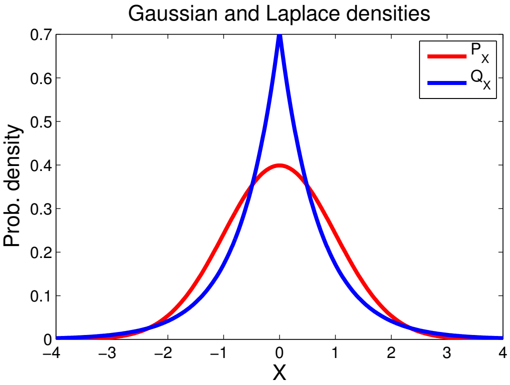
- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$





Feature mean difference

- Gaussian and Laplace distributions
- Same mean *and* same variance
- Difference in means using **higher order features**...RKHS



Infinitely many features using kernels

Kernels: dot products of features

Feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in \ell_2$$

For positive definite k,

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Infinitely many features $\varphi(x)$, dot product in closed form!

Infinitely many features using kernels

Kernels: dot products of features

Feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in \ell_2$$

For positive definite k,

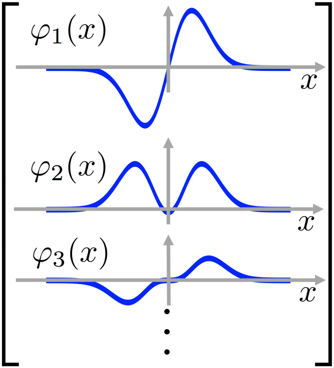
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Infinitely many features $\varphi(x)$, dot product in closed form!

Exponentiated quadratic kernel

$$k(x, x') = \exp\left(-\gamma \left\|x - x'\right\|^2\right)$$

$$\varphi(x) =$$



Features: Gaussian Processes for Machine learning, Rasmussen and Williams, Ch. 4.

Infinitely many features of *distributions*

Given P a Borel **probability measure** on $\mathcal{X}$, define feature map of probability P,

$$\mu_P = [\ldots \mathbf{E}_P [\varphi_i(X)] \ldots]$$

For positive definite $k(x, x')$,

$$\langle \mu_P, \mu_Q \rangle_{\mathcal{F}} = \mathbf{E}_{P,Q} k(x, y)$$

for $x \sim P$ and $y \sim Q$.

Fine print: feature map $\varphi(x)$ must be Bochner integrable for all probability measures considered. Always true if kernel bounded.

Infinitely many features of *distributions*

Given P a Borel **probability measure** on $\mathcal{X}$, define feature map of probability P,

$$\mu_P = [\ldots \mathbf{E}_P [\varphi_i(X)] \ldots]$$

For positive definite $k(x, x')$,

$$\langle \mu_P, \mu_Q \rangle_{\mathcal{F}} = \mathbf{E}_{P,Q} k(x, y)$$

for $x \sim P$ and $y \sim Q$.

Fine print: feature map $\varphi(x)$ must be Bochner integrable for all probability measures considered. Always true if kernel bounded.

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} \\
&= \underbrace{\mathbf{E}_P k(X, X')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(Y, Y')}_{\text{(a)}} - 2\underbrace{\mathbf{E}_{P,Q} k(X, Y)}_{\text{(b)}}
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

$$\begin{aligned} MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\ &= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} \\ &= \underbrace{\mathbf{E}_P k(X, X')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(Y, Y')}_{\text{(a)}} - 2\underbrace{\mathbf{E}_{P,Q} k(X, Y)}_{\text{(b)}} \end{aligned}$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

$$\begin{aligned} MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\ &= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} \\ &= \underbrace{\mathbf{E}_P k(X, X')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(Y, Y')}_{\text{(a)}} - 2\underbrace{\mathbf{E}_{P,Q} k(X, Y)}_{\text{(b)}} \end{aligned}$$

(a)= within distrib. similarity, (b)= cross-distrib. similarity.

Illustration of MMD

- Dogs ($= P$) and fish ($= Q$) example revisited
- Each entry is one of $k(\text{dog}_i, \text{dog}_j)$, $k(\text{dog}_i, \text{fish}_j)$, or $k(\text{fish}_i, \text{fish}_j)$

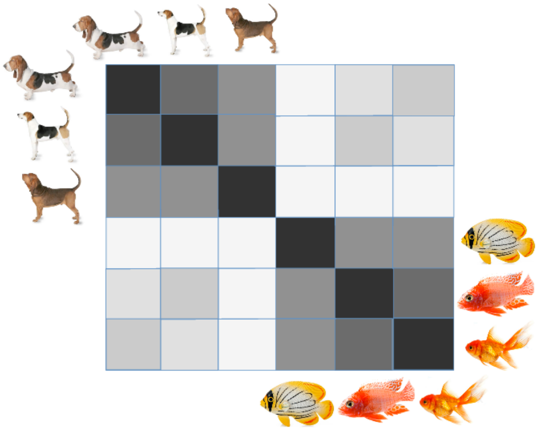
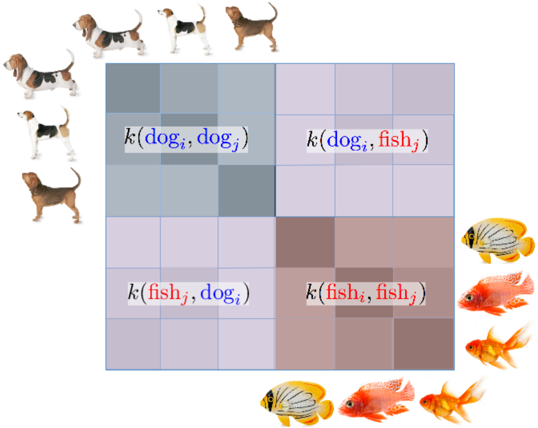


Illustration of MMD

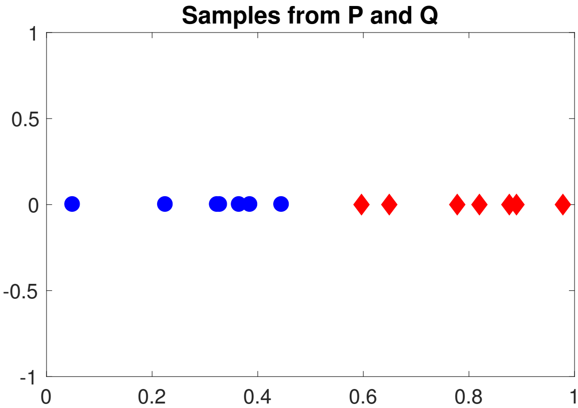
The maximum mean discrepancy:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\text{dog}_i, \text{dog}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\text{fish}_i, \text{fish}_j) - \frac{2}{n^2} \sum_{i,j} k(\text{dog}_i, \text{fish}_j)$$



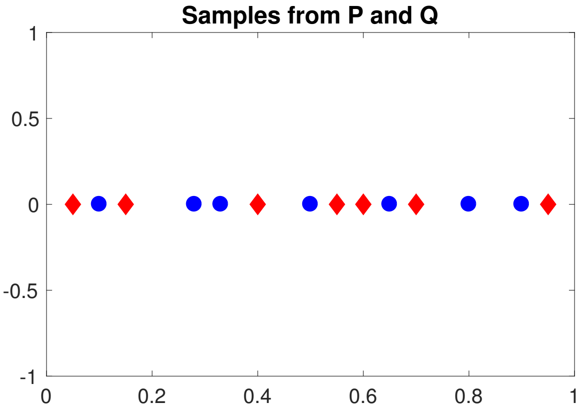
MMD as an integral probability metric

Are P and Q different?



MMD as an integral probability metric

Are P and Q different?

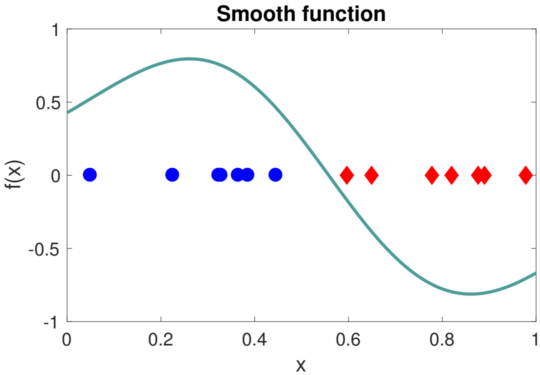


MMD as an integral probability metric

Integral probability metric:

Find a "well behaved function" $f(x)$ to maximize

$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$

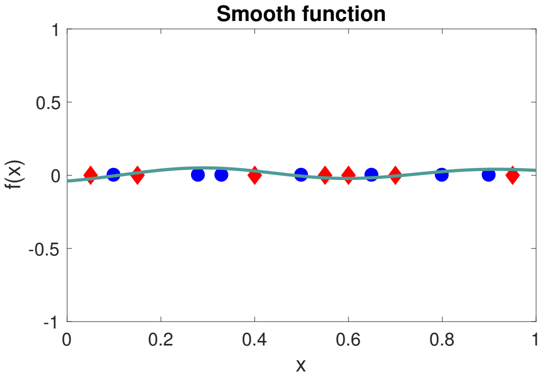


MMD as an integral probability metric

Integral probability metric:

Find a "well behaved function" $f(x)$ to maximize

$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right]$$

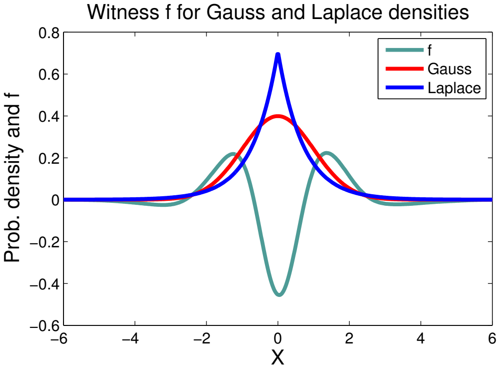
(F = unit ball in RKHS $\mathcal{F}$)

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right]$$

$$(F = \text{unit ball in RKHS } \mathcal{F})$$



MMD as an integral probability metric

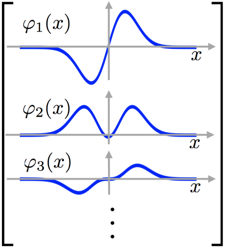
Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right]$$

$$(F = \text{unit ball in RKHS } \mathcal{F})$$

Functions are linear combinations of features:

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{F}} = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}^\top$$



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \le 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right]$$

$$(F = \text{unit ball in RKHS } \mathcal{F})$$

Expectations of functions are linear combinations of expected features

$$\mathbf{E}_P(f(X)) = \langle f, \mathbf{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right]$$

$$(F = \text{unit ball in RKHS } \mathcal{F})$$

For characteristic RKHS $\mathcal{F}$, $MMD(P, Q; F) = 0$ iff $P = Q$

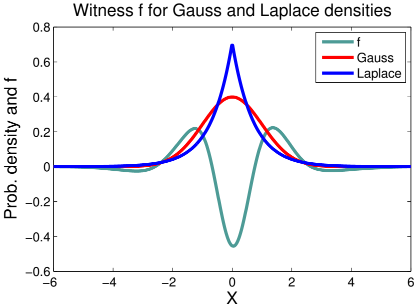
Other choices for witness function class:

- Bounded continuous [Dudley, 2002]
- Bounded varation 1 (Kolmogorov metric) [Müller, 1997]
- Bounded Lipschitz (Wasserstein distances) [Dudley, 2002]

Integral prob. metric vs feature difference

The MMD:

$$MMD(P, Q; F) = \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right]$$



Integral prob. metric vs feature difference

The MMD:

use

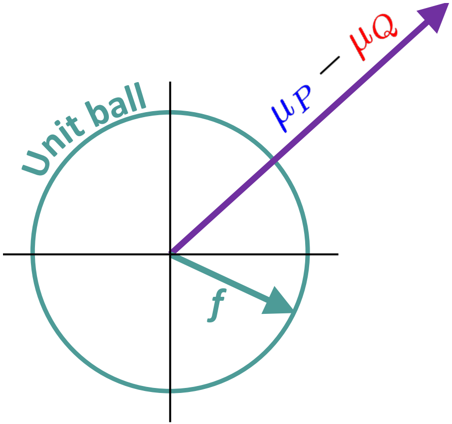
$$\mathbf{E}_P f(X) = \langle \mu_P, f \rangle_{\mathcal{F}}$$

$$MMD(P, Q; F)$$
$$= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$
$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$

Integral prob. metric vs feature difference

The MMD:

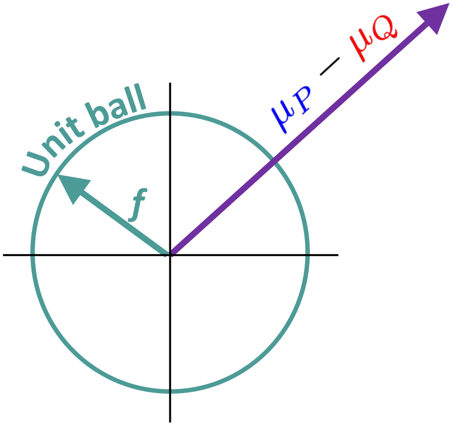
$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
&= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature difference

The MMD:

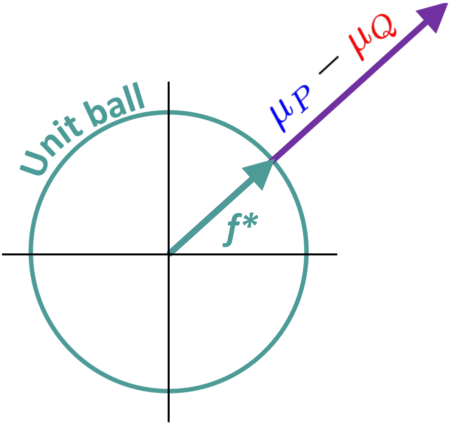
$$MMD(P, Q; F)$$
$$= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$
$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature difference

The MMD:

$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)] \\
&= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



$$f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$$

Integral prob. metric vs feature difference

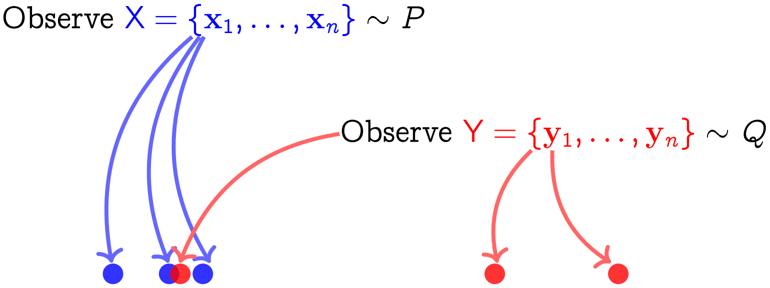
The MMD:

$$
\begin{aligned}
& MMD(P, Q; F) \\
& = \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
& = \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
& = \| \mu_P - \mu_Q \|
\end{aligned}
$$

$\boxed{\text{Function view and feature view equivalent}}$

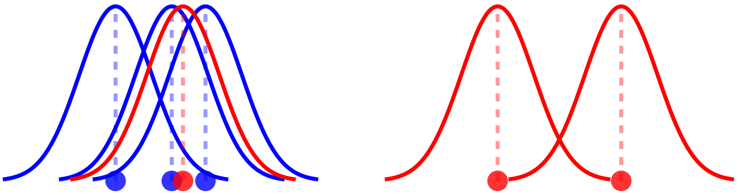
Construction of MMD witness

Construction of empirical witness function (proof: next slide!)



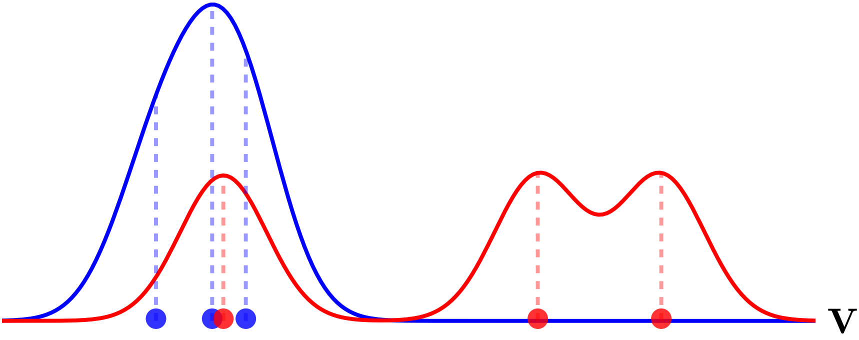
Construction of MMD witness

Construction of empirical witness function (proof: next slide!)



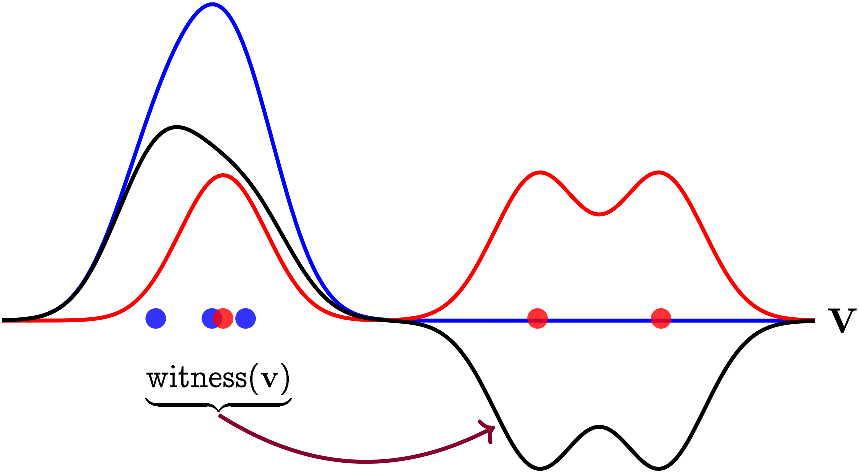
Construction of MMD witness

Construction of empirical witness function (proof: next slide!)



Construction of MMD witness

Construction of empirical witness function (proof: next slide!)



Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

$$\widehat{\mu}_P := \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

$$\widehat{\mu}_P := \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)$$

The empirical witness function at v

$$f^*(v) = \langle f^*, \varphi(v) \rangle_{\mathcal{F}}$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

$$\widehat{\mu}_P := \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)$$

The empirical witness function at v

$$\begin{aligned} f^*(v) &= \langle f^*, \varphi(v) \rangle_{\mathcal{F}} \\ &\propto \langle \widehat{\mu}_P - \widehat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \end{aligned}$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

$$\widehat{\mu}_P := \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)$$

The empirical witness function at v

$$\begin{aligned} f^*(v) &= \langle f^*, \varphi(v) \rangle_{\mathcal{F}} \\ &\propto \langle \widehat{\mu}_P - \widehat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \\ &= \frac{1}{n}\sum_{i=1}^{n} k(x_i, v) - \frac{1}{n}\sum_{i=1}^{n} k(\mathrm{y}_i, v) \end{aligned}$$

Don't need explicit feature coefficients $f^* := \begin{bmatrix} f_1^* & f_2^* & \dots \end{bmatrix}$

Two-Sample Testing

A statistical test using MMD

The empirical MMD:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{y}_i, \mathrm{y}_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, \mathrm{y}_j)$$

How does this help decide whether $P = Q$?

A statistical test using MMD

The empirical MMD:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)$$

Perspective from statistical hypothesis testing:

- **Null hypothesis** $\mathcal{H}_0$ when $P = Q$
  - should see $\widehat{MMD}^2$ "close to zero".
- **Alternative hypothesis** $\mathcal{H}_1$ when $P \neq Q$
  - should see $\widehat{MMD}^2$ "far from zero"

A statistical test using MMD

The empirical MMD:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{y}_i, \mathrm{y}_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, \mathrm{y}_j)$$

Perspective from statistical hypothesis testing:

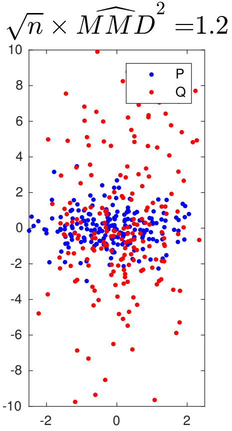
- **Null hypothesis $\mathcal{H}_0$ when $P = Q$**
  - should see $\widehat{MMD}^2$ "close to zero".
- **Alternative hypothesis $\mathcal{H}_1$ when $P \neq Q$**
  - should see $\widehat{MMD}^2$ "far from zero"

Want Threshold c_α for $\widehat{MMD}^2$ to get false positive rate α

Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

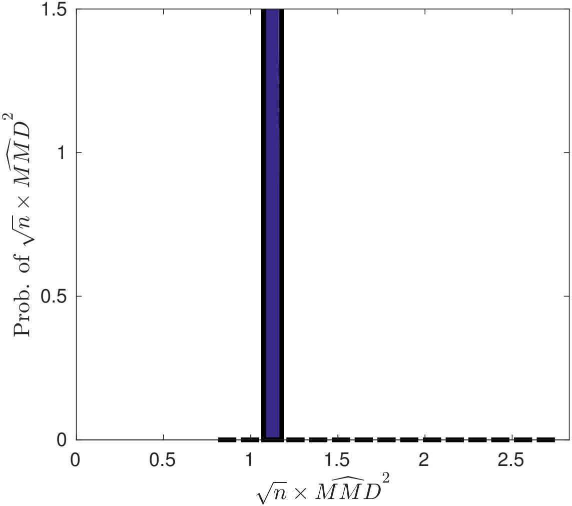
Draw $n = 200$ i.i.d samples from P and Q

- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.2$

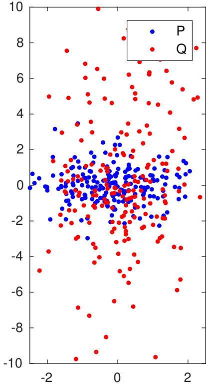


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Number of MMDs: 1



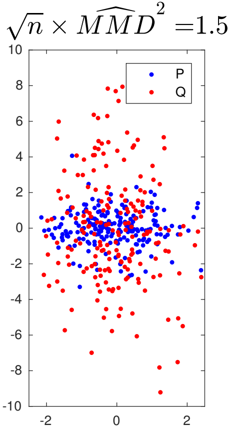
$\sqrt{n} \times \widehat{MMD}^2 = 1.2$



Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

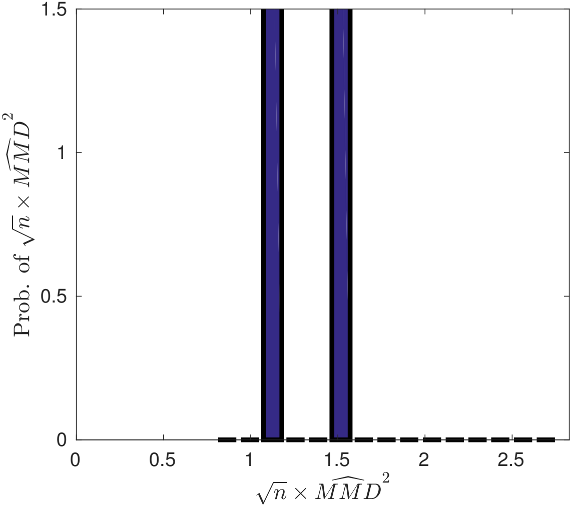
Draw $n = 200$ **new** samples from P and Q

- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.5$

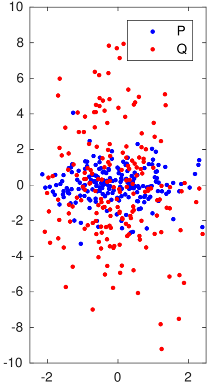


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Number of MMDs: 2

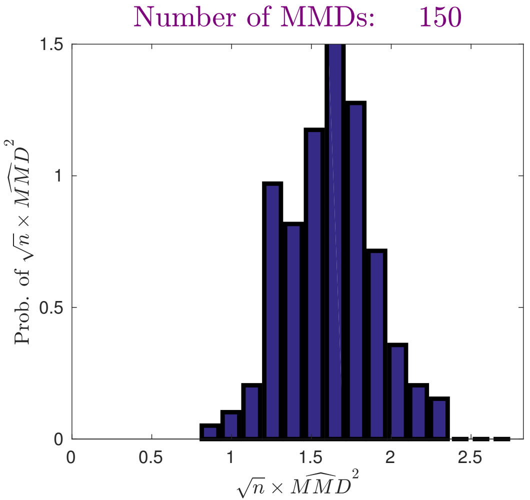


$\sqrt{n} \times \widehat{MMD}^2 = 1.5$



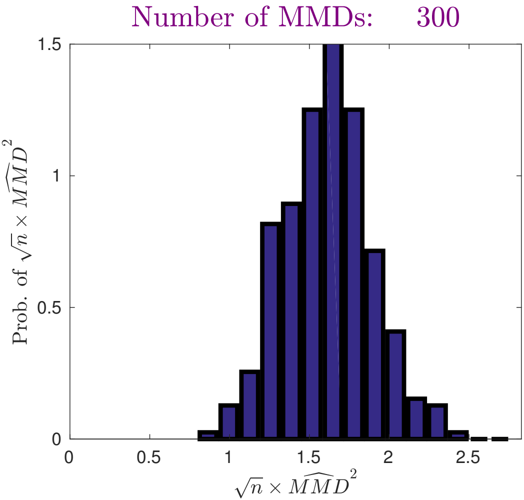
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 150 times . . .



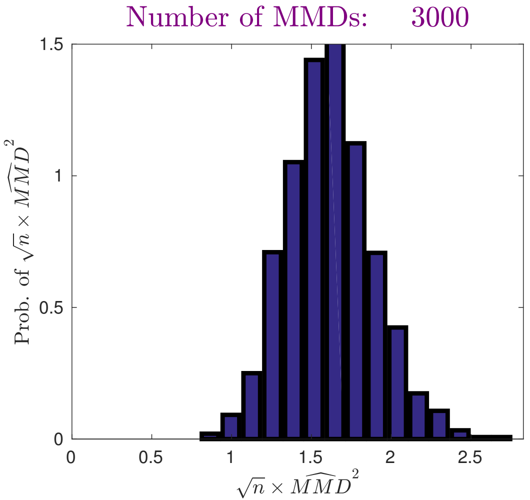
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 300 times ...



Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 3000 times ...

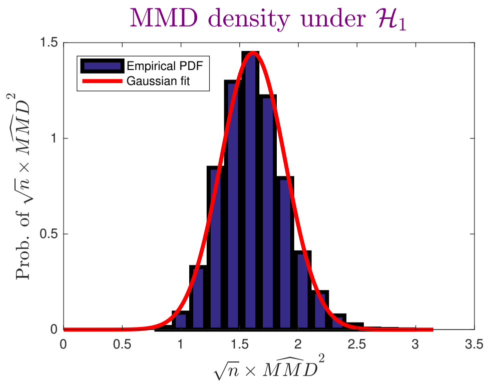


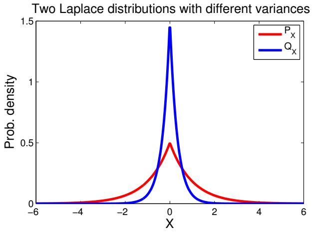
Asymptotics of $\widehat{MMD}^2$ when $P \neq Q$

When $P \neq Q$, statistic is asymptotically normal,

$$\frac{\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}(P, Q)}{\sqrt{V_n(P, Q)}} \xrightarrow{D} \mathcal{N}(0, 1),$$

where variance $V_n(P, Q) = O\left(n^{-1}\right)$.



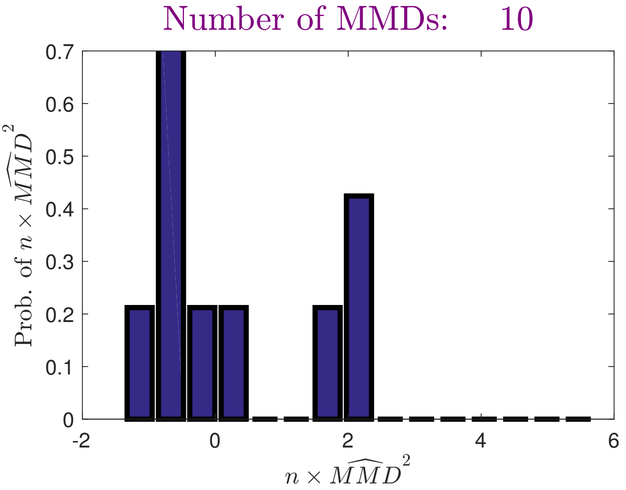


Behaviour of $\widehat{MMD}^2$ when $P = Q$

What happens when P and Q are the same?

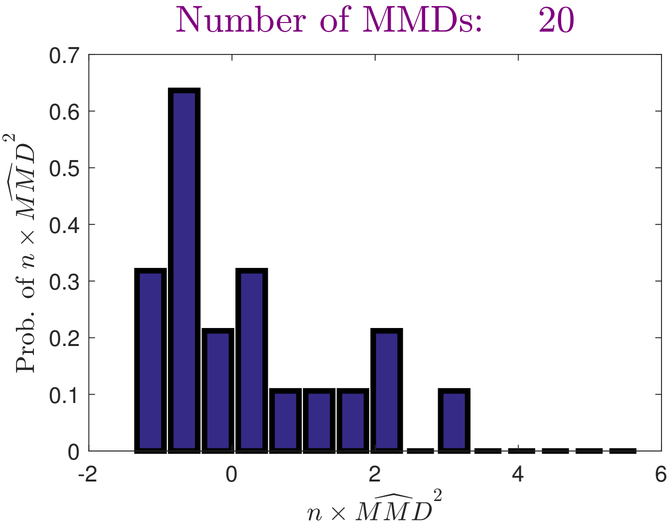
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0, 1)$



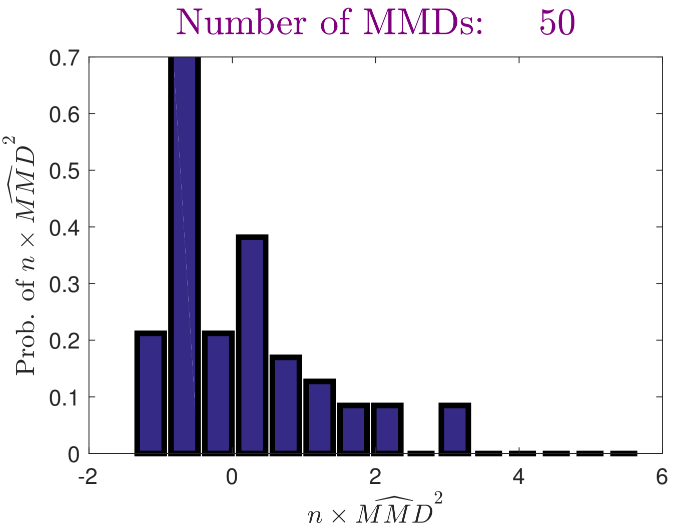
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0,1)$



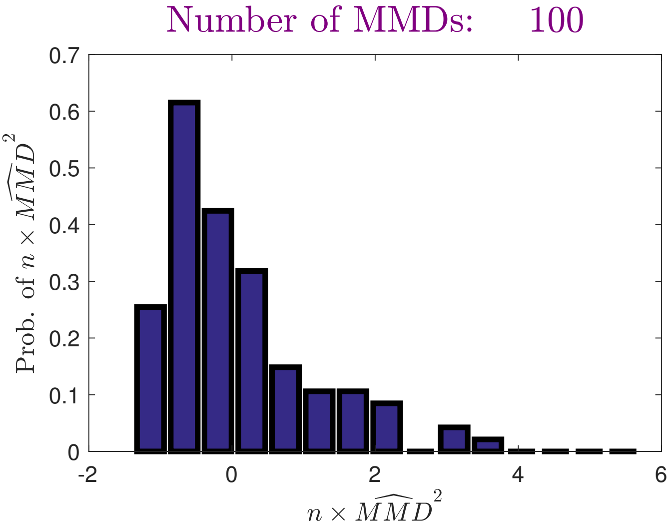
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0, 1)$



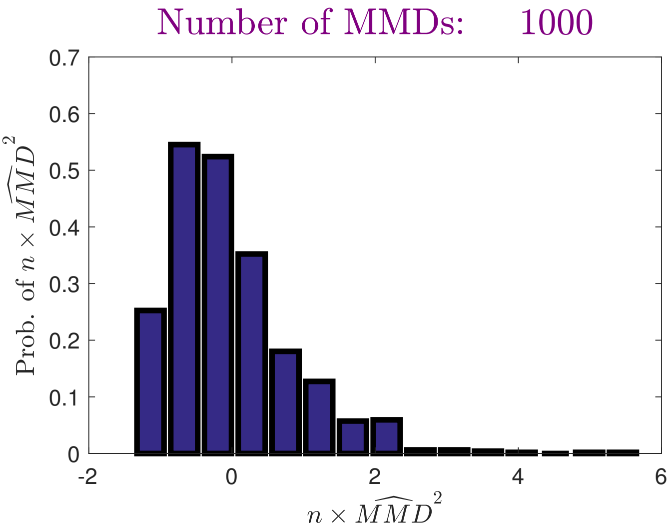
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0, 1)$



Behaviour of $\widehat{MMD}^2$ when $P = Q$

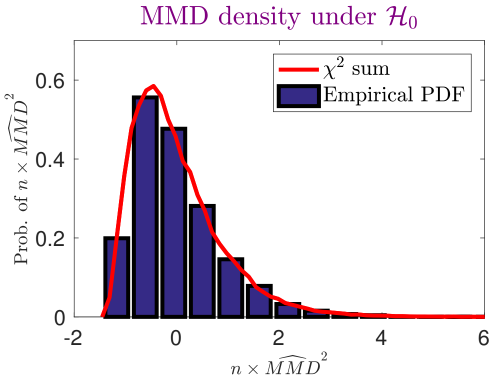
- Case of $P = Q = \mathcal{N}(0,1)$



Asymptotics of $\widehat{MMD}^2$ when $P = Q$

Where $P = Q$, statistic has asymptotic distribution

$$n\widehat{\mathrm{MMD}}^2 \sim \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2 \right]$$



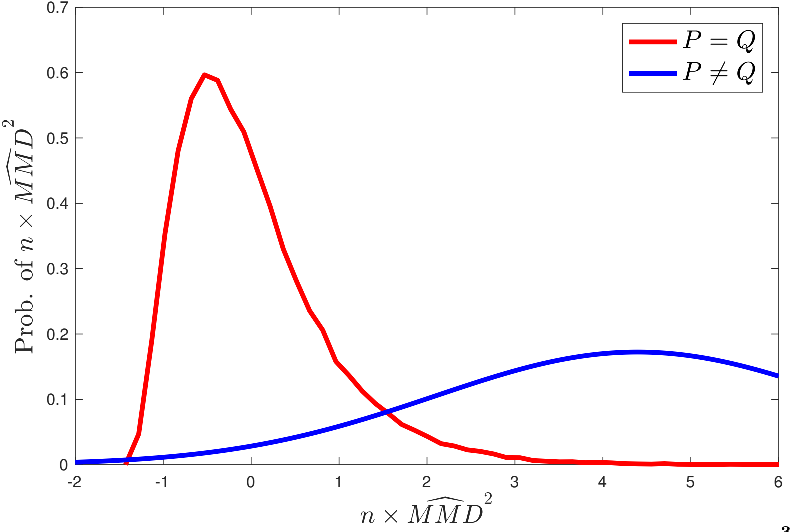
where

$$\lambda_i \psi_i(x') = \int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) dP(x)$$

$$z_l \sim \mathcal{N}(0, 2) \quad \text{i.i.d.}$$

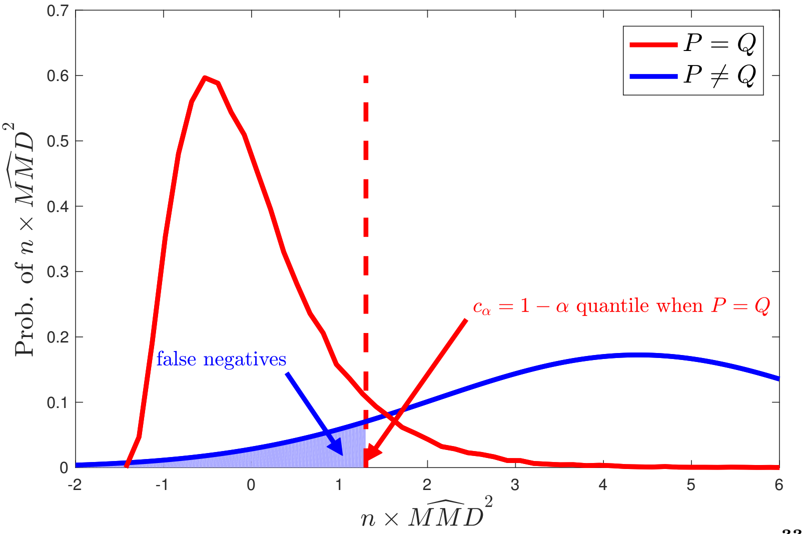
A statistical test

A summary of the asymptotics:



A statistical test

Test construction: (G., Borgwardt, Rasch, Schoelkopf, and Smola, JMLR 2012)



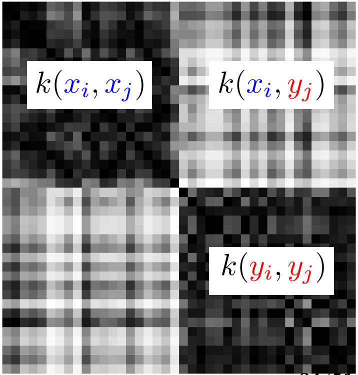
How do we get test threshold c_α?

Original empirical MMD for dogs and fish:

$X = [$  $\ldots]$

$Y = [$  $\ldots]$

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)$$



How do we get test threshold c_α?

Permuted dog and fish samples (**merdogs**):

$\widetilde{X} = [$  $\dots]$

$\widetilde{Y} = [$  $\dots]$

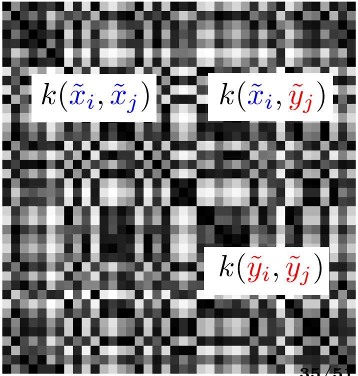
How do we get test threshold c_α?

Permuted dog and fish samples (**merdogs**):

$\widetilde{X} = [$  $\ldots\ ]$

$\widetilde{Y} = [$  $\ldots\ ]$

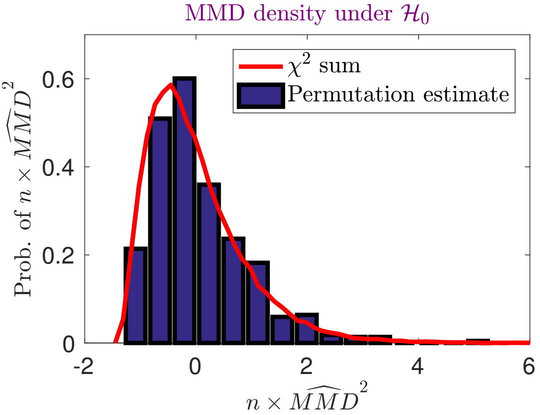
$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{x}_i, \tilde{x}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{y}_i, \tilde{y}_j) - \frac{2}{n^2} \sum_{i,j} k(\tilde{x}_i, \tilde{y}_j)$$



Permutation simulates
$P = Q$

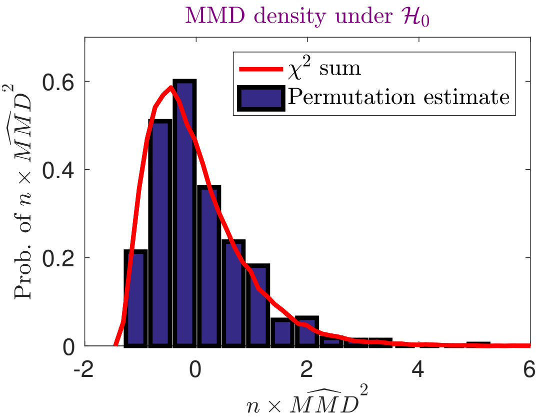
Demonstration of permutation estimate of null

- Null distribution estimated from 500 permutations
- $P = Q = \mathcal{N}(0, 1)$



Demonstration of permutation estimate of null

- Null distribution estimated from 500 permutations
- $P = Q = \mathcal{N}(0,1)$



Use $1-\alpha$ quantile of permutation distribution for test threshold c_α

How to choose the best kernel

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$
$$\to \Phi\left(\frac{n\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{\sqrt{V_n(P,Q)}}\right)$$

where

- Φ is the CDF of the standard normal distribution.
- $\hat{c}_\alpha$ is an estimate of c_α test threshold.

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$
$$\to \Phi\left(\underbrace{\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}}_{O(n^{1/2})} - \underbrace{\frac{c_\alpha}{n\sqrt{V_n(P,Q)}}}_{O(n^{-1/2})}\right)$$

Variance under $\mathcal{H}_1$ decreases as $\sqrt{V_n(P,Q)} \sim O(n^{-1/2})$

For large n, second term negligible!

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right) \to \Phi\left(\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P,Q)}}\right)$$

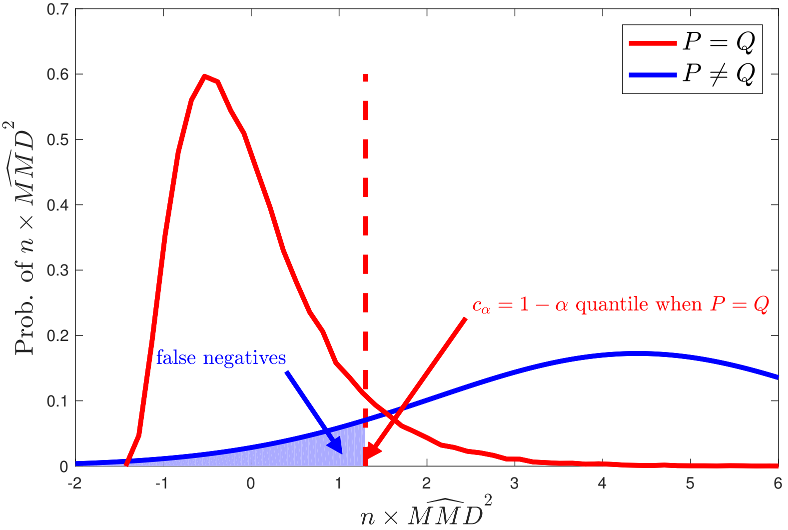
To maximize test power, maximize

$$\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}$$

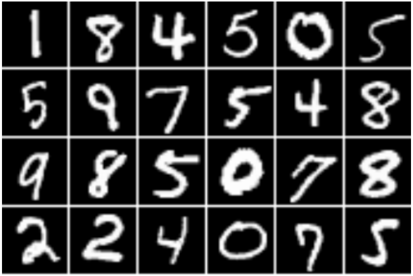
(Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017)
Code: github.com/dougalsutherland/opt-mmd

Graphical illustration

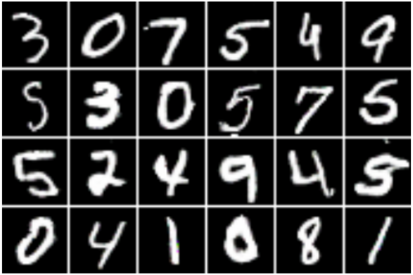
- Reminder: maximising test power same as minimizing false negatives



Troubleshooting for generative adversarial networks



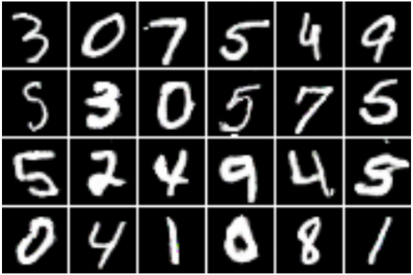
MNIST samples



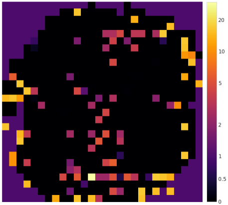
Samples from a GAN

Troubleshooting for generative adversarial networks

MNIST samples



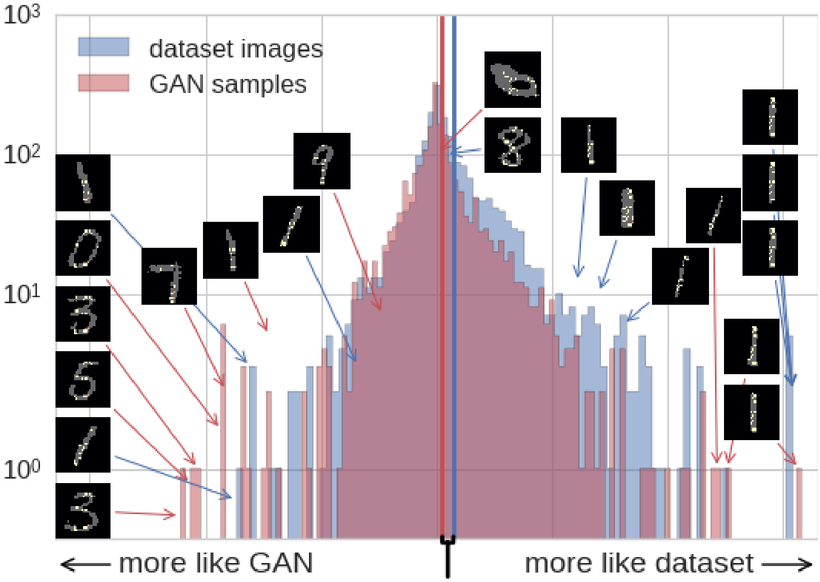
Samples from a GAN



ARD map

- Power for **optimzed ARD kernel**: 1.00 at $\alpha = 0.01$
- Power for optimized RBF kernel: 0.57 at $\alpha = 0.01$

Troubleshooting generative adversarial networks



$\mathrm{MMD}^2 = 0.0001$

MMD for GAN critic

Can you use MMD as a critic to train GANs?

Can you train convolutional features as input to the MMD critic?

From ICML 2015:

Generative Moment Matching Networks

Yujia Li[1] YUJIALI@CS.TORONTO.EDU
Kevin Swersky[1] KSWERSKY@CS.TORONTO.EDU
Richard Zemel[1,2] ZEMEL@CS.TORONTO.EDU
[1]Department of Computer Science, University of Toronto, Toronto, ON, CANADA
[2]Canadian Institute for Advanced Research, Toronto, ON, CANADA

From UAI 2015:

Training generative neural networks via Maximum Mean Discrepancy optimization

Gintare Karolina Dziugaite
University of Cambridge

Daniel M. Roy
University of Toronto

Zoubin Ghahramani
University of Cambridge

MMD for GAN critic: 2017 update

Wasserstein

arXiv.org > stat > arXiv:1701.07875

Statistics > Machine Learning

Wasserstein GAN ICML 2017

Martin Arjovsky, Soumith Chintala, Léon Bottou

arXiv.org > cs > arXiv:1704.00028

Computer Science > Learning

Improved Training of Wasserstein GANs

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville

(Submitted on 31 Mar 2017 (v1), last revised 29 May 2017 (this version, v2))

arXiv.org > stat > arXiv:1507.01972

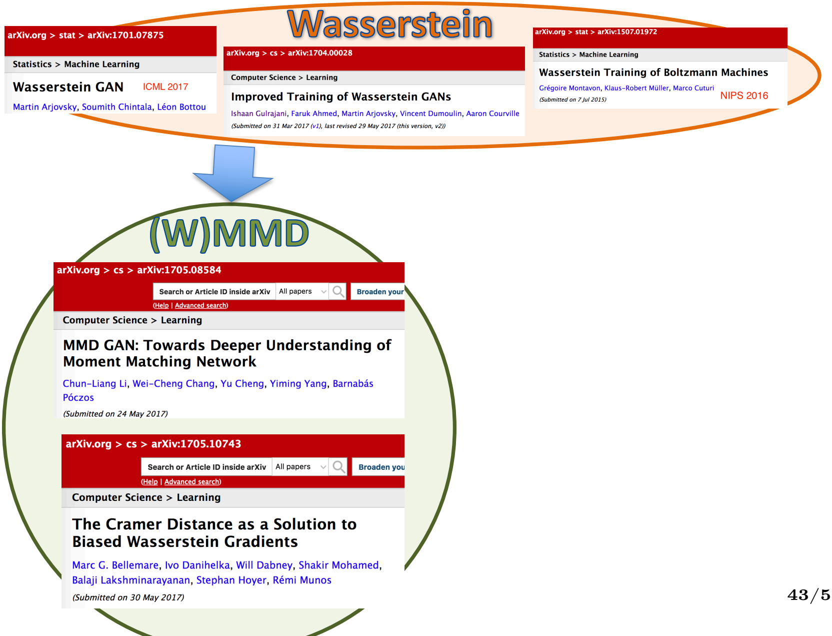
Statistics > Machine Learning

Wasserstein Training of Boltzmann Machines

Grégoire Montavon, Klaus-Robert Müller, Marco Cuturi NIPS 2016

(Submitted on 7 Jul 2015)

MMD for GAN critic: 2017 update



MMD for GAN critic: 2017 update

Wasserstein

arXiv.org > stat > arXiv:1701.07875

Statistics > Machine Learning

Wasserstein GAN ICML 2017

Martin Arjovsky, Soumith Chintala, Léon Bottou

arXiv.org > cs > arXiv:1704.00028

Computer Science > Learning

Improved Training of Wasserstein GANs

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville

(Submitted on 31 Mar 2017 (v1), last revised 29 May 2017 (this version, v2))

arXiv.org > stat > arXiv:1507.01972

Statistics > Machine Learning

Wasserstein Training of Boltzmann Machines

Grégoire Montavon, Klaus-Robert Müller, Marco Cuturi NIPS 2016

(Submitted on 7 Jul 2015)

(W)MMD

arXiv.org > cs > arXiv:1705.08584

Search or Article ID inside arXiv | All papers | Broaden your

(Help | Advanced search)

Computer Science > Learning

MMD GAN: Towards Deeper Understanding of Moment Matching Network

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, Barnabás Póczos

(Submitted on 24 May 2017)

arXiv.org > cs > arXiv:1705.10743

Search or Article ID inside arXiv | All papers | Broaden you

(Help | Advanced search)

Computer Science > Learning

The Cramer Distance as a Solution to Biased Wasserstein Gradients

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, Rémi Munos

(Submitted on 30 May 2017)

"Variance Incorporating"

arXiv.org > cs > arXiv:1702.08398

Search or Article ID inside arXiv | All papers | Broa

(Help | Advanced search)

Computer Science > Learning

McGan: Mean and Covariance Feature Matching GAN ICML 2017

Youssef Mroueh, Tom Sercu, Vaibhava Goel

arXiv.org > cs > arXiv:1705.09675

Computer Science > Learning

Fisher GAN

Youssef Mroueh, Tom Sercu

(Submitted on 26 May 2017 (v1), last revised 1 Aug 2017 (this version, v2))

"Other"

arXiv.org > cs > arXiv:1706.09549

Search or Article ID inside arXiv | All papers | Broaden

(Help | Advanced search)

Computer Science > Learning

Distributional Adversarial Networks

Chengtao Li, David Alvarez-Melis, Keyulu Xu, Stefanie Jegelka, Suvrit Sra

(Submitted on 29 Jun 2017 (v1), last revised 9 Jul 2017 (this version, v3))

MMD for GAN critic: 2017 update

Wasserstein

arXiv.org > stat > arXiv:1701.07875
Statistics > Machine Learning
Wasserstein GAN ICML 2017
Martin Arjovsky, Soumith Chintala, Léon Bottou

arXiv.org > cs > arXiv:1704.00028
Computer Science > Learning
Improved Training of Wasserstein GANs
Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville
(Submitted on 31 Mar 2017 (v1), last revised 29 May 2017 (this version, v2))

arXiv.org > stat > arXiv:1507.01972
Statistics > Machine Learning
Wasserstein Training of Boltzmann Machines
Grégoire Montavon, Klaus-Robert Müller, Marco Cuturi
(Submitted on 7 Jul 2015) NIPS 2016

(W)MMD

arXiv.org > cs > arXiv:1705.08584
Search or Article ID inside arXiv | All papers | Broaden your
(Help | Advanced search)
Computer Science > Learning
MMD GAN: Towards Deeper Understanding of Moment Matching Network
Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, Barnabás Póczos
(Submitted on 24 May 2017)

arXiv.org > cs > arXiv:1705.10743
Search or Article ID inside arXiv | All pa
(Help | Advanced search)
Computer Science > Learning
The Cramer Distance as a Sol Biased Wasserstein Gra
Marc G. Bellemare, Ivo Danihelka, Will Dabney, Balaji Lakshminarayanan, Stephan Hoyer, Rémi Munos
(Submitted on 30 May 2017)

On the Two-Sample Statistic Approach to Generative Adversarial Networks
"Energy Distance Kernel"
Lydia Tingruo Liu
Advisor: Professor Han Liu

"Variance Incorporating"

arXiv.org > cs > arXiv:1702.08398
Search or Article ID inside arXiv | All papers | Broa
(Help | Advanced search)
Computer Science > Learning
McGan: Mean and Covariance Feature Matching GAN ICML 2017
Youssef Mroueh, Tom Sercu, Vaibhava Goel

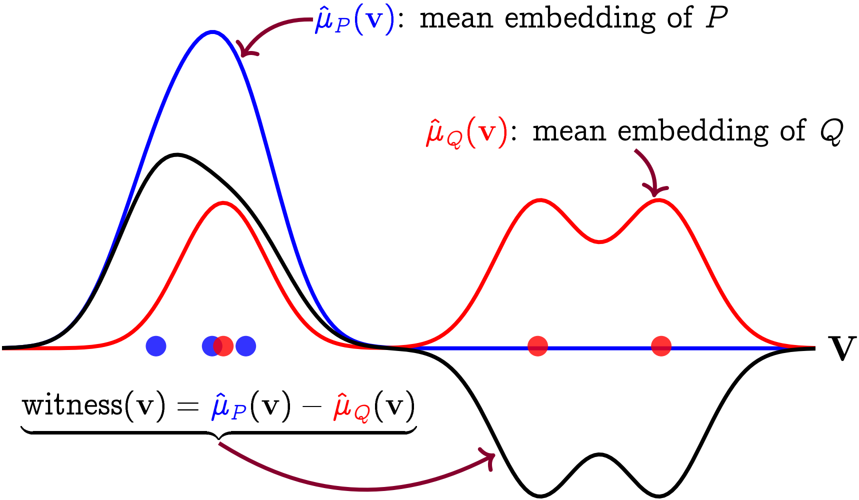
arXiv.org > cs > arXiv:1705.09675
Computer Science > Learning
Fisher GAN
Youssef Mroueh, Tom Sercu
(Submitted on 26 May 2017 (v1), last revised 1 Aug 2017 (this version, v2))

"Other"

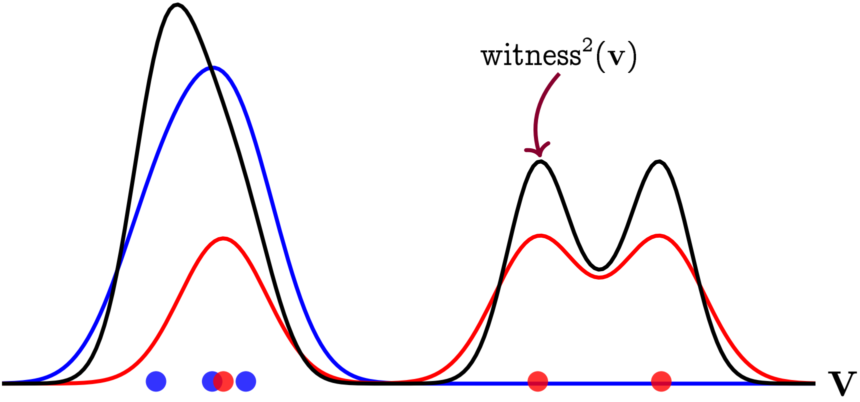
cs > arXiv:1706.09549
Search or Article ID inside arXiv | All papers | Broaden
(Help | Advanced search)
Computer Science > Learning
Distributional Adversarial Networks
Chengtao Li, David Alvarez-Melis, Keyulu Xu, Stefanie Jegelka, Suvrit Sra
(Submitted on 29 Jun 2017 (v1), last revised 9 Jul 2017 (this version, v3))

An adaptive, linear time distribution metric

Reminder: witness function for MMD

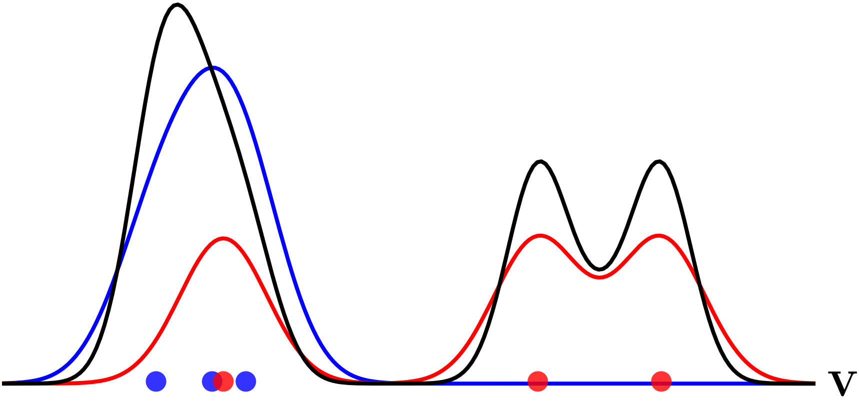


Distinguishing Feature(s)



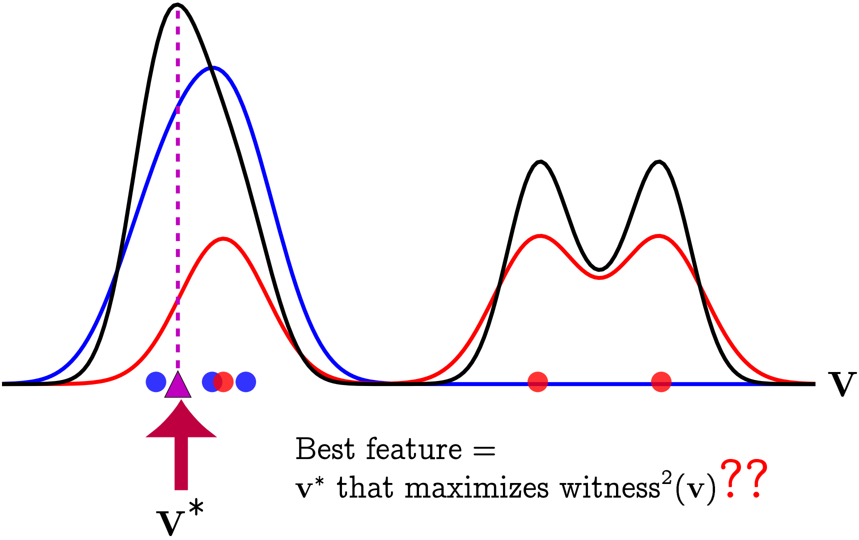
Take square of witness (only worry about amplitude)

Distinguishing Feature(s)



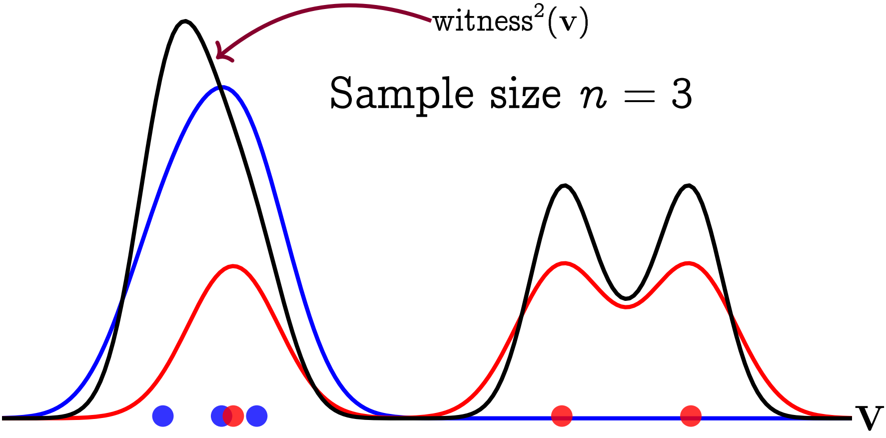
- New test statistic: witness^2 at a single $\mathbf{v}^*$;
- Linear time in number n of samples
- ....but how to choose best feature $\mathbf{v}^*$?

Distinguishing Feature(s)



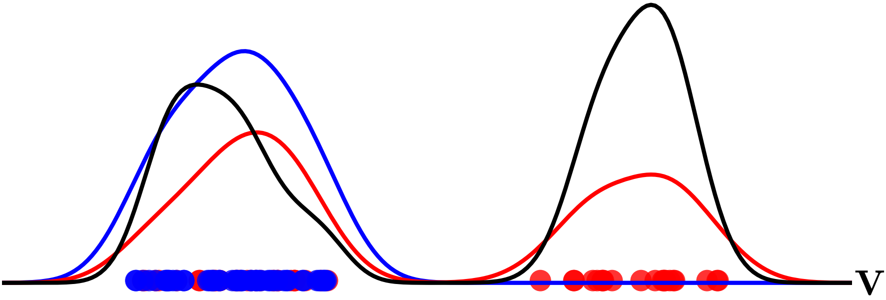
Best feature =
v^* that maximizes witness^2(v)??

Distinguishing Feature(s)

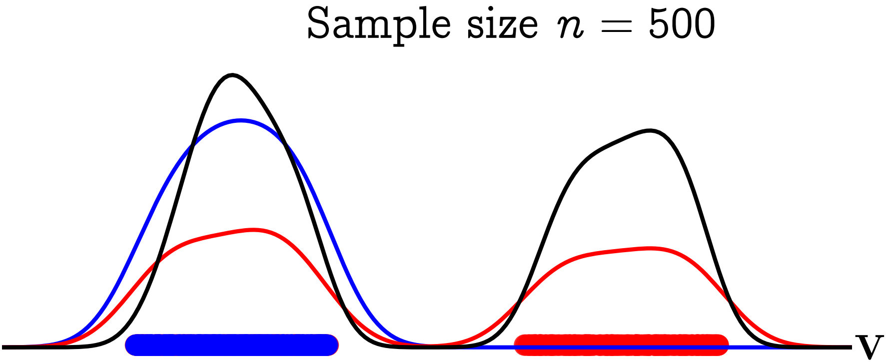


Distinguishing Feature(s)

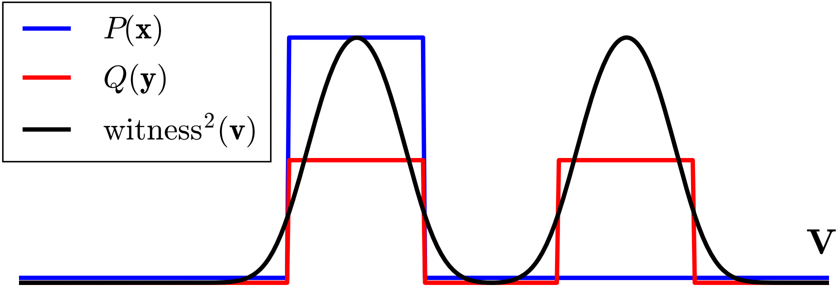
Sample size $n = 50$



Distinguishing Feature(s)

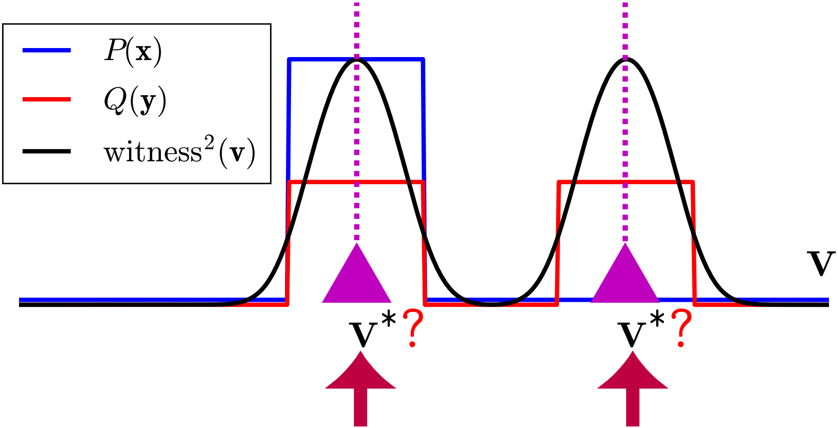


Distinguishing Feature(s)



Population witness^2 function

Distinguishing Feature(s)



Variance of witness function

- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016

Variance of witness function

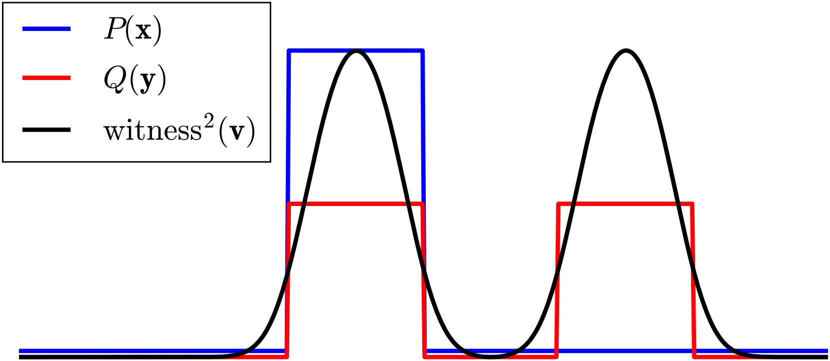
- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016

Variance of witness function

- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

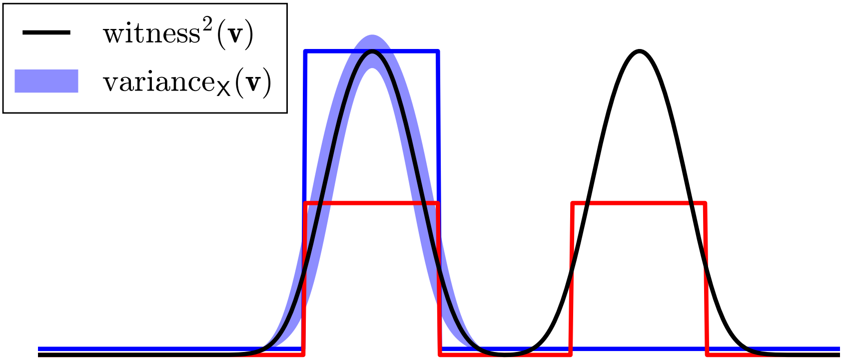
Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016



Variance of witness function

- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

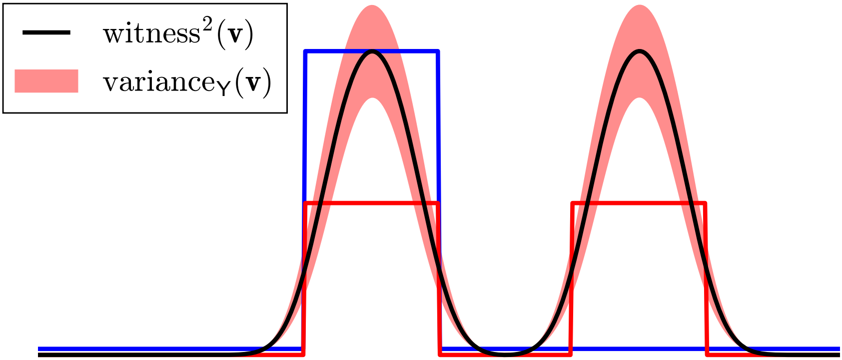
Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016



Variance of witness function

- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

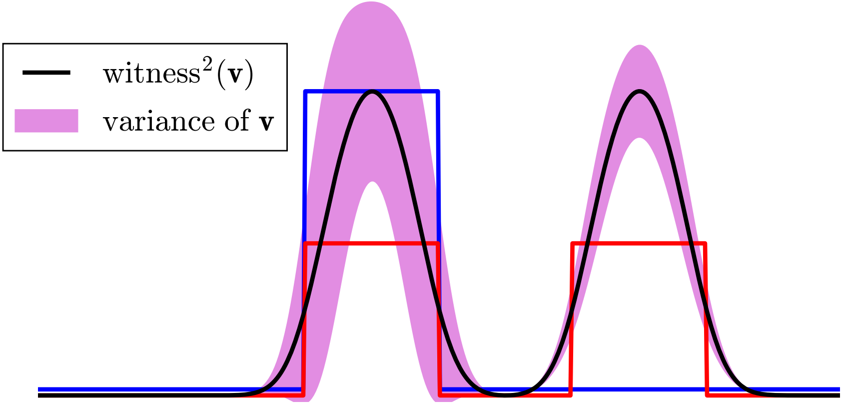
Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016



Variance of witness function

- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

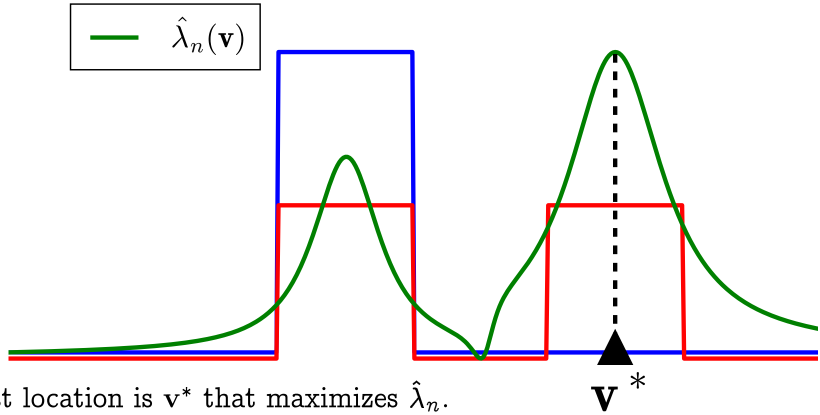
Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016



Variance of witness function

- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016



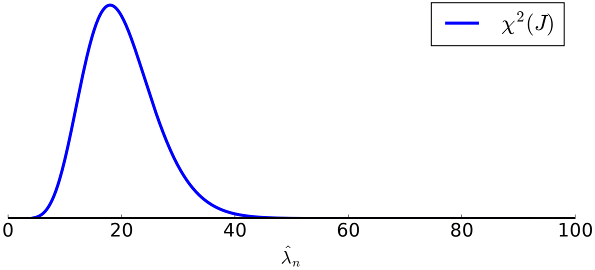
- Best location is $\mathbf{v}^*$ that maximizes $\hat{\lambda}_n$.
- Improve performance using multiple locations $\{\mathbf{v}_j^*\}_{j=1}^J$

Properties of the ME Test

- Can use J features $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1-\alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting H_0 when H_1 is true.

Properties of the ME Test

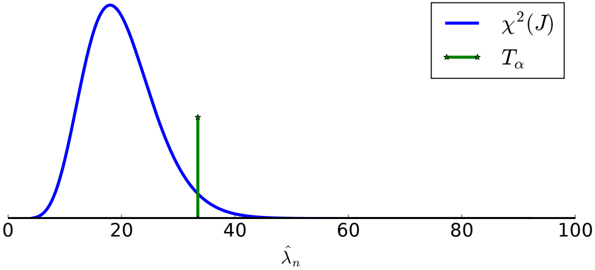
- Can use J features $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1-\alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting H_0 when H_1 is true.



- Runtime: $\mathcal{O}(n)$ for both testing and optimization.

Properties of the ME Test

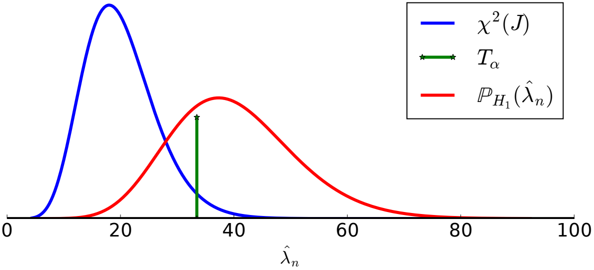
- Can use J features $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1-\alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting H_0 when H_1 is true.



- Runtime: $\mathcal{O}(n)$ for both testing and optimization.

Properties of the ME Test

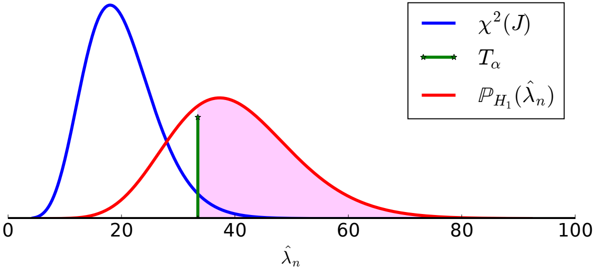
- Can use J features $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1-\alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting H_0 when H_1 is true.



- Runtime: $\mathcal{O}(n)$ for both testing and optimization.

Properties of the ME Test

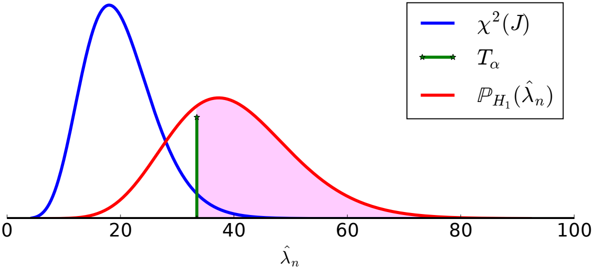
- Can use J features $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1-\alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting H_0 when H_1 is true.



- Runtime: $\mathcal{O}(n)$ for both testing and optimization.

Properties of the ME Test

- Can use J features $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1-\alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting H_0 when H_1 is true.

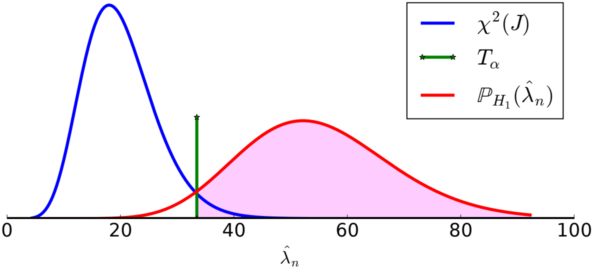


Theorem: Under H_1, optimization of $\mathcal{V}$ (by maximizing $\hat{\lambda}_n$) increases the (lower bound of) test power.

- Runtime: $\mathcal{O}(n)$ for both testing and optimization.

Properties of the ME Test

- Can use J features $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting H_0 when H_1 is true.

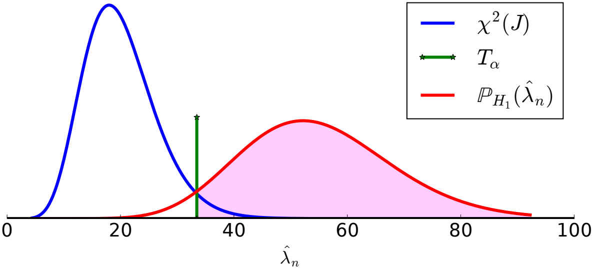


Theorem: Under H_1, optimization of $\mathcal{V}$ (by maximizing $\hat{\lambda}_n$) increases the (lower bound of) test power.

- Runtime: $\mathcal{O}(n)$ for both testing and optimization.

Properties of the ME Test

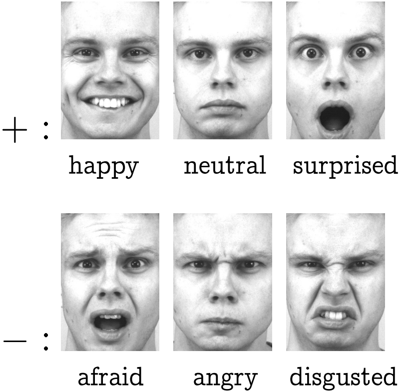
- Can use J features $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1-\alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting H_0 when H_1 is true.



Theorem: Under H_1, optimization of $\mathcal{V}$ (by maximizing $\hat{\lambda}_n$) increases the (lower bound of) test power.

- Runtime: $\mathcal{O}(n)$ for both testing and optimization.

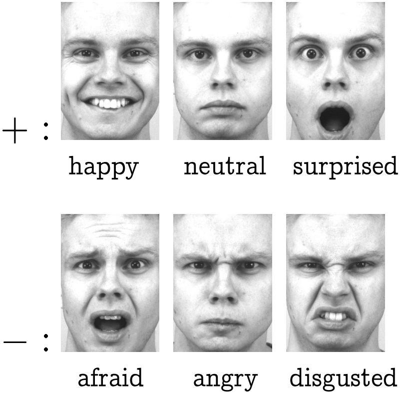
Distinguishing Positive/Negative Emotions

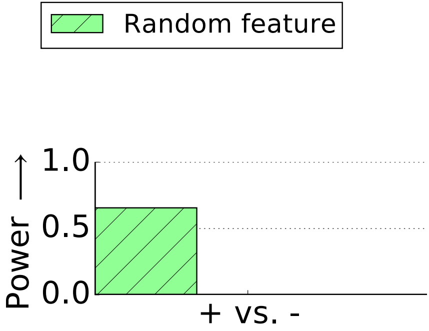


- 35 females and 35 males (Lundqvist et al., 1998).
- $48 \times 34 = 1632$ dimensions. Pixel features.
- Sample size: 402.

- The proposed test achieves **maximum test power** in **time** $O(n)$.
- **Informative features**: differences at the nose, and smile lines.

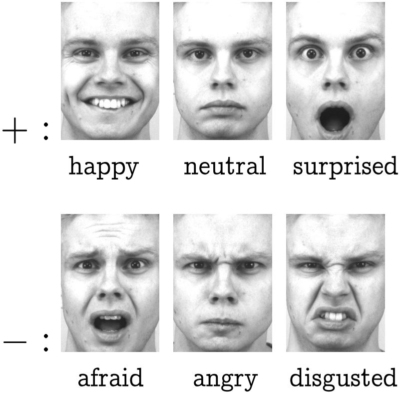
Distinguishing Positive/Negative Emotions

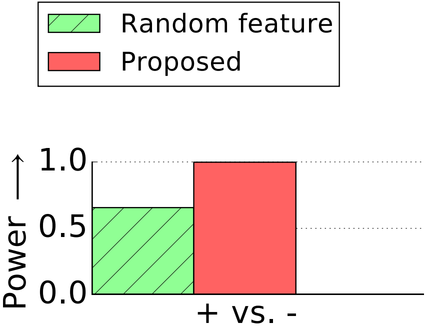




- The proposed test achieves **maximum test power** in **time** $O(n)$.
- **Informative features**: differences at the nose, and smile lines.

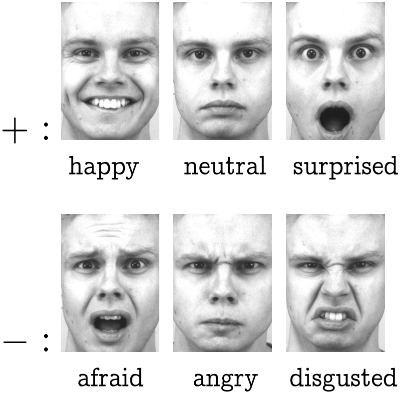
Distinguishing Positive/Negative Emotions

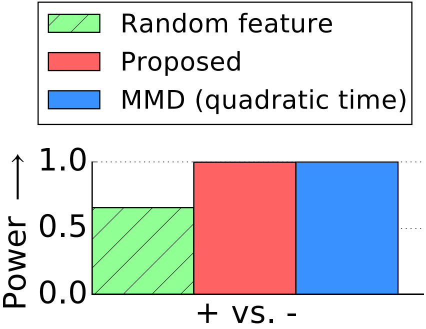




- The proposed test achieves **maximum test power** in **time $O(n)$**.
- Informative features: differences at the nose, and smile lines.

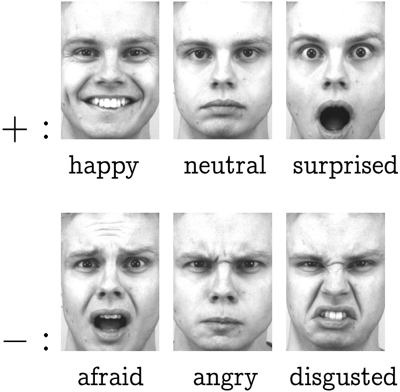
Distinguishing Positive/Negative Emotions

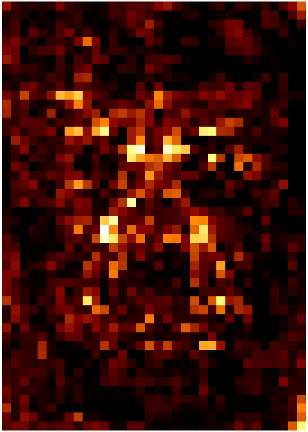




- The proposed test achieves **maximum test power** in **time $O(n)$**.
- Informative features: differences at the nose, and smile lines.

Distinguishing Positive/Negative Emotions

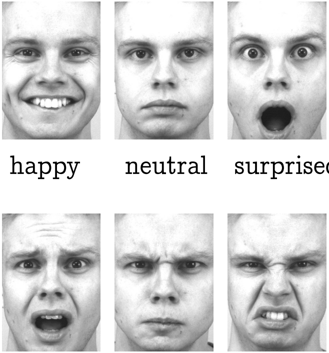




Learned feature

- The proposed test achieves **maximum test power** in **time $O(n)$**.
- **Informative features**: differences at the nose, and smile lines.

Distinguishing Positive/Negative Emotions



Learned feature

- The proposed test achieves **maximum test power** in **time $O(n)$**.
- **Informative features**: differences at the nose, and smile lines.

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016

Code: https://github.com/wittawatj/interpretable-test

Co-authors

From Gatsby:

- Kacper Chwialkowski
- Wittawat Jitkrittum
- Bharath Sriperumbudur
- Heiko Strathmann
- Dougal Sutherland
- Zoltan Szabo
- Wenkai Xu

External collaborators:

- Kenji Fukumizu
- Bernhard Schoelkopf
- Alex Smola

Questions?