The maximum mean discrepancy and Generative Adversarial Networks

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

LOD, 2019

A motivation: comparing two samples

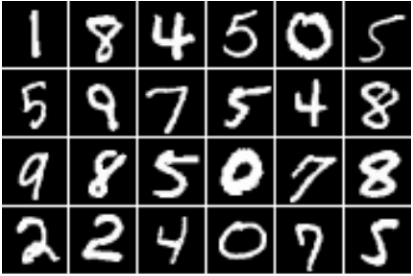
- **Given:** Samples from unknown distributions P and Q.
- **Goal:** do P and Q differ?

$$\sim P$$

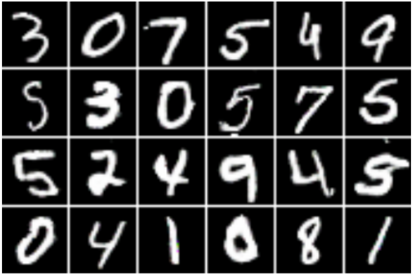
$$\sim Q$$

A real-life 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



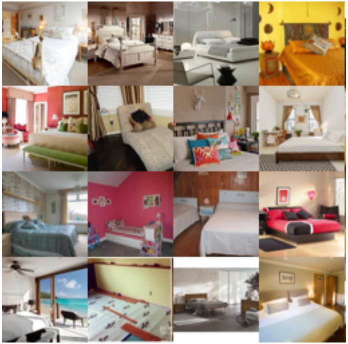
Samples from a GAN

Significant difference in GAN and MNIST?

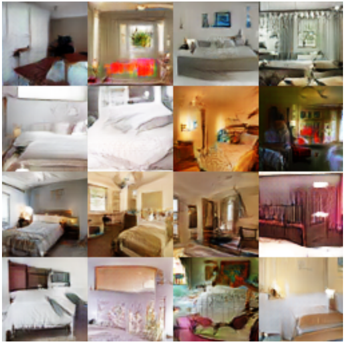
T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, Xi Chen, NeurIPS 2016
Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.

Training implicit generative models

- **Have:** One collection of samples X from unknown distribution P.
- **Goal:** generate samples Q that look like P



LSUN bedroom samples P

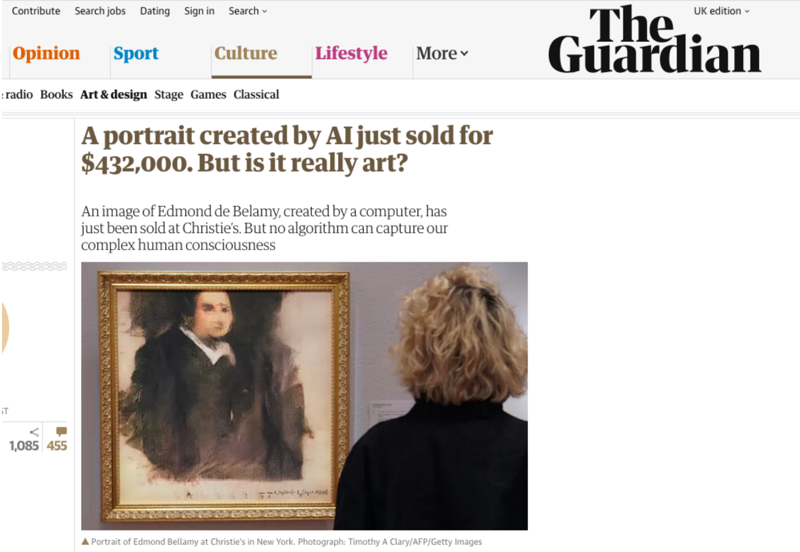


Generated Q, MMD GAN

Using a critic $D(P, Q)$ to train a GAN

(Binkowski, Sutherland, Arbel, G., ICLR 2018),
(Arbel, Sutherland, Binkowski, G., NeurIPS 2018)

Training generative models



A portrait created by AI just sold for $432,000. But is it really art?

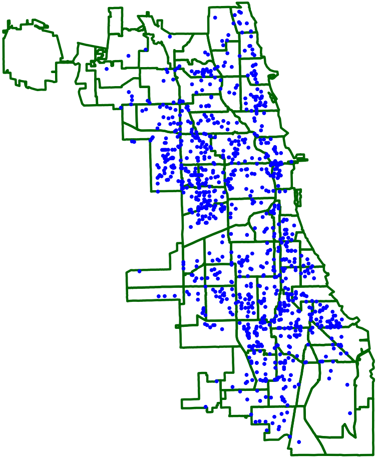
An image of Edmond de Belamy, created by a computer, has just been sold at Christie's. But no algorithm can capture our complex human consciousness

▲ Portrait of Edmond Bellamy at Christie's in New York. Photograph: Timothy A Clary/AFP/Getty Images

Testing goodness of fit

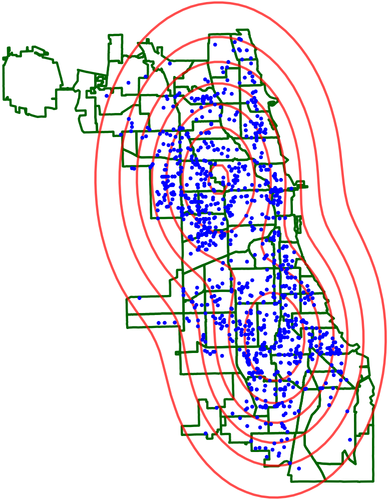
- **Given:** **A model** P and samples and Q.
- **Goal:** is P a good fit for Q?

Chicago crime data



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. Is this a good model?

Testing statistical dependence

- **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

- Measures of distance between distributions...
  - Difference in feature means
  - Integral probability metrics (not just a technicality!)

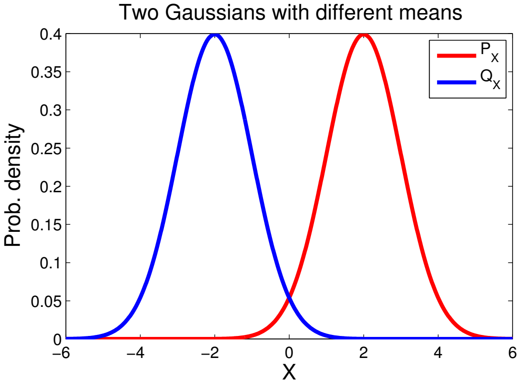
- Statistical testing to compare samples from P and Q

- GAN critic design (if time)
  - Gradient regularisation and data adaptivity

Differences in distributions

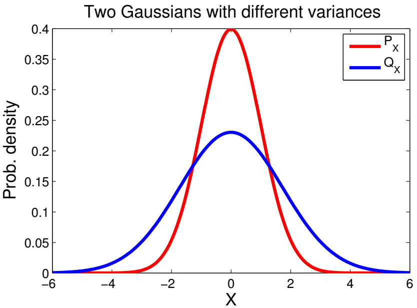
Feature mean difference

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



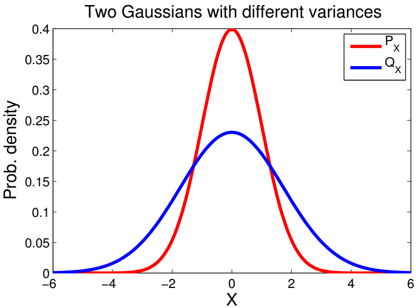
Feature mean difference

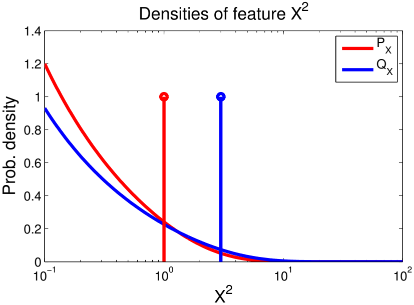
- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

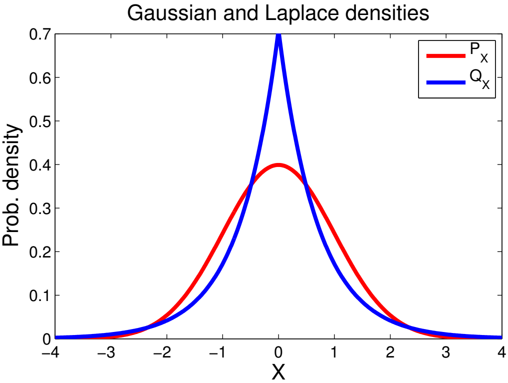
- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$





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,

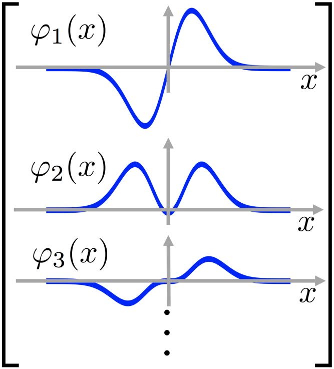
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

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,

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

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,

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

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**:

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

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 $k(\text{fish}_i, \text{fish}_j)$

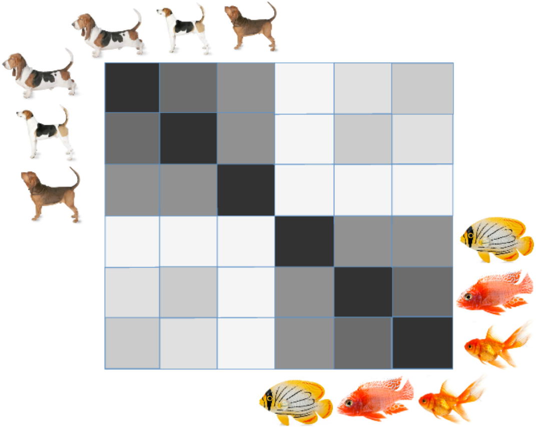
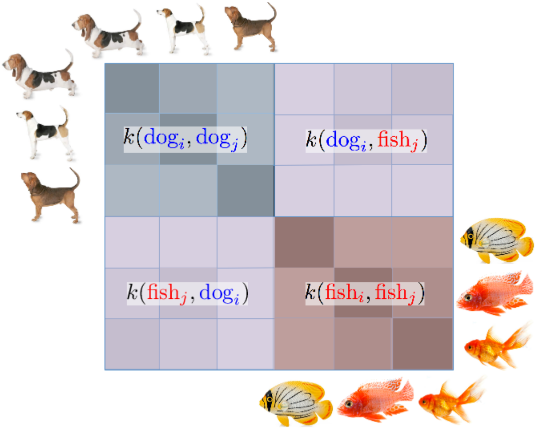


Illustration of MMD

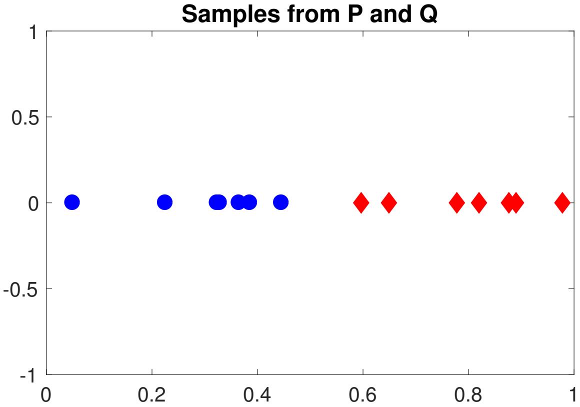
The maximum mean discrepancy:

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


$k(\text{dog}_i, \text{dog}_j)$
$k(\text{dog}_i, \text{fish}_j)$
$k(\text{fish}_j, \text{dog}_i)$
$k(\text{fish}_i, \text{fish}_j)$

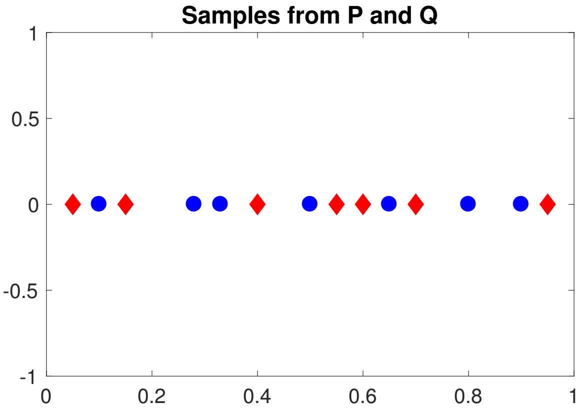
Integral probability metrics

Are P and Q different?



Integral probability metrics

Are P and Q different?

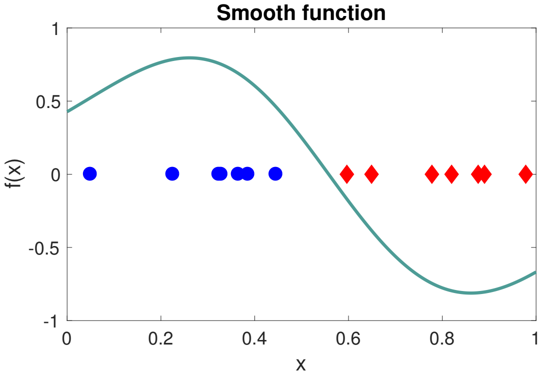


Integral probability metrics

Integral probability metric:

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

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

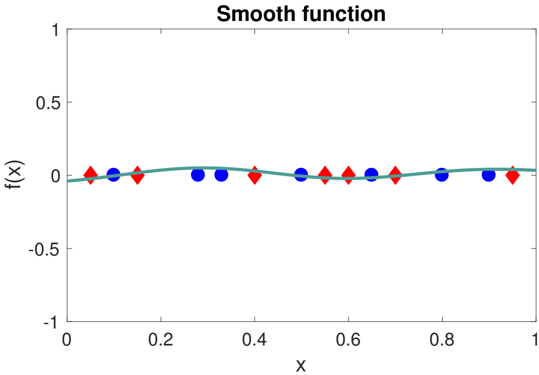


MMD as an integral probability metric

Integral probability metric:

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

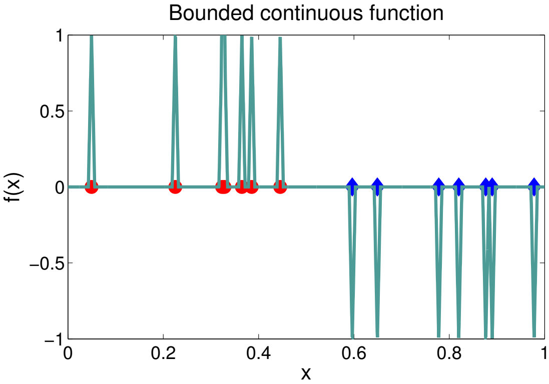
$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$



MMD as an integral probability metric

What if the function is not well behaved?

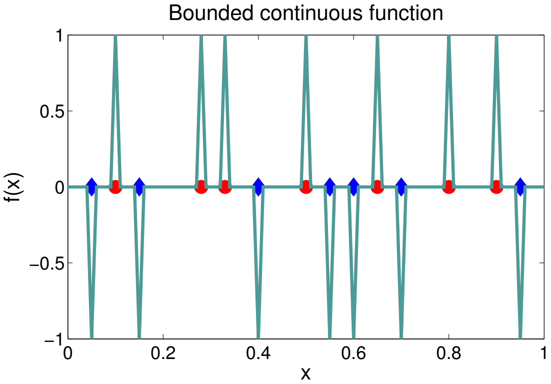
$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$



MMD as an integral probability metric

What if the function is not well behaved?

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



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

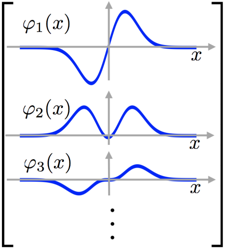
Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

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



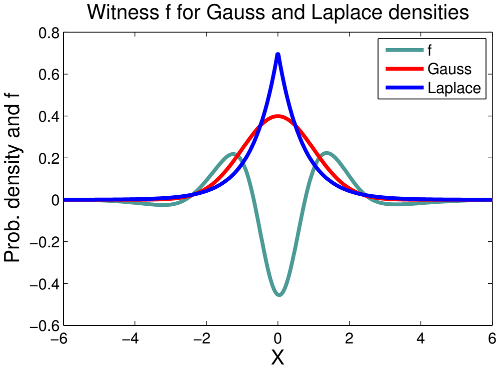
$$\|f\|_{\mathcal{F}}^2 := \sum_{i=1}^{\infty} f_i^{\,2} \leq 1$$

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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



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 $\mathcal{F})$

Expectations of functions are linear combinations of expected features

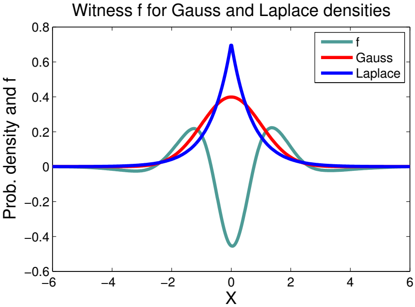
$$\mathbf{E}_P(f(X)) = \langle f, \mathbf{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

Integral prob. metric vs feature difference

The MMD:

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



Integral prob. metric vs feature difference

The MMD:

use

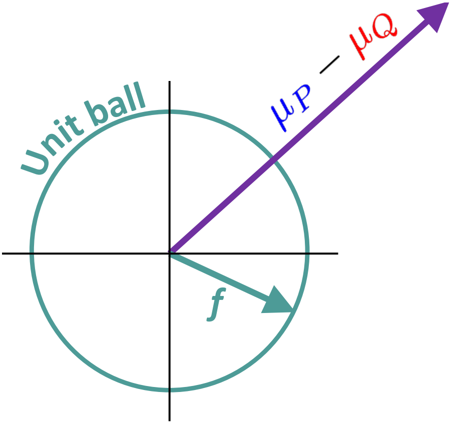
$$\begin{aligned} & MMD(P, Q; F) \\ & = \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\ & = \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \end{aligned}$$

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

Integral prob. metric vs feature difference

The MMD:

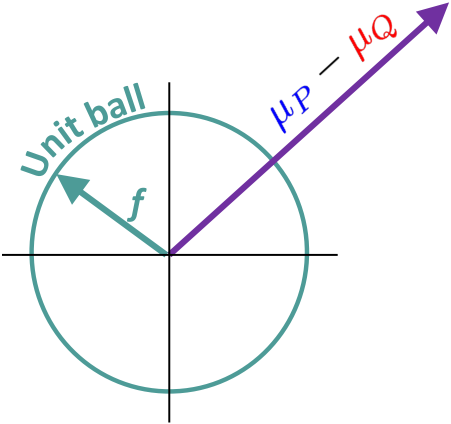
$$MMD(P, Q; F)$$
$$= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$
$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature difference

The MMD:

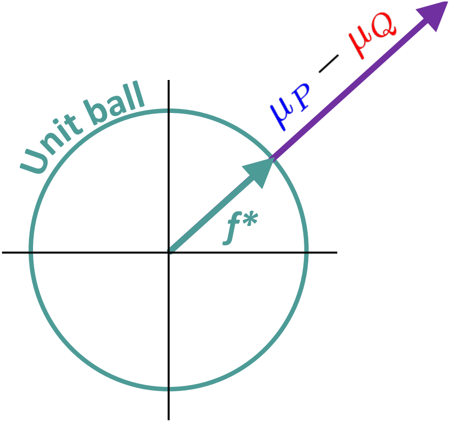
$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
&= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature difference

The MMD:

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



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

Integral prob. metric vs feature difference

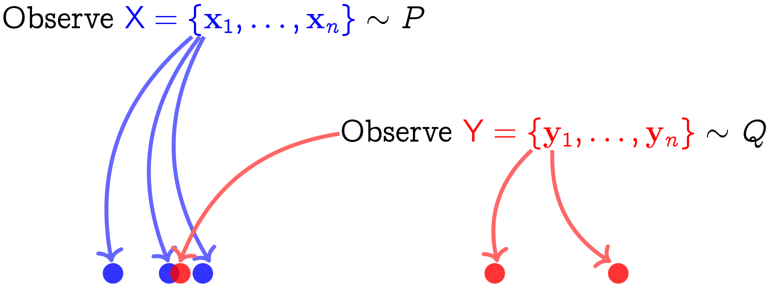
The MMD:

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

Function view and feature view equivalent
(kernel case only)

Construction of MMD witness

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

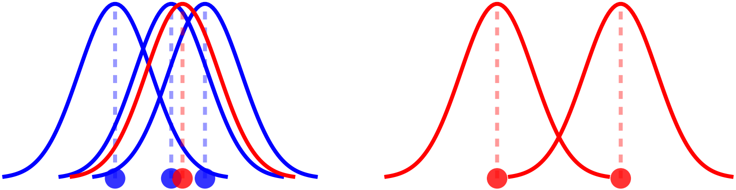


Observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$

Observe $\mathsf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim Q$

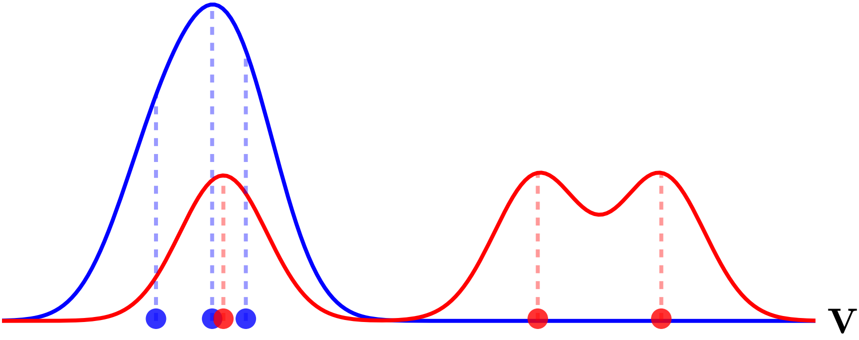
Construction of MMD witness

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



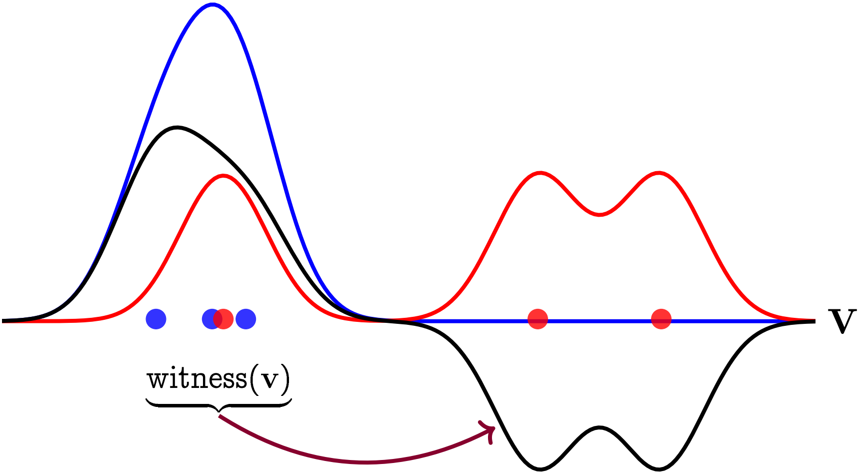
Construction of MMD witness

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

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

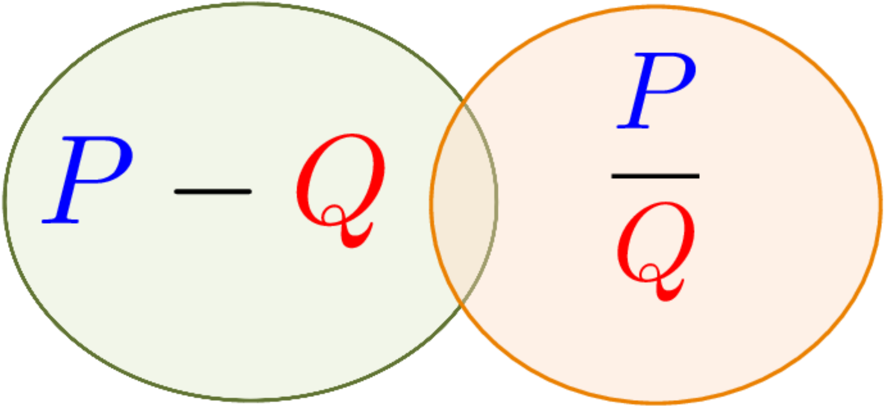
The empirical witness function at v

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

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

Interlude: divergence measures

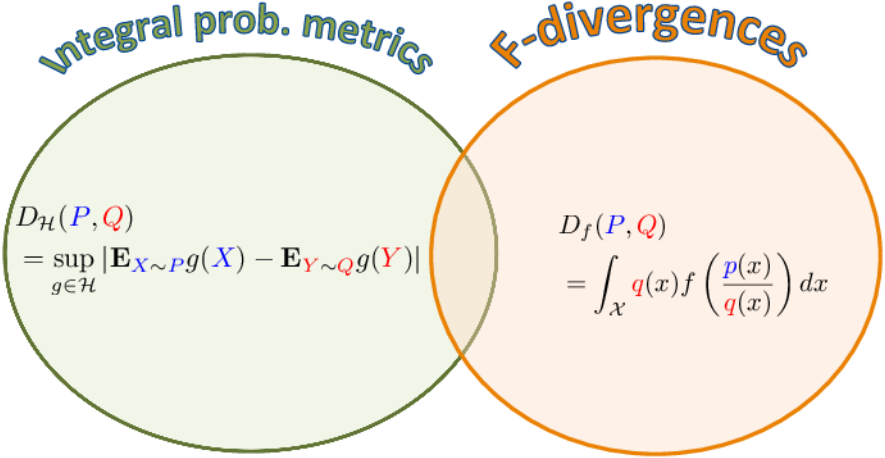
Divergences



$P - Q$

$$\frac{P}{Q}$$

Divergences



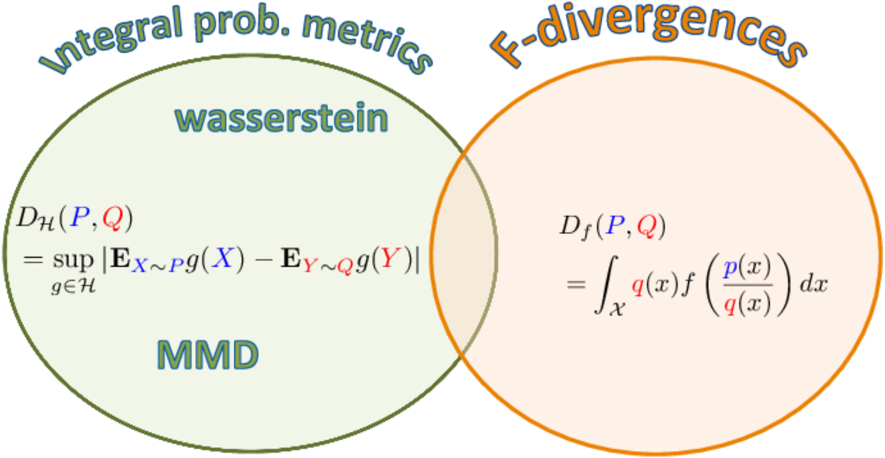
Integral prob. metrics

F-divergences

$$D_{\mathcal{H}}(P,Q) = \sup_{g \in \mathcal{H}} |\mathbf{E}_{X \sim P} g(X) - \mathbf{E}_{Y \sim Q} g(Y)|$$

$$D_f(P,Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

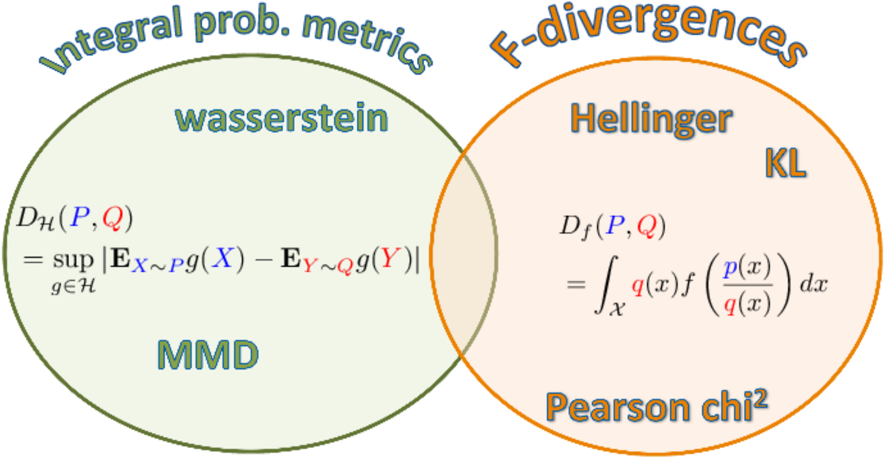
Divergences



$$D_{\mathcal{H}}(P, Q) = \sup_{g \in \mathcal{H}} |\mathbf{E}_{X \sim P} g(X) - \mathbf{E}_{Y \sim Q} g(Y)|$$

$$D_f(P, Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

Divergences



Integral prob. metrics

wasserstein

$$D_{\mathcal{H}}(P, Q) = \sup_{g \in \mathcal{H}} \left|\mathbf{E}_{X \sim P}\, g(X) - \mathbf{E}_{Y \sim Q}\, g(Y)\right|$$

MMD

F-divergences

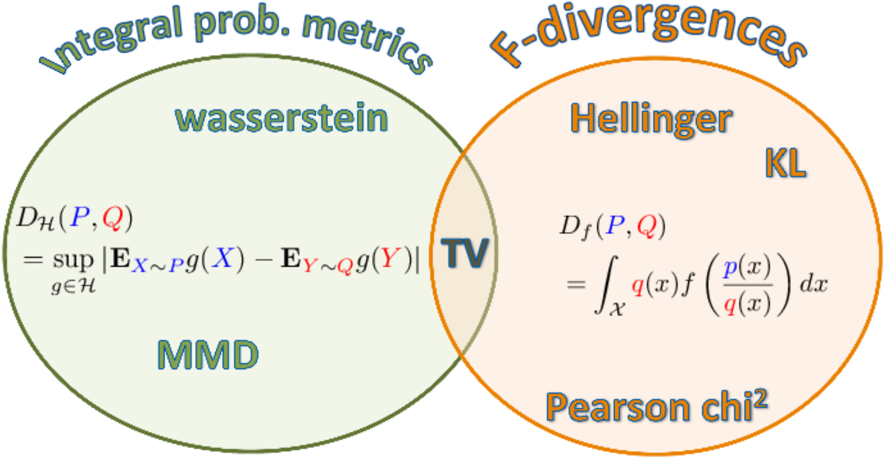
Hellinger

KL

$$D_f(P, Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

Pearson chi²

Divergences


Integral prob. metrics
F-divergences
wasserstein
Hellinger
KL
$D_{\mathcal{H}}(P,Q) = \sup_{g\in\mathcal{H}} |\mathbf{E}_{X\sim P}g(X) - \mathbf{E}_{Y\sim Q}g(Y)|$
TV
$D_f(P,Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$
MMD
Pearson chi²

Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (2012)

Two-Sample Testing with MMD

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:

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

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"

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"

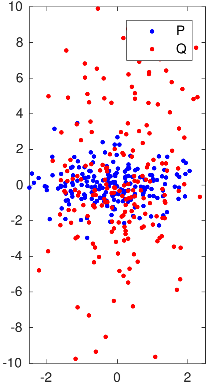
Want Threshold c_α for $\widehat{MMD}^2$ to get false positive rate α

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

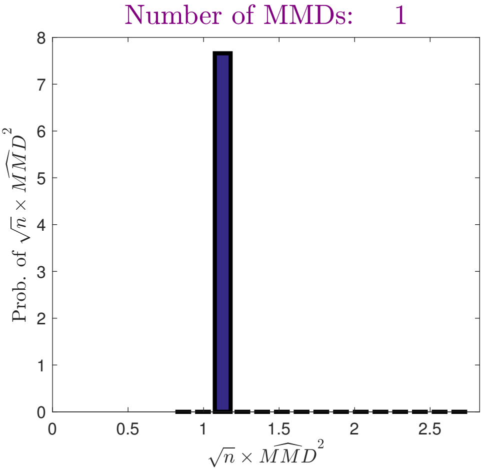
$\sqrt{n} \times \widehat{MMD}^2 = 1.2$

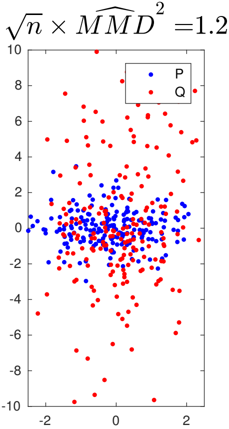


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

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

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

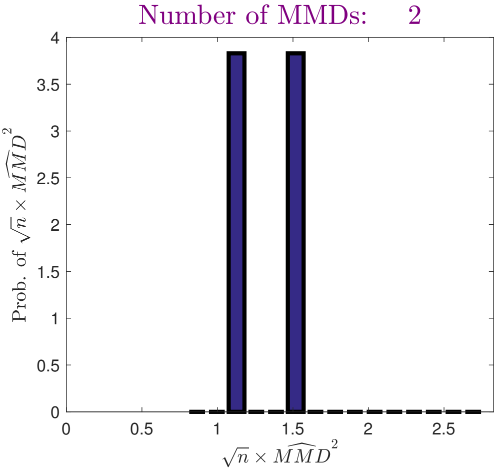


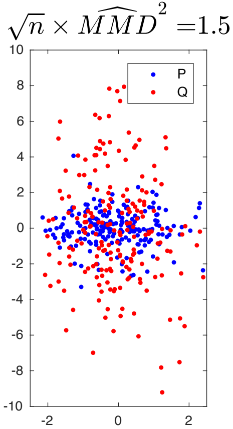


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

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

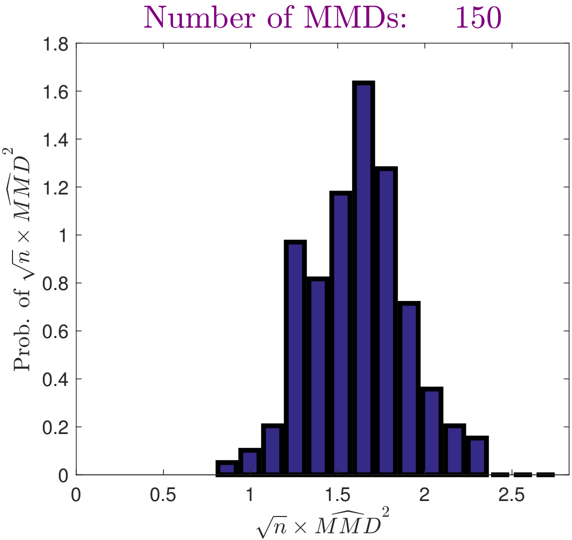
- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.5$





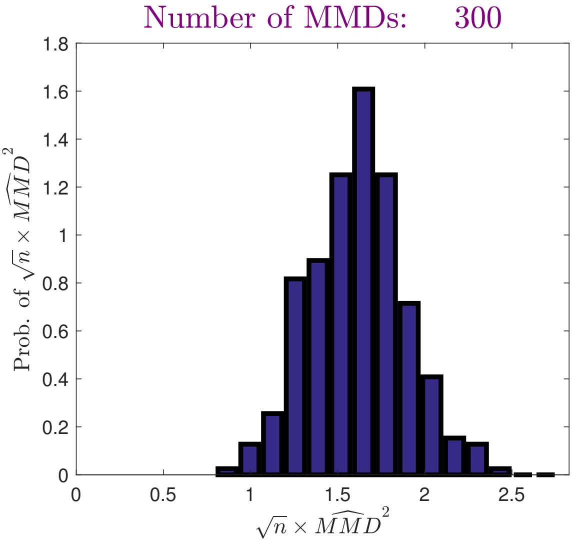
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 150 times ...



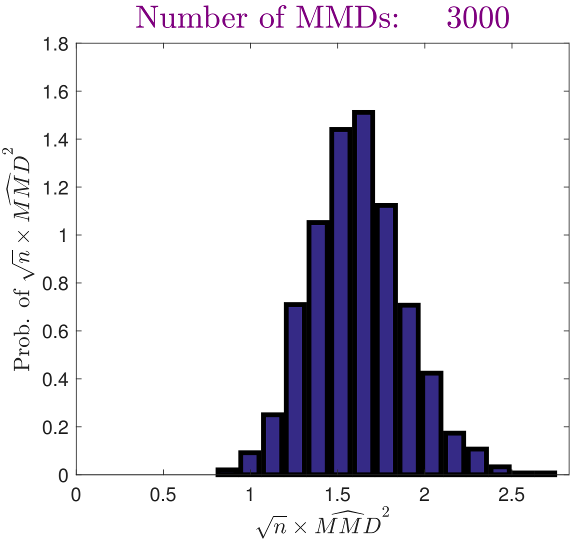
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 300 times . . .



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

Repeat this 3000 times ...

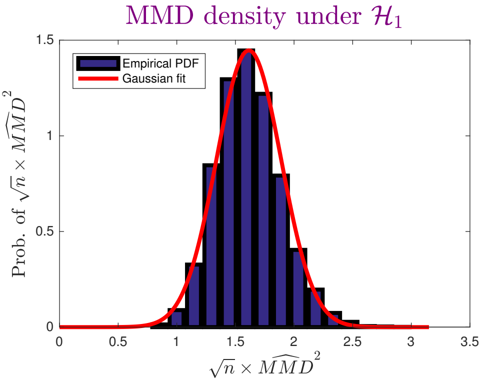


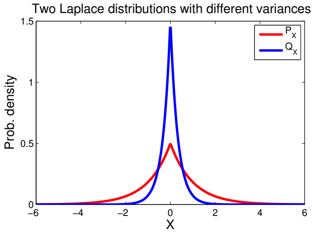
Asymptotics of $\widehat{MMD}^2$ when $P \neq Q$

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


MMD density under $\mathcal{H}_1$

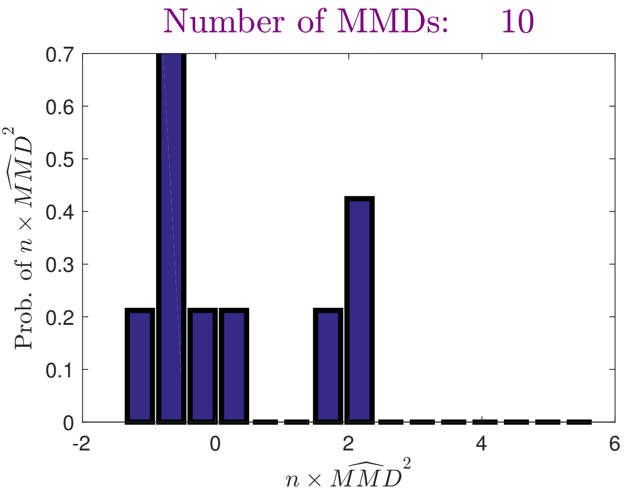


Behaviour of $\widehat{MMD}^2$ when $P = Q$

What happens when P and Q are the same?

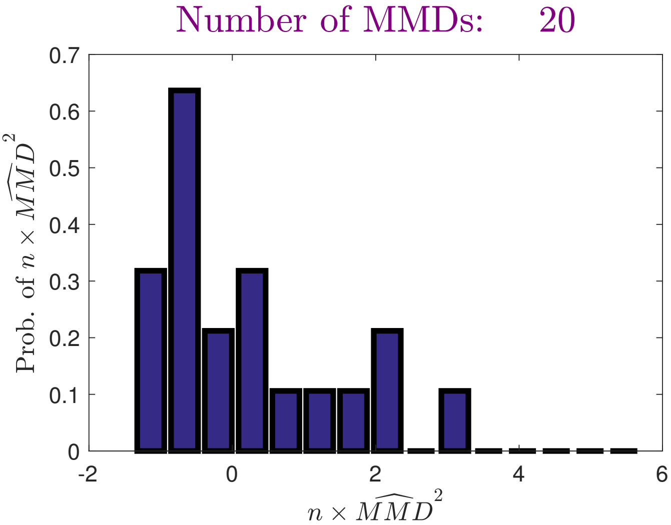
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



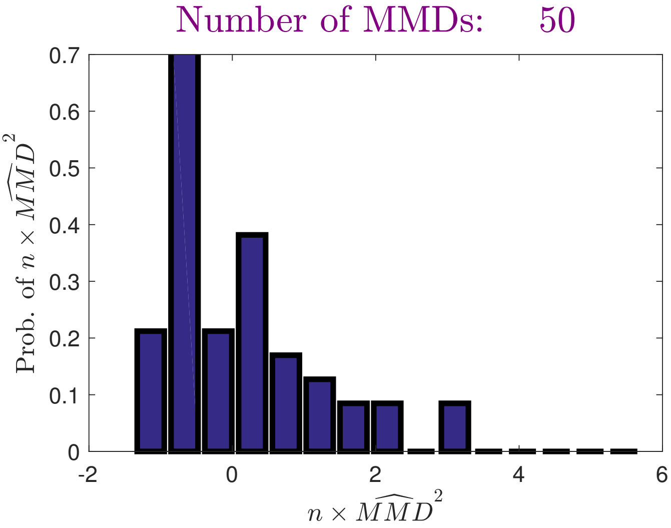
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



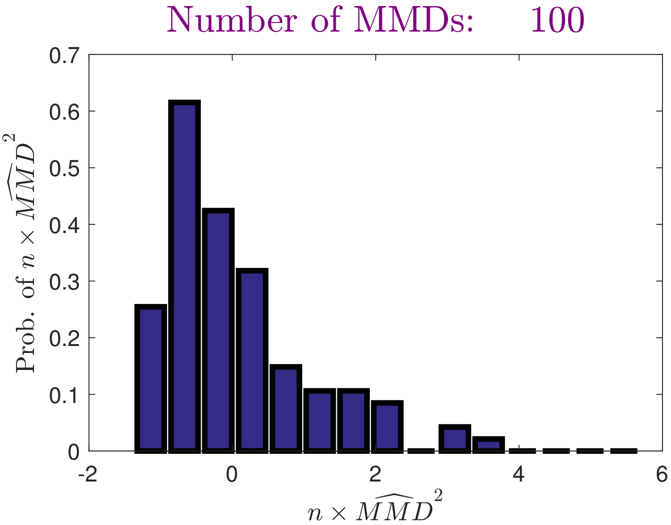
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



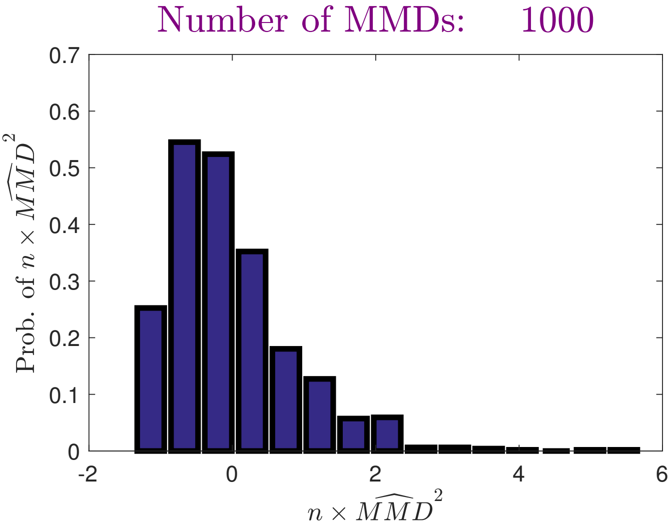
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



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

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

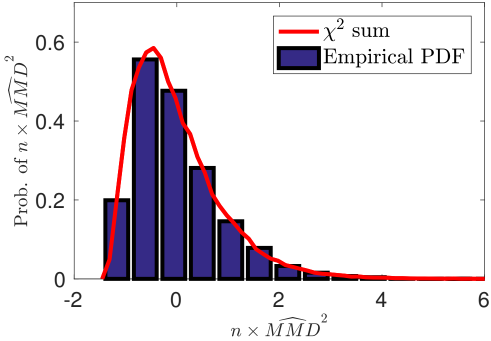


Asymptotics of $\widehat{MMD}^2$ when $P = Q$

Where $P = Q$, statistic has asymptotic distribution

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

MMD density under $\mathcal{H}_0$



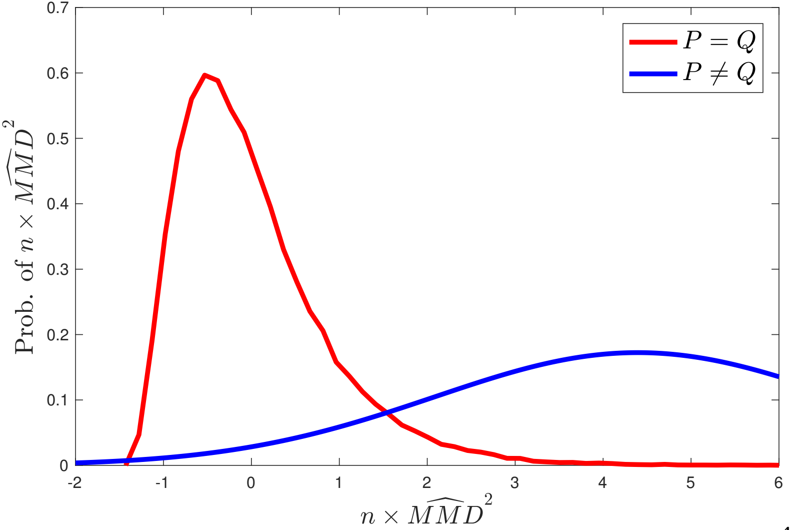
where

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

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

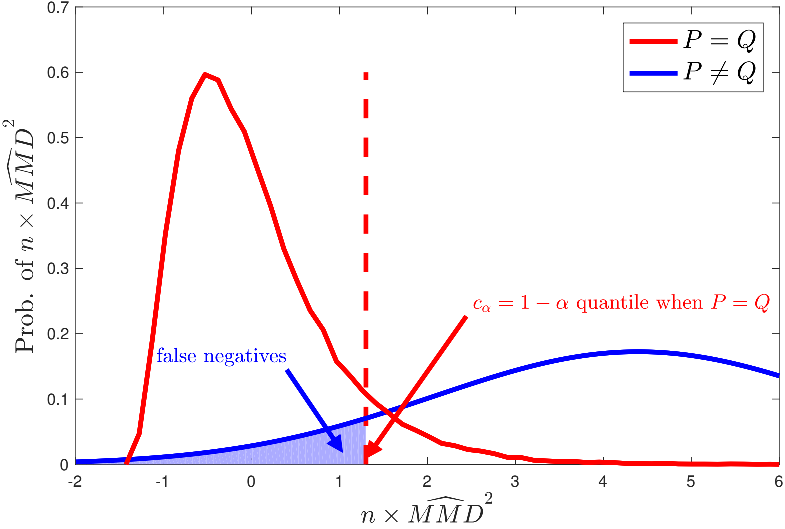
A statistical test

A summary of the asymptotics:



A statistical test

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



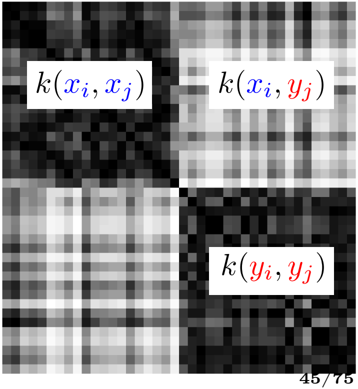
How do we get test threshold c_α?

Original empirical MMD for dogs and fish:

$X = [$ $\dots]$

$Y = [$ $\dots]$

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



How do we get test threshold c_α?

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

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

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

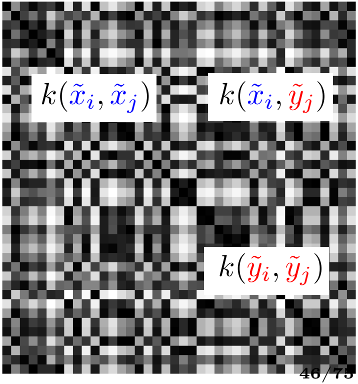
How do we get test threshold c_α?

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

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

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

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



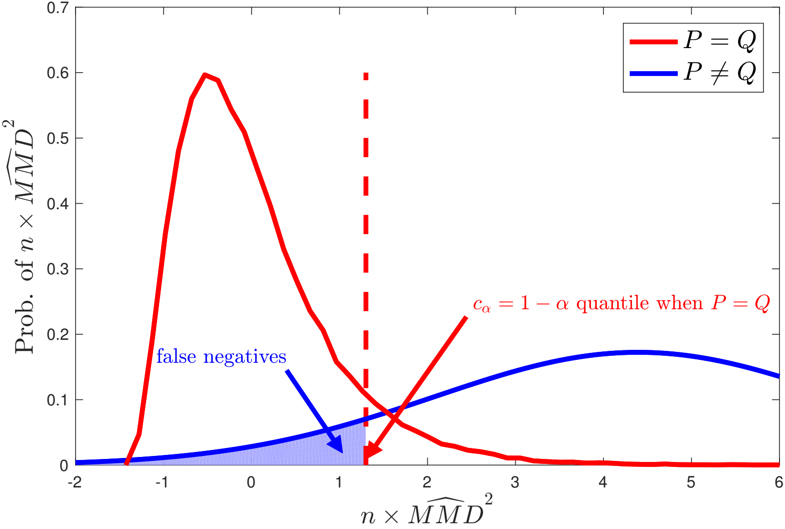
Permutation simulates

$P = Q$

Application: GAN quality evaluation

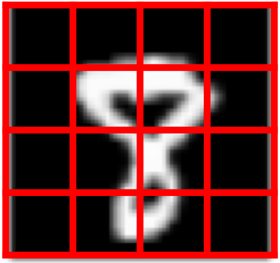
Maximising test power: graphical illustration

- Maximising test power same as minimizing false negatives



The ARD kernel

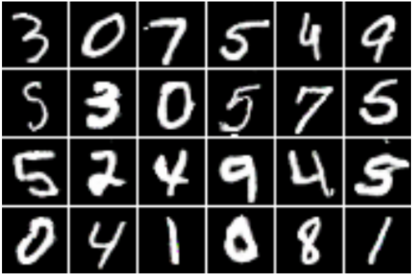
σ_1	σ_2	σ_3
σ_i	σ_{i+1}	σ_{i+2}



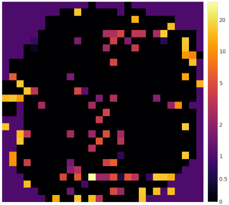
$$k(\boxed{4}, \boxed{2}) = \prod_{i=1}^{D} \exp\left(\frac{-(\boxed{4}[i] - \boxed{2}[i])^2}{\sigma_i^2}\right)$$

Troubleshooting for generative adversarial networks

MNIST samples



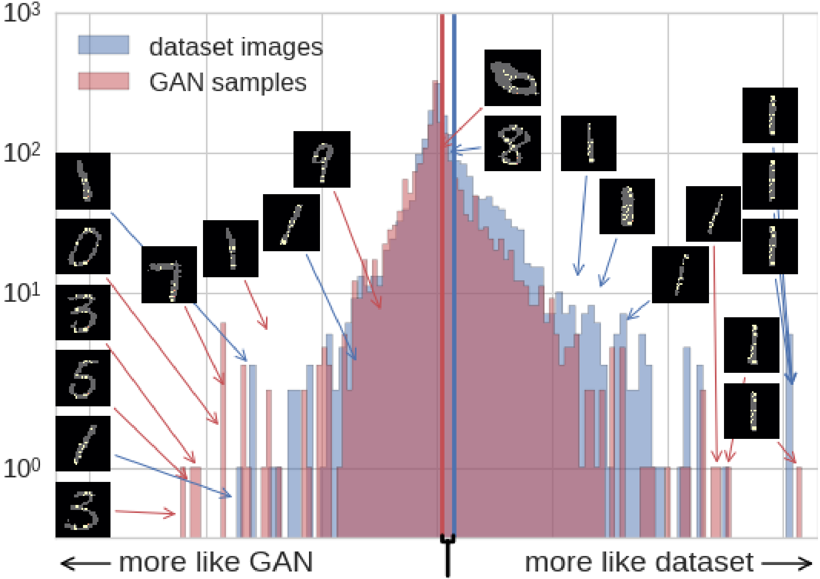
Samples from a GAN



ARD map

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

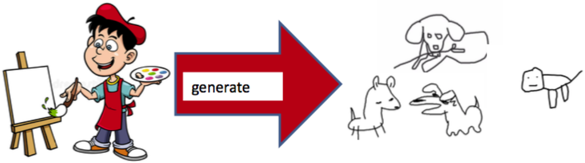
Troubleshooting generative adversarial networks



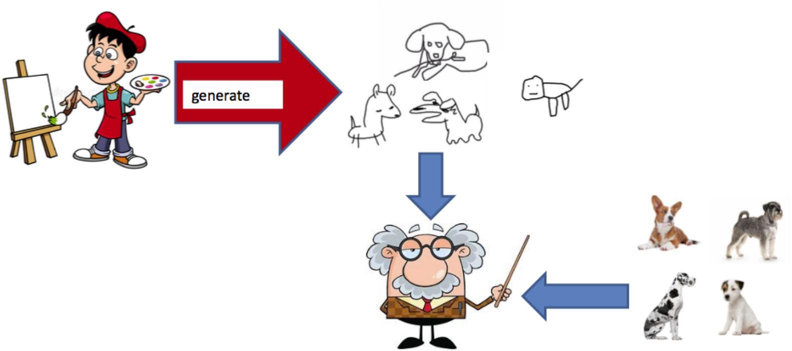
$\mathrm{MMD}^2 = 0.0001$

Training Generative Adversarial Networks

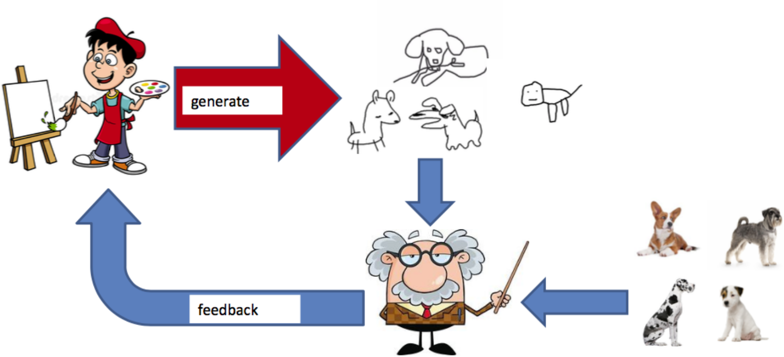
Reminder: GAN setting



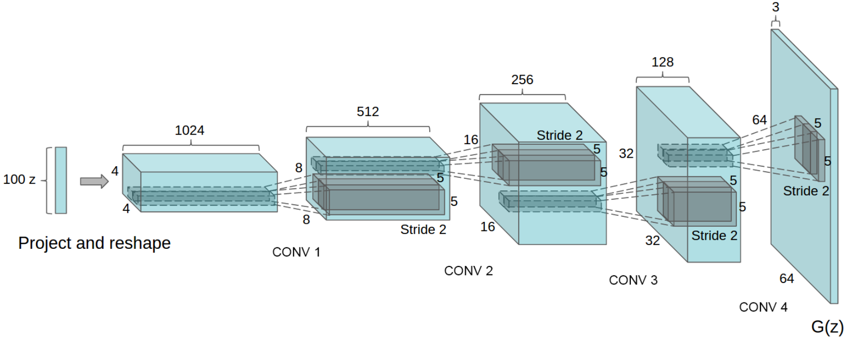
Reminder: GAN setting



Reminder: GAN setting

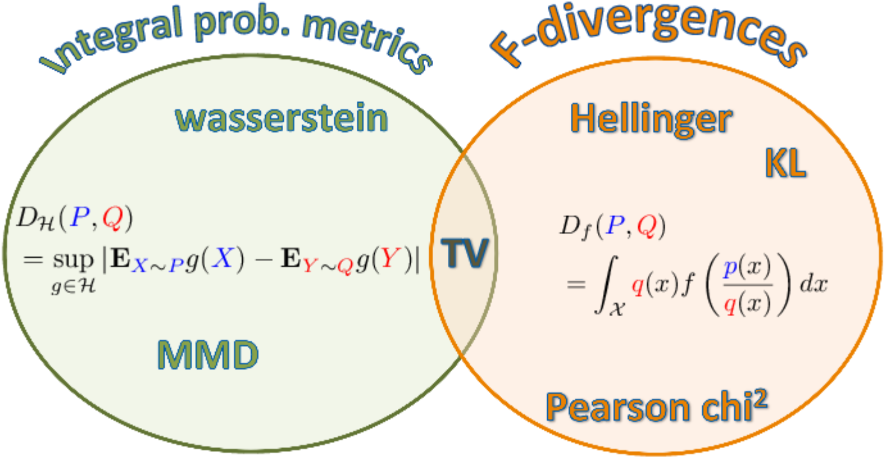


What I *won't* cover: the generator



Radford, Metz, Chintala, ICLR 2016

Choices of critic



$$D_{\mathcal{H}}(P,Q) = \sup_{g\in\mathcal{H}} |\mathbf{E}_{X\sim P}\, g(X) - \mathbf{E}_{Y\sim Q}\, g(Y)|$$

$$D_f(P,Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

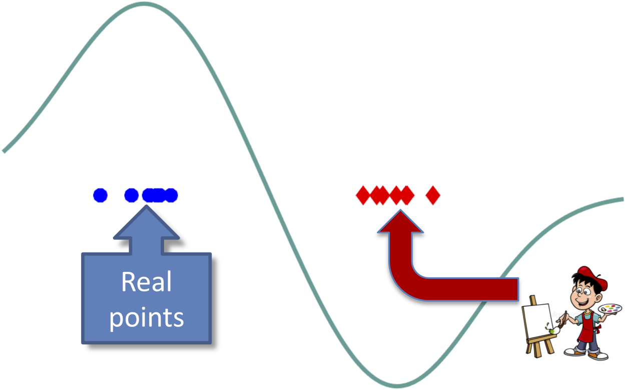
Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (2012)

MMD as critic

A helpful critic witness:

$$MMD(P, Q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} E_P f(X) - E_Q f(Y).$$

MMD=1.8

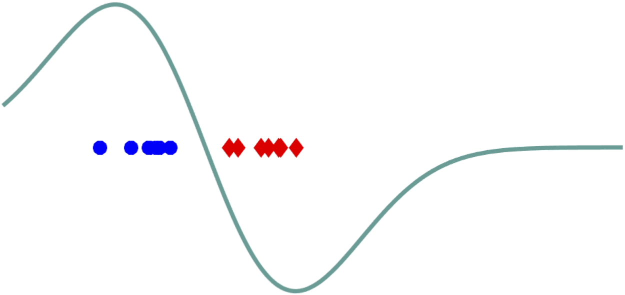


MMD as critic

A helpful critic witness:

$$MMD(P, Q) = \sup_{||f||_{\mathcal{F}} \leq 1} E_P f(X) - E_Q f(Y)$$

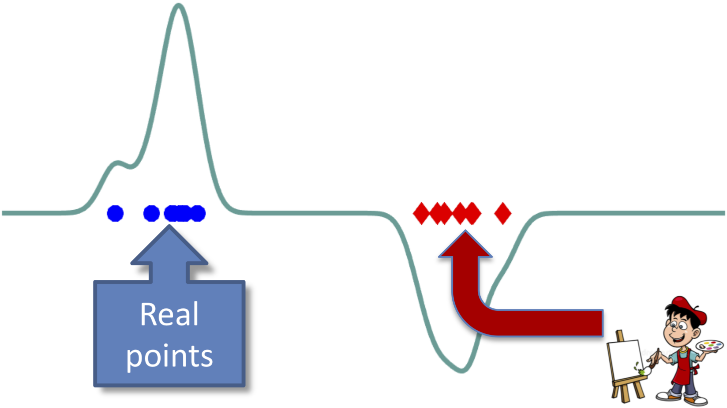
MMD=1.1



MMD as critic

An unhelpful critic witness:
$MMD(P, Q)$ with a narrow kernel.

MMD=0.64

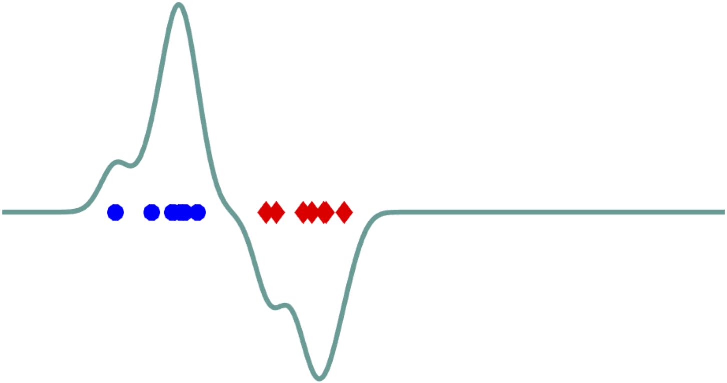


Real points

MMD as critic

An unhelpful critic witness:
$MMD(P, Q)$ with a narrow kernel.

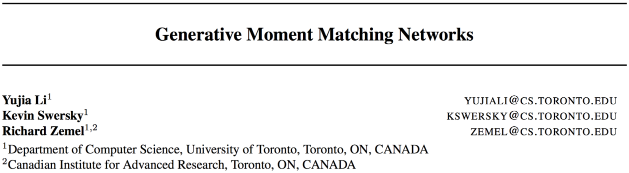
MMD=0.64



MMD for GAN critic

Can you use MMD as a critic to train GANs?

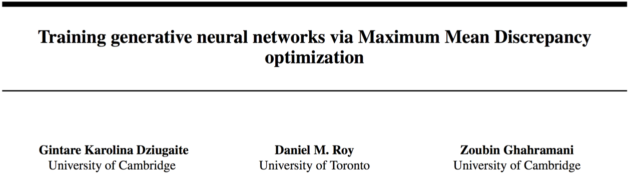
From ICML 2015:



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

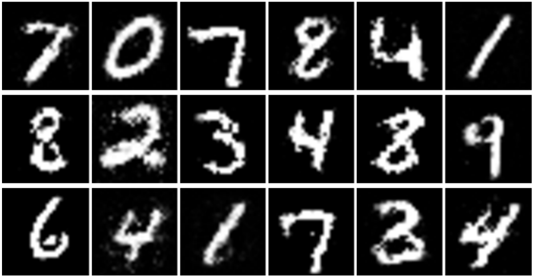
Gintare Karolina Dziugaite
University of Cambridge

Daniel M. Roy
University of Toronto

Zoubin Ghahramani
University of Cambridge

MMD for GAN critic

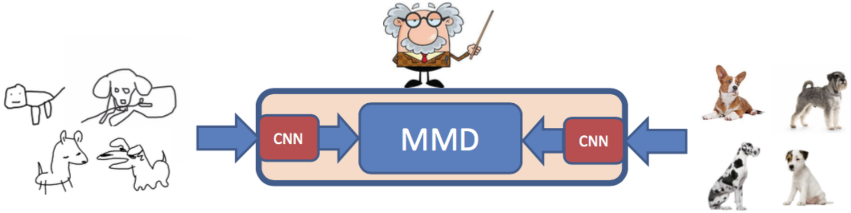
Can you use MMD as a critic to train GANs?



Need better image features.

CNN features for an MMD witness

- Add convolutional features!
- The critic (teacher) also needs to be trained.



$\mathfrak{K}(x, y) = {h_\psi}^\top(x) h_\psi(y)$
where $h_\psi(x)$ is a CNN map:

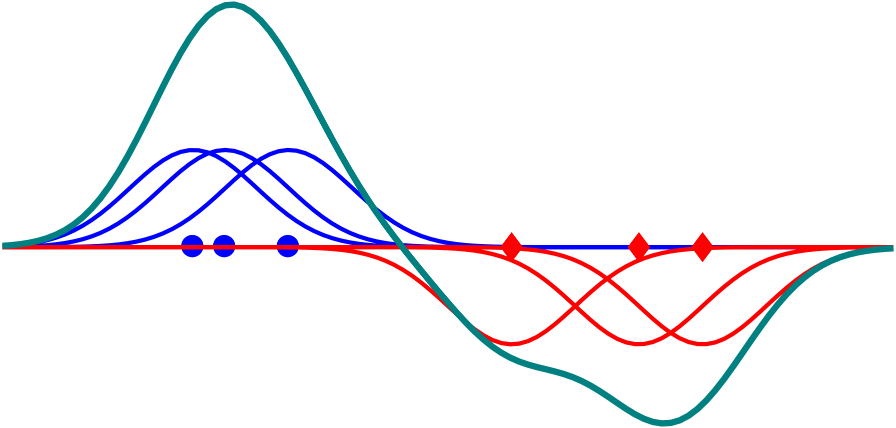
- Wasserstein GAN Arjovsky et al. [ICML 2017]
- WGAN-GP Gulrajani et al. [NeurIPS 2017]

$\mathfrak{K}(x, y) = k(h_\psi(x), h_\psi(y))$
where $h_\psi(x)$ is a CNN map,
k is e.g. an exponentiated quadratic kernel
MMD Li et al., [NeurIPS 2017]
Cramer Bellemare et al. [2017]
Coulomb Unterthiner et al., [ICLR 2018]
Demystifying MMD GANs Binkowski, Sutherland, Arbel, G., [ICLR 2018]

Witness function, kernels on deep features

Reminder: witness function,

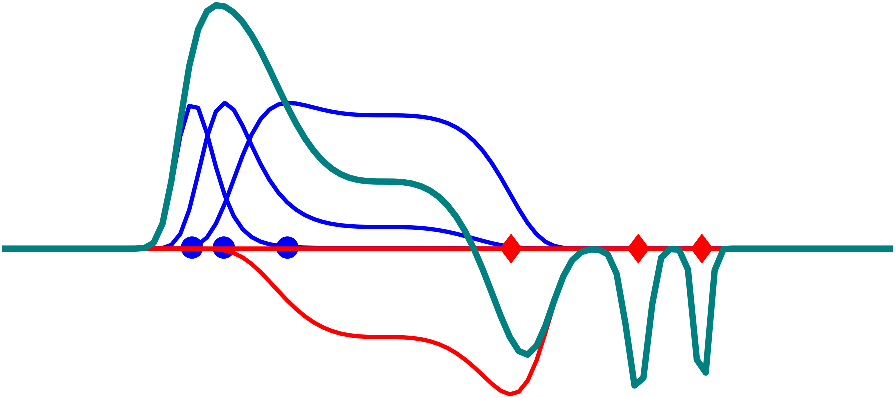
$k(x, y)$ is exponentiated quadratic



Witness function, kernels on deep features

Reminder: witness function,

$k(h_\psi(x), h_\psi(y))$ with neural network h_ψ and exp. quadratic k



Challenges for learned critic features

Learned critic features:

MMD with kernel $k(h_\psi(x), h_\psi(y))$ must give useful gradient to generator.

Challenges for learned critic features

Learned critic features:

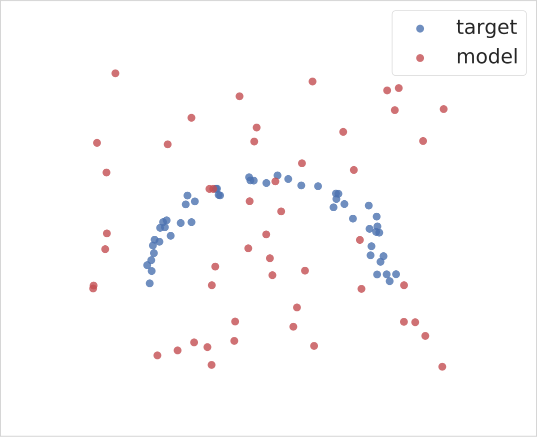
MMD with kernel $k(h_\psi(x), h_\psi(y))$ must give useful gradient to generator.

Relation with test power?

If the MMD with kernel $k(h_\psi(x), h_\psi(y))$ gives a powerful test, will it be a good critic?

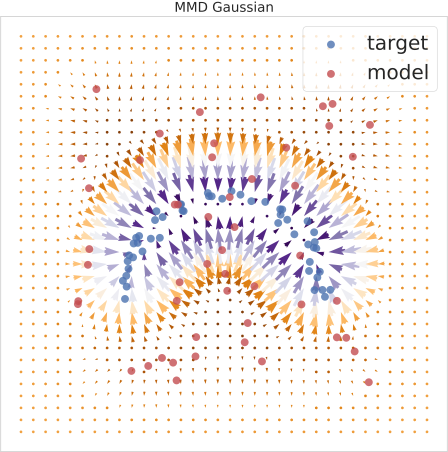
A simple 2-D example

Samples from target P and model Q



A simple 2-D example

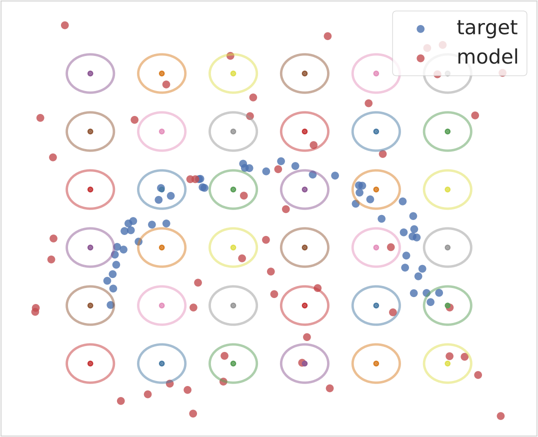
Witness gradient, MMD with exp. quad. kernel $k(x, y)$



A simple 2-D example

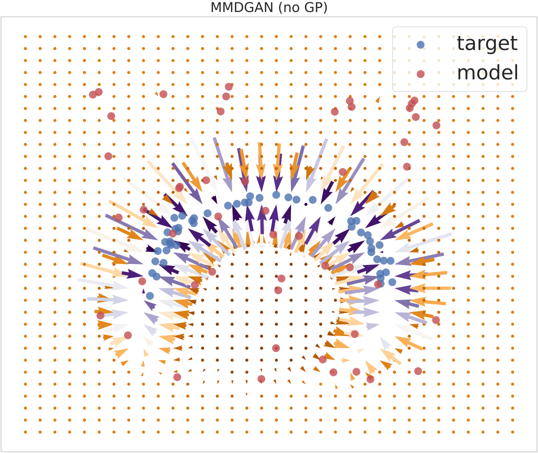
What the kernels $k(x, y)$ look like

MMD Gaussian



A simple 2-D example

Witness gradient, maximise MMD to learn $h_\psi(x)$ for $k(h_\psi(x), h_\psi(y))$

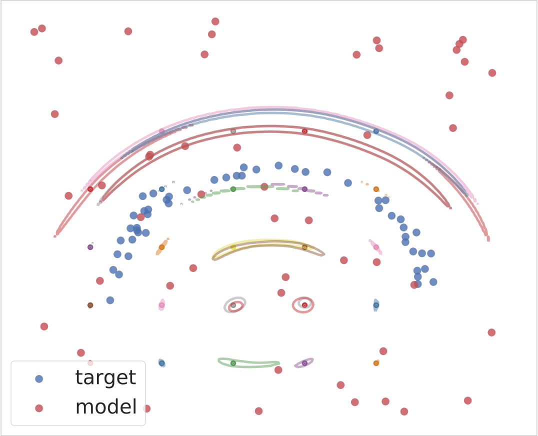


(4 layer, fully connected, RELU, skipthrough 1-4, early stage)

A simple 2-D example

What the kenels $k(h_\psi(x), h_\psi(y))$ look like

MMDGAN (no GP)

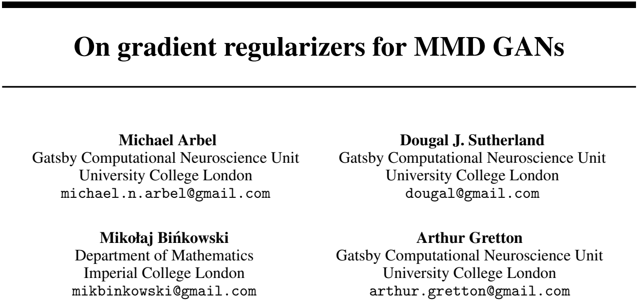


(4 layer, fully connected, RELU, skipthrough 1-4, early stage)

A simple 2-D example

A data-adaptive gradient penalty

- New gradient regulariser Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Also related to Sobolev GAN Mroueh et al. [ICLR 2018]



On gradient regularizers for MMD GANs

Michael Arbel
Gatsby Computational Neuroscience Unit
University College London
michael.n.arbel@gmail.com

Dougal J. Sutherland
Gatsby Computational Neuroscience Unit
University College London
dougal@gmail.com

Mikołaj Bińkowski
Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
arthur.gretton@gmail.com

A data-adaptive gradient penalty

- New gradient regulariser Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Also related to Sobolev GAN Mroueh et al. [ICLR 2018]

Maximise scaled MMD over critic features:

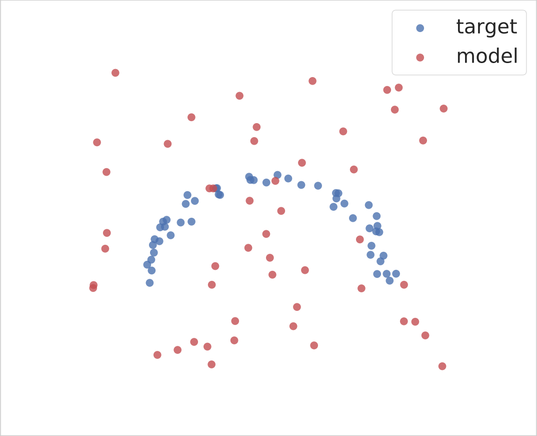
$$SMMD(P, \lambda) = \sigma_{P,\lambda}\ MMD$$

where

$$\sigma^2_{P,\lambda} = \lambda + \int k(h_\psi(x), h_\psi(x))\, dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(h_\psi(x), h_\psi(x))\ dP(x)$$

Simple 2-D example revisited

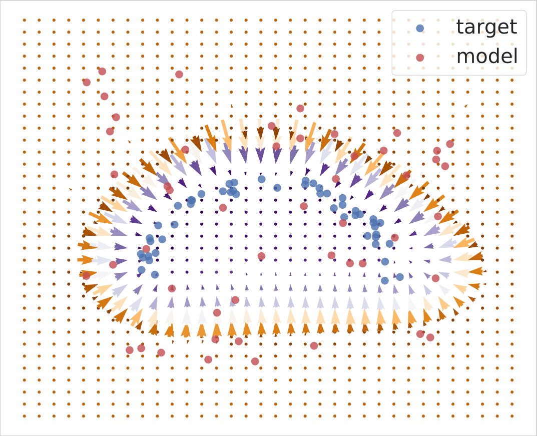
Samples from target P and model Q



Simple 2-D example revisited

Witness gradient, maximise $SMMD(P, \lambda)$
to learn $h_\psi(x)$ for $k(h_\psi(x), h_\psi(y))$

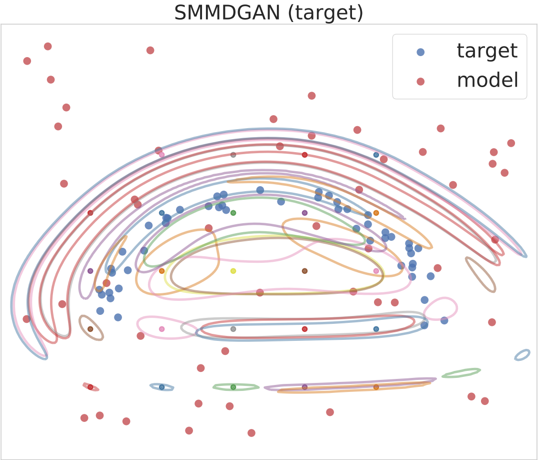
SMMDGAN (target)



(early stage of critic optimisation)

Simple 2-D example revisited

What the kenels $k(h_\psi(x), h_\psi(y))$ look like

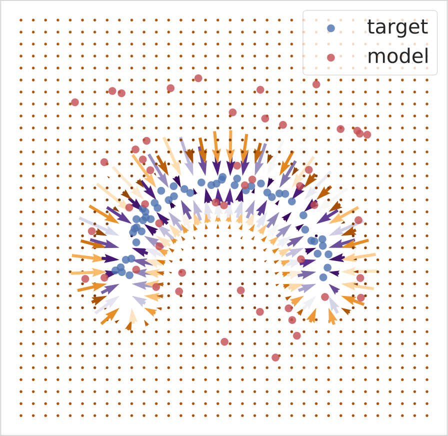


(early stage of critic optimisation)

Simple 2-D example revisited

Witness gradient, maximise $SMMD(P, \lambda)$
to learn $h_\psi(x)$ for $k(h_\psi(x), h_\psi(y))$

SMMDGAN (target)

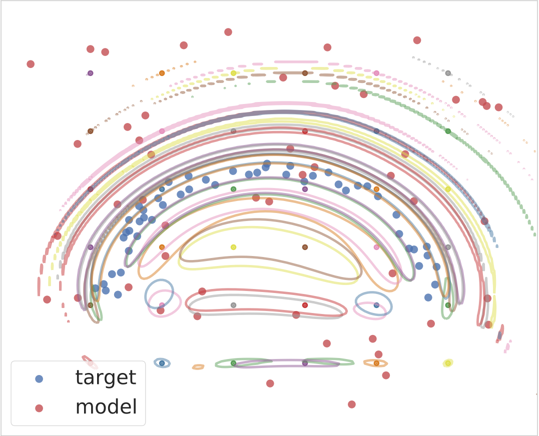


(late stage of critic optimisation)

Simple 2-D example revisited

What the kenels $k(h_\psi(x), h_\psi(y))$ look like

SMMDGAN (target)



(late stage of critic optimisation)

Our empirical observations

Data-adaptive critic loss:

- Witness function class for $SMMD(P, \lambda)$ depends on P.
  - Without data-dependent regularisation, maximising MMD over features h_ψ of kernel $k(h_\psi(x), h_\psi(y))$ is unhelpful.

Our empirical observations

Data-adaptive critic loss:

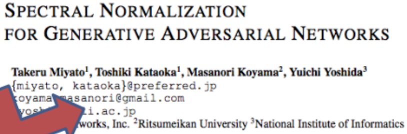
- Witness function class for $SMMD(P, \lambda)$ depends on P.
  - Without data-dependent regularisation, maximising MMD over features h_ψ of kernel $k(h_\psi(x), h_\psi(y))$ is unhelpful.

Alternate critic and generator training:

- Weaker critics can give better signals to poor (early stage) generators.

Evaluation and experiments

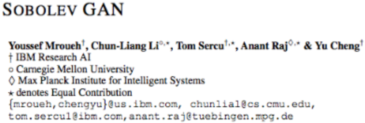
Benchmarks for comparison (all from ICLR 2018)



SPECTRAL NORMALIZATION
FOR GENERATIVE ADVERSARIAL NETWORKS

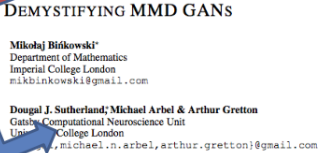
Takeru Miyato[1], Toshiki Kataoka[1], Masanori Koyama[2], Yuichi Yoshida[3]
{miyato, kataoka}@preferred.jp
koyama asanori@gmail.com
yos i.ac.jp
works, Inc. [2]Ritsumeikan University [3]National Institute of Informatics

We combine with scaled MMD



SOBOLEV GAN

Youssef Mroueh[†], Chun-Liang Li[○,⋆], Tom Sercu[†,⋆], Anant Raj[◊,⋆] & Yu Cheng[†]
† IBM Research AI
○ Carnegie Mellon University
◊ Max Planck Institute for Intelligent Systems
⋆ denotes Equal Contribution
{mroueh,chengyu}@us.ibm.com, chunlial@cs.cmu.edu,
tom.sercu1@ibm.com,anant.raj@tuebingen.mpg.de

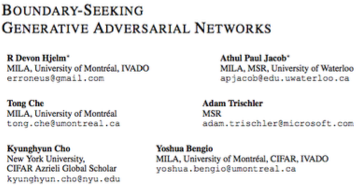


DEMYSTIFYING MMD GANS

Mikołaj Bińkowski*
Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Dougal J. Sutherland*, Michael Arbel & Arthur Gretton
Gatsby Computational Neuroscience Unit
Uni College London
,michael.n.arbel,arthur.gretton}@gmail.com

Our ICLR 2018 paper



BOUNDARY-SEEKING
GENERATIVE ADVERSARIAL NETWORKS

R Devon Hjelm*
MILA, University of Montréal, IVADO
erroneus@gmail.com

Athul Paul Jacob*
MILA, MSR, University of Waterloo
apjacob@edu.uwaterloo.ca

Tong Che
MILA, University of Montréal
tong.che@umontreal.ca

Adam Trischler
MSR
adam.trischler@microsoft.com

Kyunghyun Cho
New York University,
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

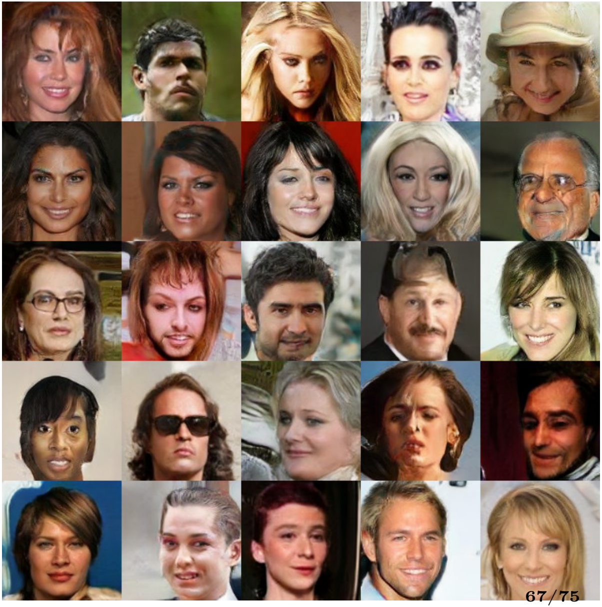
Yoshua Bengio
MILA, University of Montréal, CIFAR, IVADO
yoshua.bengio@umontreal.ca

Results: celebrity faces 160×160

KID scores:

- Sobolev GAN: 14
- SN-GAN: 18
- Old MMD GAN: 13
- SMMD GAN: 6

202 599 face images, resized and cropped to 160 × 160



Results: unconditional imagenet 64×64

KID scores:

- BGAN:
  47
- SN-GAN:
  44
- SMMD GAN:
  35

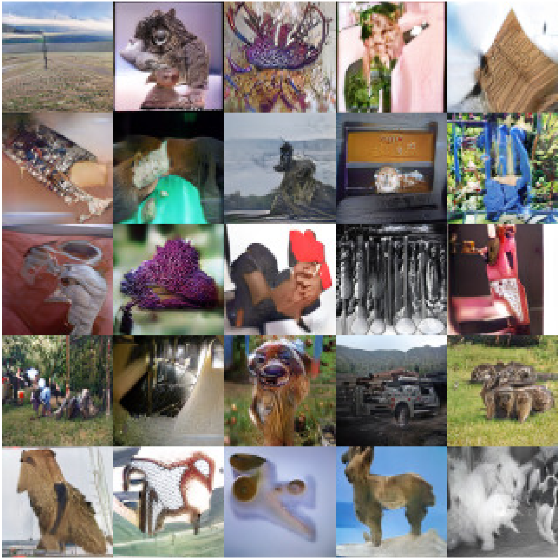
ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. 1000 classes.

Results: unconditional imagenet 64×64

KID scores:

- BGAN: 47
- SN-GAN: 44
- SMMD GAN: 35

ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. 1000 classes.



Results: unconditional imagenet 64×64

KID scores:

- BGAN: 47
- SN-GAN: 44
- SMMD GAN: 35

ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. 1000 classes.

Summary

- MMD critic gives state-of-the-art performance for GAN training (FID and KID)
  - use convolutional input features
  - train with new gradient regulariser
- Faster training, simpler critic network
- Reasons for good performance:
  - Unlike WGAN-GP, MMD loss still a valid critic when features not optimal
  - Kernel features do some of the "work", so simpler h_ψ features possible.
  - Better gradient/feature regulariser gives better critic

"Demystifying MMD GANs," including KID score, ICLR 2018:

https://github.com/mbinkowski/MMD-GAN

Gradient regularised MMD, NeurIPS 2018:

https://github.com/MichaelArbel/Scaled-MMD-GAN

Co-authors

From Gatsby:

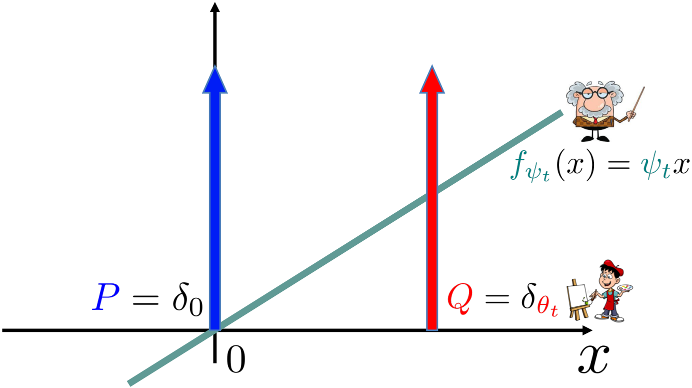
- Michael Arbel
- Mikolaj Binkowski
- Heiko Strathmann
- Dougal Sutherland

External collaborators:

- Soumyajit De
- Aaditya Ramdas
- Bernhard Schoelkopf
- Alex Smola
- Hsiao-Yu Tung

Questions?

THE KERNEL BREWERY LONDON
BALTIC PORTER
7.3% ABV
THE KERNEL
EXPORT STOUT
LONDON 1890

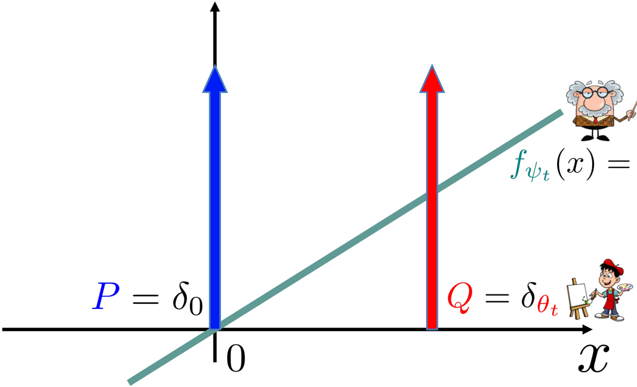


$$
\begin{aligned}
D(P, Q; \psi_t) &= \mathbf{E}_Q f_{\psi_t}(Y) - \mathbf{E}_P f_{\psi_t}(X) \\
&= \psi_t \theta_t
\end{aligned}
$$

Mescheder et al. [ICML 2018]

Optimization: simple example

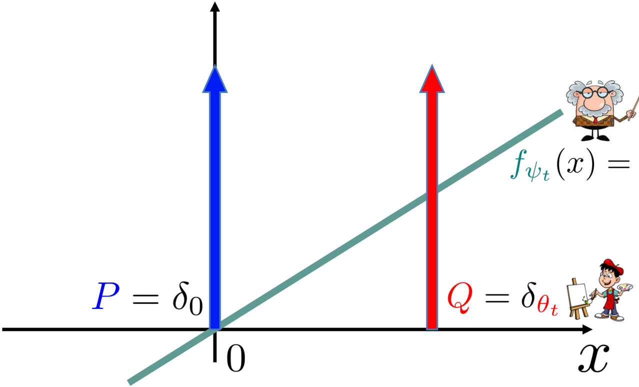
Gradient descent on generator:



$$\frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \frac{\partial}{\partial \theta} \psi_t \theta_t = \psi_t$$

Optimization: simple example

Gradient descent on generator:

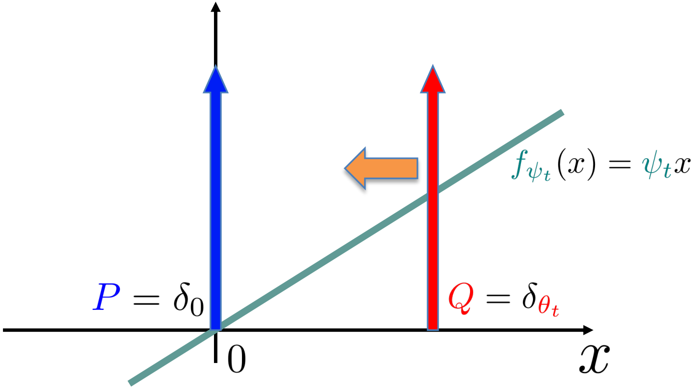


$$\frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \frac{\partial}{\partial \theta} \psi_t \theta_t = \psi_t$$

$$\theta_{t+1} = \theta_t - \gamma \frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \theta_t - \gamma \psi_t$$

Optimization: simple example

Gradient descent on generator:

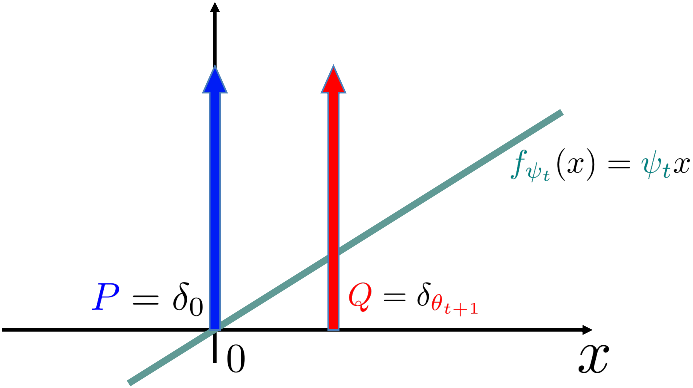


$$\frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \frac{\partial}{\partial \theta} \psi_t \theta_t = \psi_t$$

$$\theta_{t+1} = \theta_t - \gamma \frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \theta_t - \gamma \psi_t$$

Optimization: simple example

Gradient descent on generator:

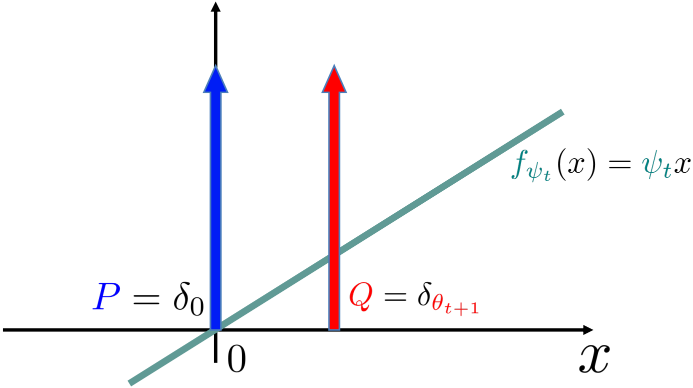


$$\frac{\partial}{\partial \boldsymbol{\theta}} D(P, Q; \psi_t) = \frac{\partial}{\partial \boldsymbol{\theta}} \psi_t \boldsymbol{\theta}_t = \psi_t$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma \frac{\partial}{\partial \boldsymbol{\theta}} D(P, Q; \psi_t) = \boldsymbol{\theta}_t - \gamma \psi_t$$

Optimization: simple example

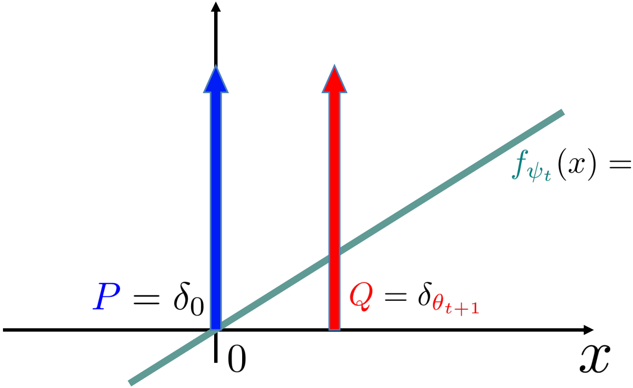
Gradient ascent on critic:



$$\frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \theta_{t+1}$$

Optimization: simple example

Gradient **ascent** on critic:

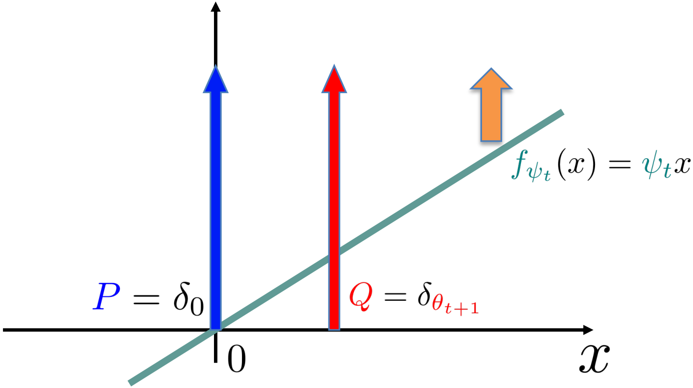


$$\frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \theta_{t+1}$$

$$\psi_{t+1} = \psi_t + \lambda \frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \psi_t + \lambda \theta_{t+1}$$

Optimization: simple example

Gradient ascent on critic:

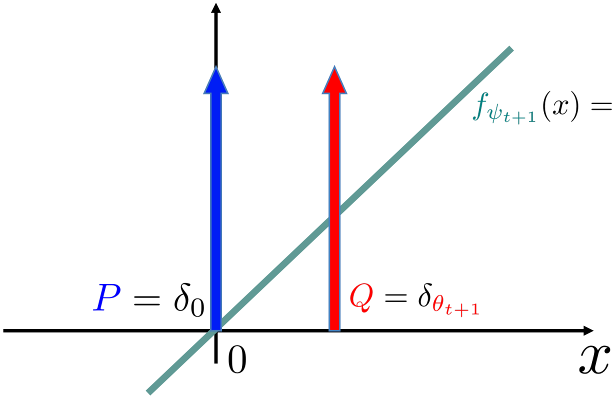


$$\frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \theta_{t+1}$$

$$\psi_{t+1} = \psi_t + \lambda \frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \psi_t + \lambda \theta_{t+1}$$

Optimization: simple example

Gradient ascent on critic:



$f_{\psi_{t+1}}(x) = \psi_{t+1} x$

$P = \delta_0$

$Q = \delta_{\theta_{t+1}}$

0

x

$$\frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \theta_{t+1}$$

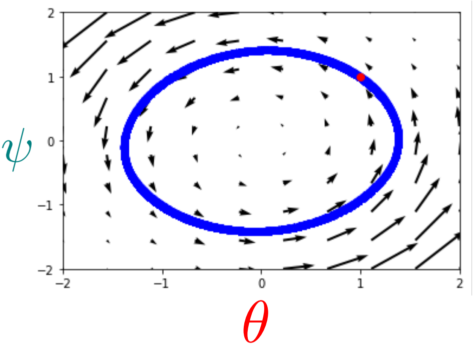
$$\psi_{t+1} = \psi_t + \lambda \frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \psi_t + \lambda \theta_{t+1}$$

Optimization: simple example

Idealised continuous system (infinitely small learning rate)

$$\begin{bmatrix} \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} -\nabla_{\psi} D(P, Q; \psi) \\ \nabla_{\theta} D(P, Q; \psi) \end{bmatrix}$$

Every integral curve $(\psi(t), \theta(t))$ of the gradient vector field satisfies $\psi^2(t) + \theta^2(t) = c$ for all $t \in [0, \infty)$.



Mescheder et al. [ICML 2018, Lemma 2.3]

Optimization: simple example

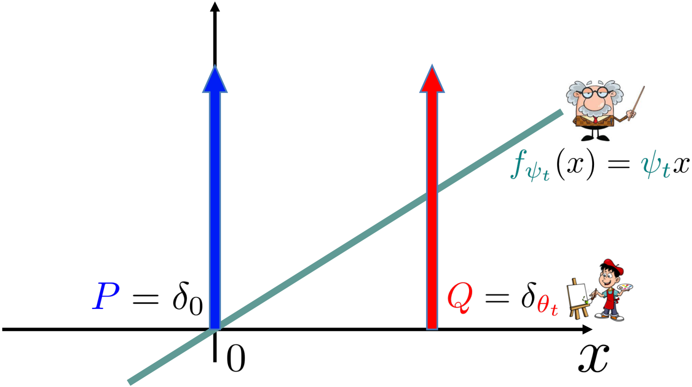
Idealised continuous system (infinitely small learning rate)

$$\begin{bmatrix} \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} -\nabla_{\psi} D(P, Q; \psi) \\ \nabla_{\theta} D(P, Q; \psi) \end{bmatrix}$$

Every integral curve $(\psi(t), \theta(t))$ of the gradient vector field satisfies $\psi^2(t) + \theta^2(t) = c$ for all $t \in [0, \infty)$.

A solution: control witness gradient

Mescheder et al. [ICML 2018, Lemma 2.3]



$$\begin{aligned} D(P, Q; \psi_t) &= \mathbf{E}_Q f_{\psi_t}(Y) - \mathbf{E}_P f_{\psi_t}(X) \\ &= \psi_t \theta_t \end{aligned}$$

Mescheder et al. [ICML 2018]

Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

Nagarajan and Kolter [NeurIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

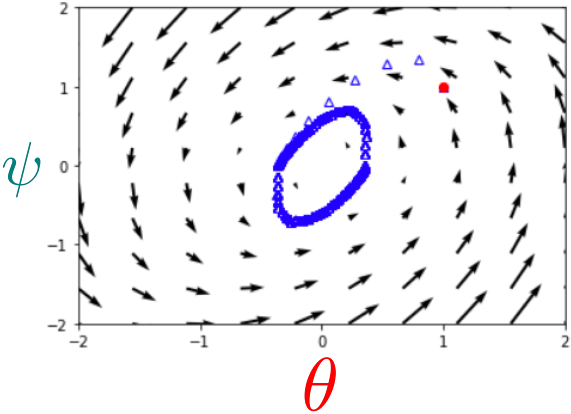


Figure from Mescheder et al. [ICML 2018]