# The Maximum Mean Discrepancy
# for Training Generative Adversarial Networks
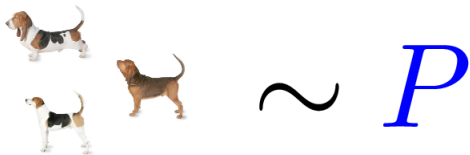
**Arthur Gretton**

Gatsby Computational Neuroscience Unit,
University College London

Cardiff, 2018

# A motivation: comparing two samples

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

# 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, NIPS 2016
Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.

# *Training* generative models

## *Training* generative models

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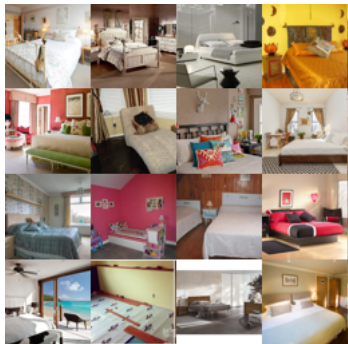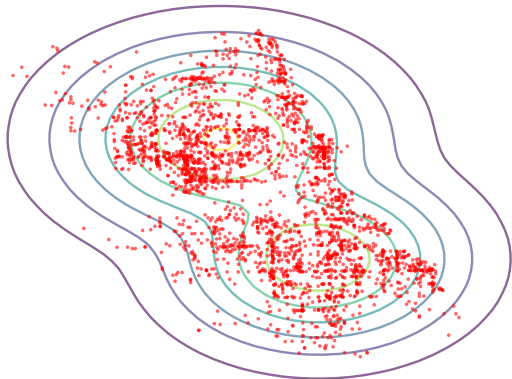


LSUN bedroom samples $P$        Generated $Q$, MMD GAN

## Using MMD to train a GAN

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

# 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

- Maximum Mean Discrepancy (MMD)...
  - ...as a difference in feature means
  - ...as an integral probability metric (not just a technicality!)

- A statistical test based on the MMD

- Training generative adversarial networks with MMD
  - Gradient regularisation and data adaptivity
  - Evaluating GAN performance? Problems with Inception and FID.

# Maximum Mean Discrepancy

# Feature mean difference

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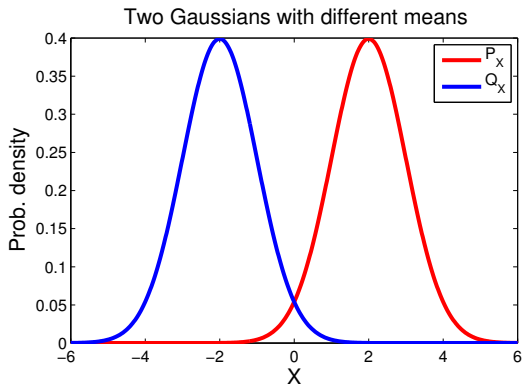


Two Gaussians with different means

# 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 $\varphi(x) = x^2$



Two Gaussians with different variances
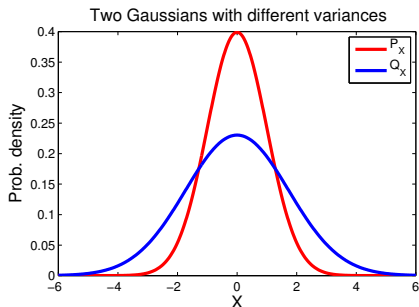
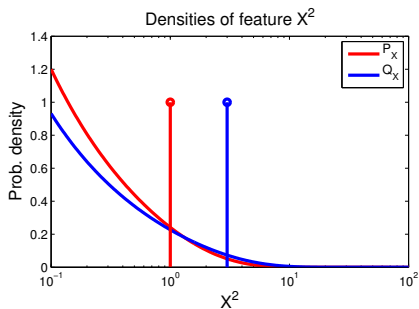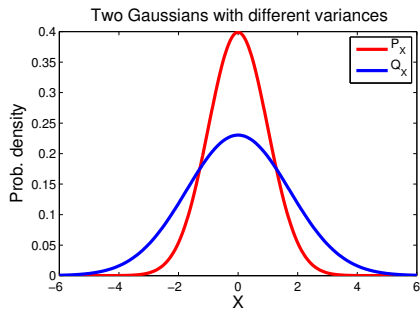# Feature mean difference

- 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
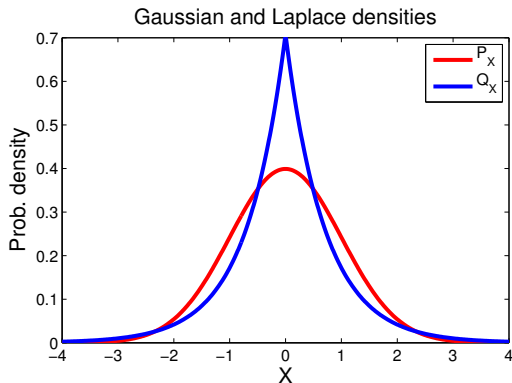

Gaussian and Laplace densities

# 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 $k$,

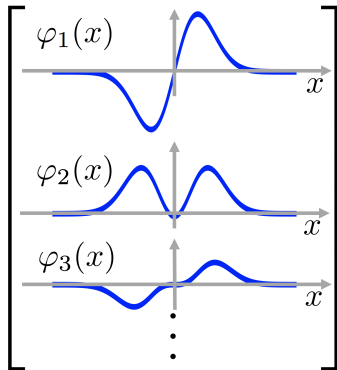$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$

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

**Exponentiated quadratic kernel**

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



$$\varphi(x) = \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \\ \vdots \end{bmatrix}$$

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 $P$,

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

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 $P$,

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

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**:

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

# The maximum mean discrepancy

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

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

# The maximum mean discrepancy

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

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

(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 $k(\text{fish}_i, \text{fish}_j)$
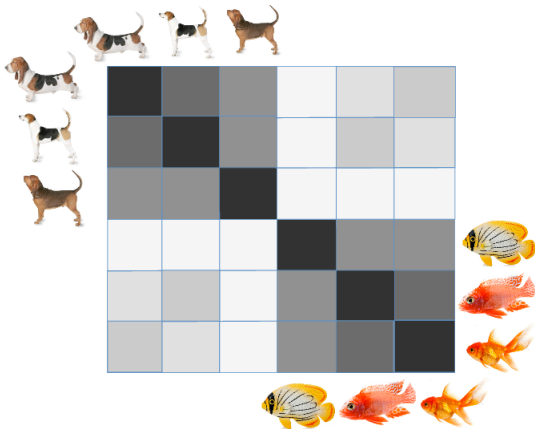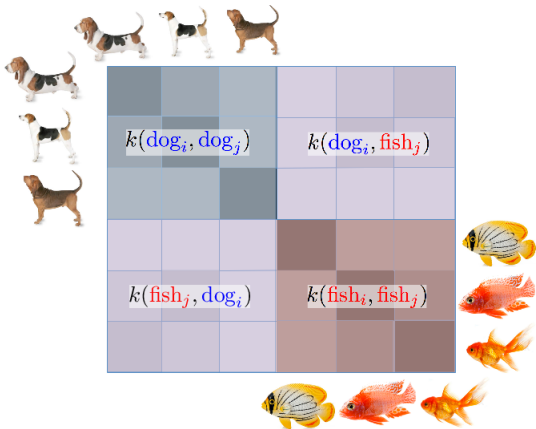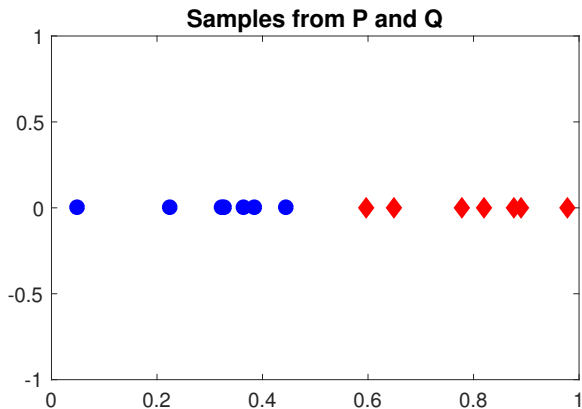
# Illustration of MMD

**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?



**Samples from P and Q**

# MMD as an integral probability metric

Are $P$ and $Q$ different?



**Samples from P and Q**

# MMD as an integral probability metric

Integral probability metric:

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

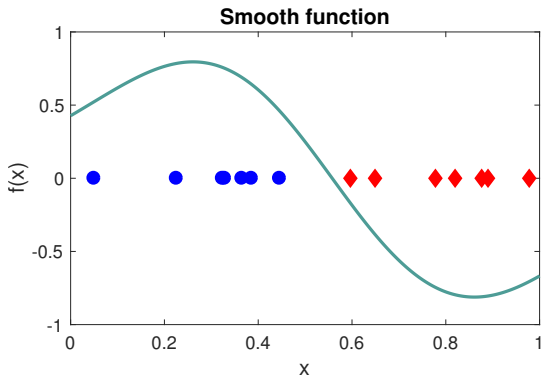$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$
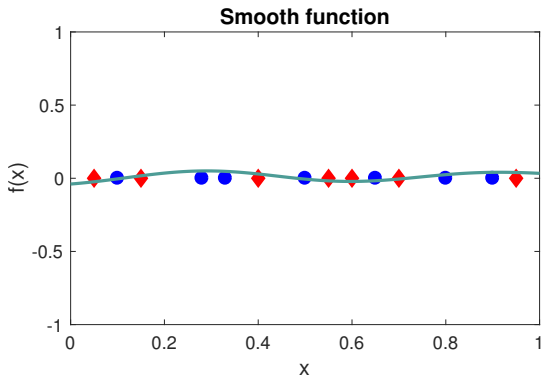
# MMD as an integral probability metric

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 $Q$

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

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

# 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:

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{F}} = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}^{\top} \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \\ \vdots \end{bmatrix}$$

# MMD as an integral probability metric

**Maximum mean discrepancy**: smooth function for $P$ vs $Q$

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

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



Witness f for Gauss and Laplace densities
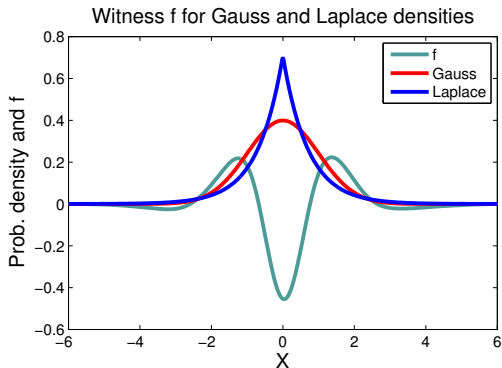
# MMD as an integral probability metric

**Maximum mean discrepancy**: smooth function for $P$ vs $Q$

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

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

**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 $Q$

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

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

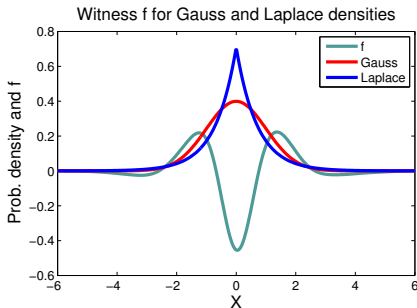For characteristic RKHS $\mathcal{F}$, $MMD(P, Q; F) = 0$ iff $P = Q$

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:**

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



Witness f for Gauss and Laplace densities

# Integral prob. metric vs feature difference

**The MMD:**

use

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

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

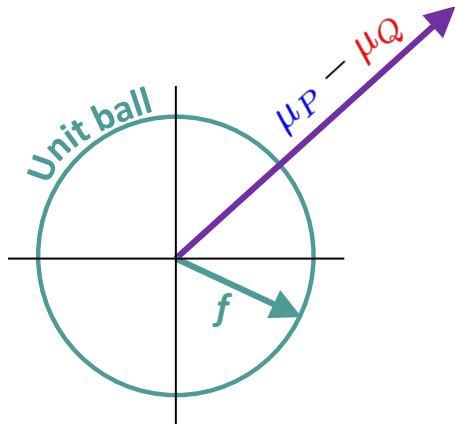# Integral prob. metric vs feature difference

**The MMD:**

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

# Integral prob. metric vs feature difference

**The MMD:**

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

# Integral prob. metric vs feature difference

**The MMD:**

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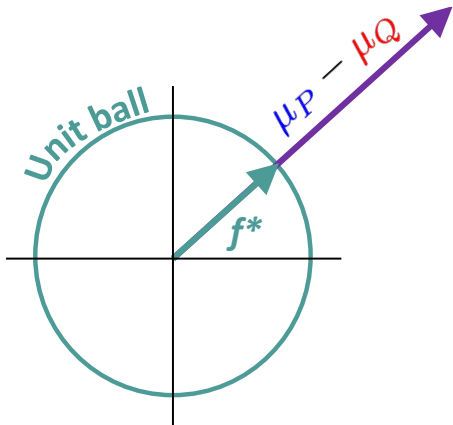$$= \sup_{f \in F} \left[ \mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right]$$
$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



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

**The MMD:**

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

Function view and feature view equivalent

# Construction of MMD witness

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



Observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$
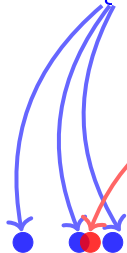
Observe $\mathsf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim Q$
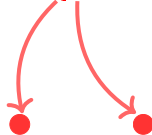
Construction of empirical witness function (proof: next slide!)
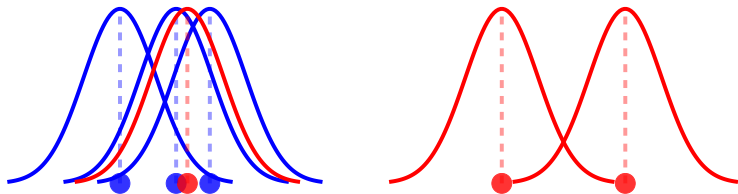
Construction of empirical witness function (proof: next slide!)
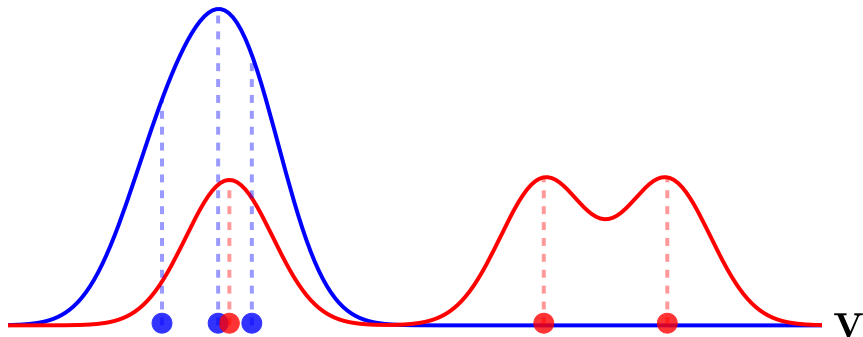
# Construction of MMD witness

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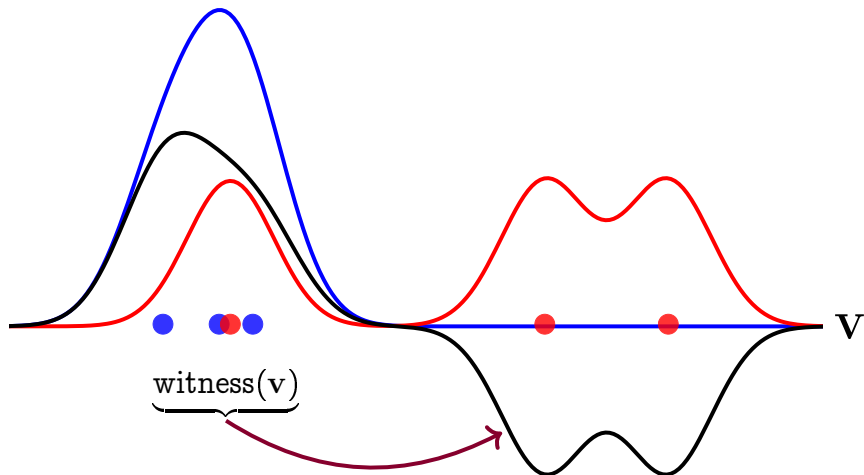
# Derivation of empirical witness function

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 $v$

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

# Derivation of empirical witness function

Recall the witness function expression

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

The empirical feature mean for $P$

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

The empirical witness function at $v$

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

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

Interlude: divergence measures

# Divergences



$$P - Q \qquad \frac{P}{Q}$$
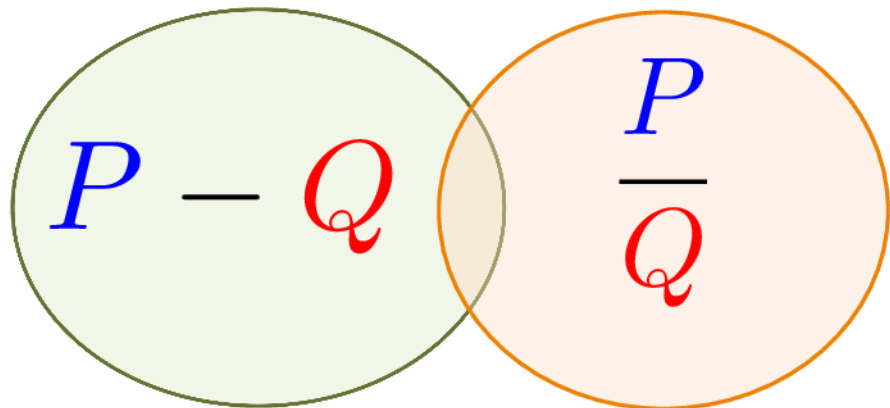
# Divergences



Integral prob. metrics

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

F-divergences

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

# Divergences



**Integral prob. metrics**

**wasserstein**

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

**MMD**

**F-divergences**

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

# Divergences



Integral prob. metrics

wasserstein

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

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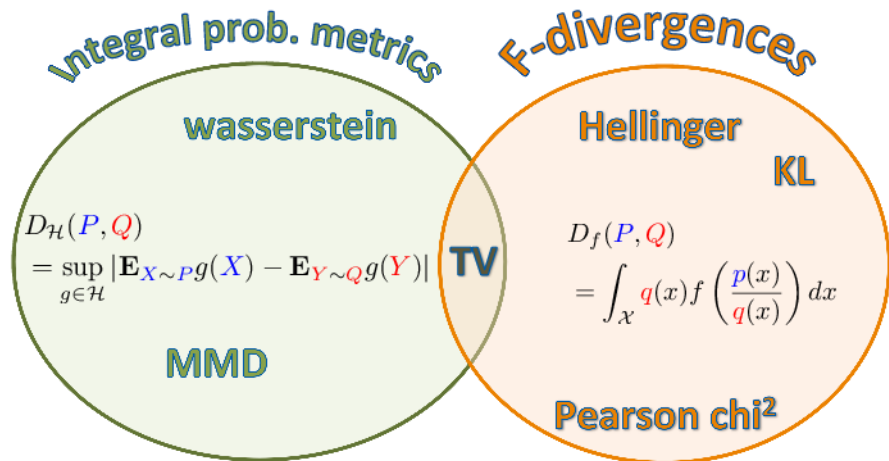
Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (2012)

# Two-Sample Testing with MMD

# A statistical test using MMD

The empirical MMD:

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

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:

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

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"

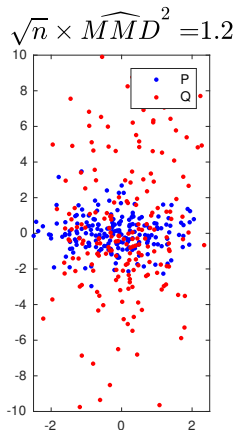Want Threshold $c_\alpha$ for $\widehat{MMD}^2$ to get false positive rate $\alpha$

# 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 $Q$

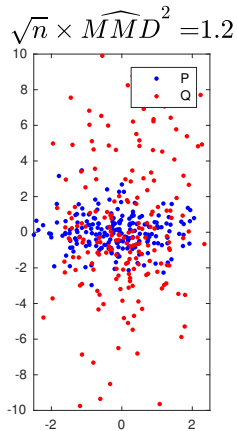- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.2$



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 $Q$

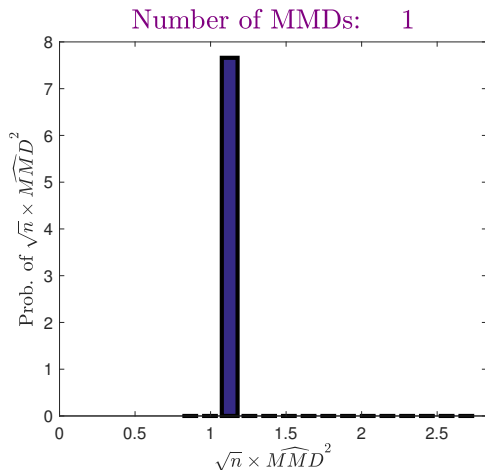- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.5$



Number of MMDs: 2

$\sqrt{n} \times \widehat{MMD}^2 = 1.5$
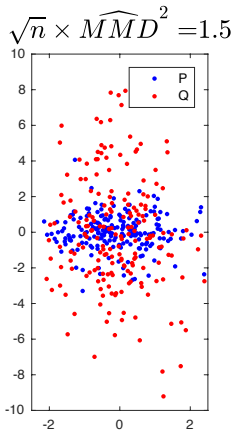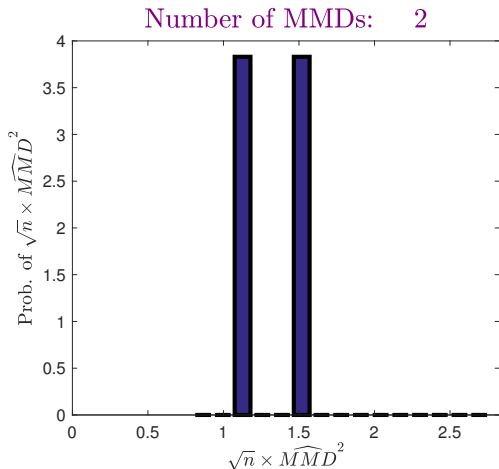
# Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 150 times ...
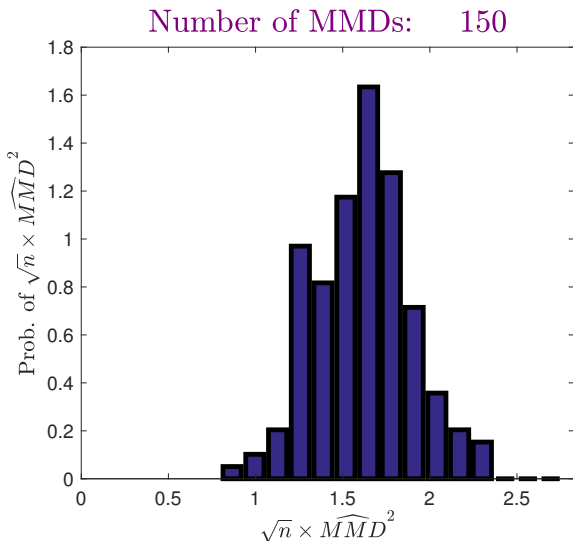


Number of MMDs:     150

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

Repeat this 300 times . . .
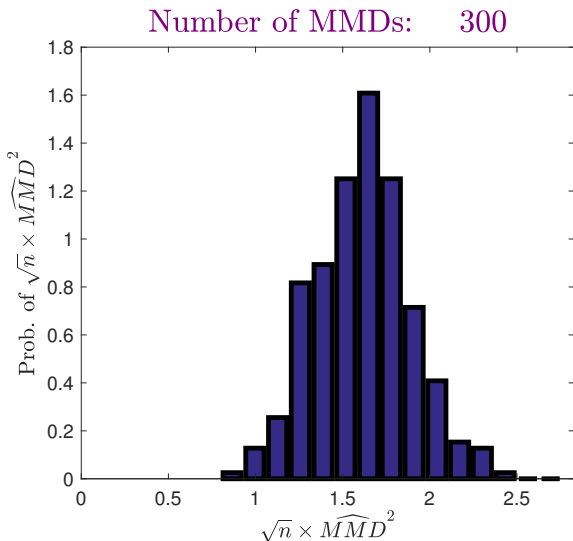


Number of MMDs: 300

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

Repeat this 3000 times ...



Number of MMDs:     3000
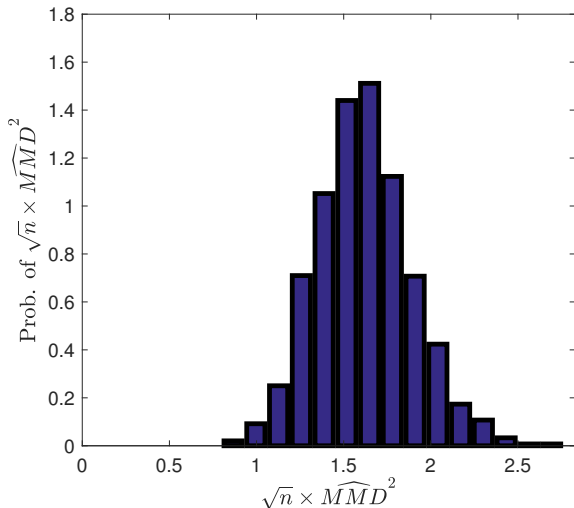
# Asymptotics of $\widehat{MMD}^2$ when $P \neq Q$

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

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

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



MMD density under $\mathcal{H}_1$



Two Laplace distributions with different variances
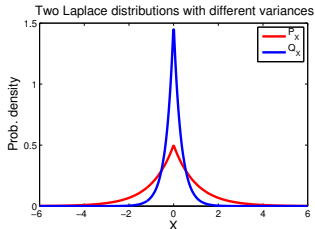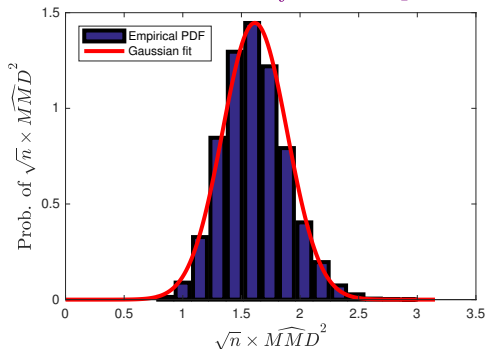
What happens when $P$ and $Q$ are the same?

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

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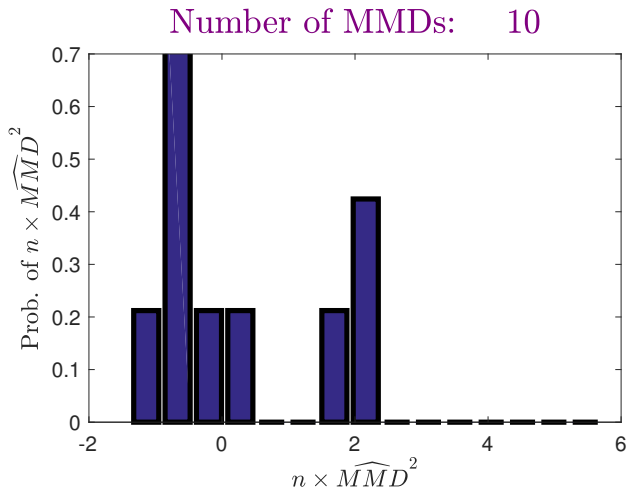


Number of MMDs:     10

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

■ Case of $P = Q = \mathcal{N}(0, 1)$



Number of MMDs:    20
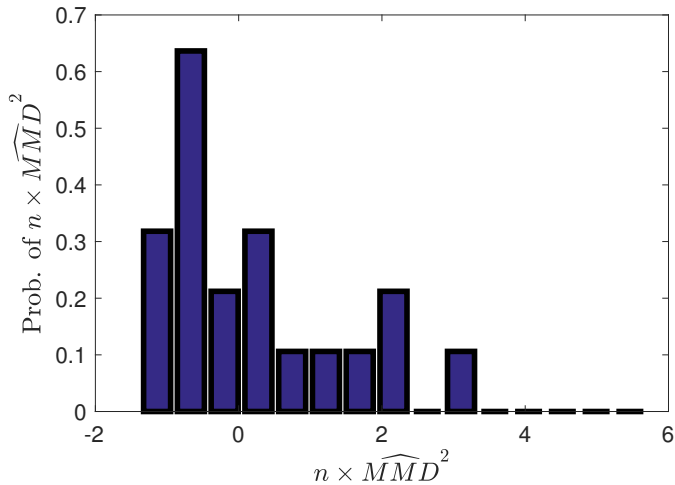
# Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



Number of MMDs:     50
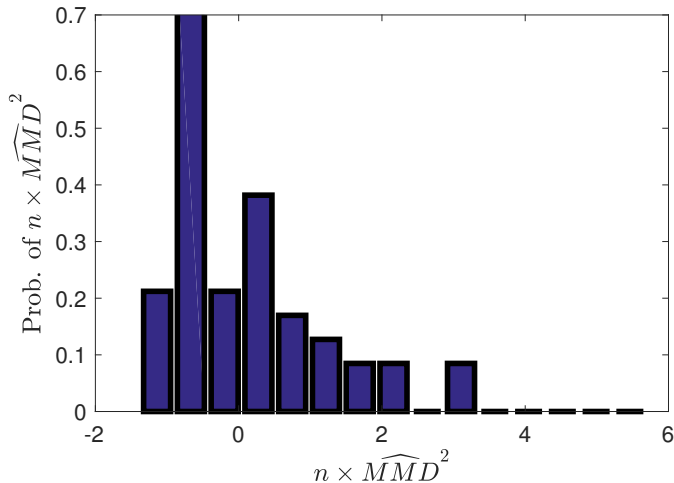
# Behaviour of $\widehat{MMD}^2$ when $P = Q$

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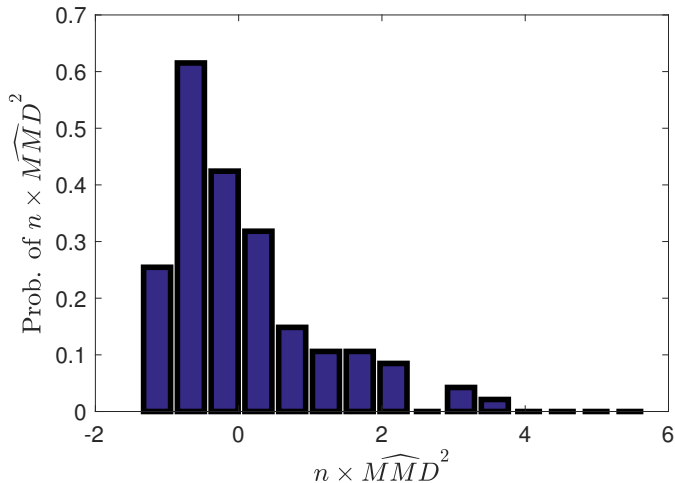


Number of MMDs:    100

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

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



Number of MMDs:    1000
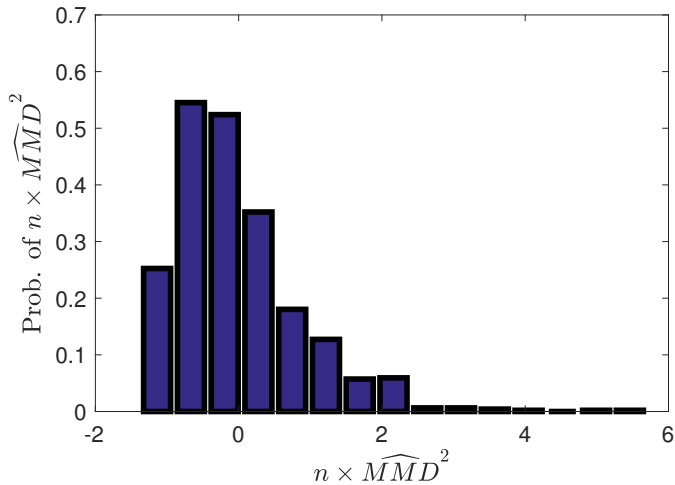
# Asymptotics of $\widehat{MMD}^2$ when $P = Q$

Where $P = Q$, statistic has asymptotic distribution

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



MMD density under $\mathcal{H}_0$

where

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

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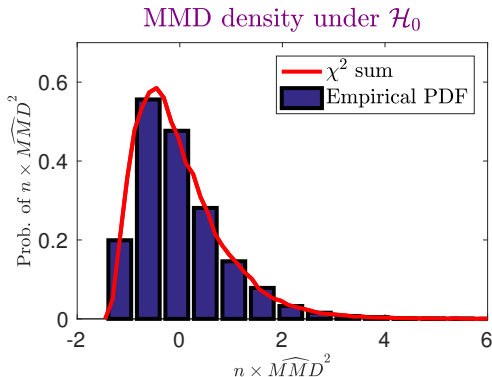
# A statistical test

A summary of the asymptotics:
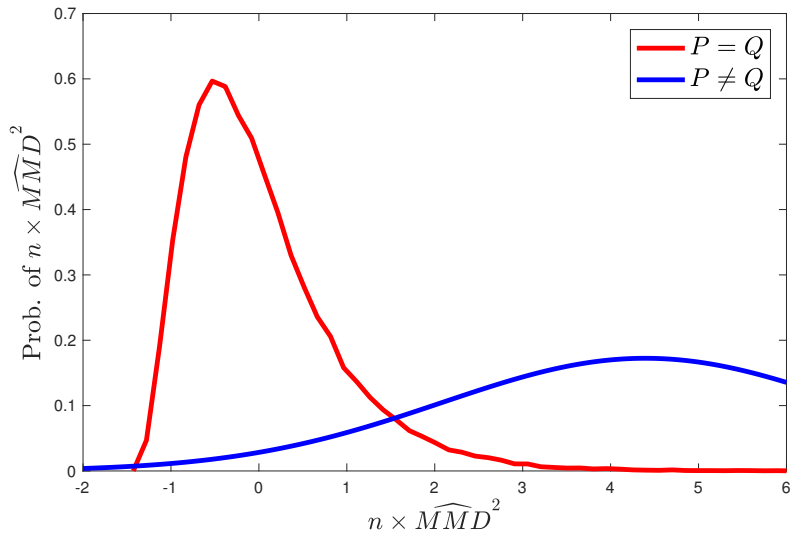
# A statistical test

**Test construction:** (G., Borgwardt, Rasch, Schoelkopf, and Smola, JMLR 2012)
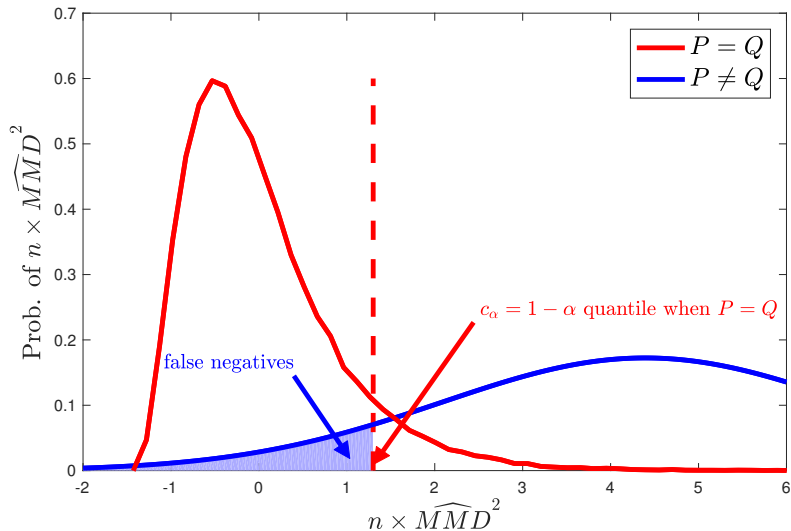
# How do we get test threshold $c_\alpha$?

Original empirical MMD for dogs and fish:

$$X = \left[ \begin{array}{cccc} \text{🐕} & \text{🐕} & \text{🐕} & \dots \end{array} \right]$$

$$Y = \left[ \begin{array}{cccc} \text{🐠} & \text{🐟} & \text{🐠} & \dots \end{array} \right]$$

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



$k(x_i, x_j)$   $k(x_i, y_j)$

$k(y_i, y_j)$

# How do we get test threshold $c_\alpha$?

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

$$\widetilde{X} = \left[\ \text{}\ \ldots\ \right]$$

$$\widetilde{Y} = \left[\ \text{}\ \ldots\ \right]$$

# How do we get test threshold $c_\alpha$?

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

$$\widetilde{X} = \left[ \begin{array}{cccc} & & & \cdots \end{array} \right]$$

$$\widetilde{Y} = \left[ \begin{array}{cccc} & & & \cdots \end{array} \right]$$

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

$$+ \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{y}_i, \tilde{y}_j)$$

$$- \frac{2}{n^2} \sum_{i,j} k(\tilde{x}_i, \tilde{y}_j)$$

Permutation simulates
$P = Q$



$k(\tilde{x}_i, \tilde{x}_j)$   $k(\tilde{x}_i, \tilde{y}_j)$

$k(\tilde{y}_i, \tilde{y}_j)$
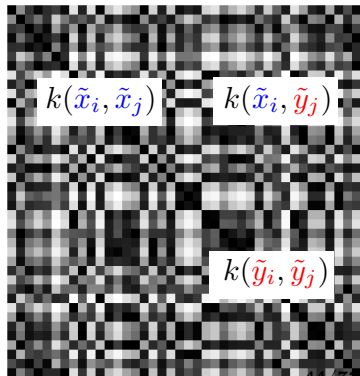
# How to choose the best kernel: optimising the kernel parameters

# Graphical illustration

■ Maximising test power same as minimizing false negatives
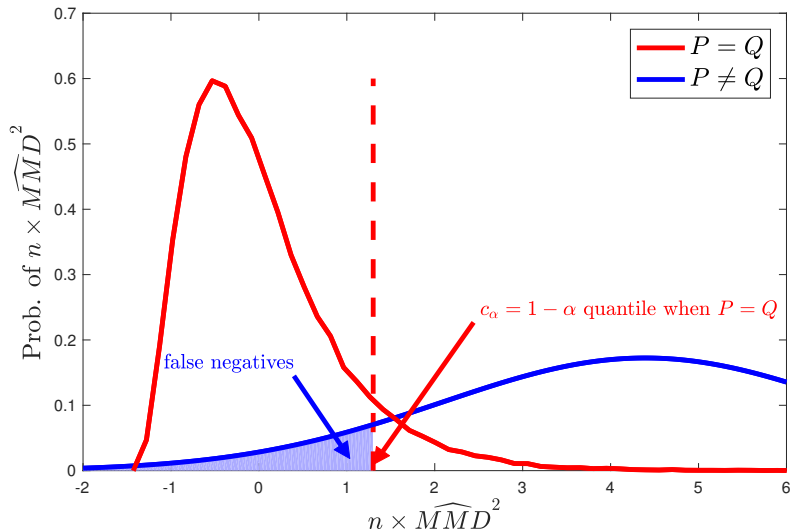
# Optimizing kernel for test power

The power of our test ($\text{Pr}_1$ denotes probability under $P \neq Q$):

$$\text{Pr}_1\left(n\widehat{\text{MMD}}^2 > \hat{c}_\alpha\right)$$

# Optimizing kernel for test power

The power of our test ($\mathrm{Pr}_1$ denotes probability under $P \neq Q$):

$$\mathrm{Pr}_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

$$\rightarrow \Phi\left(\frac{n\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{\sqrt{V_n(P,Q)}}\right)$$

where

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

# Optimizing kernel for test power

The power of our test ($\text{Pr}_1$ denotes probability under $P \neq Q$):

$$\text{Pr}_1\left(n\widehat{\text{MMD}}^2 > \hat{c}_\alpha\right)$$

$$\rightarrow \Phi\left(\underbrace{\frac{\text{MMD}^2(P, Q)}{\sqrt{V_n(P, Q)}}}_{O(n^{1/2})} - \underbrace{\frac{c_\alpha}{n\sqrt{V_n(P, Q)}}}_{O(n^{-1/2})}\right)$$

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 ($\mathrm{Pr}_1$ denotes probability under $P \neq Q$):

$$\mathrm{Pr}_1 \left( n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha \right)$$

$$\rightarrow \Phi \left( \frac{\mathrm{MMD}^2(P, Q)}{\sqrt{V_n(P, Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P, Q)}} \right)$$

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

# Troubleshooting for generative adversarial networks
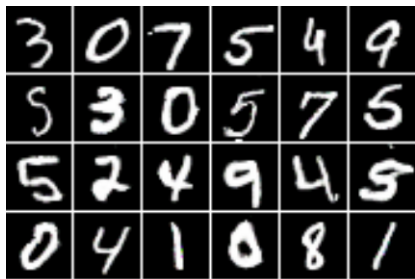


MNIST samples

Samples from a GAN

# Troubleshooting for generative adversarial networks



MNIST samples



Samples from a GAN
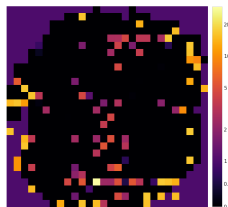


ARD map

- 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



$$MMD^2 = 0.0001$$
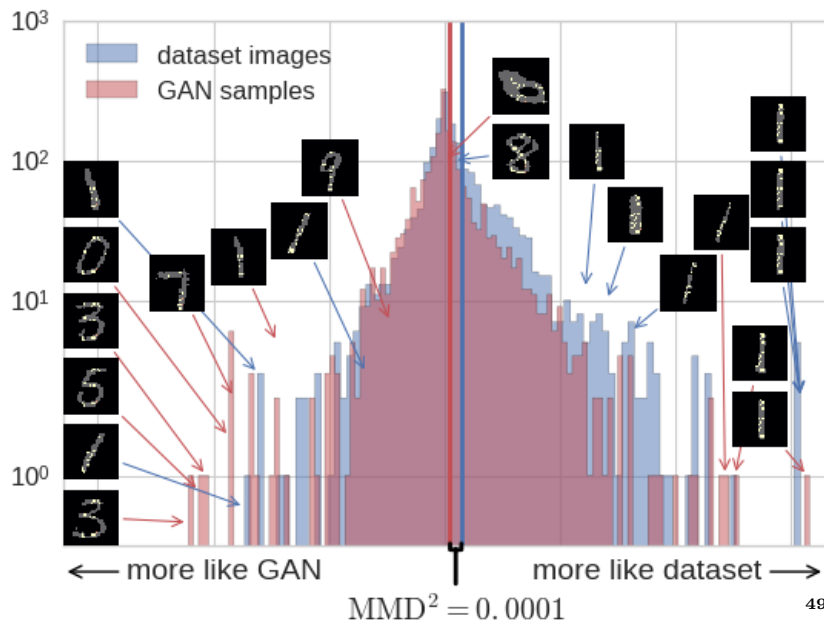
# Training GANs with MMD

# What is a Generative Adversarial Network (GAN)?

- **Generator** (student)

- **Critic** (teacher)



- Task: critic must teach generator to draw images (here dogs)
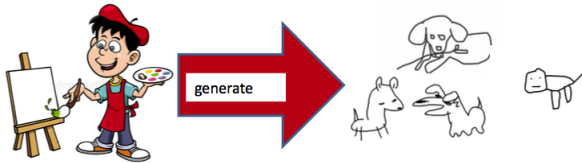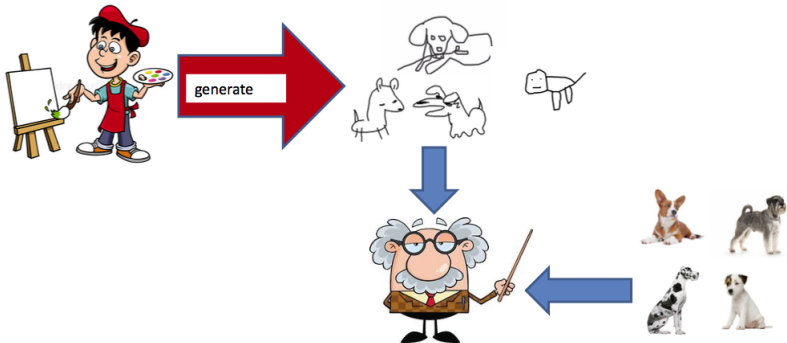
# What is a Generative Adversarial Network (GAN)?
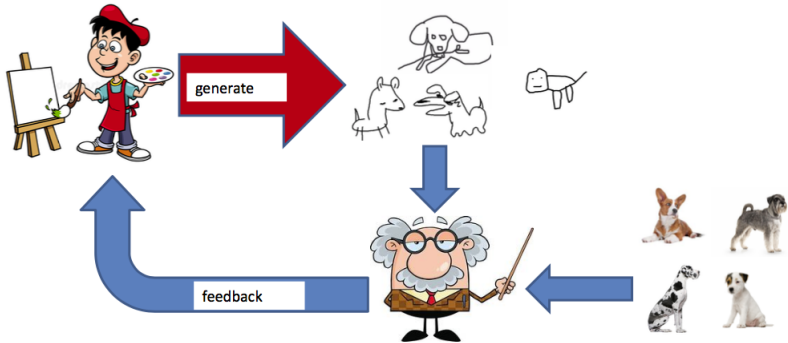
**Classification not enough!**
**Need to compare sets**
(otherwise student can just produce the same dog over and over)

# 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          Daniel M. Roy          Zoubin Ghahramani
University of Cambridge              University of Toronto       University of Cambridge

# MMD for GAN critic

Can you use MMD as a critic to train GANs?



Need better image features.

# How to improve the critic witness

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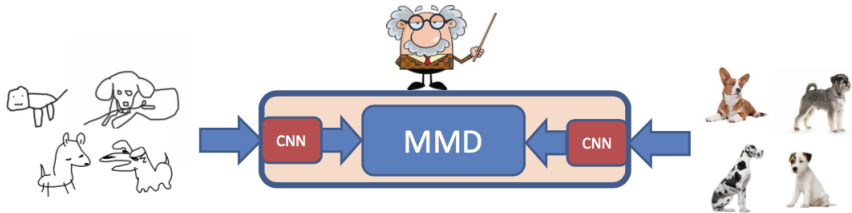


MMD GAN Li et al., [NIPS 2017]
Coulomb GAN Unterthiner et al., [ICLR 2018]

# WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]
WGAN-GP Gukrajani et al. [NIPS 2017]
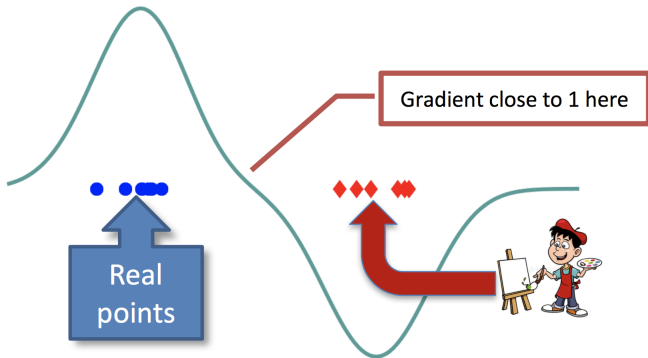


Gradient close to 1 here

Real points

# WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]
WGAN-GP Gukrajani et al. [NIPS 2017]

- Given a generator $G_\theta$ with parameters $\theta$ to be trained. Samples $Y \sim G_\theta(Z)$ where $Z \sim R$

- Given critic features $h_\psi$ with parameters $\psi$ to be trained. $f_\psi$ a linear function of $h_\psi$.

# WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]
WGAN-GP Gukrajani et al. [NIPS 2017]

- Given a generator $G_\theta$ with parameters $\theta$ to be trained. Samples $Y \sim G_\theta(Z)$ where $Z \sim R$

- Given critic features $h_\psi$ with parameters $\psi$ to be trained. $f_\psi$ a linear function of $h_\psi$.

WGAN-GP gradient penalty:

$$\max_\psi \mathbf{E}_{X \sim P} f_\psi(X) - \mathbf{E}_{Z \sim R} f_\psi(G_\theta(Z)) + \lambda \mathbf{E}_{\widetilde{X}} \left( \left\| \nabla_{\widetilde{X}} f_\theta(\widetilde{X}) \right\| - 1 \right)^2$$

where

$$\widetilde{X} = \gamma x_i + (1 - \gamma) G_\psi(z_j)$$
$$\gamma \sim \mathcal{U}([0, 1]) \quad x_i \in \{x_\ell\}_{\ell=1}^m \quad z_j \in \{z_\ell\}_{\ell=1}^n$$

# The (W)MMD

Train MMD critic features with the witness function gradient penalty

Binkowski, Sutherland, Arbel, G. [ICLR 2018], Bellemare et al. [2017] for energy distance:

$$\max_\psi MMD^2(h_\psi(X), h_\psi(G_\theta(Z))) + \lambda \mathbf{E}_{\widetilde{X}} \left( \left\| \nabla_{\widetilde{X}} f_\psi(\widetilde{X}) \right\| - 1 \right)^2$$

where

$$f_\psi(\cdot) = \frac{1}{m} \sum_{i=1}^{m} k(h_\psi(x_i), \cdot) - \frac{1}{n} \sum_{j=1}^{n} k(h_\psi(G_\theta(z_j)), \cdot)$$

**New**

$$\widetilde{X} = \gamma x_i + (1 - \gamma) G_\psi(z_j)$$
$$\gamma \sim \mathcal{U}([0,1]) \quad x_i \in \{x_\ell\}_{\ell=1}^{m} \quad z_j \in \{z_\ell\}_{\ell=1}^{n}$$

Remark by Bottou et al. (2017): gradient penalty modifies the function class. So critic is not an MMD in RKHS $\mathcal{F}$.

# MMD for GAN critic: revisited

## 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
University College London
{dougal,michael.n.arbel,arthur.gretton}@gmail.com

# MMD for GAN critic: revisited
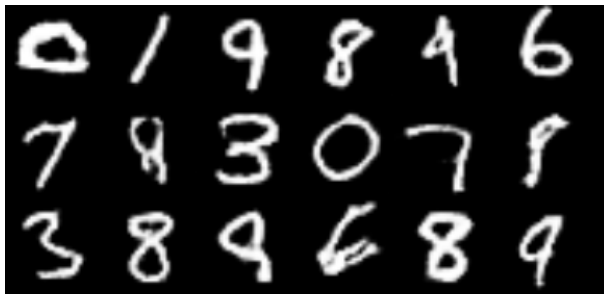


Samples are better!

# MMD for GAN critic: revisited
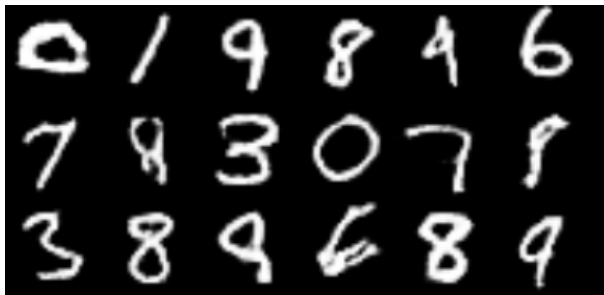


Samples are better!

Can we do better still?

# Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty *may not converge near solution*

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

The Dirac-GAN

$$P = \delta_0 \qquad Q = \delta_\theta \qquad f_\psi(x) = \psi \cdot x$$



Figure from Mescheder et al. [ICML 2018]
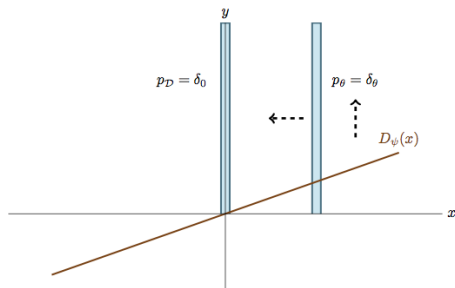
# Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

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

The Dirac-GAN

$$P = \delta_0 \qquad Q = \delta_\theta \qquad f_\psi(x) = \psi \cdot x$$

# A better gradient penalty

- **New MMD GAN witness regulariser (NIPS 2018)**
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

arXiv.org > stat > arXiv:1805.11565

Statistics > Machine Learning

**On gradient regularizers for MMD GANs**

Michael Arbel, Dougal J. Sutherland, Mikołaj Bińkowski, Arthur Gretton

*(Submitted on 29 May 2018)*

# A better gradient penalty

- **New MMD GAN witness regulariser (NIPS 2018)**
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]



Should be flat here

Can be steep here

Real points

# A better gradient penalty

- **New MMD GAN witness regulariser (NIPS 2018)**
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

$$\widetilde{MMD} := \sup_{\boxed{\|f\|_S \leq 1}} [\mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)]$$

where

$$\|f\|_S^2 = \|f\|_{L_2(P)}^2 + \|\nabla f\|_{L_2(P)}^2 + \lambda \|f\|_k^2$$

L$_2$ norm control  Gradient control  RKHS smoothness

# A better gradient penalty

- New MMD GAN witness regulariser (NIPS 2018)
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

$$\widetilde{MMD} := \sup_{\boxed{\|f\|_S \leq 1}} \left[ \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \right]$$

where

$$\|f\|_S^2 = \|f\|_{L_2(P)}^2 + \|\nabla f\|_{L_2(P)}^2 + \lambda \|f\|_k^2$$

| L₂ norm control | Gradient control | RKHS smoothness |

Problem: not computationally feasible: $O(n^3)$ per iteration.

# A better gradient penalty

- New MMD GAN witness regulariser (NIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

$$SMMD = \sigma_{k,P,\lambda} \ MMD$$

where

$$\sigma_{k,P,\lambda} = \left( \ \lambda + \int k(x,x)dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x,x) \ dP(x) \right)^{-1/2}$$

Replace expensive constraint with cheap upper bound:

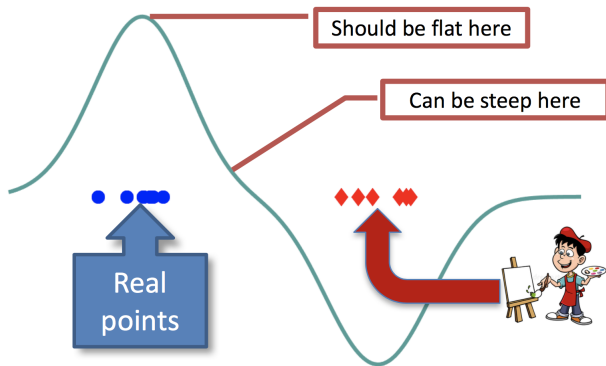$$\|f\|_S^2 \leq \sigma_{k,P,\lambda}^{-1} \ \|f\|_k^2$$

# A better gradient penalty

- **New MMD GAN witness regulariser (NIPS 2018)**
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

$$SMMD = \sigma_{k,P,\lambda} \; MMD$$

where

$$\sigma_{k,P,\lambda} = \left( \lambda + \int k(x,x) dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x,x) \; dP(x) \right)^{-1/2}$$

Replace expensive constraint with cheap upper bound:

$$\|f\|_S^2 \leq \sigma_{k,P,\lambda}^{-1} \; \|f\|_k^2$$

Idea: rather than regularise the critic or witness function, regularise features directly

# Evaluation and experiments

# Evaluation of GANs

The inception score? <small>Salimans et al. [NIPS 2016]</small>

Based on the classification output $p(y|x)$ of the inception model <small>Szegedy et al. [ICLR 2014]</small>,

$$E_X \exp KL(P(y|X)\|P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

# Evaluation of GANs

The inception score? <span style="font-size:smaller">Salimans et al. [NIPS 2016]</span>

Based on the classification output $p(y|x)$ of the inception model <span style="font-size:smaller">Szegedy et al. [ICLR 2014]</span>,

$$E_X \exp KL(P(y|X) \| P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

Problem: relies on a trained classifier! Can't be used on new categories (celeb, bedroom...)

# Evaluation of GANs

The Frechet inception distance? <span style="font-size:small">Heusel et al. [NIPS 2017]</span>

Fits Gaussians to features in the inception architecture (pool3 layer):

$$FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}(\Sigma_P) + \text{tr}(\Sigma_Q) - 2\text{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)$$

where $\mu_P$ and $\Sigma_P$ are the feature mean and covariance of $P$

# Evaluation of GANs

The Frechet inception distance? <span>Heusel et al. [NIPS 2017]</span>

Fits Gaussians to features in the inception architecture (pool3 layer):

$$FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}(\Sigma_P) + \text{tr}(\Sigma_Q) - 2\text{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)$$

where $\mu_P$ and $\Sigma_P$ are the feature mean and covariance of $P$

Problem: **bias**. For finite samples can consistently give incorrect answer.

- Bias demo, CIFAR-10 train vs test

# Evaluation of GANs

The FID can give the wrong answer in theory.

Assume $m$ samples from $P$ and $n \to \infty$ samples from $Q$.
Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \qquad P_2 \sim \mathcal{N}(0, 1) \qquad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given $m$ samples from $P_1$ and $P_2$,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

# Evaluation of GANs

The FID can give the wrong answer in theory.

Assume $m$ samples from $P$ and $n \to \infty$ samples from $Q$.

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \qquad P_2 \sim \mathcal{N}(0, 1) \qquad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given $m$ samples from $P_1$ and $P_2$,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

# Evaluation of GANs

The FID can give the wrong answer in theory.

Assume $m$ samples from $P$ and $n \to \infty$ samples from $Q$.

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \qquad P_2 \sim \mathcal{N}(0, 1) \qquad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given $m$ samples from $P_1$ and $P_2$,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

# Evaluation of GANs

The FID can give the wrong answer in theory.

Assume $m$ samples from $P$ and $n \to \infty$ samples from $Q$.

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \qquad P_2 \sim \mathcal{N}(0, 1) \qquad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given $m$ samples from $P_1$ and $P_2$,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

# Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \qquad Q = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d} C C^T$, with $C$ a $d \times d$ matrix with iid standard normal entries.

For a random draw of $C$:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P_1}, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P_2}, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of $C$.

# Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \qquad Q = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d} C C^T$, with $C$ a $d \times d$ matrix with iid standard normal entries.

For a random draw of $C$:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P_1}, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P_2}, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.
This behavior is similar for other random draws of $C$.

# Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2 I_d)) \quad Q = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d} CC^T$, with $C$ a $d \times d$ matrix with iid standard normal entries.

For a random draw of $C$:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P_1}, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P_2}, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.
This behavior is similar for other random draws of $C$.

# Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \quad Q = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d}CC^T$, with $C$ a $d \times d$ matrix with iid standard normal entries.

For a random draw of $C$:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P_1}, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P_2}, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.
This behavior is similar for other random draws of $C$.

# The kernel inception distance (KID)

The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left( \frac{1}{d} x^\top y + 1 \right)^3.$$



- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test

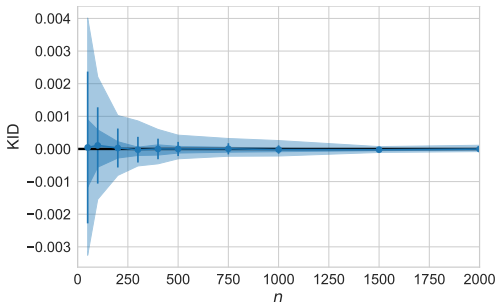# The kernel inception distance (KID)

The Kernel inception distance  Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left( \frac{1}{d} x^\top y + 1 \right)^3 .$$



- Checks match for feature means, variances, skewness

- Unbiased : eg CIFAR-10 train/test

…"but isn't KID is computationally costly?"

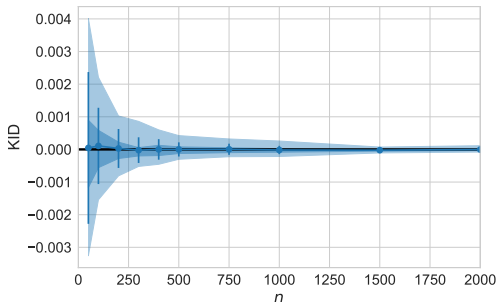# The kernel inception distance (KID)

The Kernel inception distance <span style="color:gray">Binkowski, Sutherland, Arbel, G. [ICLR 2018]</span>

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d}x^\top y + 1\right)^3.$$



- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test

..."but isn't KID is computationally costly?"

"Block" KID implementation is cheaper than FID: see paper (or use Tensorflow implementation)!

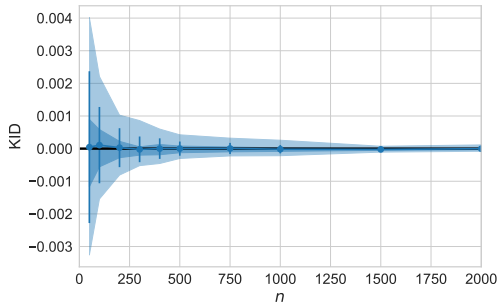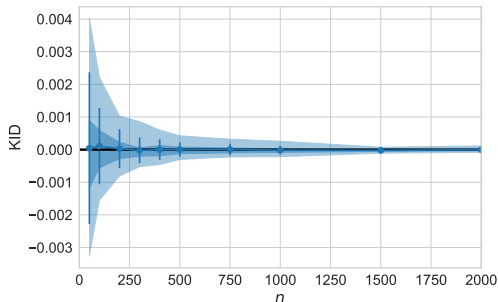# The kernel inception distance (KID)

The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left( \frac{1}{d} x^\top y + 1 \right)^3.$$



- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test

Also used for automatic learning rate adjustment: if $KID(\widehat{P}_{t+1}, Q)$ not significantly better than $KID(\widehat{P}_t, Q)$ then reduce learning rate.
[Bounliphone et al. ICLR 2016]

Related: "An empirical study on evaluation metrics of generative adversarial networks", Xu et al. [arxiv, June 2018]

## SPECTRAL NORMALIZATION FOR GENERATIVE ADVERSARIAL NETWORKS

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

## SOBOLEV GAN

Youssef Mroueh[1], Chun-Liang Li[◊,*], Tom Sercu[1,*], Anant Raj[◊,*] & Yu Cheng[1]
† 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
Univer...y College London
{...y, michael.n.arbel, arthur.gretton}@gmail.com

## BOUNDARY-SEEKING GENERATIVE ADVERSARIAL NETWORKS

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

Athul Paul Jacob[*]
MILA, 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

We combine with scaled MMD

Our ICLR 2018 paper

# Results: what does MMD buy you?

- **Critic** features from **DCGAN**: an $f$-filter critic has $f$, $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN $64 \times 64$.
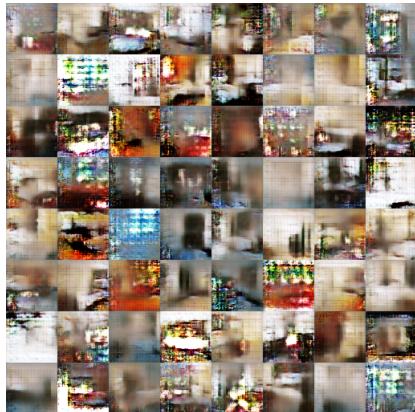
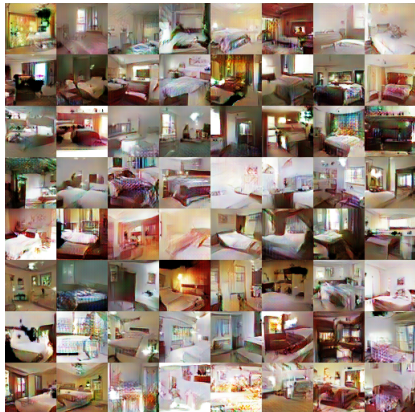

MMD GAN samples, $f = 64$,
KID=3



WGAN samples, $f = 64$,
KID=4

# Results: what does MMD buy you?

■ Critic features from DCGAN: an $f$-filter critic has $f$, $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN $64 \times 64$.
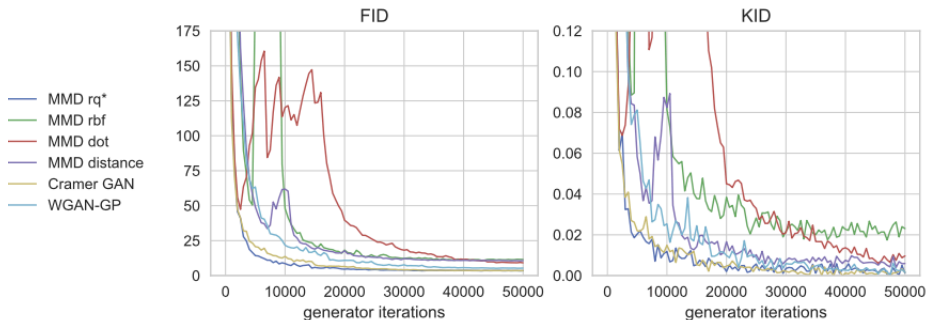


MMD GAN samples, $f = 16$,
KID=9

WGAN samples, $f = 16$,
$f = 64$, KID=37

# The kernel inception distance (KID)

**Faster training:** performance scores vs generator iterations on MNIST

# Results: celebrity faces 160×160

KID scores:

- Sobolev GAN:
  14

- SN-GAN:
  18

- Old MMD
  GAN:
  13

- SMMD GAN:
  6

202 599 face images, re-
sized and cropped to 160
× 160

# Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
  47

- SN-GAN:
  44

- SMMD GAN:
  35

ILSVRC2012 (ImageNet)
dataset, 1 281 167 im-
ages, resized to 64 × 64.
Around 20 000 classes.

# Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
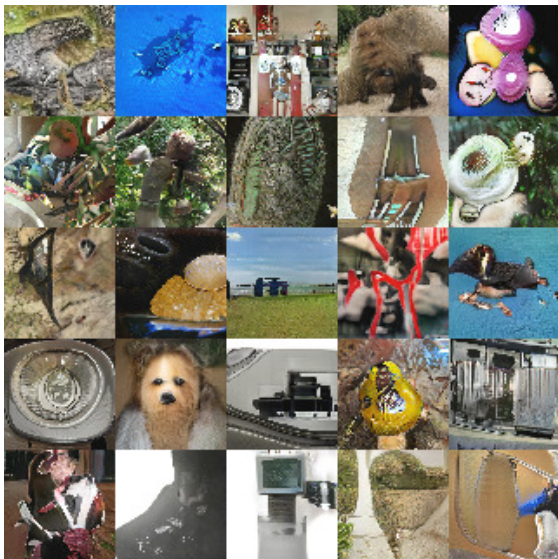  47

- SN-GAN:
  44

- SMMD GAN:
  35

ILSVRC2012 (ImageNet)
dataset, 1 281 167 im-
ages, resized to 64 × 64.
Around 20 000 classes.
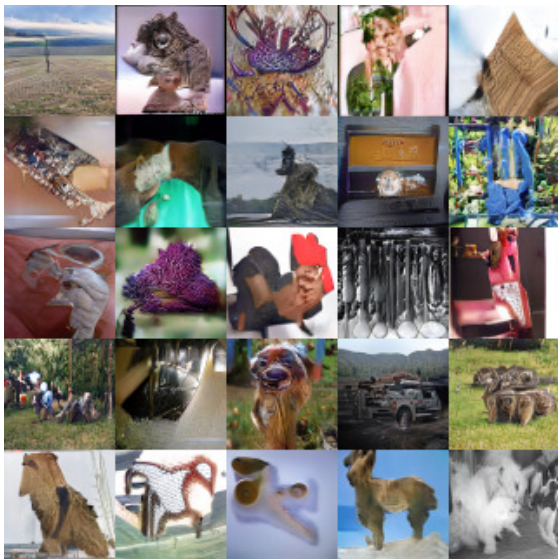
# Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
  47

- SN-GAN:
  44

- SMMD GAN:
  35

ILSVRC2012 (ImageNet)
dataset, 1 281 167 im-
ages, resized to 64 × 64.
Around 20 000 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 $h_\psi$ features possible.
  - Better gradient/feature regulariser gives better critic

Code for "Demystifying MMD GANs," ICLR 2018, including KID score: https://github.com/mbinkowski/MMD-GAN

Code for new SMMD:
https://github.com/MichaelArbel/Scaled-MMD-GAN

# Co-authors

## From Gatsby:

- Mikolaj Binkowski
- Kacper Chwialkowski
- Wittawat Jitkrittum
- Heiko Strathmann
- Dougal Sutherland
- Wenkai Xu

## External collaborators:

- Kenji Fukumizu
- Bernhard Schoelkopf
- Dino Sejdinovic
- Bharath Sriperumbudur
- Alex Smola
- Zoltan Szabo

# Questions?