The Maximum Mean Discrepancy for Training Generative Adversarial Networks

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Paris, 2019

A motivation: comparing two samples

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

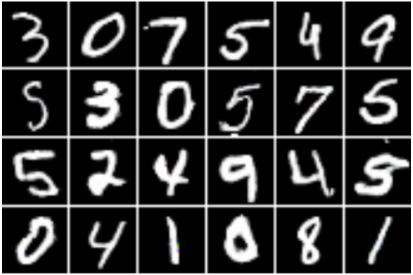
$$\sim P$$

$$\sim Q$$

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

Training generative models

Contribute Search jobs Dating Sign in Search UK edition

Opinion Sport Culture Lifestyle More

The Guardian

radio Books **Art & design** Stage Games Classical

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

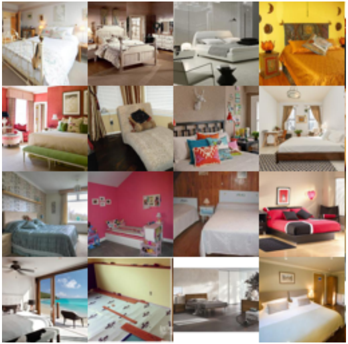
An image of Edmond de Belamy, created by a computer, has just been sold at Christie's. But no algorithm can capture our complex human consciousness

1,085 455

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

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

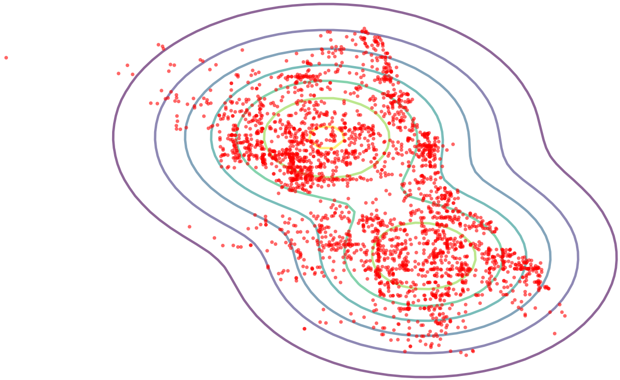
Generated Q, MMD GAN

Using MMD to train a GAN

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

Part 2: 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.

Part 2: 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!)

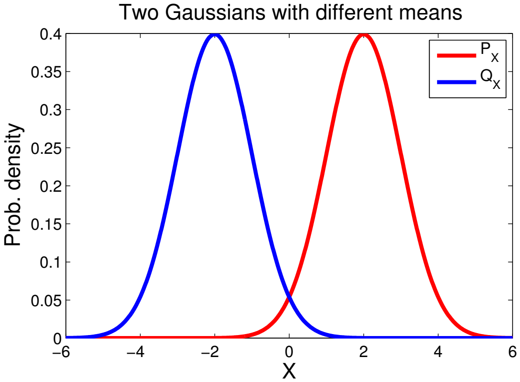
- A statistical test based on the MMD

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

Maximum Mean Discrepancy

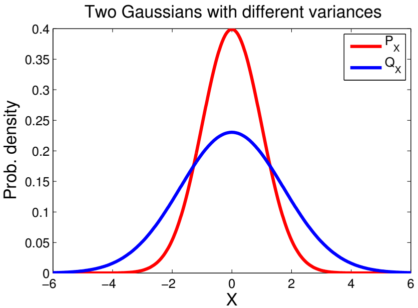
Feature mean difference

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



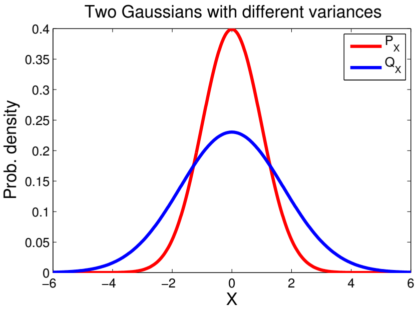
Feature mean difference

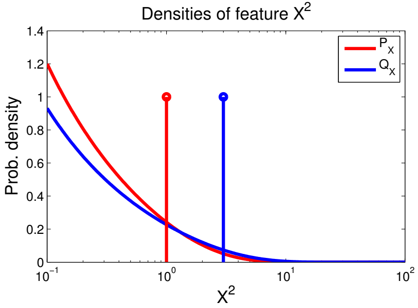
- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

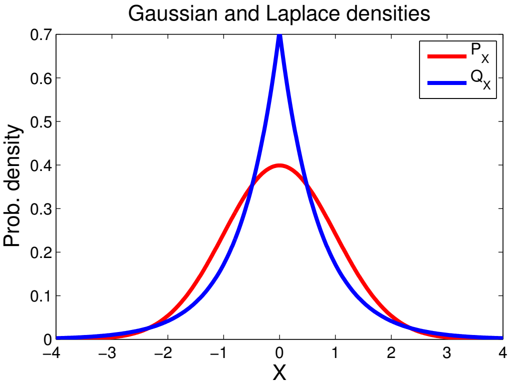
- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$





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,

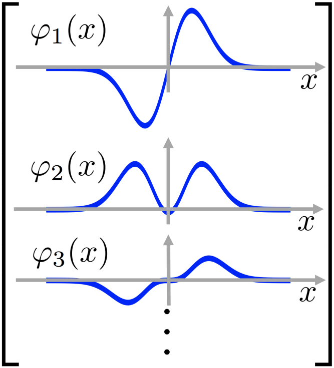
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

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

Exponentiated quadratic kernel

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

$$\varphi(x) =$$



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

Feature space construction: details

Consider (truncated) Gaussian density on $\mathcal{X} \subset \mathbb{R}$,

$$p(x) \propto \exp\left(-x^2\right) \mathbb{I}_{\mathcal{X}}(x)$$

Define the eigenexpansion of $k(x, x')$ wrt this density:

$$\lambda_\ell e_\ell(x) = \int_{\mathcal{X}} k(x, x') e_\ell(x') p(x') dx' \qquad \int_{\mathcal{X}} e_i(x) e_j(x) p(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x') = \sum_{\ell=1}^{\infty} \underbrace{\left(\sqrt{\lambda_\ell} e_\ell(x)\right)}_{\varphi_\ell(x)} \underbrace{\left(\sqrt{\lambda_\ell} e_\ell(x')\right)}_{\varphi_\ell(x')}$$

which converges in $L_2(p)$.

Warning: for RKHS, need absolute and uniform convregence, eg via Mercer's theorem for compact $\mathcal{X}$.

Feature space construction: details

Consider (truncated) Gaussian density on $\mathcal{X} \subset \mathbb{R}$,

$$p(x) \propto \exp\left(-x^2\right) \mathbb{I}_{\mathcal{X}}(x)$$

Define the eigenexpansion of $k(x, x')$ wrt this density:

$$\lambda_\ell e_\ell(x) = \int_{\mathcal{X}} k(x, x') e_\ell(x') p(x') dx' \qquad \int_{\mathcal{X}} e_i(x) e_j(x) p(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x') = \sum_{\ell=1}^{\infty} \underbrace{\left(\sqrt{\lambda_\ell} e_\ell(x)\right)}_{\varphi_\ell(x)} \underbrace{\left(\sqrt{\lambda_\ell} e_\ell(x')\right)}_{\varphi_\ell(x')}$$

which converges in $L_2(p)$.

Warning: for RKHS, need absolute and uniform convregence, eg via Mercer's theorem for compact $\mathcal{X}$.

Infinitely many features of *distributions*

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

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

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

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

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

Fine print: is this allowed for infinite feature spaces?

Infinitely many features of *distributions*

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

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

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

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

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

Fine print: is this allowed for infinite feature spaces?

Does the feature space mean exist?

Does there exist an element $\mu_P \in \mathcal{F}$ such that

$$\mathbf{E}_P f(x) = \langle f, \mu_P \rangle_{\mathcal{F}} \qquad \forall f \in \mathcal{F}$$

We recall the concept of a bounded operator: a linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$|Af| \leq \lambda_A \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

Riesz representation theorem: In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$

Does the feature space mean exist?

Does there exist an element $\mu_P \in \mathcal{F}$ such that

$$\mathbf{E}_P f(x) = \langle f, \mu_P \rangle_{\mathcal{F}} \qquad \forall f \in \mathcal{F}$$

We recall the concept of a bounded operator: a linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$|Af| \leq \lambda_A \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

Riesz representation theorem: In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$

Does the feature space mean exist?

Does there exist an element $\mu_P \in \mathcal{F}$ such that

$$\mathbf{E}_P f(x) = \langle f, \mu_P \rangle_{\mathcal{F}} \qquad \forall f \in \mathcal{F}$$

We recall the concept of a bounded operator: a linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$|Af| \leq \lambda_A ||f||_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

Riesz representation theorem: In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathbf{E}_P\sqrt{k(x,x)} = \mathbf{E}_P\|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathbf{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$\begin{aligned}
|T_P f| &= |\mathbf{E}_P f(x)|\,. \\
&\leq \mathbf{E}_P |f(x)| \\
&= \mathbf{E}_P |\langle f, \varphi(x)\rangle_{\mathcal{F}}| \\
&\leq \mathbf{E}_P \left(\sqrt{k(x,x)}\,\|f\|_{\mathcal{F}}\right)
\end{aligned}$$

Hence by Riesz (with $\lambda_{T_P} = \mathbf{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P\rangle_{\mathcal{F}}\,.$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathbf{E}_P\sqrt{k(x,x)} = \mathbf{E}_P\|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathbf{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$\begin{aligned}
|T_P f| &= |\mathbf{E}_P f(x)|. \\
&\leq \mathbf{E}_P |f(x)| \\
&= \mathbf{E}_P |\langle f, \varphi(x) \rangle_{\mathcal{F}}| \\
&\leq \mathbf{E}_P \left(\sqrt{k(x,x)}\, \|f\|_{\mathcal{F}} \right)
\end{aligned}$$

Hence by Riesz (with $\lambda_{T_P} = \mathbf{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P \rangle_{\mathcal{F}}.$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathbf{E}_P\sqrt{k(x,x)} = \mathbf{E}_P\,\|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathbf{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$\begin{aligned}
|T_P f| &= |\mathbf{E}_P f(x)|\,. \\
&\le \mathbf{E}_P\,|f(x)| \\
&= \mathbf{E}_P\,|\langle f, \varphi(x)\rangle_{\mathcal{F}}| \\
&\le \mathbf{E}_P\left(\sqrt{k(x,x)}\,\|f\|_{\mathcal{F}}\right)
\end{aligned}$$

Hence by Riesz (with $\lambda_{T_P} = \mathbf{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P\rangle_{\mathcal{F}}\,.$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathbf{E}_P\sqrt{k(x,x)} = \mathbf{E}_P \|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathbf{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$
\begin{aligned}
|T_P f| &= |\mathbf{E}_P f(x)|. \\
&\leq \mathbf{E}_P |f(x)| \\
&= \mathbf{E}_P |\langle f, \varphi(x)\rangle_{\mathcal{F}}| \\
&\leq \mathbf{E}_P \left(\sqrt{k(x,x)}\, \|f\|_{\mathcal{F}}\right)
\end{aligned}
$$

Hence by Riesz (with $\lambda_{T_P} = \mathbf{E}_P\sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P \rangle_{\mathcal{F}}.$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathbf{E}_P \sqrt{k(x,x)} = \mathbf{E}_P \|\varphi(x)\|_{\mathcal{F}} < \infty$ then $\exists \mu_P \in \mathcal{F}$.

Proof:

The linear operator $T_P f := \mathbf{E}_P f(x)$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$\begin{aligned}
|T_P f| &= |\mathbf{E}_P f(x)| . \\
&\leq \mathbf{E}_P |f(x)| \\
&= \mathbf{E}_P |\langle f, \varphi(x) \rangle_{\mathcal{F}}| \\
&\leq \mathbf{E}_P \left(\sqrt{k(x,x)} \|f\|_{\mathcal{F}} \right)
\end{aligned}$$

Hence by Riesz (with $\lambda_{T_P} = \mathbf{E}_P \sqrt{k(x,x)}$), $\exists \mu_P \in \mathcal{F}$ such that

$$T_P f = \langle f, \mu_P \rangle_{\mathcal{F}} .$$

The maximum mean discrepancy

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

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

The maximum mean discrepancy

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

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

The maximum mean discrepancy

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

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

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

Illustration of MMD

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

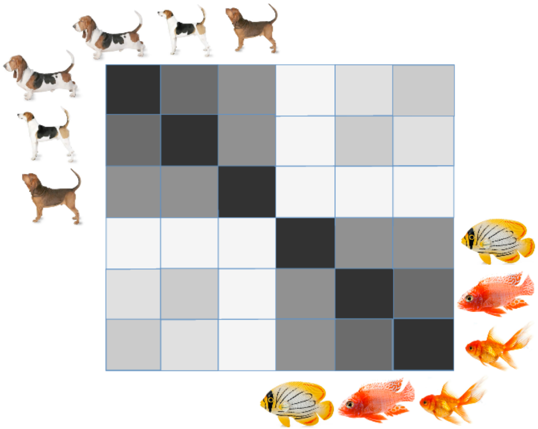
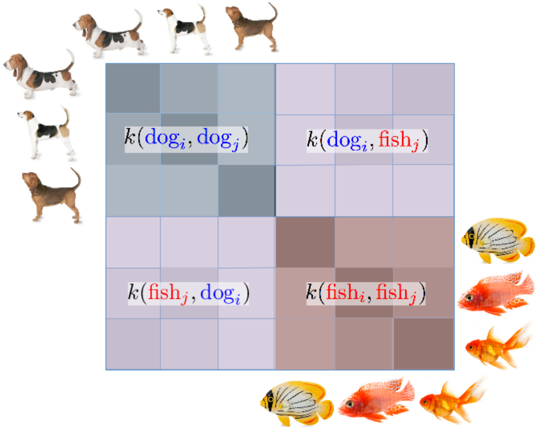


Illustration of MMD

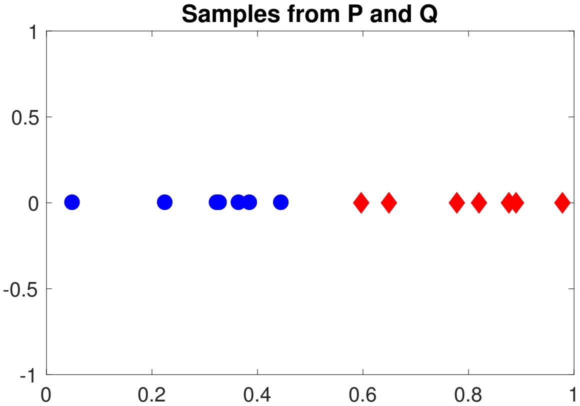
The maximum mean discrepancy:

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



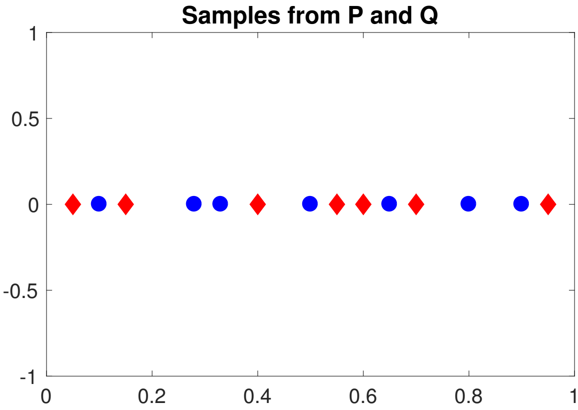
MMD as an integral probability metric

Are P and Q different?



MMD as an integral probability metric

Are P and Q different?

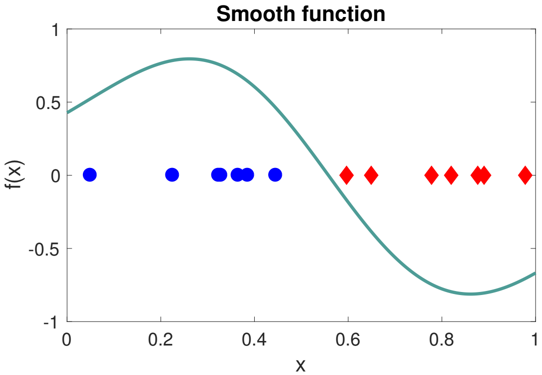


MMD as an integral probability metric

Integral probability metric:

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

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

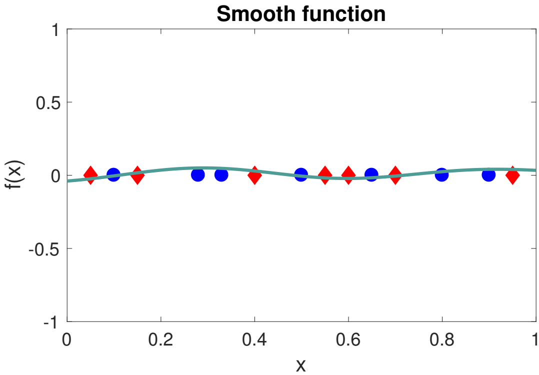


MMD as an integral probability metric

Integral probability metric:

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

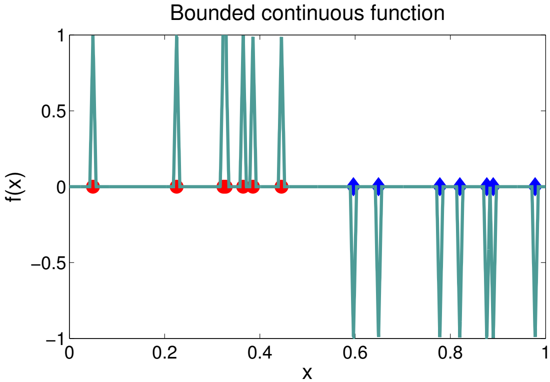
$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$



MMD as an integral probability metric

What if the function is not smooth?

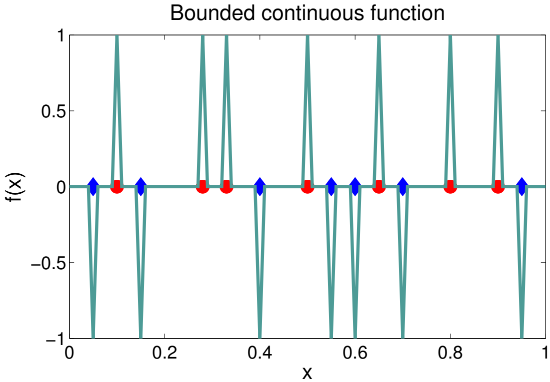
$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$



MMD as an integral probability metric

What if the function is not smooth?

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

MMD as an integral probability metric

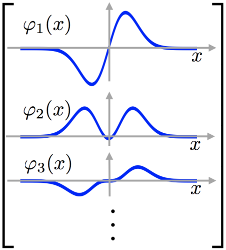
Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

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



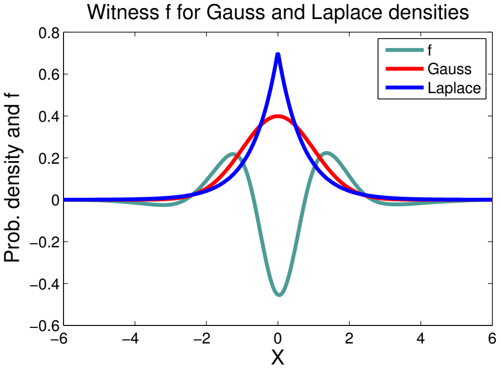
$$\|f\|_{\mathcal{F}}^2 := \sum_{i=1}^{\infty} f_i^2 \leq 1$$

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

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]

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

Reminder for next slide: expectations of functions are linear combinations of expected features

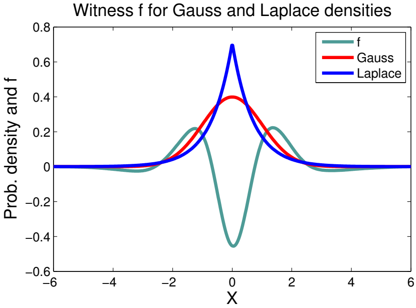
$$\mathbf{E}_P(f(X)) = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

Integral prob. metric vs feature difference

The MMD:

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



Integral prob. metric vs feature difference

The MMD:

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

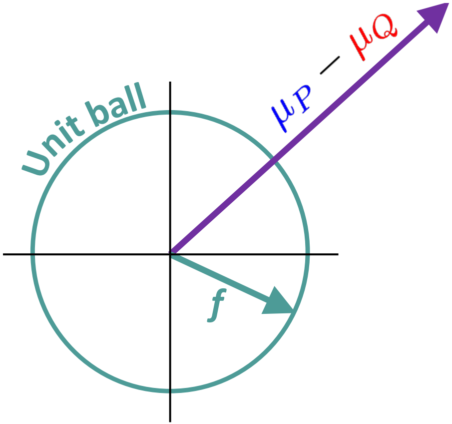
use

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

Integral prob. metric vs feature difference

The MMD:

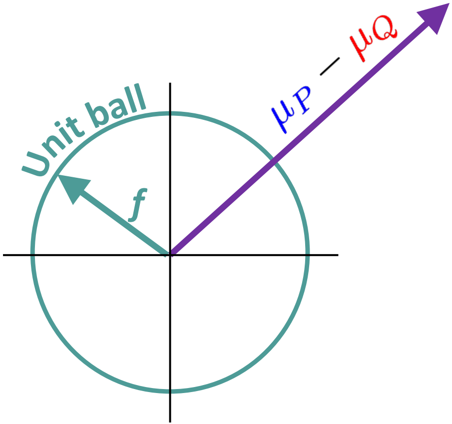
$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
&= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature difference

The MMD:

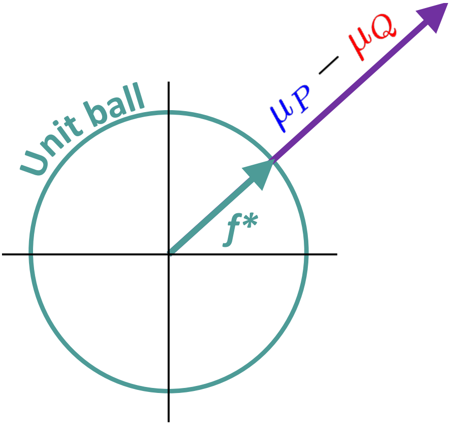
$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right] \\
&= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature difference

The MMD:

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



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

Integral prob. metric vs feature difference

The MMD:

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

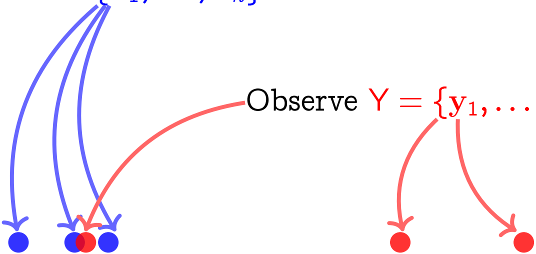
$$\boxed{\text{Function view and feature view equivalent}}$$

Construction of MMD witness

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

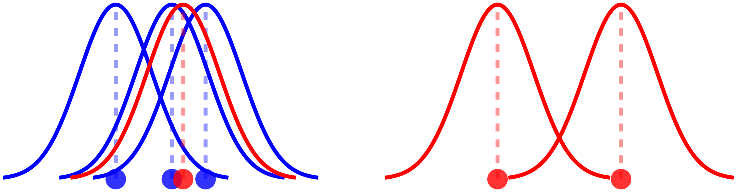
Observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$

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



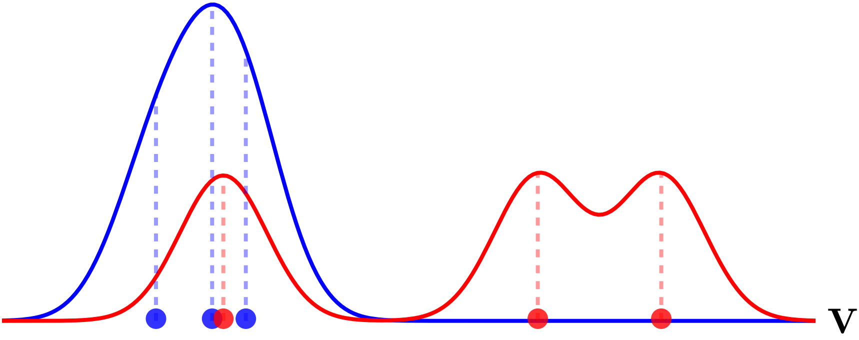
Construction of MMD witness

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



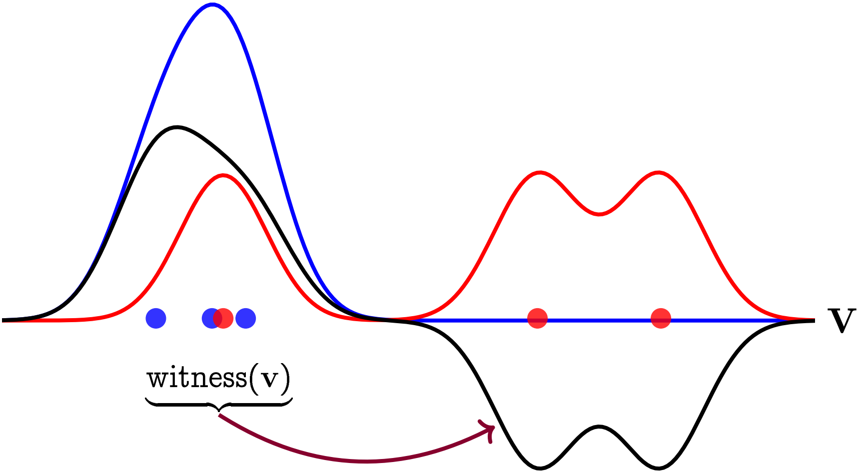
Construction of MMD witness

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

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

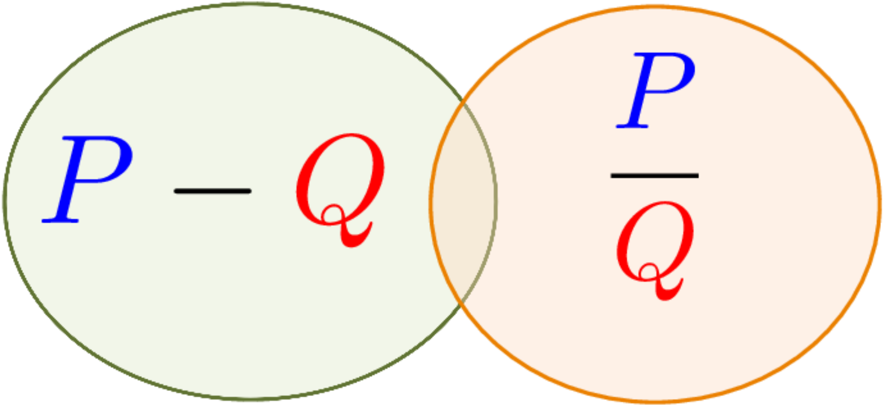
The empirical witness function at v

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

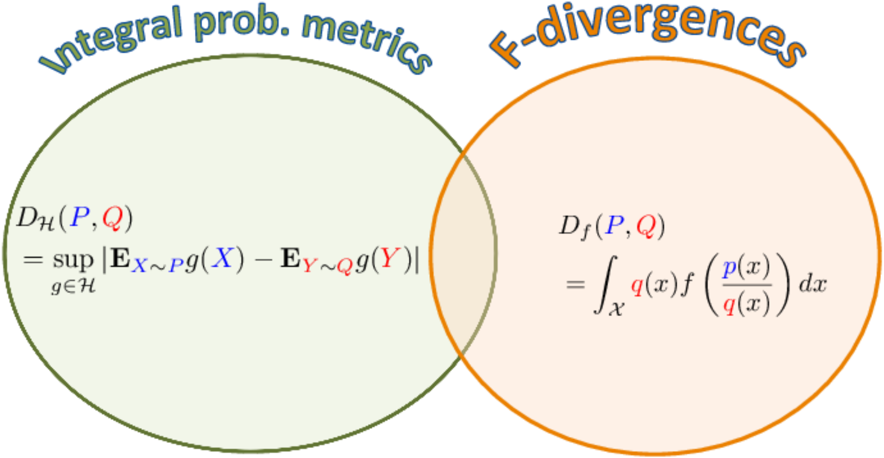
Don't need explicit feature coefficients $f^* := \begin{bmatrix} f_1^* & f_2^* & \dots \end{bmatrix}$

Interlude: divergence measures

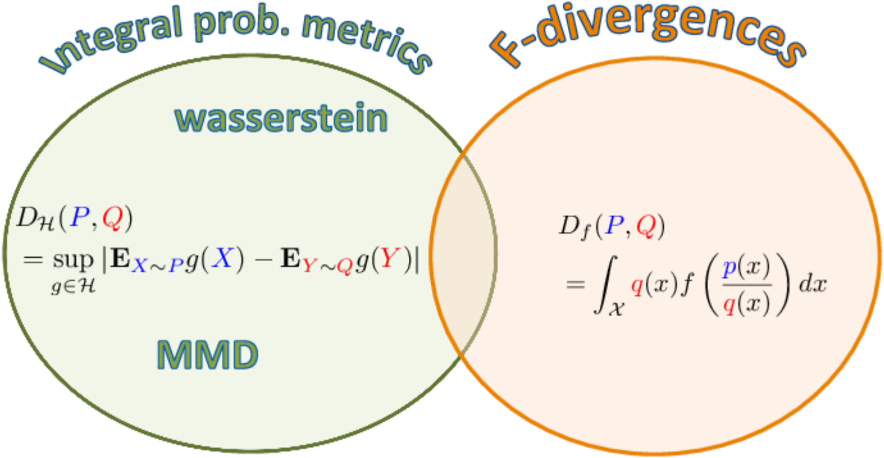
Divergences



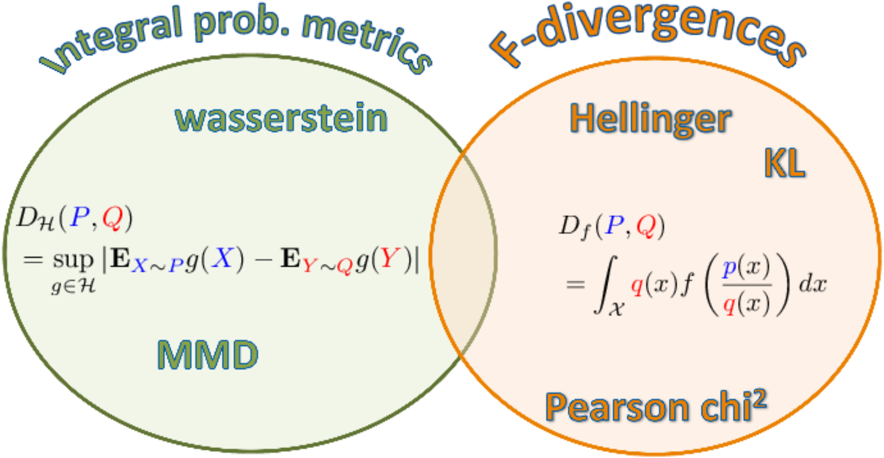
Divergences



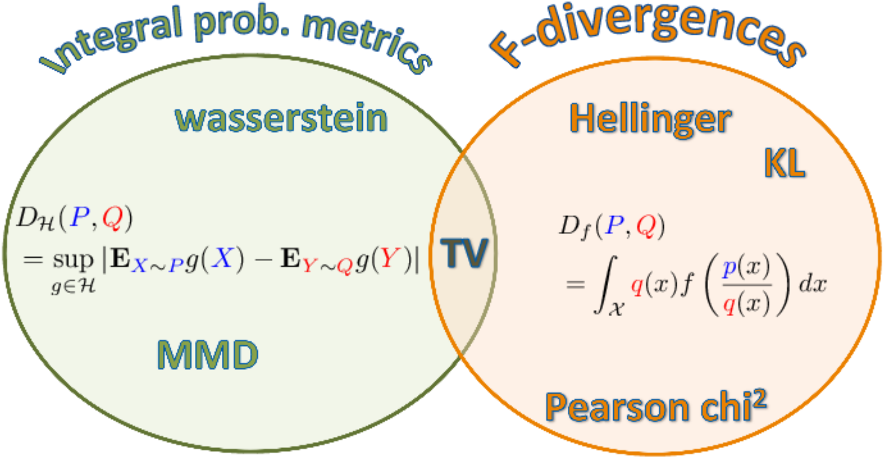
Divergences



Divergences



Divergences



Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (EJS, 2012, Theorem A.1)

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:

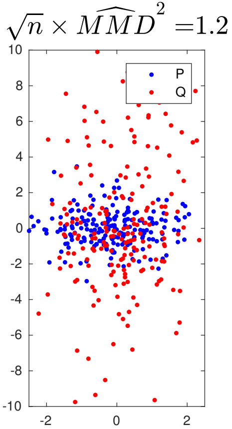
- **Null hypothesis** $\mathcal{H}_0$ when $P = Q$
  - should see $\widehat{MMD}^2$ "close to zero".
- **Alternative hypothesis** $\mathcal{H}_1$ when $P \neq Q$
  - should see $\widehat{MMD}^2$ "far from zero"

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

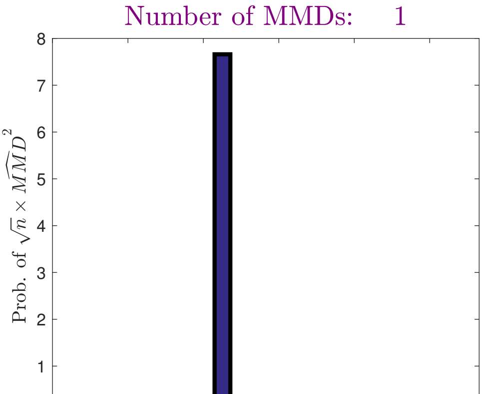
- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.2$



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

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

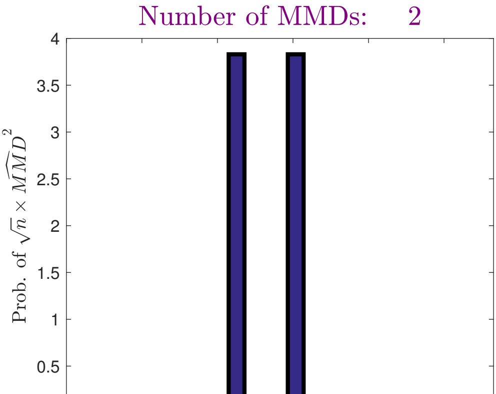
- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.2$



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

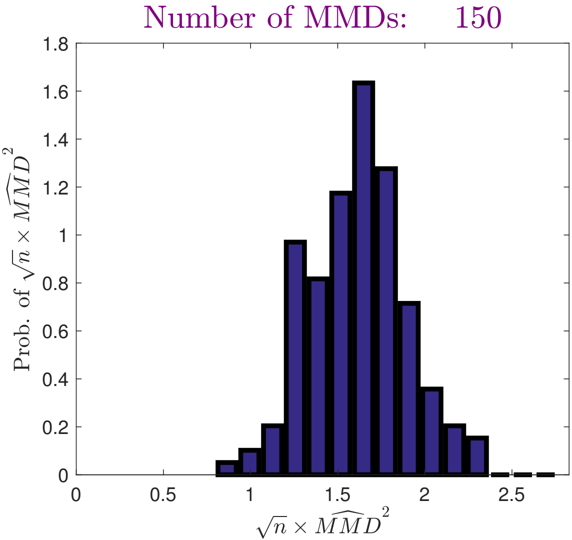
Draw $n = 200$ **new** samples from P and Q

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



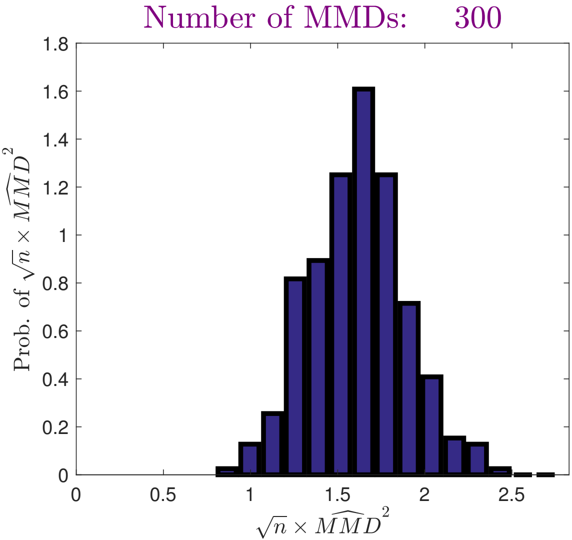
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 150 times ...



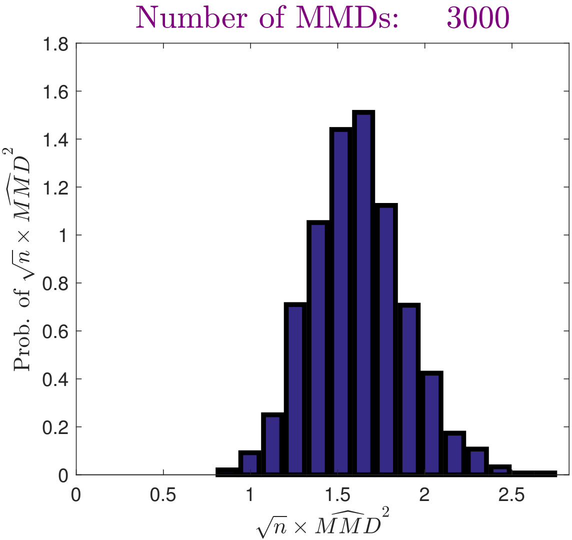
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 300 times ...



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

Repeat this 3000 times ...

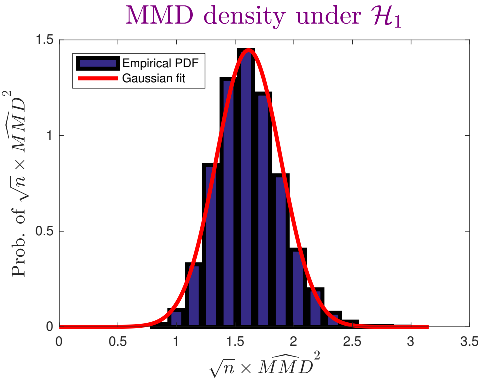


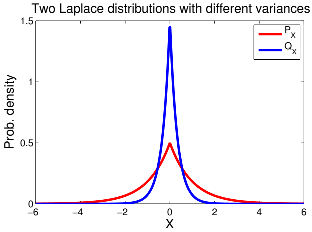
Asymptotics of $\widehat{MMD}^2$ when $P \neq Q$

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

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

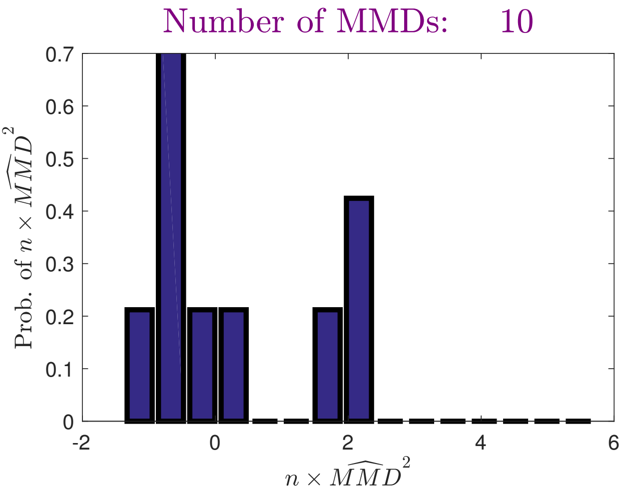
where variance $V_n(P, Q) = O\left(n^{-1}\right)$.





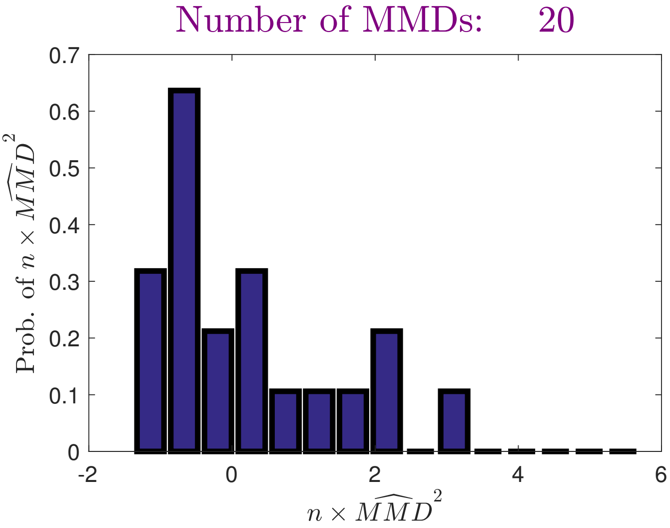
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



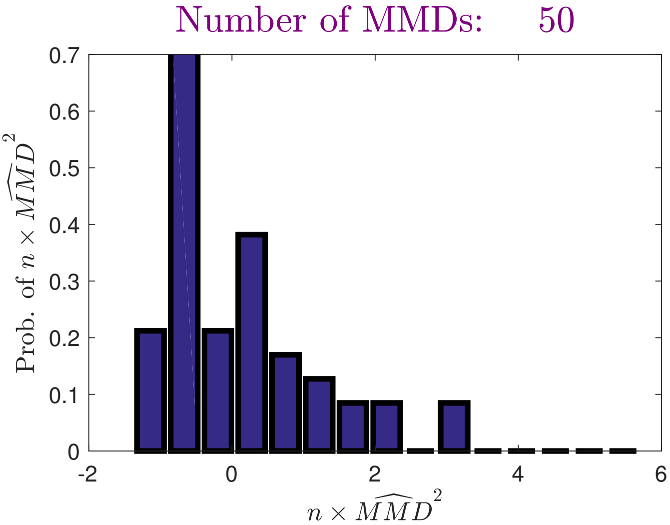
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



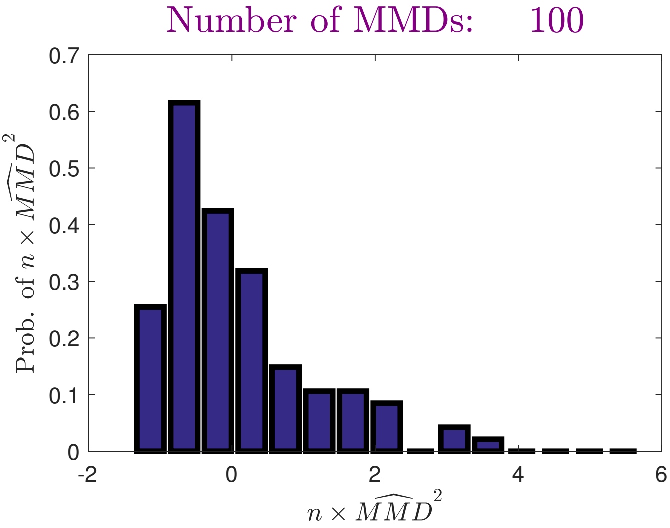
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



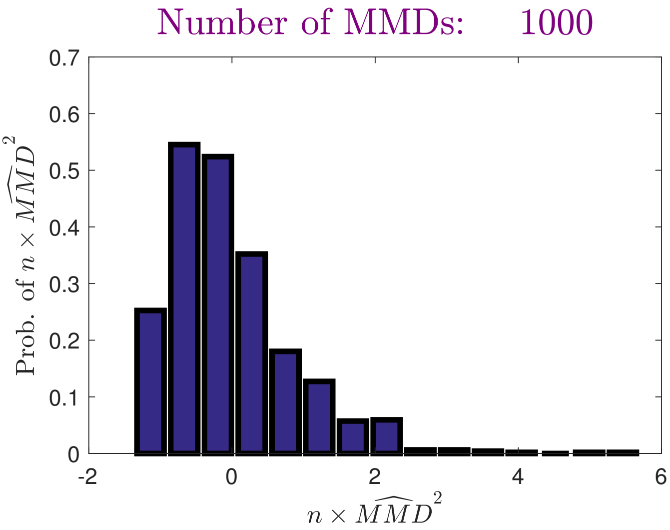
Behaviour of $\widehat{MMD}^2$ when $P = Q$

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



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

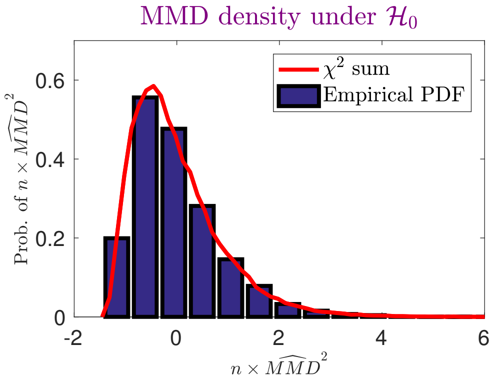
- Case of $P = Q = \mathcal{N}(0,1)$



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

Where $P = Q$, statistic has asymptotic distribution

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



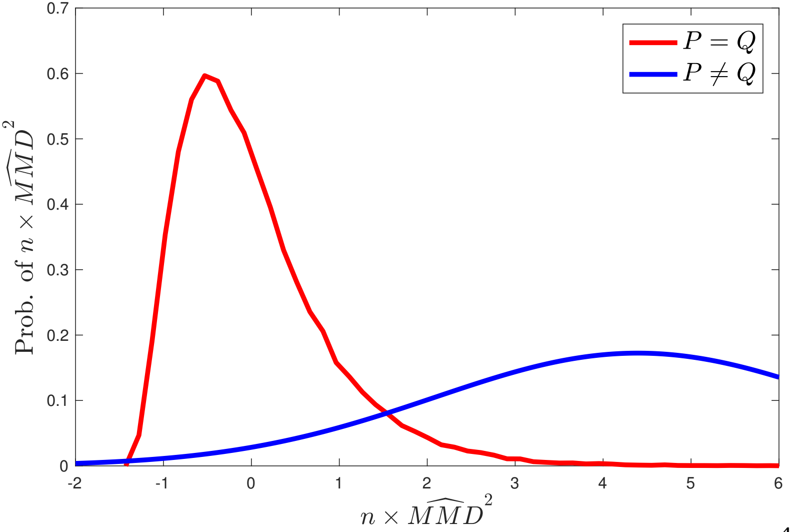
where

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

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

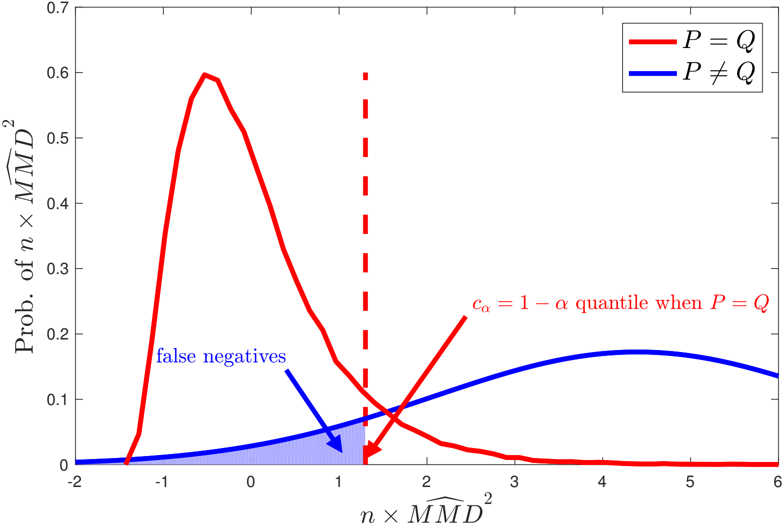
A statistical test

A summary of the asymptotics:



A statistical test

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



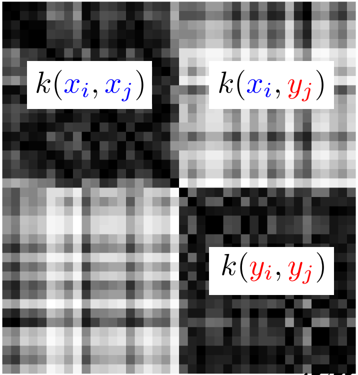
How do we get test threshold c_α?

Original empirical MMD for dogs and fish:

$X = [$  $\ldots]$

$Y = [$  $\ldots]$

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



How do we get test threshold c_α?

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

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

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

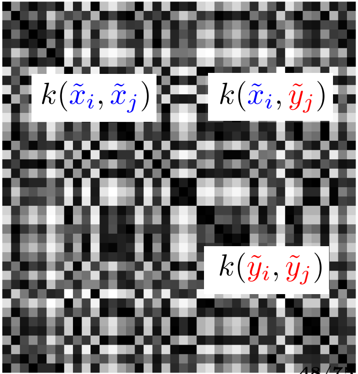
How do we get test threshold c_α?

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

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

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

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



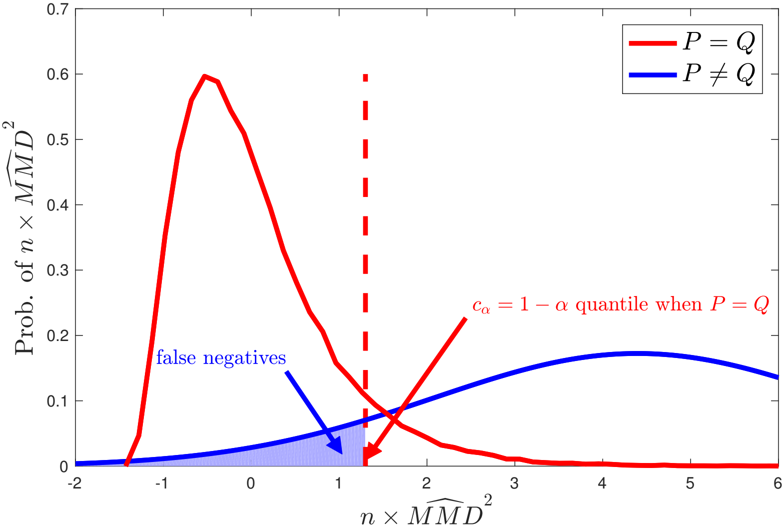
Permutation simulates

$P = Q$

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 ($\Pr_1$ denotes probability under $P \neq Q$):

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

Optimizing kernel for test power

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

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

where

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

Optimizing kernel for test power

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

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

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

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

For large n, second term negligible!

Optimizing kernel for test power

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

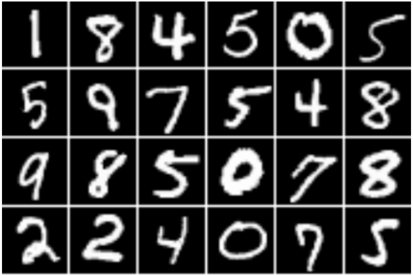
$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$
$$\to \Phi\left(\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P,Q)}}\right)$$

To maximize test power, maximize

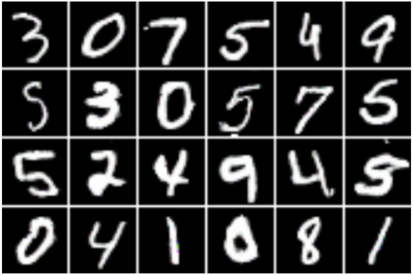
$$\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}$$

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

Troubleshooting for generative adversarial networks



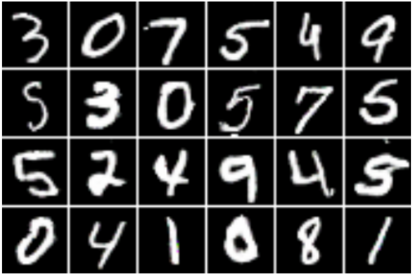
MNIST samples



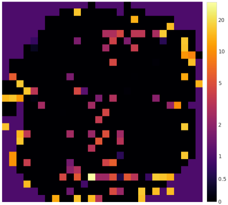
Samples from a GAN

Troubleshooting for generative adversarial networks

MNIST samples



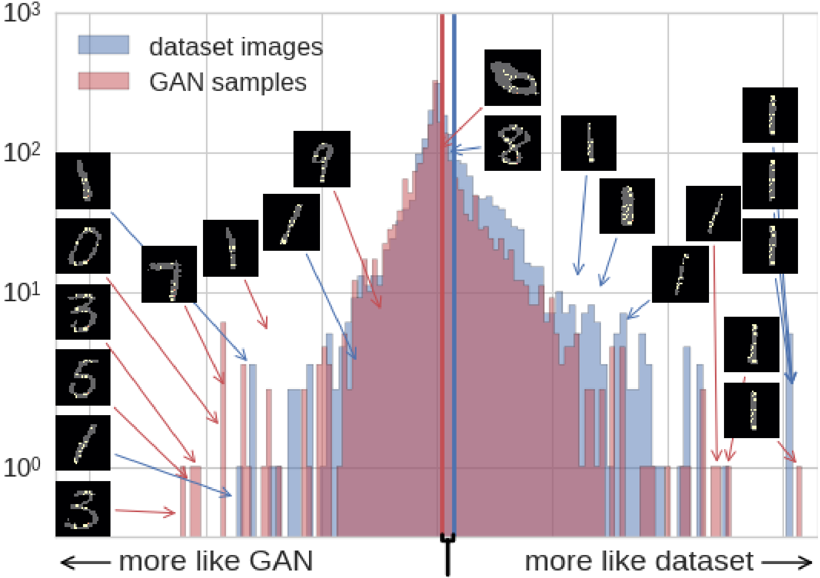
Samples from a GAN



ARD map

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

Troubleshooting generative adversarial networks



$\mathrm{MMD}^2 = 0.0001$

Training GANs with MMD

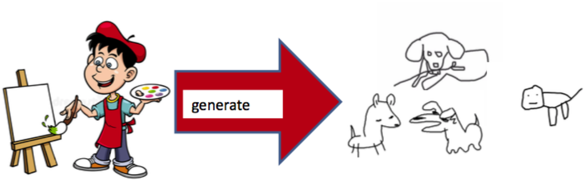
What is a Generative Adversarial Network (GAN)?

- Generator (student)

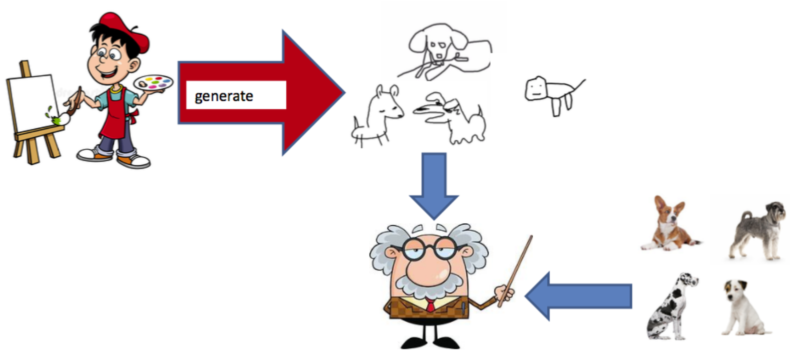
- Task: critic must teach generator to draw images (here dogs)

- Critic (teacher)

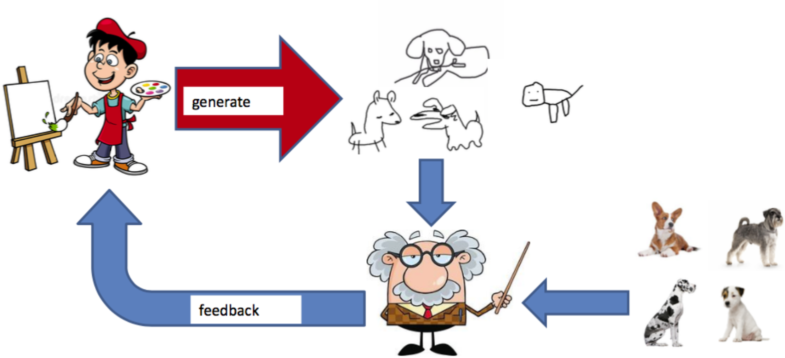
What is a Generative Adversarial Network (GAN)?



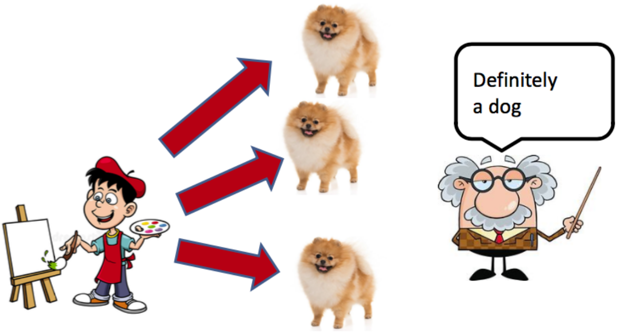
What is a Generative Adversarial Network (GAN)?



What is a Generative Adversarial Network (GAN)?



Why is classification not enough?



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
University of Cambridge

Daniel M. Roy
University of Toronto

Zoubin Ghahramani
University of Cambridge

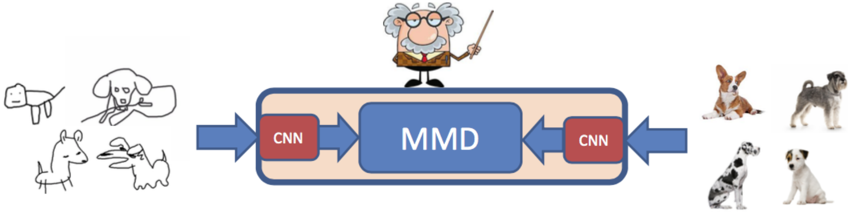
MMD for GAN critic

Can you use MMD as a critic to train GANs?

Need better image features.

How to improve the critic witness

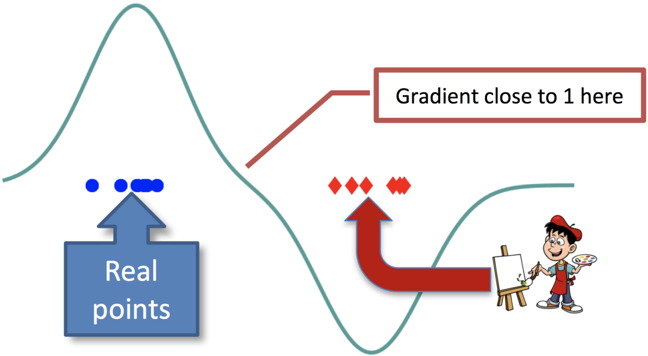
- Add convolutional features!
- The critic (teacher) also needs to be trained.
- How to regularise?



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. [NeurIPS 2017]



WGAN-GP

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

-  Given a generator G_θ with parameters θ to be trained. Samples $Y \sim G_\theta(Z)$ where $Z \sim R$
-  Given critic features h_ψ with parameters ψ to be trained. f_ψ a linear function of h_ψ.

WGAN-GP

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

-  Given a generator G_θ with parameters θ to be trained. Samples $Y \sim G_\theta(Z)$ where $Z \sim R$
-  Given critic features h_ψ with parameters ψ to be trained. f_ψ a linear function of h_ψ.

WGAN-GP gradient penalty:

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

where

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

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

$$\widetilde{X} = \gamma x_i + (1-\gamma) G_\theta(z_j)$$

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

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

From ICLR 2018:

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

The Dirac-GAN

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

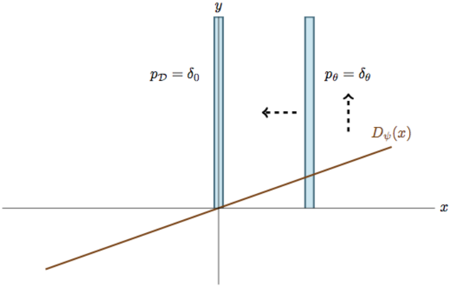


Figure from Mescheder et al. [ICML 2018]

Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

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

The Dirac-GAN

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

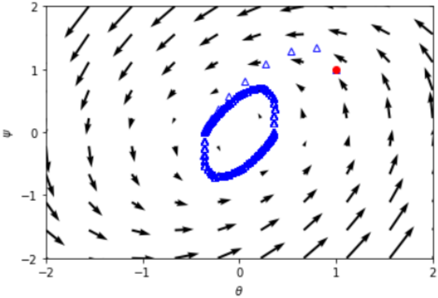


Figure from Mescheder et al. [ICML 2018]

A better gradient penalty

- New MMD GAN witness regulariser (NeurIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NeurIPS 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

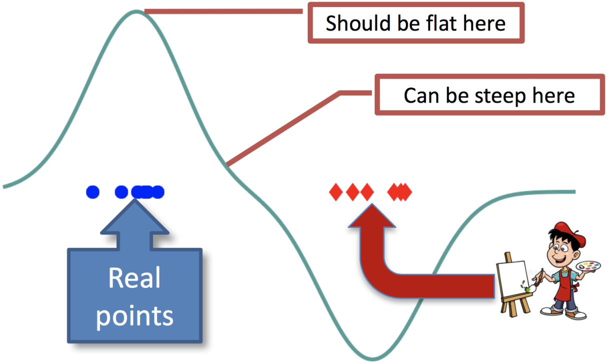
Michael Arbel, Dougal J. Sutherland, Mikołaj Bińkowski, Arthur Gretton

(Submitted on 29 May 2018)

A better gradient penalty

- New MMD GAN witness regulariser (NeurIPS 2018)

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



A better gradient penalty

- New MMD GAN witness regulariser (NeurIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NeurIPS 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

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

L_2 norm control

Gradient control

RKHS smoothness

A better gradient penalty

- New MMD GAN witness regulariser (NeurIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NeurIPS 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

$$\|f\|_S^2 = \underbrace{\|f\|^2_{L_2(P)}}_{\text{L}_2\text{ norm control}} + \underbrace{\|\nabla f\|^2_{L_2(P)}}_{\text{Gradient control}} + \underbrace{\lambda \|f\|_k^2}_{\text{RKHS smoothness}}$$

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

A better gradient penalty

- New MMD GAN witness regulariser (NeurIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NeurIPS 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 (NeurIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NeurIPS 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? Salimans et al. [NeurIPS 2016]

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

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

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? Salimans et al. [NeurIPS 2016]

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

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

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? Heusel et al. [NeurIPS 2017]

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

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

where μ_P and Σ_P are the feature mean and covariance of P

Evaluation of GANs

The Frechet inception distance? Heusel et al. [NeurIPS 2017]

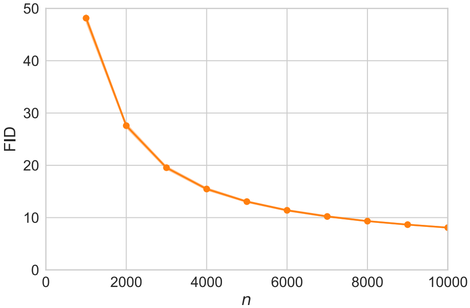
Fits Gaussians to features in the inception architecture (pool3 layer):

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

where μ_P and Σ_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 practice.

Let $d = 2048$, and define

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

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

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

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

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

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

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

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)

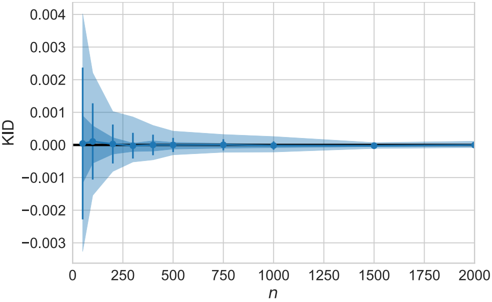
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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)

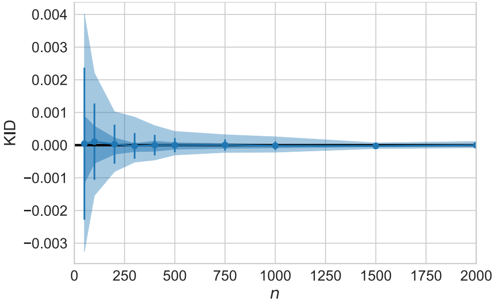
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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)

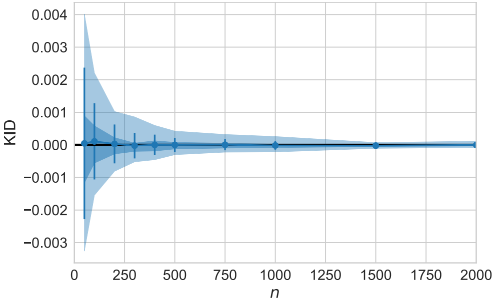
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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?"

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

The kernel inception distance (KID)

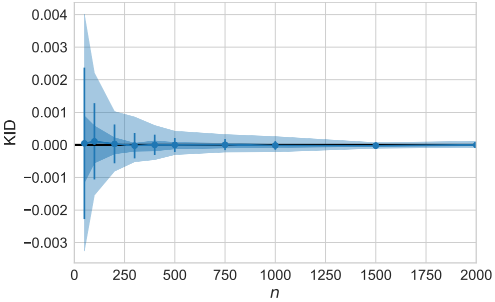
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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]

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 masanori@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: what does MMD buy you?

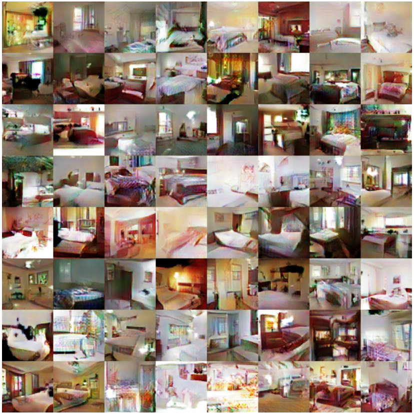
- Critic features from DCGAN: an f-filter critic has f, $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64.

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

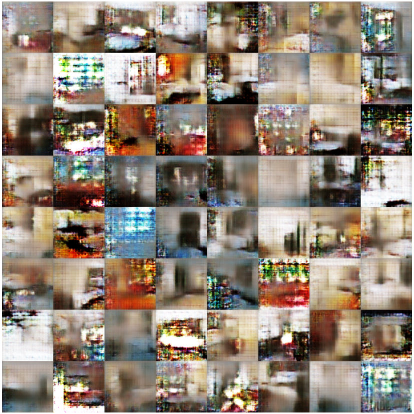
WGAN samples, $f = 64$,
KID=4

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 64×64.



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



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

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

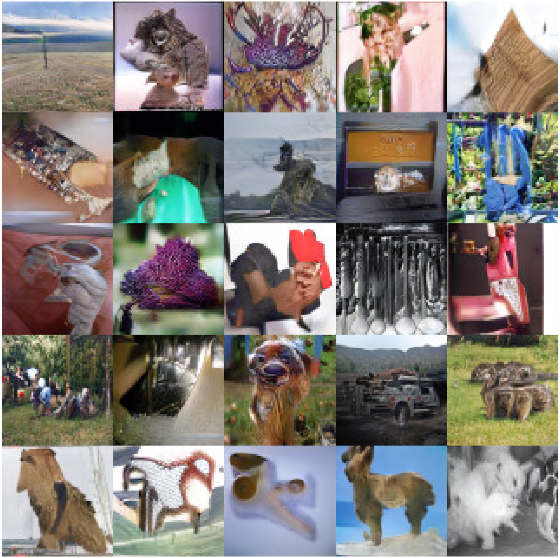
ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. Around 20 000 classes.

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. Around 20 000 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. 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_ψ features possible.
  - Better gradient/feature regulariser gives better critic

"Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy," ICLR 2017 https://github.com/dougalsutherland/opt-mmd

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

"On gradient regularizers for MMD GANs", 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
- Alex Smola
- Hsiao-Yu Tung

Questions?