The Maximum Mean Discrepancy for Training Generative Adversarial Networks

Arthur Gretton

Gatsby Computational Neuroscience Unit, University College London
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?**

Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.
Training generative models

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

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

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

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

Using MMD to train a GAN

(Binkowski, Sutherland, Arbel, G., ICLR 2018),
(Arbel, Sutherland, Binkowski, G., arXiv 2018)
Not covered: 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.
Not covered: testing independence

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

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Dog Image]</td>
<td>A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.</td>
</tr>
<tr>
<td>![Beagle Image]</td>
<td>Their noses guide them through life, and they're never happier than when following an interesting scent.</td>
</tr>
<tr>
<td>![Cat Image]</td>
<td>A responsive, interactive pet, one that will blow in your ear and follow you everywhere.</td>
</tr>
</tbody>
</table>

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**

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

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 $P$,

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

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

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

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 $P$,

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

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

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

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 is the distance between feature means:

\[
MMD^2(P, Q) = \|\mu_P - \mu_Q\|^2_F = \langle \mu_P, \mu_P \rangle_F + \langle \mu_Q, \mu_Q \rangle_F - 2 \langle \mu_P, \mu_Q \rangle_F \\
= \mathbb{E}_P k(X, X') + \mathbb{E}_Q k(Y, Y') - 2\mathbb{E}_{P,Q} k(X, Y)
\]

(a) (a) (b)
The maximum mean discrepancy is the distance between feature means:

\[
MMD^2(P, Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2\langle \mu_P, \mu_Q \rangle_{\mathcal{F}} \\
= \underbrace{\mathbb{E}_P k(X, X')}_{(a)} + \underbrace{\mathbb{E}_Q k(Y, Y')}_{(a)} - 2\underbrace{\mathbb{E}_{P,Q} k(X, Y)}_{(b)}
\]
The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

\[
MMD^2(P, Q) = \|\mu_P - \mu_Q\|^2_F
\]

\[
= \langle \mu_P, \mu_P \rangle_F + \langle \mu_Q, \mu_Q \rangle_F - 2 \langle \mu_P, \mu_Q \rangle_F
\]

\[
= \mathbb{E}_P k(X, X') + \mathbb{E}_Q k(Y, Y') - 2 \mathbb{E}_{P,Q} k(X, Y)
\]

(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 $k(\text{fish}_i, \text{fish}_j)$
The maximum mean discrepancy:

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

\[- \frac{2}{n^2} \sum_{i,j} k(\text{dog}_i, \text{fish}_j) \]
MMD as an integral probability metric

Are $P$ and $Q$ different?
MMD as an integral probability metric

Are $P$ and $Q$ different?

![Graph showing samples from P and Q](image-url)
MMD as an integral probability metric

Integral probability metric:
Find a "well behaved function" $f(x)$ to maximize

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

Integral probability metric:
Find a "well behaved function" $f(x)$ to maximize

$$E_P f(X) - E_Q f(Y)$$
MMD as an integral probability metric

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

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[ E_P f(X) - E_Q f(Y) \right]$$
\((F = \text{unit ball in RKHS } \mathcal{F})\)
MMD as an integral probability metric

Maximum mean discrepancy: smooth function for \( P \) vs \( Q \)

\[
MMD(P, Q; F) := \sup_{\|f\| \leq 1} [E_P f(X) - E_Q f(Y)]
\]

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

Functions are linear combinations of features:

\[
f(x) = \langle f, \varphi(x) \rangle_{\mathcal{F}} = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x) = \begin{bmatrix} f_1 \\
 f_2 \\
 f_3 \\
 \vdots 
\end{bmatrix}^T \begin{bmatrix} \varphi_1(x) \\
 \varphi_2(x) \\
 \varphi_3(x) \\
 \vdots 
\end{bmatrix}
\]
MMD as an integral probability metric

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

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

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

[Graph showing witness function for Gauss and Laplace densities]
MMD as an integral probability metric

Maximum mean discrepancy: smooth function for \( P \) vs \( Q \)

\[
MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[ E_P f(X) - E_Q f(Y) \right]
\]

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

**Expectations of functions are linear combinations of expected features**

\[
E_P(f(X)) = \langle f, E_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}
\]

(always true if kernel is bounded)
MMD as an integral probability metric

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

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

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

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

Other choices for witness function class:

- Bounded continuous [Dudley, 2002]
- Bounded variation 1 (Kolmogorov metric) [Müller, 1997]
- Bounded Lipschitz (Wasserstein distances) [Dudley, 2002]
The MMD:

$$MMD(P, Q; F) = \sup_{f \in F} \left[ \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \right]$$
Integral prob. metric vs feature difference

The MMD:

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

use

\[
\mathbb{E}_P f(X) = \langle \mu_P, f \rangle_F
\]
The MMD:

$$MMD(P, Q; F') = \sup_{f \in F} \left[ E_P f(X) - E_Q f(Y) \right]$$

$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_F$$
The MMD:

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

$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_F$$
Integral prob. metric vs feature difference

The MMD:

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

$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_F$$

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

Unit ball
Integral prob. metric vs feature difference

The MMD:

\[
MMD(P, Q; F) = \sup_{f \in F} \left[ \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \right] = \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_F = \|\mu_P - \mu_Q\|
\]
Construction of MMD witness

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

Observe $X = \{x_1, \ldots, x_n\} \sim P$

Observe $Y = \{y_1, \ldots, y_n\} \sim Q$
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 \( P \)

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

The empirical witness function at \( \nu \)

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

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

\[ \propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \]
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 \( v \)

\[
\begin{align*}
    f^*(v) &= \langle f^*, \varphi(v) \rangle_F \\
    &\propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_F \\
    &= \frac{1}{n} \sum_{i=1}^{n} k(x_i, v) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, v)
\end{align*}
\]

Don’t need explicit feature coefficients \( f^* := \begin{bmatrix} f_1^* & f_2^* & \ldots \end{bmatrix} \)
Interlude: divergence measures
Divergences

\[ P \quad Q \quad \frac{P}{Q} \]
Divergences

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

\[ D_f(P, Q) = \int_X q(x) f \left( \frac{p(x)}{q(x)} \right) dx \]
Divergences

Integral prob. metrics

\( D_H(P, Q) = \sup_{g \in \mathcal{H}} |E_{X \sim P} g(X) - E_{Y \sim Q} g(Y)| \)

F-divergences

\( D_f(P, Q) = \int_X q(x) f \left( \frac{p(x)}{q(x)} \right) dx \)
Divergences

**Integral prob. metrics**

- Wasserstein

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

- MMD

**F-divergences**

- Hellinger

\[
D_f(P, Q) = \int_X q(x) f \left( \frac{p(x)}{q(x)} \right) dx
\]

- Pearson chi\(^2\)

- KL
Divergences

\[
D_\mathcal{H}(P, Q) = \sup_{g \in \mathcal{H}} |E_{X \sim P} g(X) - E_{Y \sim Q} g(Y)|
\]

\[
D_f(P, Q) = \int \chi q(x) f \left( \frac{p(x)}{q(x)} \right) dx
\]

Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (2012)
Two-Sample Testing with MMD
A statistical test using MMD

The empirical MMD:

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

How does this help decide whether \( P = Q \)?
A statistical test using MMD

The empirical MMD:

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

Perspective from statistical hypothesis testing:

- **Null hypothesis** $\mathcal{H}_0$ when $P = \mathcal{Q}$
  - should see $\hat{\text{MMD}}^2$ “close to zero”.
- **Alternative hypothesis** $\mathcal{H}_1$ when $P \neq \mathcal{Q}$
  - should see $\hat{\text{MMD}}^2$ “far from zero”
A statistical test using MMD

The empirical MMD:

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

Perspective from statistical hypothesis testing:

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

Want Threshold $c_\alpha$ for $\hat{\text{MMD}}^2$ to get false positive rate $\alpha$
Behaviour of $\widehat{\text{MMD}}^2$ when $P \neq Q$

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

- Laplace with different $y$-variance.
- $\sqrt{n} \times \widehat{\text{MMD}}^2 = 1.2$
Behaviour of $\overline{\text{MMD}}^2$ when $P \neq Q$

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

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

Number of MMDs: 1

\[
\sqrt{n} \times \overline{\text{MMD}}^2
\]
Behaviour of $\hat{MMD}^2$ when $P \neq Q$

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

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

Number of MMDs: 2
Behaviour of $MMD^2$ 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 $\sqrt{n} \times MMD^2$ when $P \neq Q$

Repeat this 3000 times …

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

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

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

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

MMD density under $\mathcal{H}_1$
Behaviour of $\widehat{\text{MMD}}^2$ when $P = Q$

What happens when $P$ and $Q$ are the same?
Behaviour of $\overline{\text{MMD}}^2$ when $P = Q$

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

Number of MMDs: 10
Behaviour of $\tilde{MMD}^2$ when $P = Q$

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

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

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

Number of MMDs: 50

![Histogram showing the distribution of $n \times \widehat{\text{MMD}}^2$ with 50 counts. The x-axis represents $n \times \widehat{\text{MMD}}^2$, and the y-axis represents the probability of each bin. The histogram peaks around 0 with decreasing frequencies as we move away from 0.]
Behaviour of $\text{MMD}^2$ when $P = Q$

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

Number of MMDs: 100
Behaviour of $\hat{MMD}^2$ when $P = Q$

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

Number of MMDs: 1000

![Histogram](chart.png)
Asymptotics of $MMD^2$ when $P = Q$

Where $P = Q$, statistic has asymptotic distribution

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

where

$$\lambda_i \psi_i(x') = \int_{\chi} \tilde{k}(x, x') \psi_i(x) dP(x)$$

is centred.

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

A summary of the asymptotics:

-2 -1 0 1 2 3 4 5 6

0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

\[ \frac{42}{73} \]
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:

$$X = [\text{dogs} \ldots]$$

$$Y = [\text{fish} \ldots]$$

$$\hat{\text{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j)$$

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

$$- \frac{2}{n^2} \sum_{i, j} k(x_i, y_j)$$
How do we get test threshold $c_\alpha$?

Permutated dog and fish samples (merdogs):

$$\tilde{X} = \begin{bmatrix}
\text{fish} & \text{dog} & \ldots
\end{bmatrix}$$

$$\tilde{Y} = \begin{bmatrix}
\text{dog} & \text{fish} & \ldots
\end{bmatrix}$$
How do we get test threshold $c_\alpha$?

Permuted dog and fish samples (merdogs):

\[ \tilde{X} = [\text{dog, fish, ...}] \]
\[ \tilde{Y} = [\text{dog, fish, ...}] \]

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

Permutation simulates
\[ P = Q \]
How to choose the best kernel: optimising the kernel parameters
Maximising test power same as minimizing false negatives
Optimizing kernel for test power

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

$$\Pr_1 \left( \hat{nMMD^2} > \hat{c}_\alpha \right)$$
Optimizing kernel for test power

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

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

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

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

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

$$\to \Phi \left( \frac{\overline{MMD}^2(P, Q)}{\sqrt{\overline{V}_n(P, Q)}} - \frac{c_\alpha}{n \sqrt{\overline{V}_n(P, Q)}} \right)$$

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

For large $n$, second term negligible!
Optimizing kernel for test power

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

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

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

To maximize test power, maximize

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

(Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017)

Code: github.com/dougalsutherland/opt-mmd
Troubleshooting for generative adversarial networks

| 1 8 4 5 0 5 |
| 5 9 7 5 4 8 |
| 9 8 5 0 7 8 |
| 2 2 4 0 7 5 |

**MNIST samples**

| 3 0 7 5 4 9 |
| 5 3 0 5 7 5 |
| 5 2 4 9 4 5 |
| 0 4 1 0 8 1 |

**Samples from a GAN**
Troubleshooting for generative adversarial networks

MNIST samples

Samples from a GAN

- Power for optimized ARD kernel: 1.00 at $\alpha = 0.01$
- Power for optimized RBF kernel: 0.57 at $\alpha = 0.01$
Troubleshooting generative adversarial networks

The diagram illustrates the comparison between dataset images and GAN samples. The x-axis represents the MMD$^2$ value, with MMD$^2 = 0.0001$ indicating a low discrepancy between the GAN-generated samples and the dataset images. The y-axis shows the scale from $10^0$ to $10^3$. The graph indicates that the samples are more likely to be classified as GAN-generated towards the left, and more likely to be classified as dataset images towards the right.
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)

Definitely a dog
Can you use MMD as a critic to train GANs?

From ICML 2015:

**Generative Moment Matching Networks**

Yujia Li
Kevin Swersky
Richard Zemel

1Department of Computer Science, University of Toronto, Toronto, ON, CANADA
2Canadian 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]
Wasserstein GAN  Arjovsky et al. [ICML 2017]
WGAN-GP  Gukrajani et al. [NIPS 2017]
Wasserstein GAN  Arjovsky et al. [ICML 2017]
WGAN-GP  Gulrajani 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:

\[
\max_{\psi} \mathbb{E}_{X \sim P} f_\psi(X) - \mathbb{E}_{Z \sim R} f_\psi(G_\theta(Z)) + \lambda \mathbb{E}_{\tilde{X}} \left( \left\| \nabla_{\tilde{X}} f_\psi(\tilde{X}) \right\| - 1 \right)^2
\]

where

\[
\tilde{X} = \gamma x_i + (1 - \gamma) G_\theta(z_j)
\]

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

\[
\max_{\psi} \text{MMD}^2(h_{\psi}(X), h_{\psi}(G_{\theta}(Z))) + \lambda E_{\tilde{X}} \left(\left\| \nabla_{\tilde{X}} f_{\psi}(\tilde{X}) \right\| - 1 \right)^2
\]

where

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

\[
\tilde{X} = \gamma x_i + (1 - \gamma) G_{\theta}(z_j)
\]

\[
\gamma \sim \mathcal{U}([0, 1]) \quad x_i \in \{x_l\}_{l=1}^{m} \quad z_j \in \{z_l\}_{l=1}^{n}
\]

Remark by Bottou et al. (2017): gradient penalty modifies the function class. So critic 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 [NIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

The Dirac-GAN

\[ P = \delta_0 \quad Q = \delta_{\theta} \quad f_{\psi}(x) = \psi \cdot x \]
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

\[ P = \delta_0 \quad Q = \delta_\theta \quad f_\psi(x) = \psi \cdot x \]

Figure from Mescheder et al. [ICML 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]
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]
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:

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

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
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:

$$\widehat{MMD} := \sup_{\|f\|_{S} \leq 1} \left[ \mathbb{E}_{P} f(X) - \mathbb{E}_{Q} f(Y) \right]$$

where

$$\|f\|_{S}^{2} = \|f\|_{L_{2}(P)}^{2} + \|\nabla f\|_{L_{2}(P)}^{2} + \lambda \|f\|_{k}^{2}$$

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
The inception score? Salimans et al. [NIPS 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. [NIPS 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. [NIPS 2017]

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? Heusel et al. [NIPS 2017]

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:

\[ P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1). \]

Clearly,

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

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

\[ FID(\hat{P}_1, Q) < FID(\hat{P}_2, Q). \]
The FID can give the wrong answer in theory.
Assume \( m \) samples from \( P \) and \( n \to \infty \) samples from \( Q \).
Given two alternatives:

\[
P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).
\]

Clearly,

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

Given \( m \) samples from \( P_1 \) and \( P_2 \),

\[
FID(\hat{P}_1, Q) < FID(\hat{P}_2, Q).
\]
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:

\[
P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).
\]

Clearly,

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

Given \( m \) samples from \( P_1 \) and \( P_2 \),

\[
FID(\hat{P}_1, Q) < FID(\hat{P}_2, Q).
\]
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:

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

Clearly,

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

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

$$FID(\hat{P}_1, Q) < FID(\hat{P}_2, Q).$$
The FID can give the wrong answer in practice.

Let $d = 2048$, and define

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

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 $m = 50000$ samples,

$$FID(\hat{P}_1, Q) \approx 1133.7 < 1136.2 \approx FID(\hat{P}_2, Q)$$

At $m = 100000$ samples, the ordering of the estimates is correct. This behavior is similar for other random draws of $C$. 

---

64/73
Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

\[
P_1 = \text{relu}(\mathcal{N}(0, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(1, 0.8\Sigma + 0.2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d))
\]

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

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

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 $m = 50000$ samples,

$$FID(\hat{P}_1, Q) \approx 1133.7 < 1136.2 \approx FID(\hat{P}_2, Q)$$

At $m = 100000$ 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

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

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 $m = 50000$ samples,

$$FID(\widehat{P}_1, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P}_2, Q)$$

At $m = 100000$ 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)!

65/73
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 \( KID(\hat{P}_{t+1}, Q) \) not significantly better than \( KID(\hat{P}_t, Q) \) 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 Miyato¹, Toshiki Kataoka¹, Masanori Koyama², Yuichi Yoshida³
¹miyato, kataoka}@preferred.jp
²koyama, masanori1@gmail.com
³yoshida@nii.ac.jp

SOBOLEV GAN

Youssef Mroueh¹, Chun-Liang Li², Tom Sercu¹, Anant Raj³ & Yu Cheng¹
¹IBM Research AI
²Carnegie Mellon University
³Max Planck Institute for Intelligent Systems
* denotes Equal Contribution
{mroueh, chengyu}@us.ibm.com, chunlial@cs.cmu.edu, tom.sercul@ibm.com, anant.raj@tuebingen.mpg.de

DEMystifying MMD GANS

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

BOUNdary-SEEking GENERATIVE ADVERSARIAL NETWORKS

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

Boundary-Seeking Generative Adversarial Networks

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

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

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

Boundary-Seeking Generative Adversarial Networks

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, $f = 64$, \text{KID}=3

WGAN samples, $f = 64$, \text{KID}=4

67/73
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 67/73
The kernel inception distance (KID)

Faster training: performance scores vs generator iterations on MNIST
Results: celebrity faces 160x160

KID scores:

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

202,599 face images, resized and cropped to 160 x 160
Results: imagenet 64×64

KID (FID) scores:

- **BGAN:**
  47

- **SN-GAN:**
  44

- **SMMD GAN:**
  35

ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. Around 20 000 classes.
KID (FID) 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: imagenet $64 \times 64$

KID (FID) scores:

- **BGAN:**
  - 47

- **SN-GAN:**
  - 44

- **SMMD GAN:**
  - 35

ILSVRC2012 (ImageNet) dataset, 1,281,167 images, resized to $64 \times 64$. 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

“Demystifying MMD GANs,” including KID score, ICLR 2018: https://github.com/mbinkowski/MMD-GAN
Gradient regularised MMD, NIPS 2018: 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