Representing and comparing probabilities

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

UAI, 2017
Comparing two samples

- **Given:** Samples from unknown distributions $P$ and $Q$.
- **Goal:** do $P$ and $Q$ differ?
An example: two-sample tests

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

MNIST samples

Samples from a GAN

Significant difference in GAN and MNIST?

Testing goodness of fit

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

Chicago crime data
Model is Gaussian mixture with two components.
Testing independence

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

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://via.placeholder.com/150" alt="Dog" /></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><img src="https://via.placeholder.com/150" alt="Dog" /></td>
<td>Their noses guide them through life, and they're never happier than when following an interesting scent.</td>
</tr>
<tr>
<td><img src="https://via.placeholder.com/150" alt="Cat" /></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: part 1

Two sample testing

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

- Statistical testing with the MMD

- Troubleshooting GANs with MMD
Outline: part 2

Goodness of fit testing
- The kernel Stein discrepancy

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

Additional topics
Outline: part 2

Goodness of fit testing
- The kernel Stein discrepancy

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

Additional topics
Maximum Mean Discrepancy
Feature mean difference

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

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

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

- Gaussian and Laplace distributions
- Same mean \textit{and} same variance
- Difference in means using \textbf{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

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

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

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

\[ = E_P k(X, X') + E_Q k(Y, Y') - 2E_{P,Q} k(X, Y) \]

\[ \begin{array}{c}
(a) \\
(b)
\end{array} \]
The maximum mean discrepancy is the distance between feature means:

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

(a) \hspace{2cm} (a) \hspace{2cm} (b)
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 $k(dog_i, dog_j)$, $k(dog_i, fish_j)$, or $k(fish_i, fish_j)$
Illustration of MMD

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?

Samples from $P$ and $Q$
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 $Q$

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

($F =$ unit ball in RKHS $\mathcal{F}$)
MMD as an integral probability metric

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

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

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

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

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

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

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

Functions are linear combinations of features:

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

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

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

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

Expectations of functions are linear combinations of expected features

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

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

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

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

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

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]
Integral prob. metric vs feature difference

The MMD:

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

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

use

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

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

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

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

$$f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$$
Integral prob. metric vs feature difference

The MMD:

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

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

\[
= \|\mu_P - \mu_Q\|
\]

Function view and feature view equivalent
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 \( P \)

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

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

\[ f^*(\nu) = \langle f^*, \varphi(\nu) \rangle_{\mathcal{F}} \]
\[ \propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(\nu) \rangle_{\mathcal{F}} \]
\[ = \frac{1}{n} \sum_{i=1}^{n} k(x_i, \nu) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, \nu) \]

Don’t need explicit feature coefficients \( f^* := [ f_1^* \ f_2^* \ \ldots ] \)
Two-Sample Testing
A statistical test using MMD

The empirical MMD:

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

\[- \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)\]

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

The empirical MMD:

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

Perspective from statistical hypothesis testing:

- **Null hypothesis** $\mathcal{H}_0$ when $P = Q$
  - should see $\hat{MMD}^2$ "close to zero".

- **Alternative hypothesis** $\mathcal{H}_1$ when $P \neq Q$
  - should see $\hat{MMD}^2$ "far from zero"
A statistical test using MMD

The empirical MMD:

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

Perspective from statistical hypothesis testing:

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

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

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

- Laplace with different $y$-variance.

- $\sqrt{n} \times \hat{MMD}^2 = 1.2$
Behaviour of $\hat{MMD}^2$ when $P \neq Q$

Number of MMDs: 1

\[
\sqrt{n} \times \hat{MMD}^2 = 1.2
\]
Behaviour of $\hat{MMD}^2$ when $P \neq Q$

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

- Laplace with different y-variance.
- $\sqrt{n} \times \hat{MMD}^2 = 1.5$
Behaviour of $\hat{MMD}^2$ when $P \neq Q$

Number of MMDs: 2

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

Repeat this 150 times …

Number of MMDs: 150

$\sqrt{n} \times \text{MMD}^2$
Behaviour of $\hat{MMD}^2$ when $P \not= Q$

Repeat this 300 times …

Number of MMDs: 300
Repeat this 3000 times …

Number of MMDs: 3000
Asymptotics of $\sqrt{n} \times \text{MMD}^2$ when $P \neq Q$

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

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

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

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

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

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

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

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

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

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

**Number of MMDs: 50**

![Histogram](image.png)
Behaviour of $\overline{MMD}^2$ when $P = Q$

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

Number of MMDs: 100

![Histogram showing the distribution of $n \times \overline{MMD}^2$ for 100 samples with $P = Q = \mathcal{N}(0, 1)$]
Behaviour of $\hat{\text{MMD}}^2$ when $P = Q$

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

Number of MMDs: 1000

![Histogram of $n \times \hat{\text{MMD}}^2$](image)
Asymptotics of $\hat{MMD}^2$ when $P = Q$

Where $P = Q$, statistic has asymptotic distribution

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

where

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

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

A summary of the asymptotics:

![Graph showing the asymptotics of a statistical test](image-url)
A statistical test

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

\[
\text{Prob. of } n \times \hat{MMD}^2
\]

\text{false negatives}

\[c_\alpha = 1 - \alpha \text{ quantile when } P = Q\]
How do we get test threshold $c_\alpha$?

Original empirical MMD for dogs and fish:

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

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

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

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

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

Permuted dog and fish samples (merdogs):

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

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

Permutated dog and fish samples (merdogs):

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

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

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

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

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

Permutation simulates

$P = Q$
Demonstration of permutation estimate of null

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

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

Use $1 - \alpha$ quantile of permutation distribution for test threshold $c_\alpha$
How to choose the best kernel
Optimizing kernel for test power

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

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

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

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

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

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 $P \neq Q$):

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

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

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 $P \neq Q$):

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

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
Reminder: maximising test power same as minimizing false negatives

\[ c_{\alpha} = 1 - \alpha \text{ quantile when } P = Q \]
Troubleshooting for generative adversarial networks

MNIST samples

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

\[
MMD^2 = 0.0001
\]
MMD for GAN critic

Can you use MMD as a critic to train GANs?

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

From ICML 2015:

Generative Moment Matching Networks

Yujia Li
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
### Wasserstein GAN

**Wasserstein GAN**

Martin Arjovsky, Soumith Chintala, Léon Bottou

ICML 2017

### Improved Training of Wasserstein GANs

**Improved Training of Wasserstein GANs**

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

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

### Wasserstein Training of Boltzmann Machines

**Wasserstein Training of Boltzmann Machines**

Grégoire Montavon, Klaus-Robert Müller, Marco Cuturi

NIPS 2016
MMD for GAN critic: 2017 update

Wasserstein

MMD GAN: Towards Deeper Understanding of Moment Matching Network
Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, Barnabás Póczos
(Submitted on 24 May 2017)

The Cramer Distance as a Solution to Biased Wasserstein Gradients
Marc G. Bellemare, IvoDanihelka, WillDabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, Rémi Munos
(Submitted on 30 May 2017)
MMD for GAN critic: 2017 update

Wasserstein

(W)MMD

"Variance Incorporating"

"Other"
MMD for GAN critic: 2017 update

Wasserstein GAN
Martin Arjovsky, Soumith Chintala, Léon Bottou

MMD GAN: Towards Deeper Understanding of Moment Matching Network
Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, Barnabás Póczos
(Submitted on 24 May 2017)

The Cramer Distance as a Sol-Biased Wasserstein Grail
Marc G. Bellemare, Ivo Danihelka, Will Dabney, Balaji Lakshminarayanan, Stephan Hoyer, Rémi Munos
(Submitted on 30 May 2017)

McGAN: Mean and Covariance Feature Matching GAN
Youssef Mroueh, Tom Sercu, Vaibhava Goel
ICML 2017

Fisher GAN
Youssef Mroueh, Tom Sercu
(Submitted on 26 May 2017 (v1), last revised 1 Aug 2017 (this version, v2))

“Energy Distance Kernel”
LYDIA TINGCHU LIU
ADVISOR: PROFESSOR HAN LIU

“Variance Incorporating”

Other

Distributional Adversarial Networks
Chengtao Li, David Alvarez-Melis, Keyulu Xu, Stefanie Jegelka, Suvrit Sra
(Submitted on 29 Jun 2017 (v3), last revised 9 Jul 2017 (this version, v3))
An adaptive, linear time distribution metric
Reminder: witness function for MMD

\[ \text{witness}(\mathbf{v}) = \hat{\mu}_P(\mathbf{v}) - \hat{\mu}_Q(\mathbf{v}) \]

\( \hat{\mu}_P(\mathbf{v}) \): mean embedding of \( P \)

\( \hat{\mu}_Q(\mathbf{v}) \): mean embedding of \( Q \)
Distinguishing Feature(s)

Take square of witness (only worry about amplitude)

\[ \text{witness}^2(v) \]
Distinguishing Feature(s)

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

Best feature = $v^*$ that maximizes $\text{witness}^2(v)$?
Distinguishing Feature(s)

Sample size $n = 3$

$\text{witness}^2(v)$
Distinguishing Feature(s)

Sample size $n = 50$
Distinguishing Feature(s)

Sample size $n = 500$
Distinguishing Feature(s)

Population witness² function
Distinguishing Feature(s)

\[ P(x), \quad Q(y), \quad \text{witness}^2(v) \]
Variance of witness function

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

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016
**Variance of witness function**

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

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016
Variance of witness function

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

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016
Variance of witness function

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

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016
Variance of witness function

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

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016
Variance of witness function

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

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016
Variance of witness function

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

Best location is $v^*$ that maximizes $\hat{\lambda}_n$.

Improve performance using multiple locations $\{v^*_j\}_{j=1}^J$. 

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016
Properties of the ME Test

- Can use $J$ features $\mathcal{V} = \{v_1, \ldots, v_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $P_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting $H_0$ when $H_1$ is true.
Properties of the ME Test

- Can use $J$ features $\mathcal{V} = \{v_1, \ldots, v_J\}$.
- Under $H_0: P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.
- Under $H_1: P \neq Q$, it follows $P_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting $H_0$ when $H_1$ is true.

![Graph of $\chi^2(J)$]

- Runtime: $O(n)$ for both testing and optimization.
Properties of the ME Test

- Can use $J$ features $\mathcal{V} = \{v_1, \ldots, v_J\}$.
- Under $H_0: P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.
- Under $H_1: P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting $H_0$ when $H_1$ is true.

![Diagram showing test power and rejection threshold](image)

- Runtime: $\mathcal{O}(n)$ for both testing and optimization.
Properties of the ME Test

- Can use $J$ features $\mathcal{V} = \{v_1, \ldots, v_J\}$.
- Under $H_0: P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.
- Under $H_1: P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting $H_0$ when $H_1$ is true.

Runtime: $\mathcal{O}(n)$ for both testing and optimization.
Properties of the ME Test

- Can use $J$ features $\mathcal{V} = \{v_1, \ldots, v_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- **Test power** = probability of rejecting $H_0$ when $H_1$ is true.

![Diagram showing $\hat{\lambda}_n$, $\chi^2(J)$, $T_\alpha$, and $\mathbb{P}_{H_1}(\hat{\lambda}_n)$]
Can use $J$ features $\mathcal{V} = \{v_1, \ldots, v_J\}$.

Under $H_0: P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.

- Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.

Under $H_1: P \neq Q$, it follows $P_{H_1}(\hat{\lambda}_n)$ (unknown).

- But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.

**Test power** = probability of rejecting $H_0$ when $H_1$ is true.

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

Runtime: $\mathcal{O}(n)$ for both testing and optimization.
Properties of the ME Test

- Can use $J$ features $\mathcal{V} = \{v_1, \ldots, v_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \rightarrow \infty$. Consistent test.

**Test power** = probability of rejecting $H_0$ when $H_1$ is true.

![Diagram](image)

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

Runtime: $\mathcal{O}(n)$ for both testing and optimization.
Properties of the ME Test

- Can use $J$ features $\mathcal{V} = \{v_1, \ldots, v_J\}$.
- Under $H_0 : P = Q$, asymptotically $\hat{\lambda}_n(\mathcal{V})$ follows $\chi^2(J)$ for any $\mathcal{V}$.
  - Rejection threshold is $T_\alpha = (1 - \alpha)$-quantile of $\chi^2(J)$.
- Under $H_1 : P \neq Q$, it follows $\mathbb{P}_{H_1}(\hat{\lambda}_n)$ (unknown).
  - But, asymptotically $\hat{\lambda}_n \to \infty$. Consistent test.
- Test power = probability of rejecting $H_0$ when $H_1$ is true.

![Diagram](image)

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

- Runtime: $\mathcal{O}(n)$ for both testing and optimization.
Distinguishing Positive/Negative Emotions

+ : happy neutral surprised

− : afraid angry disgusted

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

The proposed test achieves maximum test power in time $O(n)$.
Informative features: differences at the nose, and smile lines.
Distinguishing Positive/Negative Emotions

+ : happy neutral surprised

- : afraid angry disgusted

The proposed test achieves maximum test power in time $O(n)$.

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

- Informative features: differences at the nose, and smile lines.
Distinguishing Positive/Negative Emotions

+ : happy neutral surprised
- : afraid angry disgusted

The proposed test achieves **maximum test power** in time $O(n)$.

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

- **Informative features**: differences at the nose, and smile lines.
Distinguishing Positive/Negative Emotions

+: happy neutral surprised

-: afraid angry disgusted

Learned feature

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

Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016
Code: https://github.com/wittawatj/interpretable-test
Co-authors
From Gatsby:
- Kacper Chwialkowski
- Wittawat Jitkrittum
- Bharath Sriperumbudur
- Heiko Strathmann
- Dougal Sutherland
- Zoltan Szabo
- Wenkai Xu

Questions?

External collaborators:
- Kenji Fukumizu
- Bernhard Schoelkopf
- Alex Smola