Deep kernel-based distances between distributions

Based on work with:

Michael Arbel
Arthur Gretton
Aaditya Ramdas
Hsiao-Yu (Fish) Tung

Mikołaj Bińkowski
Feng Liu
Alex Smola
Wenkai Xu

Soumyajit De
Jie Lu
Heiko Strathmann
Guangquan Zhang

\[ D(\cdot, \cdot) \]

PIHOT kick-off, 30 Jan 2021
What's a kernel again?

- Linear classifiers: $\hat{y}(x) = \text{sign}(f(x))$, $f(x) = w^T (x, 1)$
What's a kernel again?

- Linear classifiers: \( \hat{y}(x) = \text{sign}(f(x)), f(x) = w^T (x, 1) \)
What's a kernel again?

- Linear classifiers: \( \hat{y}(x) = \text{sign}(f(x)), f(x) = w^T(x, 1) \)
What's a kernel again?

- Linear classifiers: \( \hat{y}(x) = \text{sign}(f(x)), f(x) = w^T (x, 1) \)
- Use a “richer” \( x \):

\[
f(x) = w^T (x, x^2, 1) = w^T \phi(x)
\]
What's a kernel again?

- Linear classifiers: \( \hat{y}(x) = \text{sign}(f(x)), \ f(x) = w^\top (x, 1) \)
- Use a “richer” \( x \):

\[
f(x) = w^\top (x, x^2, 1) = w^\top \phi(x)
\]
What's a kernel again?

- Linear classifiers: $\hat{y}(x) = \text{sign}(f(x))$, $f(x) = w^T (x, 1)$
- Use a “richer” $x$:

$$f(x) = w^T (x, x^2, 1) = w^T \phi(x)$$
What's a kernel again?

- Linear classifiers: \( \hat{y}(x) = \text{sign}(f(x)) \), \( f(x) = w^T (x, 1) \)
- Use a "richer" \( x \):

\[
f(x) = w^T (x, x^2, 1) = w^T \phi(x)
\]
- Can avoid explicit \( \phi(x) \); instead \( k(x, y) = \langle \phi(x), \phi(y) \rangle_\mathcal{H} \)
What's a kernel again?

- Linear classifiers: \( \hat{y}(x) = \text{sign}(f(x)), f(x) = w^T (x, 1) \)
- Use a “richer” \( x \):

\[
f(x) = w^T (x, x^2, 1) = w^T \phi(x)
\]

- Can avoid explicit \( \phi(x) \); instead \( k(x, y) = \langle \phi(x), \phi(y) \rangle_\mathcal{H} \)
- “Kernelized” algorithms access data only through \( k(x, y) \)

\[
f(x) = \langle w, \phi(x) \rangle_\mathcal{H} = \sum_{i=1}^{n} \alpha_i k(X_i, x)
\]
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF

\[ k(x, y) = \exp \left( - \frac{\|x - y\|^2}{2\sigma^2} \right) \]
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF

\[ k(x, y) = \exp \left( - \frac{\|x - y\|^2}{2\sigma^2} \right) \]
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF / exponentiated quadratic / squared exponential / ...

\[ k(x, y) = \exp \left( - \frac{||x - y||^2}{2\sigma^2} \right) \]
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF / exponentiated quadratic / squared exponential / ...

\[ k(x, y) = \exp \left( - \frac{\| x - y \|^2}{2\sigma^2} \right) \]

- Reproducing property: \( \langle f, \phi(x) \rangle_\mathcal{H} = f(x) \) for \( f \in \mathcal{H} \)
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF / exponentiated quadratic / squared exponential / ...

\[ k(x, y) = \exp \left( -\frac{\|x - y\|^2}{2\sigma^2} \right) \]

- Reproducing property: \( \langle f, \phi(x) \rangle_{\mathcal{H}} = f(x) \) for \( f \in \mathcal{H} \)

- \( \mathcal{H} = \text{cl}(\{\sum_{i=1}^{n} \alpha_i \phi(X_i) \mid n \geq 0, \alpha \in \mathbb{R}^n, X_i \in \mathcal{X}\}) \)
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF / exponentiated quadratic / squared exponential / ... 

\[ k(x, y) = \exp \left( -\frac{\|x - y\|^2}{2\sigma^2} \right) \]

- Reproducing property: \( \langle f, \phi(x) \rangle_{\mathcal{H}} = f(x) \) for \( f \in \mathcal{H} \)

\[ \mathcal{H} = \text{cl}(\{ \sum_{i=1}^{n} \alpha_i \phi(X_i) \mid n \geq 0, \alpha \in \mathbb{R}^n, X_i \in \mathcal{X} \}) \]
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF / exponentiated quadratic / squared exponential / ...

\[ k(x, y) = \exp \left( - \frac{\|x - y\|^2}{2\sigma^2} \right) \]

- Reproducing property: \( \langle f, \phi(x) \rangle_H = f(x) \) for \( f \in \mathcal{H} \)

\( \mathcal{H} = \text{cl}(\{ \sum_{i=1}^{n} \alpha_i \phi(X_i) \mid n \geq 0, \alpha \in \mathbb{R}^n, X_i \in \mathcal{X} \}) \)
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF / exponentiated quadratic / squared exponential / ...

\[ k(x, y) = \exp \left( -\frac{\|x - y\|^2}{2\sigma^2} \right) \]

- Reproducing property: \( \langle f, \phi(x) \rangle_{\mathcal{H}} = f(x) \) for \( f \in \mathcal{H} \)

- \( \mathcal{H} = \text{cl}(\{ \sum_{i=1}^{n} \alpha_i \phi(X_i) \mid n \geq 0, \alpha \in \mathbb{R}^n, X_i \in \mathcal{X} \}) \)
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF / exponentiated quadratic / squared exponential / …

\[ k(x, y) = \exp \left( -\frac{\|x - y\|^2}{2\sigma^2} \right) \]

- Reproducing property: \( \langle f, \phi(x) \rangle_H = f(x) \) for \( f \in H \)

\[ \mathcal{H} = \text{cl}(\{ \sum_{i=1}^{n} \alpha_i \phi(X_i) \mid n \geq 0, \alpha \in \mathbb{R}^n, X_i \in X \}) \]

\[ \| \sum_i \alpha_i \phi(X_i) \|_H^2 = \alpha^T K \alpha, \text{ where } K_{ij} = k(X_i, X_j) \]
Reproducing Kernel Hilbert Space (RKHS)

- Ex: Gaussian RBF / exponentiated quadratic / squared exponential / ...

\[ k(x, y) = \exp \left( -\frac{\|x - y\|^2}{2\sigma^2} \right) \]

- Reproducing property: \( \langle f, \phi(x) \rangle_{\mathcal{H}} = f(x) \) for \( f \in \mathcal{H} \)

\[ \mathcal{H} = \text{cl}(\{ \sum_{i=1}^{n} \alpha_i \phi(X_i) \mid n \geq 0, \alpha \in \mathbb{R}^n, X_i \in \mathcal{X} \}) \]

\[ \| \sum_i \alpha_i \phi(X_i) \|_{\mathcal{H}}^2 = \alpha^T K \alpha, \text{ where } K_{ij} = k(X_i, X_j) \]

\[ \arg\min_{f \in \mathcal{H}} L(f(X_1), \ldots, f(X_n)) + \lambda \| f \|_{\mathcal{H}}^2 \] is in \( \{ \sum_{i=1}^{n} \alpha_i \phi(X_i) \mid \alpha \in \mathbb{R}^n \} \) - the representer theorem
Maximum Mean Discrepancy (MMD)

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}} [f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [f(Y)]$$
Maximum Mean Discrepancy (MMD)

\[
\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]
\]

\[
= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[\langle f, \varphi(X) \rangle_{\mathcal{H}}] - \mathbb{E}_{Y \sim \mathbb{Q}}[\langle f, \varphi(Y) \rangle_{\mathcal{H}}]
\]
Maximum Mean Discrepancy (MMD)

\[
\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_\mathcal{H} \leq 1} \mathbb{E}_{X \sim \mathbb{P}} [f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [f(Y)]
\]

\[
= \sup_{\|f\|_\mathcal{H} \leq 1} \mathbb{E}_{X \sim \mathbb{P}} [\langle f, \varphi(X) \rangle_{\mathcal{H}}] - \mathbb{E}_{Y \sim \mathbb{Q}} [\langle f, \varphi(Y) \rangle_{\mathcal{H}}]
\]

\[
= \sup_{\|f\|_\mathcal{H} \leq 1} \left\langle f, \mathbb{E}_{X \sim \mathbb{P}} [\varphi(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [\varphi(Y)] \right\rangle_{\mathcal{H}}
\]
Maximum Mean Discrepancy (MMD)

\[
MMD_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_\mathcal{H} \leq 1} \mathbb{E}_{X \sim \mathbb{P}} [f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [f(Y)]
\]

\[
= \sup_{\|f\|_\mathcal{H} \leq 1} \mathbb{E}_{X \sim \mathbb{P}} [\langle f, \varphi(X) \rangle_\mathcal{H}] - \mathbb{E}_{Y \sim \mathbb{Q}} [\langle f, \varphi(Y) \rangle_\mathcal{H}]
\]

\[
= \sup_{\|f\|_\mathcal{H} \leq 1} \left\langle f, \mathbb{E}_{X \sim \mathbb{P}} [\varphi(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [\varphi(Y)] \right\rangle_\mathcal{H}
\]

\[
= \sup_{\|f\|_\mathcal{H} \leq 1} \left\langle f, \mu_\mathbb{P}^k - \mu_\mathbb{Q}^k \right\rangle_\mathcal{H}
\]
Maximum Mean Discrepancy (MMD)

\[
\begin{align*}
\text{MMD}_k(P,Q) &= \sup_{\|f\|_\mathcal{H} \leq 1} \mathbb{E}_{X \sim P} [f(X)] - \mathbb{E}_{Y \sim Q} [f(Y)] \\
&= \sup_{\|f\|_\mathcal{H} \leq 1} \mathbb{E}_{X \sim P} [\langle f, \varphi(X) \rangle_\mathcal{H}] - \mathbb{E}_{Y \sim Q} [\langle f, \varphi(Y) \rangle_\mathcal{H}] \\
&= \sup_{\|f\|_\mathcal{H} \leq 1} \left\langle f, \mathbb{E}_{X \sim P} [\varphi(X)] - \mathbb{E}_{Y \sim Q} [\varphi(Y)] \right\rangle_\mathcal{H} \\
&= \sup_{\|f\|_\mathcal{H} \leq 1} \left\langle f, \mu^k_P - \mu^k_Q \right\rangle_\mathcal{H} = \|\mu^k_P - \mu^k_Q\|_\mathcal{H}
\end{align*}
\]
MMD as feature matching

$$\text{MMD}_k(\mathcal{P}, \mathcal{Q}) = \left\| \mathbb{E}_{X \sim \mathcal{P}} [\varphi(X)] - \mathbb{E}_{Y \sim \mathcal{Q}} [\varphi(Y)] \right\|_\mathcal{H}$$

- \( \varphi : X \rightarrow \mathcal{H} \) is the feature map for \( k(x, y) = \langle \varphi(x), \varphi(y) \rangle \)
MMD as feature matching

\[ \text{MMD}_k(P, Q) = \left\| \mathbb{E}_{X \sim P} [\varphi(X)] - \mathbb{E}_{Y \sim Q} [\varphi(Y)] \right\|_\mathcal{H} \]

- \( \varphi : X \rightarrow \mathcal{H} \) is the feature map for \( k(x, y) = \langle \varphi(x), \varphi(y) \rangle \)
- If \( k(x, y) = x^T y, \varphi(x) = x \); MMD is distance between means
**MMD as feature matching**

\[
\text{MMD}_k(P, Q) = \left\| \mathbb{E}_{X \sim P} [\varphi(X)] - \mathbb{E}_{Y \sim Q} [\varphi(Y)] \right\|_\mathcal{H}
\]

- \( \varphi : X \rightarrow \mathcal{H} \) is the feature map for \( k(x, y) = \langle \varphi(x), \varphi(y) \rangle \)
- If \( k(x, y) = x^T y, \varphi(x) = x \); MMD is distance between means
- Many kernels: **infinite-dimensional** \( \mathcal{H} \)
### Entropic Regularization

**Schrodinger's problem:**

\[
\min_{\mathbf{P}_1 = a, \mathbf{P}^\top_1 = b} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \epsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})
\]

\[
\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{x_i, y_j}
\]

### Sinkhorn Divergences

**Problem:** \(W^\varepsilon_p(\alpha, \alpha) \neq 0\)

\[
\min_{\alpha} W^\varepsilon_p(\alpha, \beta)
\]

\[
W^\varepsilon_p(\alpha, \beta)^p \overset{\text{def}}{=} \min_{\mathbf{P}_1 = a, \mathbf{P}^\top_1 = b} \sum_{i,j} d(x_i, y_j)^p \mathbf{P}_{i,j} + \epsilon \mathbf{P}_{i,j} \log(\mathbf{P}_{i,j})
\]

\[
\overline{W}^\varepsilon_p(\alpha, \beta)^p \overset{\text{def}}{=} W^\varepsilon_p(\alpha, \beta)^p - \frac{1}{2} W^\varepsilon_p(\alpha, \alpha)^p - \frac{1}{2} W^\varepsilon_p(\beta, \beta)^p
\]

[Randas, Garcia Trillos, Cuturi, 2017]

**Theorem:** \(W_p(\alpha, \beta)^p \xrightarrow{\varepsilon \to 0} -W^\varepsilon_p(\alpha, \beta)^p \xrightarrow{\varepsilon \to +\infty} \|\alpha - \beta\|_{d^p}^2\)

[Leonard 2012]

[Carlier et al 2017]

[Randas, Garcia Trillos, Cuturi, 2017]

**Kernel norms (MMD):**

\[
\|\xi\|_{d^p}^2 \overset{\text{def}}{=} -\int_{\mathcal{X}} d(x, y)^p d\xi(x) d\xi(y)
\]

**Proposition:** \(\|\cdot\|_{P}^p\) is a norm for \(0 < p < 2\).
Estimating MMD

\[
\text{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{X, X' \sim \mathbb{P}} [k(X, X')] + \mathbb{E}_{Y, Y' \sim \mathbb{Q}} [k(Y, Y')] - 2 \mathbb{E}_{X \sim \mathbb{P}} \mathbb{E}_{Y \sim \mathbb{Q}} [k(X, Y)]
\]
Estimating MMD

\[
\text{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{X, X' \sim \mathbb{P}}[k(X, X')] + \mathbb{E}_{Y, Y' \sim \mathbb{Q}}[k(Y, Y')] - 2 \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}}[k(X, Y)]
\]

\[
\widehat{\text{MMD}}_k^2(X, Y) = \text{mean}(K_{XX}) + \text{mean}(K_{YY}) - 2 \text{mean}(K_{XY})
\]
Estimating MMD

\[
\begin{align*}
\text{MMD}_k^2(\mathbb{P}, \mathbb{Q}) &= \mathbb{E}_{X, X' \sim \mathbb{P}} [k(X, X') ] + \mathbb{E}_{Y, Y' \sim \mathbb{Q}} [k(Y, Y') ] - 2 \mathbb{E}_{X \sim \mathbb{P}} [k(X, Y) ] \\
\hat{\text{MMD}}_k^2(X, Y) &= \text{mean}(K_{XX}) + \text{mean}(K_{YY}) - 2 \text{mean}(K_{XY})
\end{align*}
\]
Estimating MMD

\[
\text{MMD}^2_k(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{X,X' \sim \mathbb{P}} [k(X, X')] + \mathbb{E}_{Y,Y' \sim \mathbb{Q}} [k(Y, Y')] - 2 \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} [k(X, Y)]
\]

\[
\hat{\text{MMD}}^2_k(X, Y) = \text{mean}(K_{XX}) + \text{mean}(K_{YY}) - 2 \text{mean}(K_{XY})
\]

\[
K_{XX} \quad K_{YY}
\]

\[
\begin{array}{ccc}
1.0 & 0.2 & 0.6 \\
0.2 & 1.0 & 0.5 \\
0.6 & 0.5 & 1.0 \\
\end{array}
\quad
\begin{array}{ccc}
1.0 & 0.8 & 0.7 \\
0.8 & 1.0 & 0.6 \\
0.7 & 0.6 & 1.0 \\
\end{array}
\]
Estimating MMD

\[ \text{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{X,X' \sim \mathbb{P}} [k(X, X')] + \mathbb{E}_{Y,Y' \sim \mathbb{Q}} [k(Y, Y')] - 2 \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} [k(X, Y)] \]

\[ \widehat{\text{MMD}}_k^2(X, Y) = \text{mean}(K_{XX}) + \text{mean}(K_{YY}) - 2 \text{mean}(K_{XY}) \]
I: Two-sample testing

- Given samples from two unknown distributions

\[ X \sim P \quad Y \sim Q \]

- Question: is \( P = Q \)?
I: Two-sample testing

- Given samples from two unknown distributions

\[ X \sim P \quad Y \sim Q \]

- Do smokers/non-smokers get different cancers?
I: Two-sample testing

- Given samples from two unknown distributions
  \[ X \sim P, \quad Y \sim Q \]

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
I: Two-sample testing

- Given samples from two unknown distributions

Given samples from two unknown distributions:

\[ X \sim P \quad Y \sim Q \]

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
I: Two-sample testing

- Given samples from two unknown distributions
  \[ X \sim P, \quad Y \sim Q \]

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?
I: Two-sample testing

- Given samples from two unknown distributions

\[ X \sim P \quad Y \sim Q \]

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?
- Does presence of this protein affect DNA binding? [MMDiff2]
I: Two-sample testing

- Given samples from two unknown distributions

\[ X \sim p \quad Y \sim q \]

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?
- Does presence of this protein affect DNA binding? [MMDiff2]
- Do these dob and birthday columns mean the same thing?
I: Two-sample testing

- Given samples from two unknown distributions

\[ X \sim P \quad Y \sim Q \]

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?
- Does presence of this protein affect DNA binding? [MMDiff2]
- Do these dob and birthday columns mean the same thing?
- Does my generative model \( Q_\theta \) match \( P_{\text{data}} \)?
I: Two-sample testing

- Given samples from two unknown distributions

\[ X \sim P \quad Y \sim Q \]

- Do smokers/non-smokers get different cancers?
- Do Brits have the same friend network types as Americans?
- When does my laser agree with the one on Mars?
- Are storms in the 2000s different from storms in the 1800s?
- Does presence of this protein affect DNA binding? \([\text{MMDiff2}]\)
- Do these dob and birthday columns mean the same thing?
- Does my generative model \(Q_\theta\) match \(P_{\text{data}}\)?
- Independence testing: is \(P(X, Y) = P(X)P(Y)\)?
I: Two-sample testing

- Given samples from two unknown distributions
  \[ X \sim P \quad Y \sim Q \]

- Question: is \( P = Q \)?
I: Two-sample testing

- Given samples from two unknown distributions
  
  \[ X \sim P \quad Y \sim Q \]

- Question: is \( P = Q \)?

- Hypothesis testing approach:
  
  \[ H_0 : P = Q \quad H_1 : P \neq Q \]
I: Two-sample testing

- Given samples from two unknown distributions

  \[ X \sim \mathcal{P} \quad Y \sim \mathcal{Q} \]

- Question: is \( \mathcal{P} = \mathcal{Q} \)?

- Hypothesis testing approach:

  \[ H_0 : \mathcal{P} = \mathcal{Q} \quad H_1 : \mathcal{P} \neq \mathcal{Q} \]

- Reject \( H_0 \) if test statistic \( \hat{T}(X, Y) > c_\alpha \)
What's a hypothesis test again?
What's a hypothesis test again?
What's a hypothesis test again?

don't reject $H_0$  $c_\alpha$  reject $H_0$ (say $P \neq Q$)

$P = Q$  $P \neq Q$
What's a hypothesis test again?

don't reject \( H_0 \)  \( c_\alpha \)  reject \( H_0 \) (say \( P \neq Q \))

\[
\text{false rejection rate: want } \leq \alpha
\]

\[
\hat{T}(X, Y)
\]
What's a hypothesis test again?

don't reject $H_0$  $c_\alpha$  reject $H_0$ (say $P \neq Q$)

false rejection rate: want $\leq \alpha$

power: true rejection rate
Permutation testing to find $c_\alpha$

Need $\Pr_{H_0} (T(X, Y) > c_\alpha) \leq \alpha$

$X_1 X_2 X_3 X_4 X_5 Y_1 Y_2 Y_3 Y_4 Y_5$

$c_\alpha$: $1 - \alpha$th quantile of \{ }

\}
Permutation testing to find $c_\alpha$

Need $\Pr_{H_0} (T(X, Y) > c_\alpha) \leq \alpha$

$c_\alpha$: 1 - $\alpha$th quantile of $\{X_1, X_2, X_3, X_4, X_5, Y_1, Y_2, Y_3, Y_4, Y_5\}$
Permutation testing to find $c_\alpha$

Need $\Pr_{H_0} (T(X, Y) > c_\alpha) \leq \alpha$

$c_\alpha$: $1 - \alpha$th quantile of $\{\hat{T}(\tilde{X}_1, \tilde{Y}_1), \}$
Permutation testing to find $c_\alpha$

Need $\Pr_{H_0} (T(X, Y) > c_\alpha) \leq \alpha$

$c_\alpha$: $1 - \alpha$th quantile of $\left\{ \hat{T}(\tilde{X}_1, \tilde{Y}_1), \hat{T}(\tilde{X}_2, \tilde{Y}_2) \right\}$
Permutation testing to find $c_\alpha$

Need $\Pr_{H_0} (T(X, Y) > c_\alpha) \leq \alpha$

$X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad Y_1 \quad Y_2 \quad Y_3 \quad Y_4 \quad Y_5$

$c_\alpha$: $1 - \alpha$th quantile of $\left\{ \hat{T}(\tilde{X}_1, \tilde{Y}_1), \hat{T}(\tilde{X}_2, \tilde{Y}_2), \cdots \right\}$
MMD-based tests

- If $k$ is characteristic, $\text{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$
- Efficient permutation testing for $\widehat{\text{MMD}}(X, Y)$
MMD-based tests

- If $k$ is characteristic, $\text{MMD}(P, Q) = 0$ iff $P = Q$
- Efficient permutation testing for $\hat{\text{MMD}}(X, Y)$
  - $H_0$: $n\hat{\text{MMD}}^2$ converges in distribution
  - $H_1$: $\sqrt{n}(\hat{\text{MMD}}^2 - \text{MMD}^2)$ asymptotically normal
MMD-based tests

- If \( k \) is characteristic, \( \text{MMD}(\mathbb{P}, \mathbb{Q}) = 0 \) iff \( \mathbb{P} = \mathbb{Q} \)

- Efficient permutation testing for \( \widehat{\text{MMD}}(\mathbf{X}, \mathbf{Y}) \)
  - \( H_0: n\widehat{\text{MMD}}^2 \) converges in distribution
  - \( H_1: \sqrt{n}(\widehat{\text{MMD}}^2 - \text{MMD}^2) \) asymptotically normal

- Any characteristic kernel gives consistent test
MMD-based tests

• If $k$ is \textit{characteristic}, $\text{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$

• Efficient permutation testing for $\widehat{\text{MMD}}(\mathbf{X}, \mathbf{Y})$
  - $H_0$: $n\widehat{nMMD}^2$ converges in distribution
  - $H_1$: $\sqrt{n}(\widehat{\text{MMD}}^2 - \text{MMD}^2)$ asymptotically normal

• Any characteristic kernel gives consistent test...eventually
MMD-based tests

- If \( k \) is characteristic, \( \text{MMD}(\mathbb{P}, \mathbb{Q}) = 0 \) iff \( \mathbb{P} = \mathbb{Q} \)

- Efficient permutation testing for \( \widehat{\text{MMD}}(X, Y) \)
  - \( H_0 : n \widehat{\text{MMD}}^2 \) converges in distribution
  - \( H_1 : \sqrt{n}(\widehat{\text{MMD}}^2 - \text{MMD}^2) \) asymptotically normal

- Any characteristic kernel gives consistent test...eventually

- Need enormous \( n \) if kernel is bad for problem
Classifier two-sample tests

\[ T(X, Y) \] is the accuracy of \( f \) on the test set

Under \( H_0 \), classification impossible: \( \hat{T} \sim \text{Binomial}(n, \frac{1}{2}) \)
Classifier two-sample tests

- \( \hat{T}(X, Y) \) is the accuracy of \( f \) on the test set
- Under \( H_0 \), classification impossible: \( \hat{T} \sim \text{Binomial}(n, \frac{1}{2}) \)
- With \( k(x, y) = \frac{1}{4} f(x) f(y) \) where \( f(x) \in \{-1, 1\} \),
  get \( \hat{\text{MMD}}(X, Y) = \left| \hat{T}(X, Y) - \frac{1}{2} \right| \)
Optimizing test power

- Asymptotics of $\widehat{\text{MMD}}^2$ give us immediately that

$$\Pr_{H_1} \left( n\widehat{\text{MMD}}^2 > c_\alpha \right) \approx \Phi \left( \frac{\sqrt{n} \text{MMD}^2}{\sigma_{H_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{H_1}} \right)$$

$\text{MMD}, \sigma_{H_1}, c_\alpha$ are constants: first term dominates
Optimizing test power

• Asymptotics of $\hat{\text{MMD}}^2$ give us immediately that

$$\Pr_{\hat{H}_1}\left(n\hat{\text{MMD}}^2 > c_\alpha\right) \approx \Phi \left(\frac{\sqrt{n} \text{MMD}^2}{\sigma_{\hat{H}_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{\hat{H}_1}}\right)$$

$\text{MMD}$, $\sigma_{\hat{H}_1}$, $c_\alpha$ are constants: first term dominates

• Pick $k$ to maximize an estimate of $\text{MMD}^2 / \sigma_{\hat{H}_1}$
Optimizing test power

- Asymptotics of $\widehat{\text{MMD}}^2$ give us immediately that

$$\Pr_{H_1}\left(n\widehat{\text{MMD}}^2 > c_{\alpha}\right) \approx \Phi\left(\frac{\sqrt{n} \text{MMD}^2}{\sigma_{H_1}} - \frac{c_{\alpha}}{\sqrt{n}\sigma_{H_1}}\right)$$

$\text{MMD}$, $\sigma_{H_1}$, $c_{\alpha}$ are constants: first term dominates

- Pick $k$ to maximize an estimate of $\text{MMD}^2 / \sigma_{H_1}$

- Can show uniform $O_P \left(n^{-\frac{1}{3}}\right)$ convergence of estimator
Blobs dataset
Blobs kernels

Sample values at 1st dimension
**CIFAR-10 vs CIFAR-10.1**

Train on 1 000, test on 1 031, repeat 10 times. Rejection rates:

<table>
<thead>
<tr>
<th></th>
<th>ME</th>
<th>SCF</th>
<th>C2ST</th>
<th>MMD-O</th>
<th>MMD-D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.588</td>
<td>0.171</td>
<td>0.452</td>
<td>0.316</td>
<td>0.744</td>
</tr>
</tbody>
</table>
# Ablation vs classifier-based tests

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cross-entropy</th>
<th>Max power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sign</td>
<td>Lin</td>
</tr>
<tr>
<td>Blob</td>
<td>0.84</td>
<td>0.94</td>
</tr>
<tr>
<td>High-$d$ Gauss. mix.</td>
<td>0.47</td>
<td>0.59</td>
</tr>
<tr>
<td>Higgs</td>
<td>0.26</td>
<td><strong>0.40</strong></td>
</tr>
<tr>
<td>MNIST vs GAN</td>
<td>0.65</td>
<td>0.71</td>
</tr>
</tbody>
</table>
II: Training implicit generative models

Given samples from a distribution $\mathbb{P}$ over $\mathcal{X}$, we want a model that can produce new samples from $Q_\theta \approx \mathbb{P}$.

$\mathcal{X} \sim \mathbb{P}$

$\mathcal{Y} \sim Q_\theta$
II: Training implicit generative models

Given samples from a distribution over $\mathcal{X}$, we want a model that can produce new samples from $P$. We aim to approximate $P$ with $Q_\theta$. We use $Q_\theta$ to learn the implicit generative models.

thispersondoesnotexist.com
II: Training implicit generative models

Given samples from a distribution $\mathbb{P}$ over $\mathcal{X}$, we want a model that can produce new samples from $Q_\theta \approx \mathbb{P}$.

“Everybody Dance Now” [Chan et al. ICCV-19]
Generator networks

Fixed distribution of latents: $Z \sim \text{Uniform} \left([-1, 1]^{100}\right)$

Maps through a network: $G_\theta(Z) \sim Q_\theta$

DCGAN generator [Radford+ ICLR-16]
Generator networks

Fixed distribution of latents: $Z \sim \text{Uniform}(\mathbb{R}^{100})$

Maps through a network: $G_\theta(Z) \sim Q_\theta$

DCGAN generator [Radford+ ICLR-16]

How to choose $\theta$?
GANs and their flaws

- GANs [Goodfellow+ NeurIPS-14] minimize discriminator accuracy (like classifier test) between $\mathbb{P}$ and $\mathbb{Q}_\theta$

- Problem: if there's a perfect classifier, discontinuous loss, no gradient to improve it [Arjovsky/Bottou ICLR-17]

- Disjoint at init:
GANs and their flaws

- GANs [Goodfellow+ NeurIPS-14] minimize discriminator accuracy (like classifier test) between $\mathbb{P}$ and $Q_\theta$

- Problem: if there's a perfect classifier, discontinuous loss, no gradient to improve it [Arjovsky/Bottou ICLR-17]

- Disjoint at init:

- For usual $G_\theta : \mathbb{R}^{100} \rightarrow \mathbb{R}^{64 \times 64 \times 3}$, $Q_\theta$ is supported on a countable union of manifolds with dim $\leq 100$
GANs and their flaws

- GANs [Goodfellow+ NeurIPS-14] minimize discriminator accuracy (like classifier test) between $\mathbb{P}$ and $\mathcal{Q}_\theta$.

- Problem: if there's a perfect classifier, discontinuous loss, no gradient to improve it [Arjovsky/Bottou ICLR-17].

- Disjoint at init:

  For usual $G_\theta : \mathbb{R}^{100} \rightarrow \mathbb{R}^{64 \times 64 \times 3}$, $\mathcal{Q}_\theta$ is supported on a countable union of manifolds with dim $\leq 100$.

- “Natural image manifold” usually considered low-dim.
GANs and their flaws

- GANs [Goodfellow+ NeurIPS-14] minimize discriminator accuracy (like classifier test) between $P$ and $Q_\theta$

- Problem: if there's a perfect classifier, discontinuous loss, no gradient to improve it [Arjovsky/Bottou ICLR-17]

- Disjoint at init:

- For usual $G_\theta : \mathbb{R}^{100} \rightarrow \mathbb{R}^{64 \times 64 \times 3}$, $Q_\theta$ is supported on a countable union of manifolds with dim $\leq 100$

- “Natural image manifold” usually considered low-dim

- Won't align at init, so won't ever align
WGANs and MMD GANs

- Integral probability metrics with “smooth” $\mathcal{F}$ are continuous.
- WGAN: $\mathcal{F}$ a set of neural networks satisfying $\|f\|_L \leq 1$.
- WGAN-GP: instead penalize $\mathbb{E}\|\nabla_x f(x)\|$ near the data.
WGANS and MMD GANs

- Integral probability metrics with “smooth” $\mathcal{F}$ are continuous
- WGAN: $\mathcal{F}$ a set of neural networks satisfying $\|f\|_L \leq 1$
- WGAN-GP: instead penalize $\mathbb{E}\|\nabla_x f(x)\|$ near the data
- Both losses are MMD with $k_\psi(x, y) = \phi_\psi(x)\phi_\psi(y)$
WGANs and MMD GANs

- Integral probability metrics with “smooth” $\mathcal{F}$ are continuous
- WGAN: $\mathcal{F}$ a set of neural networks satisfying $\|f\|_L \leq 1$
- WGAN-GP: instead penalize $\mathbb{E}\|\nabla_x f(x)\|$ near the data
- Both losses are MMD with $k_\psi(x, y) = \phi_\psi(x)\phi_\psi(y)$

\[
\min_{\theta} \left[ \mathcal{D}_{MMD}^\Psi (\mathbb{P}, \mathbb{Q}_\theta) = \sup_{\psi \in \Psi} \text{MMD}_\psi (\mathbb{P}, \mathbb{Q}_\theta) \right]
\]
WGANs and MMD GANs

- Integral probability metrics with “smooth” $\mathcal{F}$ are continuous
- WGAN: $\mathcal{F}$ a set of neural networks satisfying $\|f\|_L \leq 1$
- WGAN-GP: instead penalize $\mathbb{E}\|\nabla_x f(x)\|$ near the data
- Both losses are MMD with $k_\psi(x, y) = \phi_\psi(x)\phi_\psi(y)$

$$\min_{\theta} \left[ D_{MMD}^\psi (\mathbb{P}, Q_\theta) = \sup_{\psi \in \Psi} \text{MMD}_\psi (\mathbb{P}, Q_\theta) \right]$$

- Some kind of constraint on $\phi_\psi$ is important!
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

$$k_{\psi}(\theta, x)$$

0 \quad x \quad \theta
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

$K_{\psi \leftarrow 2}(\Theta, x)$
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

$$k_{\psi = 2}(0, x)$$

$0 \quad x \quad \theta$

$\begin{array}{c}
\text{MMD}_2 \\
\hline
\end{array}$
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

$k_\psi=2(0, x)$

$x$

$\theta$

$\text{MMD}_2$
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

\[ k_{\psi} = _2(0, x) \]

\[ k_{\psi} = _{0.25}(0, x) \]
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

$k_{\psi=2}(0, x)$

$k_{\psi=0.25}(0, x)$

$k_{\psi=0.01}(0, x)$

$x \quad \theta$

$\begin{array}{c}
\text{MMD}_{0.25} \\
\text{MMD}_2
\end{array}$

$\begin{array}{c}
0.0 \\
0.5 \\
1.0 \\
1.5
\end{array}$
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

\[ k_\psi = 2(0, x) \quad k_\psi = 0.25(0, x) \quad k_\psi = 0.01(0, x) \]
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

\[ k_\psi = 2(0, x), \quad k_\psi = 0.25(0, x), \quad k_\psi = 0.01(0, x) \]
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

\[ k_{\psi=2}(0, x) \quad k_{\psi=0.25}(0, x) \quad k_{\psi=0.01}(0, x) \]
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

- $k_\psi = 2(0, x)$
- $k_\psi = 0.25(0, x)$
- $k_\psi = 0.01(0, x)$

\[ D_{\text{MMD}} \]
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

- Just need to stay away from tiny bandwidths $\psi$
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

- Just need to stay away from tiny bandwidths $\psi$
- ...deep kernel analogue is hard.
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

- Just need to stay away from tiny bandwidths $\psi$
- ...deep kernel analogue is hard.
- Instead, keep witness function from being too steep
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

- Just need to stay away from tiny bandwidths $\psi$
- ...deep kernel analogue is hard.
- Instead, keep witness function from being too steep
- $\sup_x \| \nabla f(x) \|$ would give Wasserstein
  - Nice distance, but hard to estimate
Non-smoothness of plain MMD GANs

Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

- Just need to stay away from tiny bandwidths $\psi$
- ...deep kernel analogue is hard.
- Instead, keep witness function from being too steep
- $\sup_{x} \| \nabla f(x) \|$ would give Wasserstein
  - Nice distance, but hard to estimate
- Control $\| \nabla f(\tilde{X}) \|$ on average, near the data
  - [Gulrajani+ NeurIPS-17 / Roth+ NeurIPS-17 / Mescheder+ ICML-18]
MMD-GAN with gradient control

- If $\Psi$ gives uniformly Lipschitz critics, $D_{\text{MMD}}^\Psi$ is smooth.
- Original MMD-GAN paper [Li+ NeurIPS-17]: box constraint.
MMD-GAN with gradient control

- If $\Psi$ gives uniformly Lipschitz critics, $D^\Psi_{\text{MMD}}$ is smooth
- Original MMD-GAN paper [Li+ NeurIPS-17]: box constraint
- We [Bińkowski+ ICLR-18] used gradient penalty on critic instead
MMD-GAN with gradient control

- If $\Psi$ gives uniformly Lipschitz critics, $D_{\text{MMD}}^{\Psi}$ is smooth.
- Original MMD-GAN paper [Li+ NeurIPS-17]: box constraint.
- We [Bińkowski+ ICLR-18] used gradient penalty on critic instead:
  - Better in practice, but doesn't fix the Dirac problem...
MMD-GAN with gradient control

- If $\Psi$ gives uniformly Lipschitz critics, $D^\Psi_{\text{MMD}}$ is smooth.
- Original MMD-GAN paper [Li+ NeurIPS-17]: box constraint.
- We [Bińkowski+ ICLR-18] used gradient penalty on critic instead:
  - Better in practice, but doesn't fix the Dirac problem...
New distance: Scaled MMD

Want to ensure $\mathbb{E}_{\tilde{X} \sim S}[\|\nabla f(\tilde{X})\|^2] \leq 1$
New distance: Scaled MMD

Want to ensure $\mathbb{E}_{\tilde{X} \sim S}[\|\nabla f(\tilde{X})\|^2] \leq 1$

Can solve with $\langle \partial_i \phi(x), f \rangle_{\mathcal{H}} = \partial_i f(x)$...but too expensive!
New distance: Scaled MMD

Want to ensure \( \mathbb{E}_{\tilde{X} \sim S} [\| \nabla f(\tilde{X}) \|^2] \leq 1 \)

Can solve with \( \langle \partial_i \phi(x), f \rangle_{\mathcal{H}} = \partial_i f(x) \ldots \) but too expensive!

Guaranteed if \( \| f \|_{\mathcal{H}} \leq \sigma_{S,k,\lambda} \)

\[
\sigma_{S,k,\lambda} := \left( \lambda + \mathbb{E}_{\tilde{X} \sim S} \left[ k(\tilde{X}, \tilde{X}) + [\nabla_1 \cdot \nabla_2 k](\tilde{X}, \tilde{X}) \right] \right)^{-\frac{1}{2}}
\]
New distance: Scaled MMD

Want to ensure $\mathbb{E}_{\tilde{X} \sim S} \left[ \| \nabla f(\tilde{X}) \| ^2 \right] \leq 1$

Can solve with $\langle \partial_i \phi(x), f \rangle_\mathcal{H} = \partial_i f(x)$...but too expensive!

Guaranteed if $\| f \|_\mathcal{H} \leq \sigma_{S,k,\lambda}$

$$\sigma_{S,k,\lambda} := \left( \lambda + \mathbb{E}_{\tilde{X} \sim S} \left[ k(\tilde{X}, \tilde{X}) + [\nabla_1 \cdot \nabla_2 k](\tilde{X}, \tilde{X}) \right] \right)^{-\frac{1}{2}}$$

Gives distance $\text{SMMD}_{S,k,\lambda}(P, Q) = \sigma_{S,k,\lambda} \text{MMD}_k(P, Q)$
New distance: Scaled MMD

Want to ensure $\mathbb{E}_{\tilde{X} \sim S} [\| \nabla f(\tilde{X}) \|^2] \leq 1$

Can solve with $\langle \partial_i \phi(x), f \rangle_{\mathcal{H}} = \partial_i f(x)$...but too expensive!

Guaranteed if $\| f \|_{\mathcal{H}} \leq \sigma_{S,k,\lambda}$

$$\sigma_{S,k,\lambda} := \left( \lambda + \mathbb{E}_{\tilde{X} \sim S} \left[ k(\tilde{X}, \tilde{X}) + [\nabla_1 \cdot \nabla_2 k](\tilde{X}, \tilde{X}) \right] \right)^{-\frac{1}{2}}$$

Gives distance $\text{SMMD}_{S,k,\lambda}(P, Q) = \sigma_{S,k,\lambda} \text{MMD}_k (P, Q)$

$D_{\text{MMD}}^\Psi$ has $\mathcal{F} = \bigcup_{\psi \in \Psi} \left\{ f : \| f \|_{\mathcal{H}_\psi} \leq 1 \right\}$

$D_{\text{SMMD}}^{S,\Psi,\lambda}$ has $\mathcal{F} = \bigcup_{\psi \in \Psi} \left\{ f : \| f \|_{\mathcal{H}_\psi} \leq \sigma_{S,k,\lambda} \right\}$
Deriving the Scaled MMD

$$\mathbb{E}_{\tilde{X} \sim \mathcal{S}} \left[ \left\| \nabla f(\tilde{X}) \right\|^2 \right] \leq 1$$
Deriving the Scaled MMD

\[
\mathbb{E}_{\tilde{X} \sim S} [f(\tilde{X})^2] + \mathbb{E}_{\tilde{X} \sim S} [\|\nabla f(\tilde{X})\|^2] + \lambda \|f\|^2_H \leq 1
\]
Deriving the Scaled MMD

\[
\mathbb{E}_{\tilde{X} \sim S} [f(\tilde{X})^2] + \mathbb{E}_{\tilde{X} \sim S} [\| \nabla f(\tilde{X}) \|^2] + \lambda \| f \|^2_{\mathcal{H}} \leq 1
\]

\[
\mathbb{E}_{\tilde{X} \sim S} [f(\tilde{X})^2] = \left\langle f, \mathbb{E}_{\tilde{X} \sim S} [k(\tilde{X}, \cdot) \otimes k(\tilde{X}, \cdot)] f \right\rangle
\]
Deriving the Scaled MMD

\[
\mathbb{E}_{\tilde{X} \sim S}[f(\tilde{X})^2] + \mathbb{E}_{\tilde{X} \sim S}[\|\nabla f(\tilde{X})\|^2] + \lambda \|f\|^2_H \leq 1
\]

\[
\mathbb{E}_{\tilde{X} \sim S}[f(\tilde{X})^2] = \left\langle f, \mathbb{E}_{\tilde{X} \sim S}[k(\tilde{X}, \cdot) \otimes k(\tilde{X}, \cdot)]f \right\rangle
\]

\[
\mathbb{E}_{\tilde{X} \sim S}[\|\nabla f(\tilde{X})\|^2] = \left\langle f, \mathbb{E}_{\tilde{X} \sim S}
\left[\sum_{i=1}^{d} \partial_i k(\tilde{X}, \cdot) \otimes \partial_i k(\tilde{X}, \cdot)\right]f \right\rangle
\]
Deriving the Scaled MMD

\[\mathbb{E}_{\tilde{X} \sim S} [f(\tilde{X})^2] + \mathbb{E}_{\tilde{X} \sim S} [\| \nabla f(\tilde{X}) \|^2] + \lambda \| f \|^2_\mathcal{H} \leq 1\]

\[\mathbb{E}_{\tilde{X} \sim S} [f(\tilde{X})^2] = \left\langle f, \mathbb{E}_{\tilde{X} \sim S} [k(\tilde{X}, \cdot) \otimes k(\tilde{X}, \cdot)] f \right\rangle\]

\[\mathbb{E}_{\tilde{X} \sim S} [\| \nabla f(\tilde{X}) \|^2] = \left\langle f, \mathbb{E}_{\tilde{X} \sim S} \left[ \sum_{i=1}^{d} \partial_i k(\tilde{X}, \cdot) \otimes \partial_i k(\tilde{X}, \cdot) \right] f \right\rangle\]

\[\langle f, D_\lambda f \rangle \leq \| D_\lambda \| \| f \|^2_\mathcal{H} \leq \sigma_{S,k,\lambda}^{-2} \| f \|^2_\mathcal{H}\]
Smoothness of $D_{\text{SMMD}}$
Smoothness of $D_{SMMD}$
Smoothness of $\mathcal{D}_{\text{SMMD}}$
Smoothness of $D_{\text{SMMD}}$
Smoothness of $\mathcal{D}_{\text{SMMD}}$
**Theorem:** $D_{SMMD}^{S, \Psi, \lambda}$ is continuous.

If $S$ has a density; $k_{\text{top}}$ is Gaussian/linear/...; $\phi_\psi$ is fully-connected, Leaky-ReLU, non-increasing width; all weights in $\Psi$ have bounded condition number; then $\mathcal{W}(Q_n, P) \to 0$ implies $D_{SMMD}^{S, \Psi, \lambda}(Q_n, P) \to 0$. 
Results on 160 × 160 CelebA

SN-SMMD-GAN

KID: 0.006

WGANGP

KID: 0.022
Training process on CelebA

KID $\times 10^3$

generator iterations $\times 10^4$

WGAN-GP
Training process on CelebA

KID×10^3

- **SN-SMMDGAN**
- **WGAN-GP**

Generator iterations vs. KID×10^3
Training process on CelebA

KID $\times 10^3$

- **SN-SMMDGAN**
- **WGAN-GP**
- **MMDGAN-GP-L2**

*generator iterations* $\times 10^4$
Training process on CelebA

KID × 10³

- SN-SMMDGAN
- SN-SWGAN
- WGAN-GP
- MMDGAN-GP-L2

Generator iterations × 10⁴
Evaluating generative models

- Human evaluation: good at precision, bad at recall
- Likelihood: hard for GANs, maybe not right thing anyway
- Two-sample tests: always reject!
Evaluating generative models

- Human evaluation: good at precision, bad at recall
- Likelihood: hard for GANs, maybe not right thing anyway
- Two-sample tests: always reject!
- Most common: Fréchet Inception Distance, FID
  - Run pretrained featurizer on model and target
  - Model each as Gaussian; compute $W_2$
Evaluating generative models

- Human evaluation: good at precision, bad at recall
- Likelihood: hard for GANs, maybe not right thing anyway
- Two-sample tests: always reject!
- Most common: Fréchet Inception Distance, FID
  - Run pretrained featurizer on model and target
  - Model each as Gaussian; compute $W_2$
  - Strong bias, small variance: very misleading
Evaluating generative models

- Human evaluation: good at precision, bad at recall
- Likelihood: hard for GANs, maybe not right thing anyway
- Two-sample tests: always reject!
- Most common: Fréchet Inception Distance, FID
  - Run pretrained featurizer on model and target
  - Model each as Gaussian; compute $W_2$
  - Strong bias, small variance: very misleading
  - Simple examples where $\text{FID}(Q_1) > \text{FID}(Q_2)$ but $\hat{\text{FID}}(\hat{Q}_1) < \hat{\text{FID}}(\hat{Q}_2)$ for reasonable sample size
Evaluating generative models

- Human evaluation: good at precision, bad at recall
- Likelihood: hard for GANs, maybe not right thing anyway
- Two-sample tests: always reject!
- Most common: Fréchet Inception Distance, FID
  - Run pretrained featurizer on model and target
  - Model each as Gaussian; compute $W_2$
  - Strong bias, small variance: very misleading
  - Simple examples where $\text{FID}(Q_1) > \text{FID}(Q_2)$ but $\hat{\text{FID}}(Q_1) < \hat{\text{FID}}(Q_2)$ for reasonable sample size
- Our KID: $\text{MMD}^2$ instead. Unbiased, asymptotically normal
Recap

Combining a deep architecture with a kernel machine that takes the higher-level learned representation as input can be quite powerful.

Recap

Combining a deep architecture with a kernel machine that takes the higher-level learned representation as input can be quite powerful.

• Two-sample testing [ICLR-17, ICML-20]
  ▪ Choose $\psi$ to maximize power criterion
  ▪ Exploit closed form of $f^*_\psi$ for permutation testing

• Generative modeling with MMD GANs [ICLR-18, NeurIPS-18]
  ▪ Need a smooth loss function for the generator
  ▪ Better gradients for generator to follow (?)
Future uses of deep kernel distances

- Selective inference to avoid train/test split? Meta-testing?
Future uses of deep kernel distances

- Selective inference to avoid train/test split? Meta-testing?
- When $\mathbb{P} \neq \mathbb{Q}$, can we tell how they're different?
  - Methods so far: some mostly for low-$d$
  - Some look at points with large critic function
Future uses of deep kernel distances

- Selective inference to avoid train/test split? Meta-testing?
- When $\mathbb{P} \neq \mathbb{Q}$, can we tell how they're different?
  - Methods so far: some mostly for low-$d$
  - Some look at points with large critic function
Future uses of deep kernel distances

- Selective inference to avoid train/test split? Meta-testing?
- When $P \neq Q$, can we tell how they're different?
  - Methods so far: some mostly for low-$d$
  - Some look at points with large critic function

- Does model $Q$ match dataset $X$ (Stein testing)?
Future uses of deep kernel distances

- Selective inference to avoid train/test split? Meta-testing?
- When $P \neq Q$, can we tell how they're different?
  - Methods so far: some mostly for low-$d$
  - Some look at points with large critic function
- Does model $Q$ match dataset $X$ (Stein testing)?
- Maximize deep dependence measure for unsupervised representation learning, as in contrastive learning
Future uses of deep kernel distances

- Selective inference to avoid train/test split? Meta-testing?
- When $\mathbb{P} \neq \mathbb{Q}$, can we tell how they're different?
  - Methods so far: some mostly for low-$d$
    - Some look at points with large critic function
- Does model $\mathbb{Q}$ match dataset $\mathbb{X}$ (Stein testing)?
- Maximize deep dependence measure for unsupervised representation learning, as in contrastive learning
- ...