Kernel distances between distributions for generative models

Dougal J. Sutherland
Gatsby Computational Neuroscience Unit, University College London

Michael Arbel
UCL

Mikołaj Bińkowski
Imperial

Arthur Gretton
UCL

$$\mathcal{D}\left(\quad,\quad\right)$$

ICIAM 2019
Distance Metrics and Mass Transfer Between High Dimensional Point Clouds

Integral probability metrics

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)]$$

$f : \mathcal{X} \to \mathbb{R}$ is a *critic function*

Integral probability metrics

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)]$$

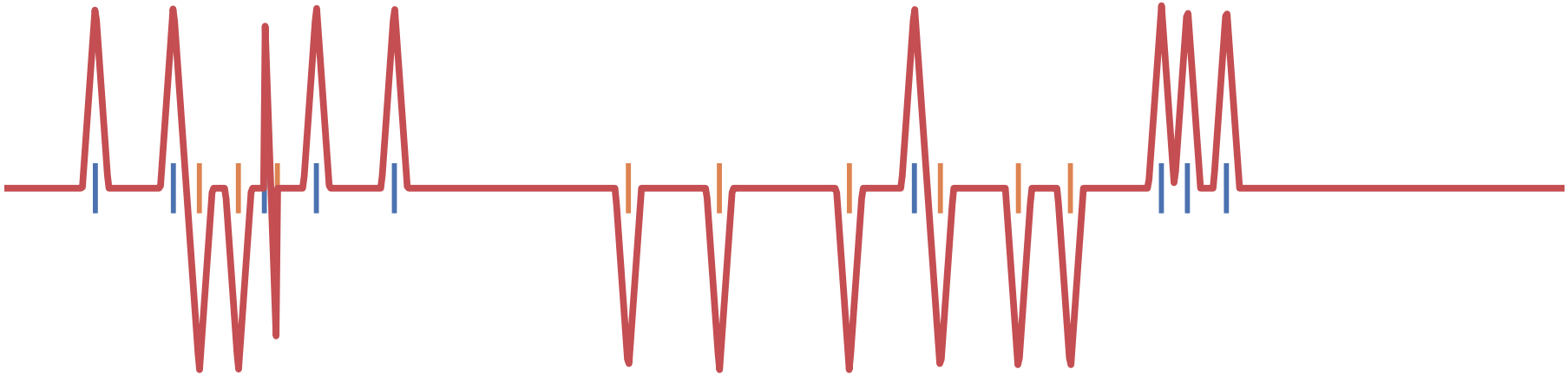
$f : \mathcal{X} \to \mathbb{R}$ is a *critic function*

Integral probability metrics

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)]$$

$f : \mathcal{X} \to \mathbb{R}$ is a *critic function*

Total variation: $\mathcal{F} = \{f : f \text{ continuous}, |f(x)| \leq 1\}$



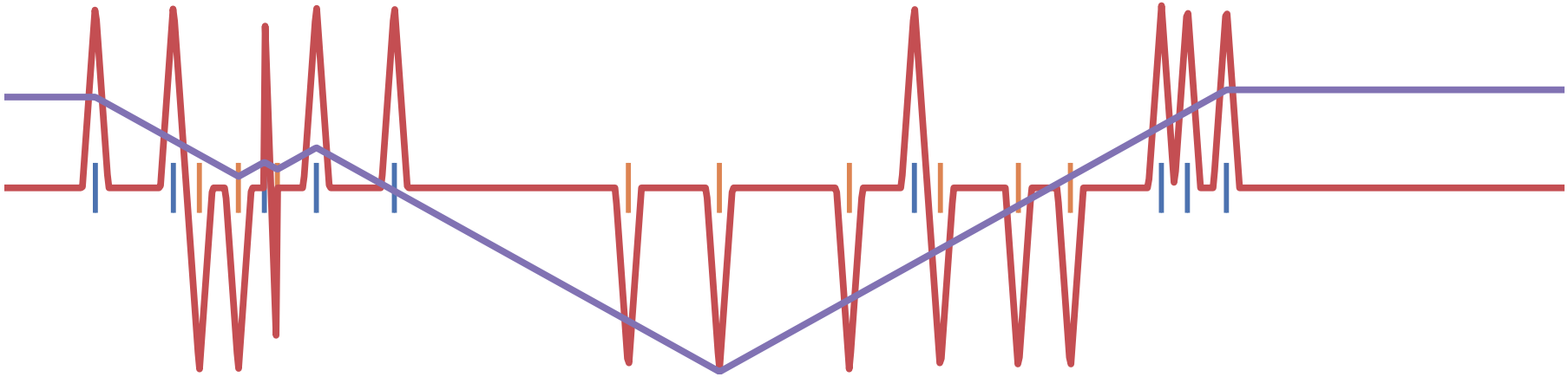
Integral probability metrics

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)]$$

$f : \mathcal{X} \to \mathbb{R}$ is a *critic function*

Total variation: $\mathcal{F} = \{f : f \text{ continuous}, |f(x)| \leq 1\}$

Wasserstein: $\mathcal{F} = \{f : \|f\|_{\text{Lip}} \leq 1\}$



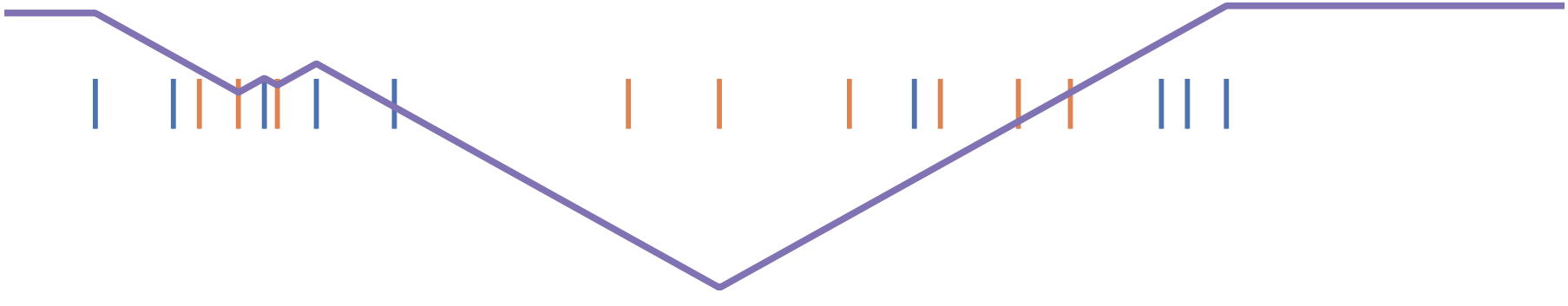
Integral probability metrics

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)]$$

$f : \mathcal{X} \to \mathbb{R}$ is a *critic function*

Total variation: $\mathcal{F} = \{f : f \text{ continuous}, |f(x)| \leq 1\}$

Wasserstein: $\mathcal{F} = \{f : \|f\|_{\text{Lip}} \leq 1\}$



Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a “similarity” function

Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a “similarity” function

Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

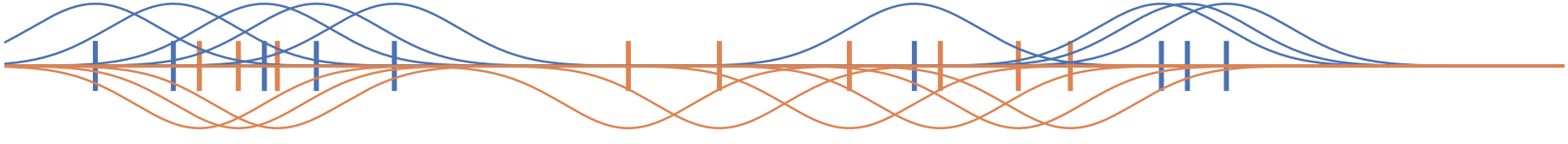
$$f^*(t) \propto \mathop{\mathbb{E}}_{X \sim \mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}} k(t, Y)$$

Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

$$f^*(t) \propto \mathop{\mathbb{E}}_{X \sim \mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}} k(t, Y)$$

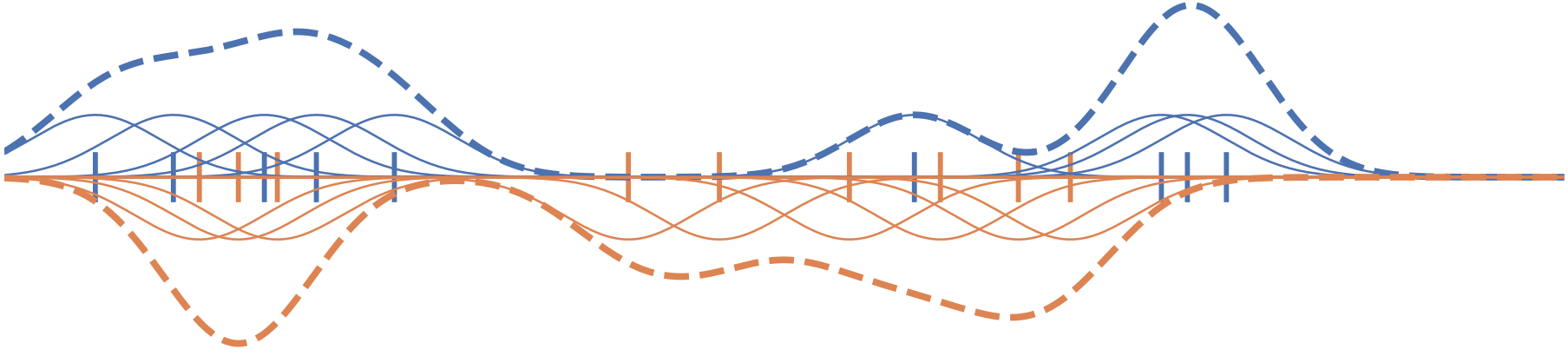


Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

$$f^*(t) \propto \mathop{\mathbb{E}}_{X \sim \mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}} k(t, Y)$$

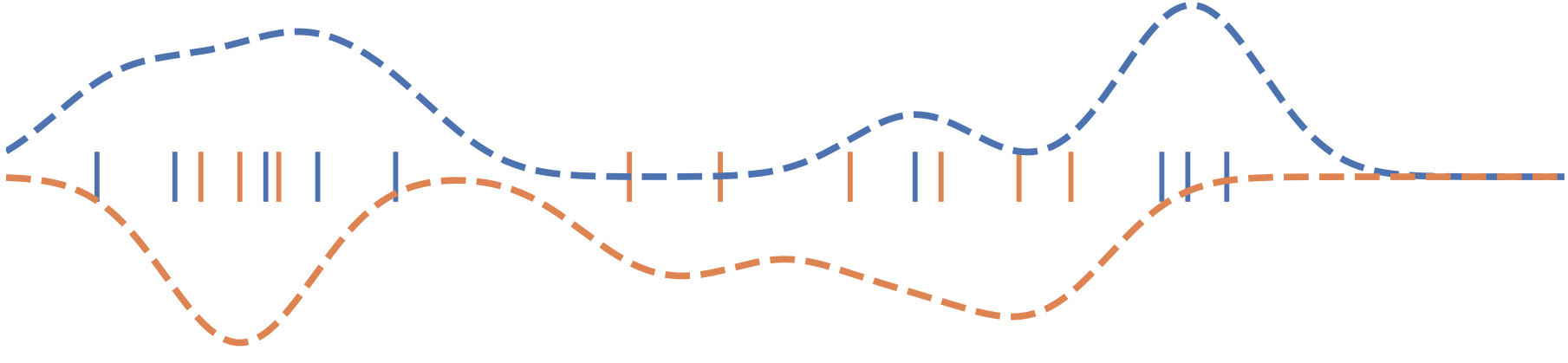


Maximum Mean Discrepancy [Gretton+ 2012]

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \le 1} \mathop{\mathbb{E}}_{X\sim\mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}[f(Y)]$$

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

$$f^*(t) \propto \mathop{\mathbb{E}}_{X\sim\mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} k(t, Y)$$

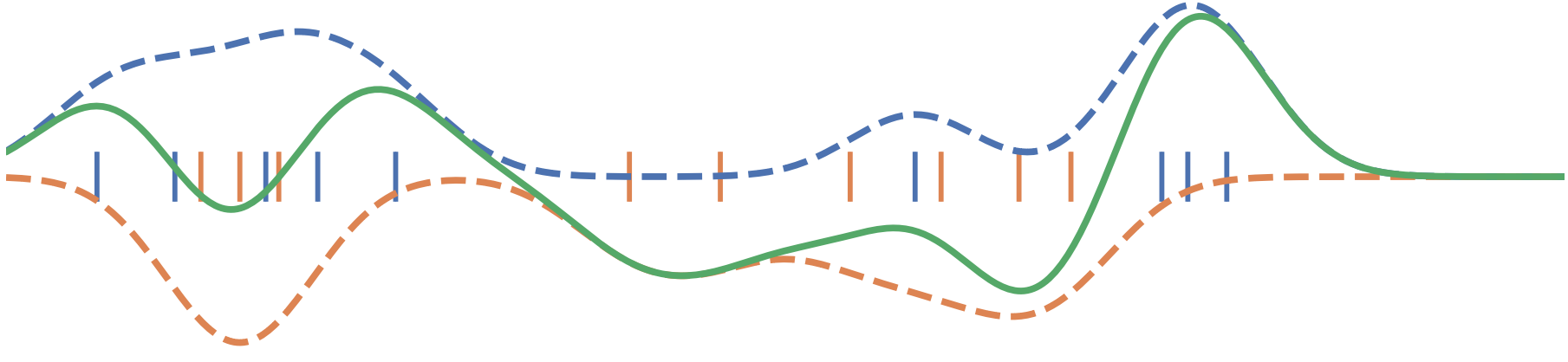


Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

$$f^*(t) \propto \mathop{\mathbb{E}}_{X \sim \mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}} k(t, Y)$$

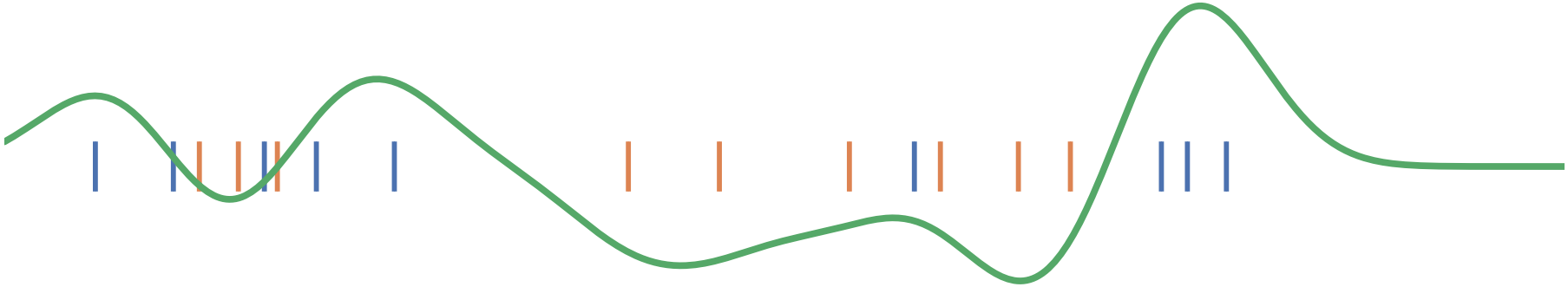


Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

$$f^*(t) \propto \mathop{\mathbb{E}}_{X \sim \mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}} k(t, Y)$$

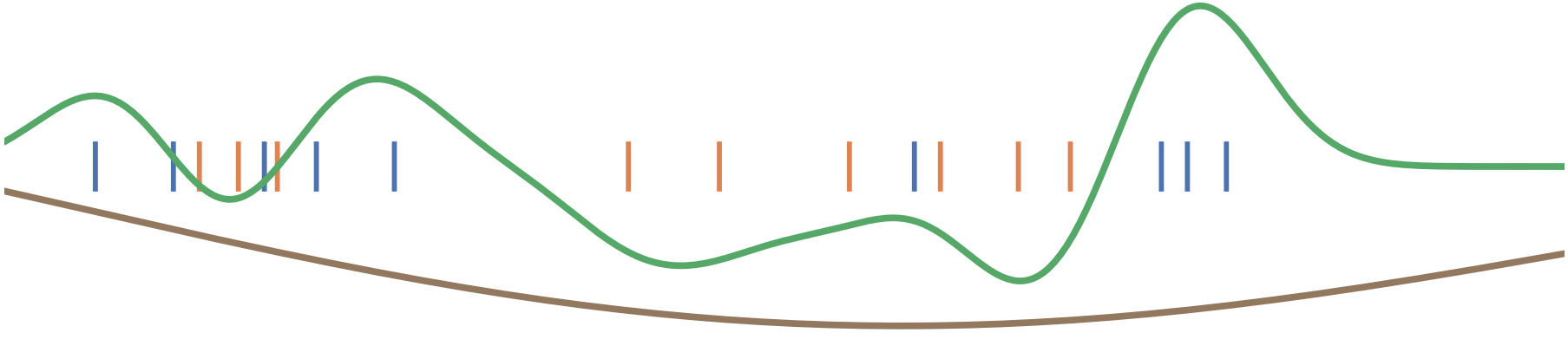


Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

$$f^*(t) \propto \mathop{\mathbb{E}}_{X \sim \mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}} k(t, Y)$$

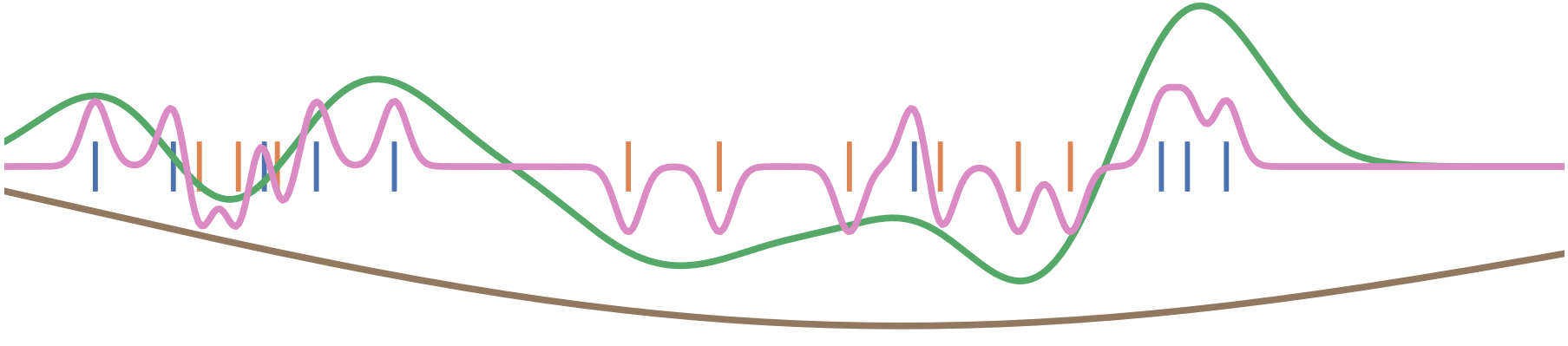


Maximum Mean Discrepancy [Gretton+ 2012]

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

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

$$f^*(t) \propto \mathop{\mathbb{E}}_{X \sim \mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}} k(t, Y)$$



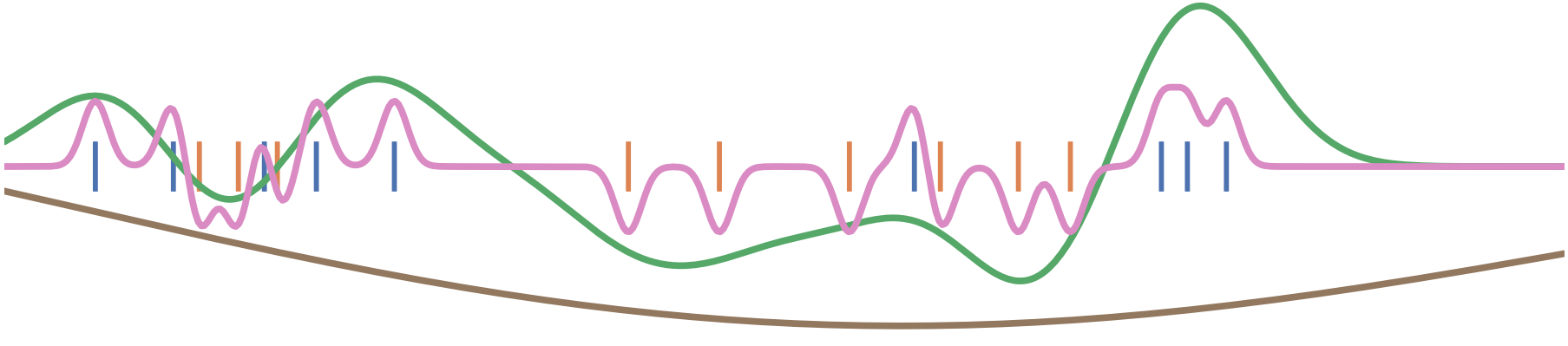
Maximum Mean Discrepancy [Gretton+ 2012]

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \le 1} \mathop{\mathbb{E}}_{X\sim\mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}[f(Y)]$$

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ – a "similarity" function

$$f^*(t) \propto \mathop{\mathbb{E}}_{X\sim\mathbb{P}} k(t, X) - \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} k(t, Y)$$

For many kernels, $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$



Derivation of MMD

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

Derivation of MMD

Reproducing property: if $f \in \mathcal{H}_k$, $f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}_k}$

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

Derivation of MMD

Reproducing property: if $f \in \mathcal{H}_k$, $f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}_k}$

$$\begin{aligned}\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) &= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)] \\ &= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\langle f, \varphi(X) \rangle_{\mathcal{H}_k}] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\langle f, \varphi(Y) \rangle_{\mathcal{H}_k}]\end{aligned}$$

Derivation of MMD

Reproducing property: if $f \in \mathcal{H}_k$, $f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}_k}$

$$\begin{aligned}
\mathbf{MMD}_k(\mathbb{P}, \mathbb{Q}) &= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\langle f, \varphi(X) \rangle_{\mathcal{H}_k}] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\langle f, \varphi(Y) \rangle_{\mathcal{H}_k}] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left\langle f, \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\varphi(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\varphi(Y)] \right\rangle_{\mathcal{H}_k}
\end{aligned}$$

Derivation of MMD

Reproducing property: if $f \in \mathcal{H}_k$, $f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}_k}$

$$\begin{aligned}
\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) &= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\langle f, \varphi(X) \rangle_{\mathcal{H}_k}] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\langle f, \varphi(Y) \rangle_{\mathcal{H}_k}] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left\langle f, \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\varphi(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\varphi(Y)] \right\rangle_{\mathcal{H}_k} \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left\langle f, \mu^k_{\mathbb{P}} - \mu^k_{\mathbb{Q}} \right\rangle_{\mathcal{H}_k}
\end{aligned}$$

Derivation of MMD

Reproducing property: if $f \in \mathcal{H}_k$, $f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}_k}$

$$\begin{aligned}
\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) &= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\langle f, \varphi(X) \rangle_{\mathcal{H}_k}] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\langle f, \varphi(Y) \rangle_{\mathcal{H}_k}] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left\langle f, \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\varphi(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\varphi(Y)] \right\rangle_{\mathcal{H}_k} \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left\langle f, \mu^k_{\mathbb{P}} - \mu^k_{\mathbb{Q}} \right\rangle_{\mathcal{H}_k} = \left\| \mu^k_{\mathbb{P}} - \mu^k_{\mathbb{Q}} \right\|_{\mathcal{H}_k}
\end{aligned}$$

Derivation of MMD

Reproducing property: if $f \in \mathcal{H}_k$, $f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}_k}$

$$\begin{aligned}
\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) &= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[f(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[f(Y)] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\langle f, \varphi(X) \rangle_{\mathcal{H}_k}] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\langle f, \varphi(Y) \rangle_{\mathcal{H}_k}] \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left\langle f, \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\varphi(X)] - \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}}[\varphi(Y)] \right\rangle_{\mathcal{H}_k} \\
&= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left\langle f, \mu^k_{\mathbb{P}} - \mu^k_{\mathbb{Q}} \right\rangle_{\mathcal{H}_k} = \left\| \mu^k_{\mathbb{P}} - \mu^k_{\mathbb{Q}} \right\|_{\mathcal{H}_k}
\end{aligned}$$

$$\langle \mu^k_{\mathbb{P}}, \mu^k_{\mathbb{Q}} \rangle_{\mathcal{H}_k} = \mathop{\mathbb{E}}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} \langle \varphi(X), \varphi(Y) \rangle_{\mathcal{H}_k} = \mathop{\mathbb{E}}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} k(X, Y)$$

MMD as feature matching

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

- $\varphi : \mathcal{X} \to \mathcal{H}_k$ is the *feature map* for $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$

MMD as feature matching

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

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

MMD as feature matching

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

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

Estimating MMD

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

Estimating MMD

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

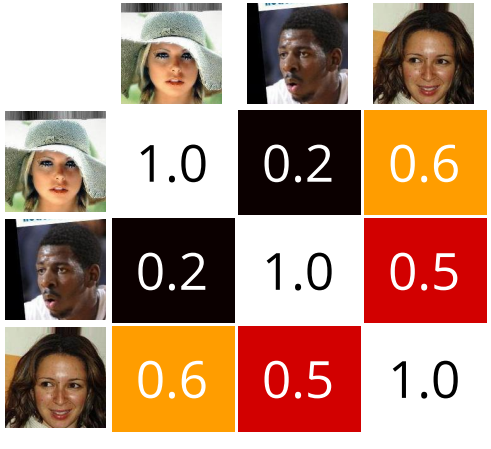
$$\widehat{\mathrm{MMD}}_k^2(X, Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2\,\mathrm{mean}(K_{XY})$$

Estimating MMD

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

$$\widehat{\mathrm{MMD}}_k^2(X, Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2\,\mathrm{mean}(K_{XY})$$

K_{XX}

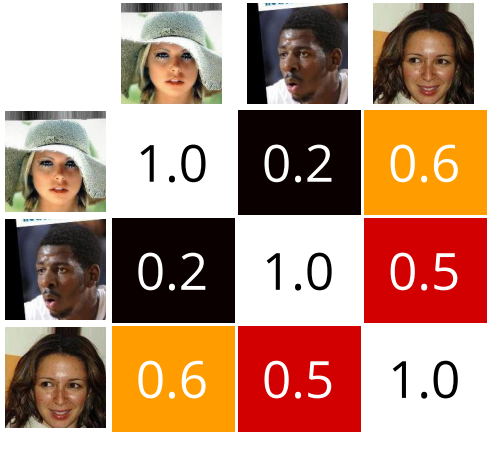


Estimating MMD

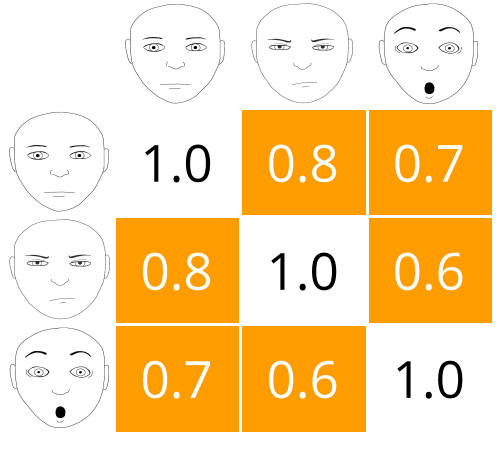
$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathop{\mathbb{E}}_{X,X' \sim \mathbb{P}}[k(X, X')] + \mathop{\mathbb{E}}_{Y,Y' \sim \mathbb{Q}}[k(Y, Y')] - 2 \mathop{\mathbb{E}}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}}[k(X, Y)]$$

$$\widehat{\mathrm{MMD}}_k^2(X, Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2\,\mathrm{mean}(K_{XY})$$

K_{XX}



K_{YY}

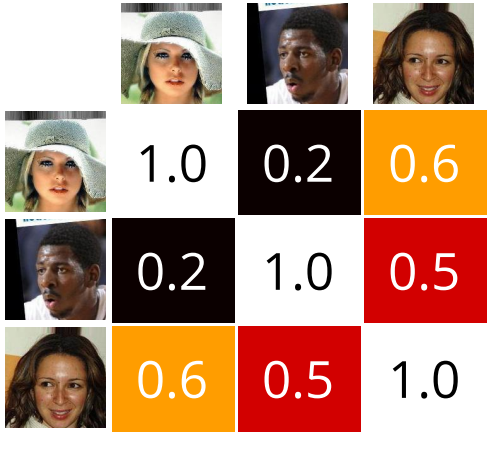


Estimating MMD

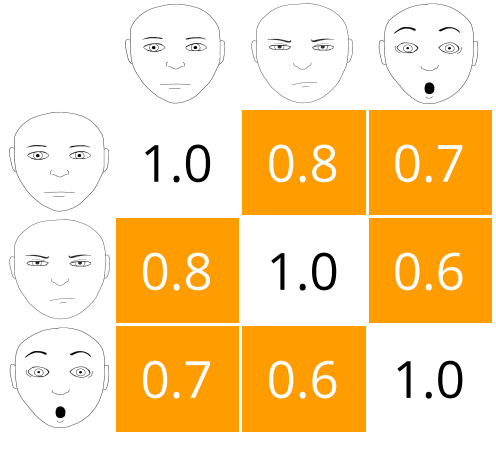
$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathop{\mathbb{E}}_{X,X' \sim \mathbb{P}}[k(X, X')] + \mathop{\mathbb{E}}_{Y,Y' \sim \mathbb{Q}}[k(Y, Y')] - 2 \mathop{\mathbb{E}}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}}[k(X, Y)]$$

$$\widehat{\mathrm{MMD}}_k^2(X, Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2\,\mathrm{mean}(K_{XY})$$

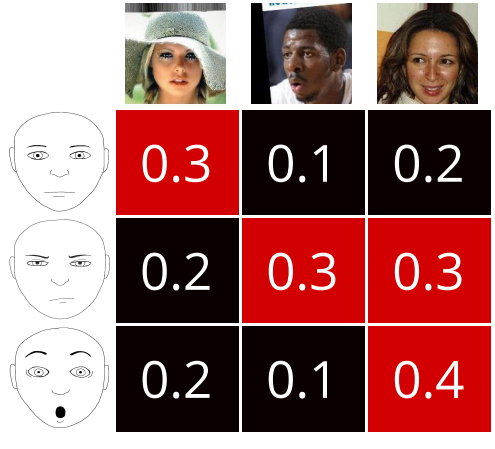
K_{XX}



K_{YY}



K_{XY}



Implicit generative models

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

$X \sim \mathbb{P}$

$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Why implicit generative models?

$X \sim \mathbb{P}$

vs

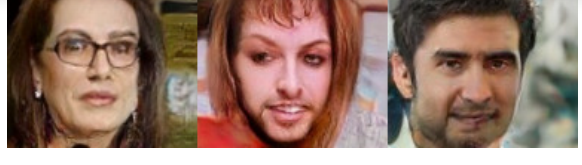
$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)

$X \sim \mathbb{P}$



$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)

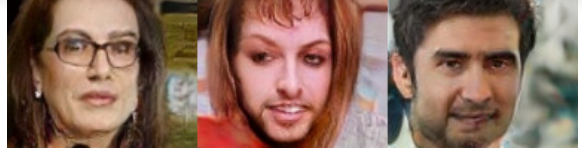
$X \sim$ [illegible] $\sim Q_\theta$

Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)

$X \sim \mathbb{P}$



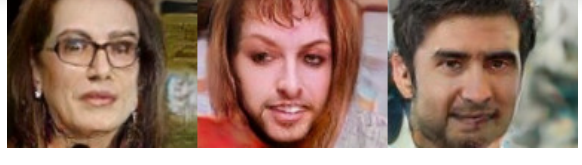
$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligraphic font style (`zi2zi`)

$X \sim \mathbb{P}$

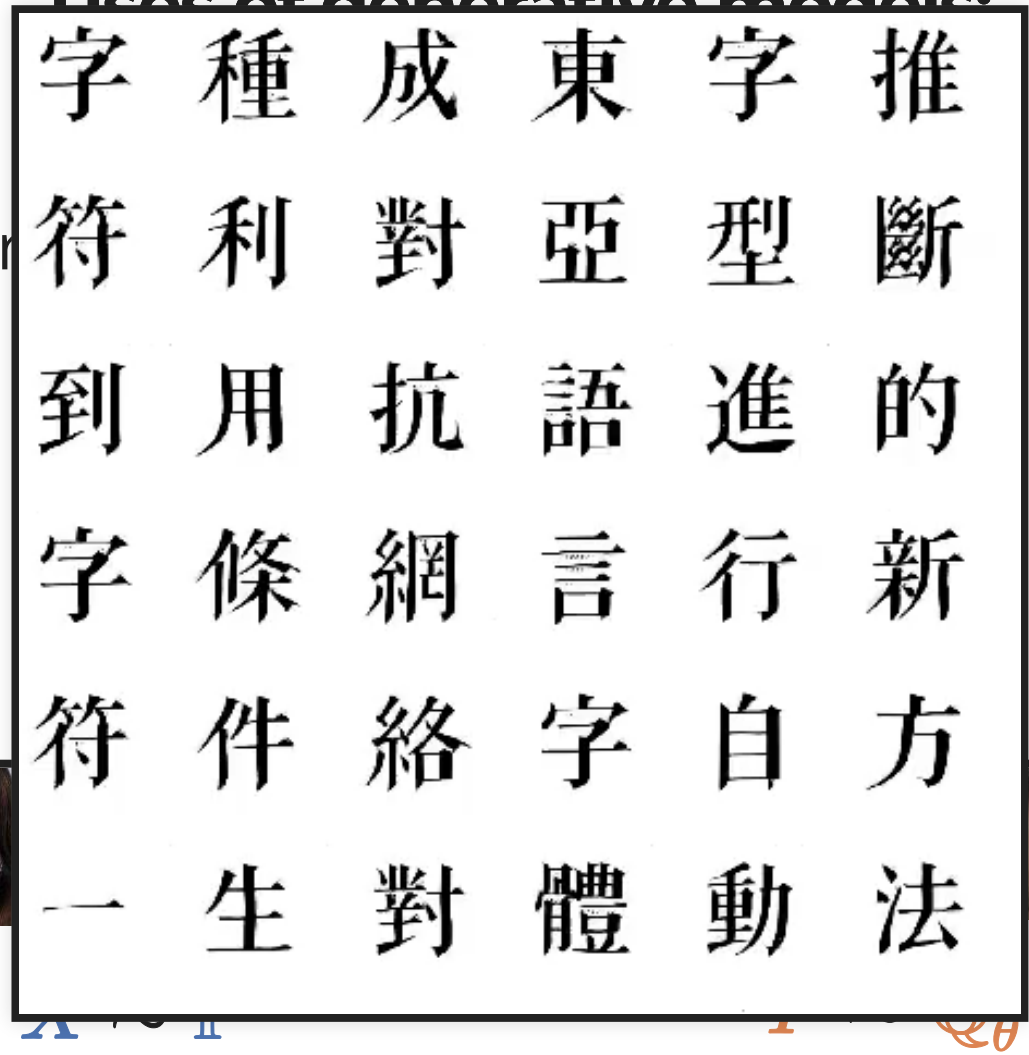


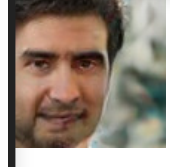
$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Uses of generative models:

- Automated
- Learn calligr



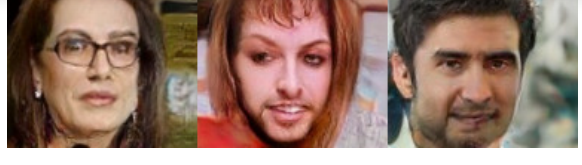


Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligraphic font style (`zi2zi`)

$X \sim \mathbb{P}$



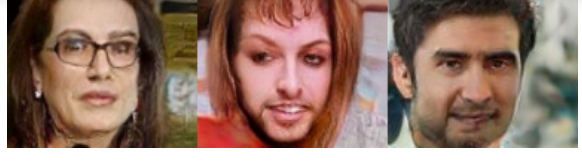
$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligraphic font style (`zi2zi`)
- Image translation (`pix2pix`, Wolf+ 17, Everybody Dance Now)

$X \sim \mathbb{P}$

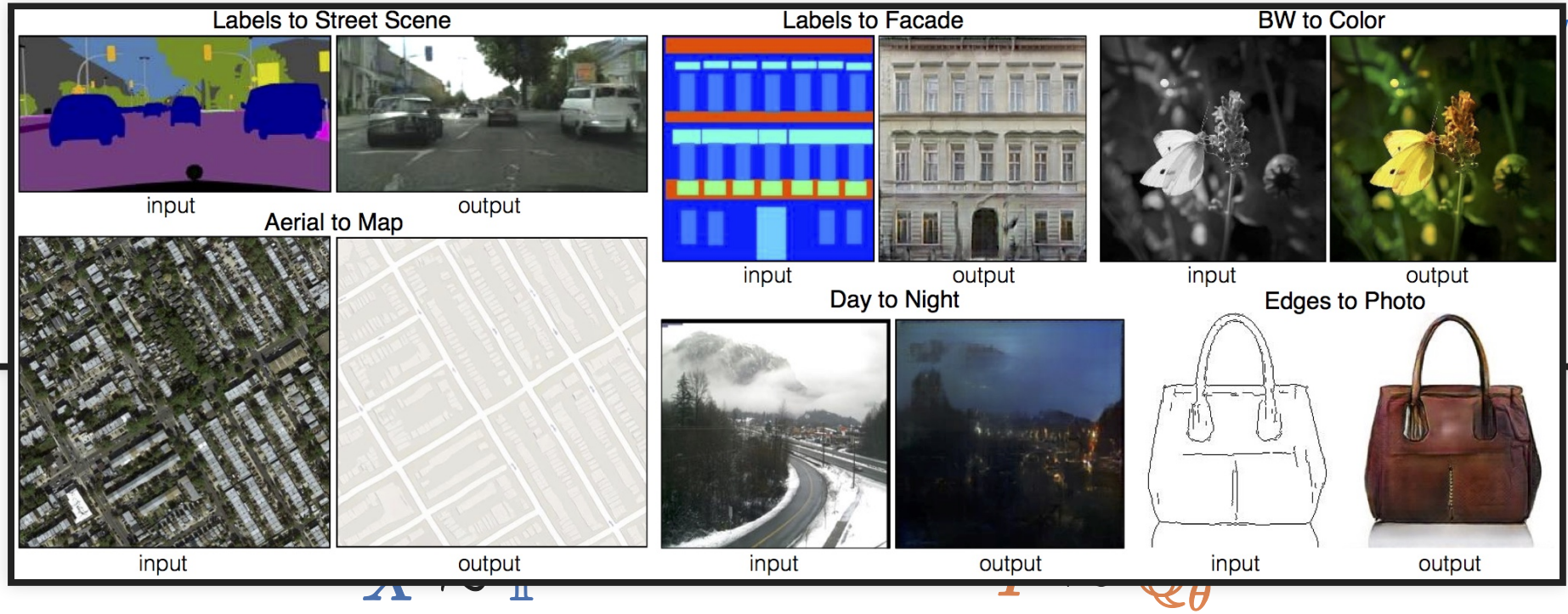


$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Uses of generative models:

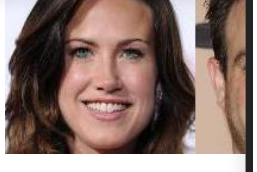
- Automated animation (anime characters)
- Learn calligraphic font style (`zi2zi`)

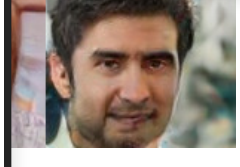


Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligrap
- Image translat body Dance Now)



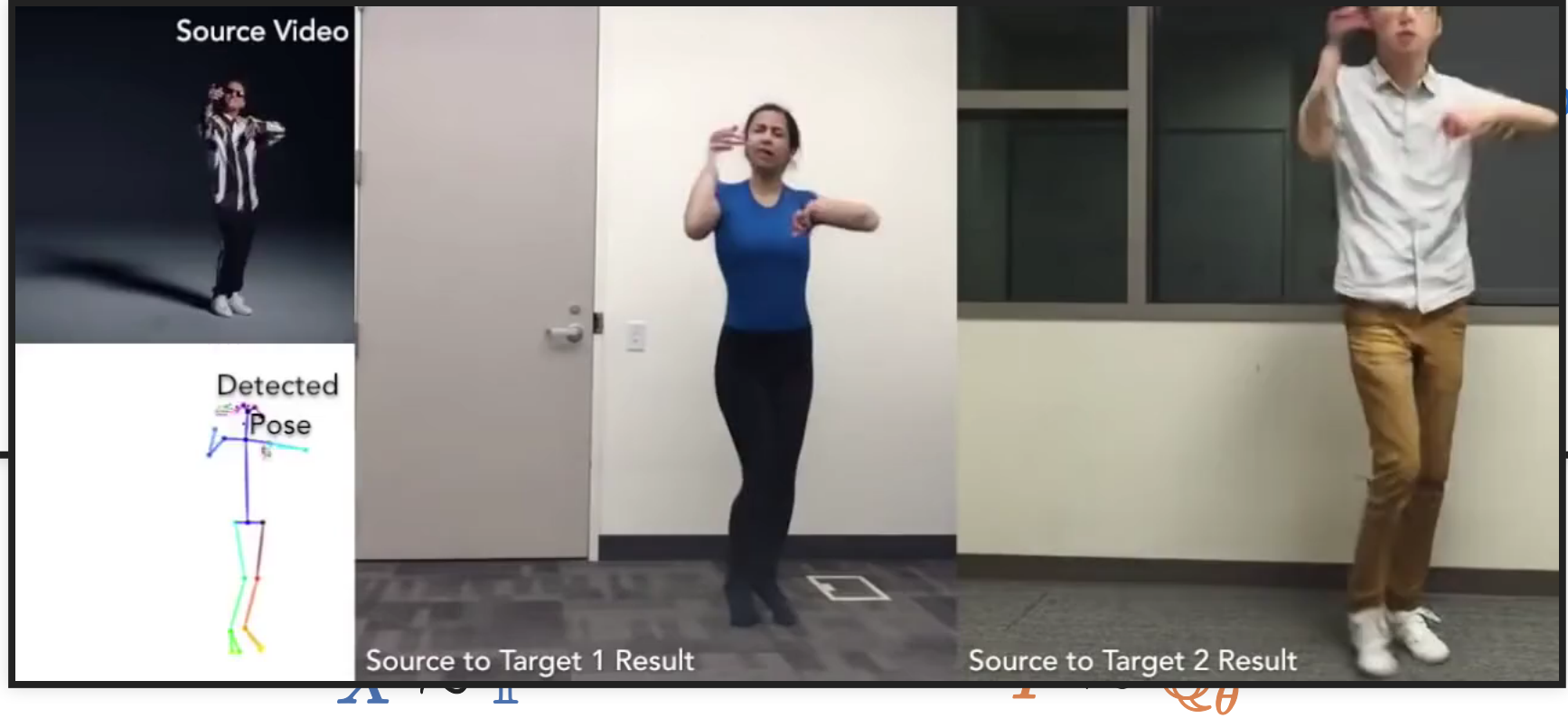


X Q_θ

Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligraphic font style (zi2zi)
- ow)

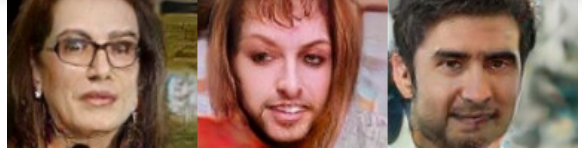


Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligraphic font style (`zi2zi`)
- Image translation (`pix2pix`, Wolf+ 17, Everybody Dance Now)

$X \sim \mathbb{P}$



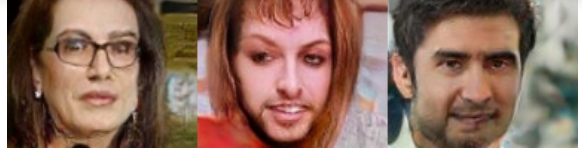
$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligraphic font style (`zi2zi`)
- Image translation (`pix2pix`, Wolf+ 17, Everybody Dance Now)
- Help musicians improvise, "natural" image editing, plan robot actions, anonymize datasets, ...

$X \sim \mathbb{P}$



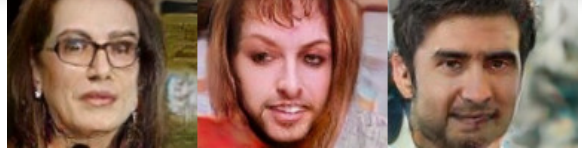
$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligraphic font style (`zi2zi`)
- Image translation (`pix2pix`, Wolf+ 17, Everybody Dance Now)
- Help musicians improvise, "natural" image editing, plan robot actions, anonymize datasets, …
- Make $432,500 selling a generated image

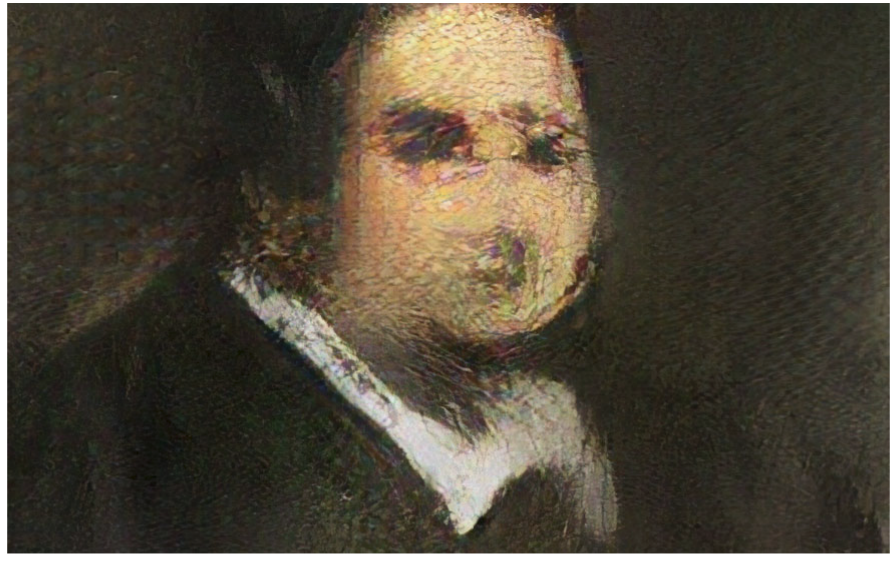
$X \sim \mathbb{P}$



$Y \sim \mathbb{Q}_\theta$

Why implicit generative models?

- Automated
- Learn callig
- Image trans ly Dance Now)
- Help musici ting, plan robot
- Make $432,



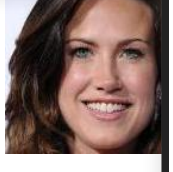
Is artificial intelligence set to become art's next medium?

AI artwork sells for $432,500 — nearly 45 times its high estimate — as Christie's becomes the first auction house to offer a work of art created by an algorithm

The portrait in its gilt frame depicts a portly gentleman, possibly French and — to judge by his dark frockcoat and plain white collar — a man of the church. The work appears unfinished: the facial features are somewhat indistinct and there are blank areas of canvas. Oddly, the whole composition is displaced slightly to the north-west. A label on the wall states that the sitter is a man named Edmond Belamy, but the giveaway clue as to the origins of the work is the artist's signature at the bottom right. In cursive Gallic script it reads:

$$\min_G \max_D \mathbb{E}_x[\log(D(x))] + \mathbb{E}_z[\log(1 - D(G(z)))]$$

Image © Obvious

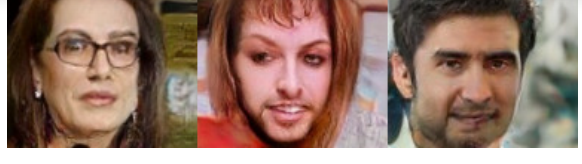


Why implicit generative models?

Uses of generative models:

- Automated animation (anime characters)
- Learn calligraphic font style (`zi2zi`)
- Image translation (`pix2pix`, Wolf+ 17, Everybody Dance Now)
- Help musicians improvise, "natural" image editing, plan robot actions, anonymize datasets, …
- Make $432,500 selling a generated image

$X \sim \mathbb{P}$

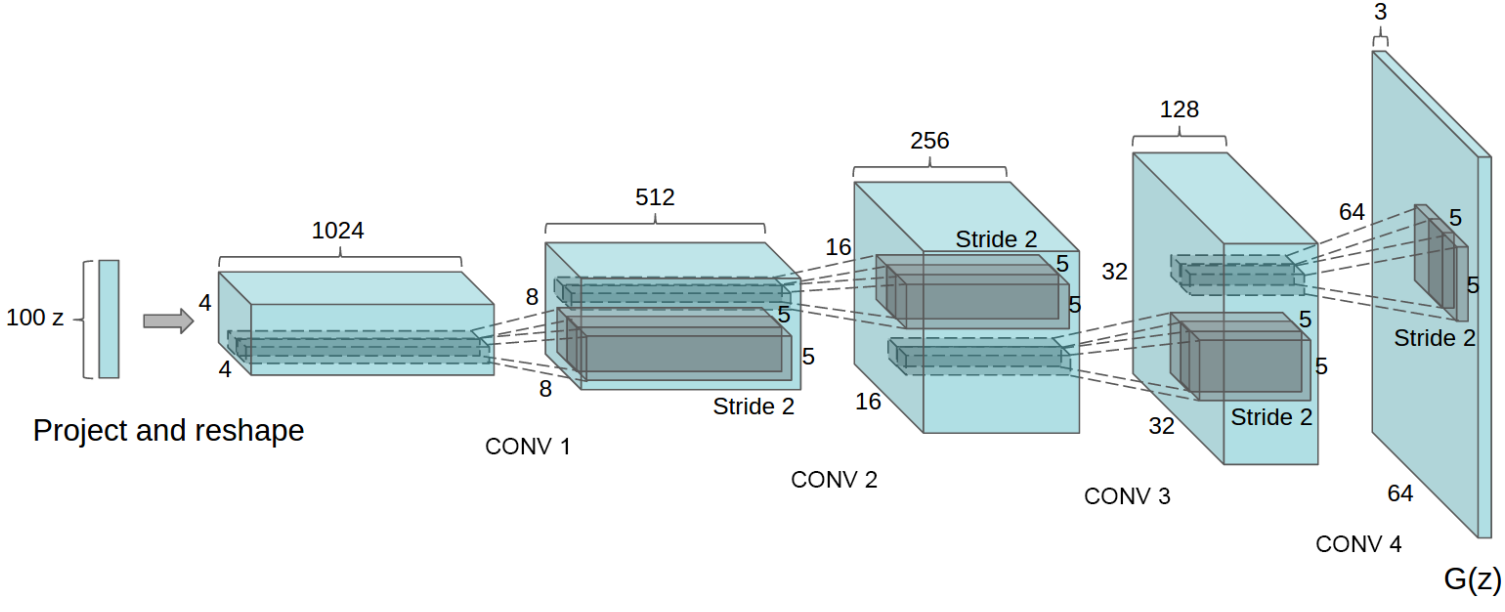


$Y \sim \mathbb{Q}_\theta$

Model: generator network

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

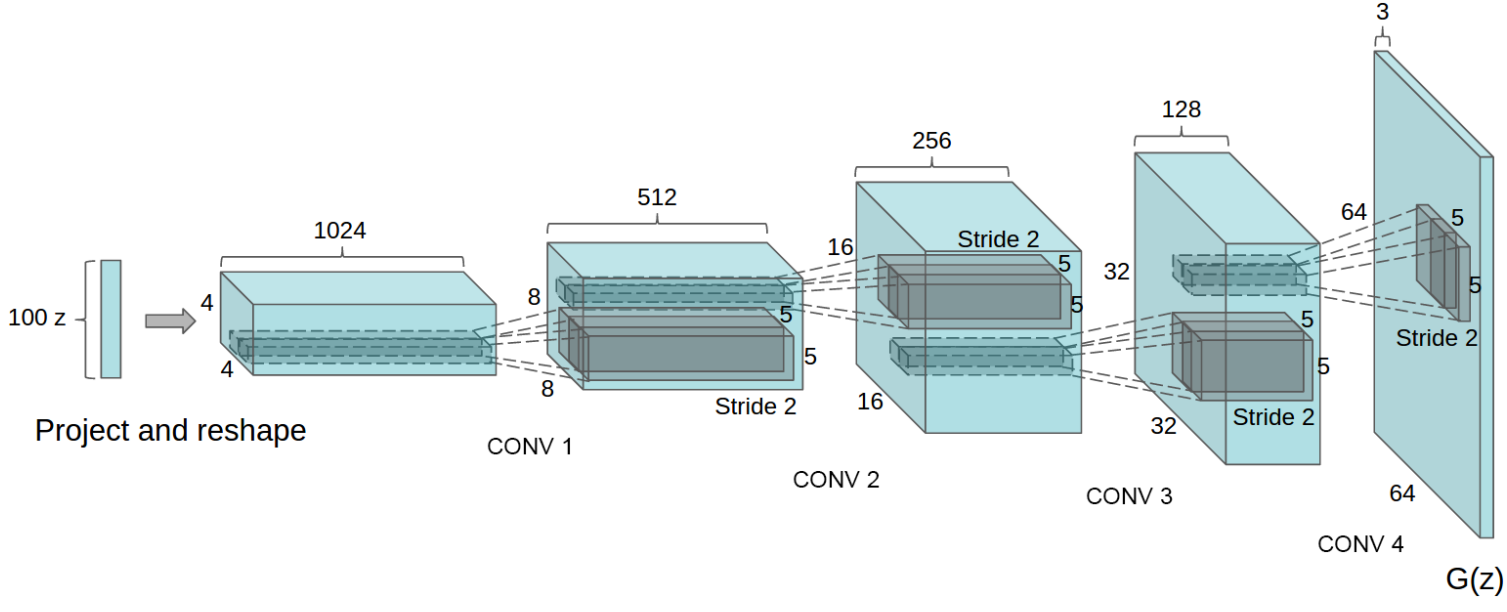
Maps through a network: $G_\theta(Z) \sim \mathbb{Q}_\theta$



Model: generator network

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

Maps through a network: $G_\theta(Z) \sim \mathbb{Q}_\theta$

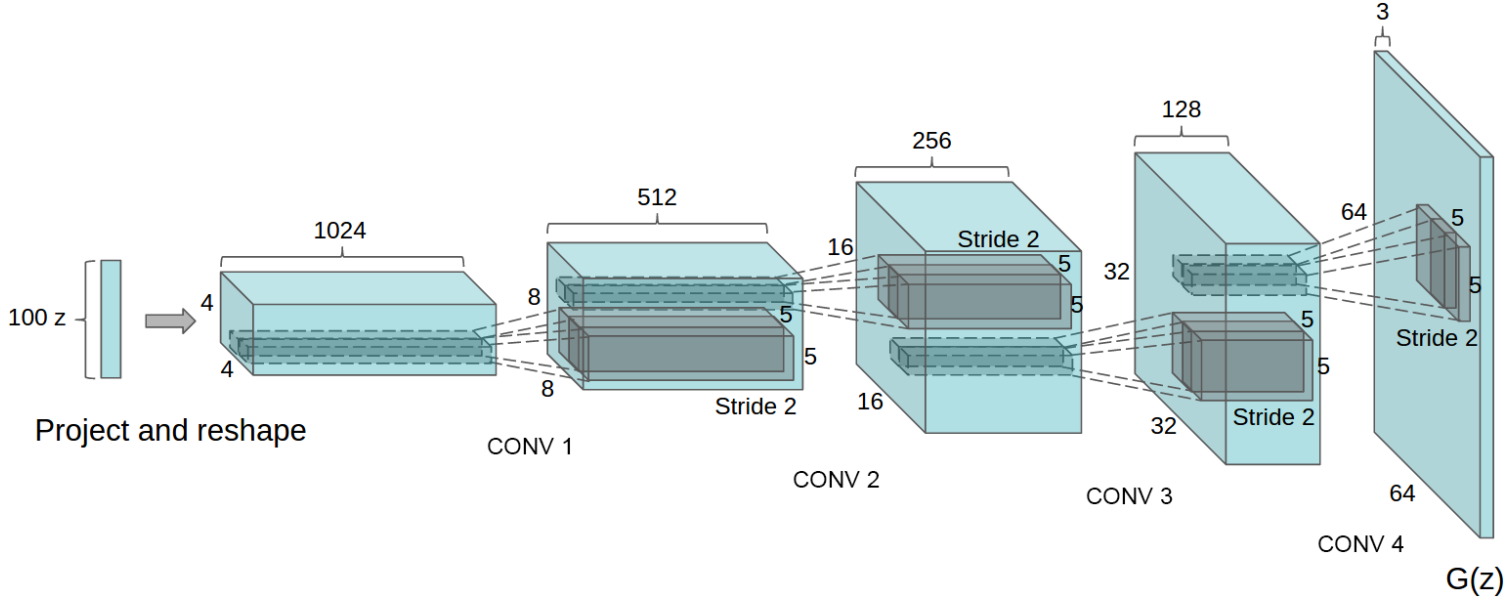


Choose θ to minimize $\mathcal{D}(\mathbb{P}_{\text{data}}, \mathbb{Q}_\theta)$

Model: generator network

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

Maps through a network: $G_\theta(Z) \sim \mathbb{Q}_\theta$



Choose θ to minimize $\mathcal{D}(\mathbb{P}_{\mathbf{data}}, \mathbb{Q}_\theta)$

Very flexible generative model

Traditional choices for $\mathcal{D}$

$$\underbrace{\operatorname{argmin}_\theta \mathrm{KL}(\mathbb{P}_{\text{data}} \| \mathbb{Q}_\theta)}_{\text{maximum likelihood}} \qquad \underbrace{\operatorname{argmin}_\theta \mathrm{JS}(\mathbb{P}_{\text{data}}, \mathbb{Q}_\theta)}_{\text{original GANs (ish)}}$$

Traditional choices for $\mathcal{D}$

$$\underbrace{\operatorname{argmin}_\theta \mathrm{KL}(\mathbb{P}_{\text{data}} \| \mathbb{Q}_\theta)}_{\text{maximum likelihood}} \qquad \underbrace{\operatorname{argmin}_\theta \mathrm{JS}(\mathbb{P}_{\text{data}}, \mathbb{Q}_\theta)}_{\text{original GANs (ish)}}$$

- Problem: $\mathbb{P}_{\text{data}}$ and $\mathbb{Q}_\theta$ don't have full-dimensional support

Traditional choices for $\mathcal{D}$

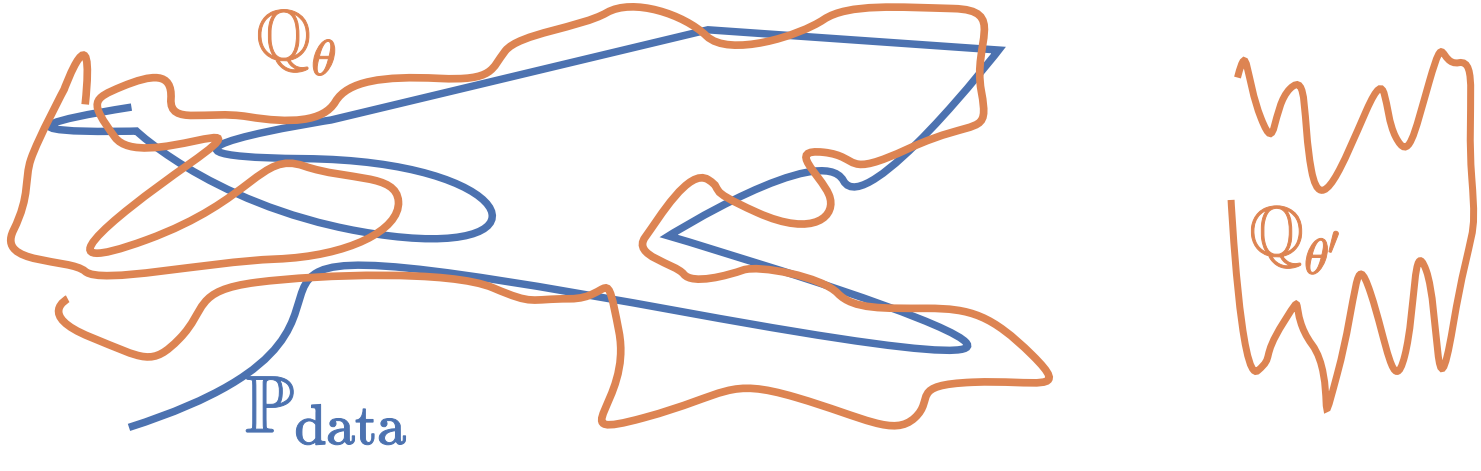
$$\underbrace{\operatorname{argmin}_\theta \mathrm{KL}(\mathbb{P}_{\text{data}} \| \mathbb{Q}_\theta)}_{\text{maximum likelihood}} \qquad \underbrace{\operatorname{argmin}_\theta \mathrm{JS}(\mathbb{P}_{\text{data}}, \mathbb{Q}_\theta)}_{\text{original GANs (ish)}}$$

- Problem: $\mathbb{P}_{\text{data}}$ and $\mathbb{Q}_\theta$ don't have full-dimensional support
- Supports almost surely don't align [Arjovsky/Bottou ICLR-17]

Traditional choices for $\mathcal{D}$

$$\underbrace{\operatorname{argmin}_\theta \mathrm{KL}(\mathbb{P}_{\text{data}} \| \mathbb{Q}_\theta)}_{\text{maximum likelihood}} \qquad \underbrace{\operatorname{argmin}_\theta \mathrm{JS}(\mathbb{P}_{\text{data}}, \mathbb{Q}_\theta)}_{\text{original GANs (ish)}}$$

- Problem: $\mathbb{P}_{\text{data}}$ and $\mathbb{Q}_\theta$ don't have full-dimensional support
- Supports almost surely don't align [Arjovsky/Bottou ICLR-17]
- No “partial credit”: loss is flat at maximum, never improves



MMD as loss [Li+ ICML-15, Dziugaite+ UAI-15]

- Does give “partial credit” to nearby points
- Can directly minimize $\widehat{\mathrm{MMD}}^2(X, G_\theta(Z))$ with SGD

MMD as loss [Li+ ICML-15, Dziugaite+ UAI-15]

- Does give "partial credit" to nearby points
- Can directly minimize $\widehat{\mathrm{MMD}}^2(X, G_\theta(Z))$ with SGD

MNIST, mix of Gaussian kernels

$\mathbb{P}_{\text{data}}$

MMD as loss [Li+ ICML-15, Dziugaite+ UAI-15]

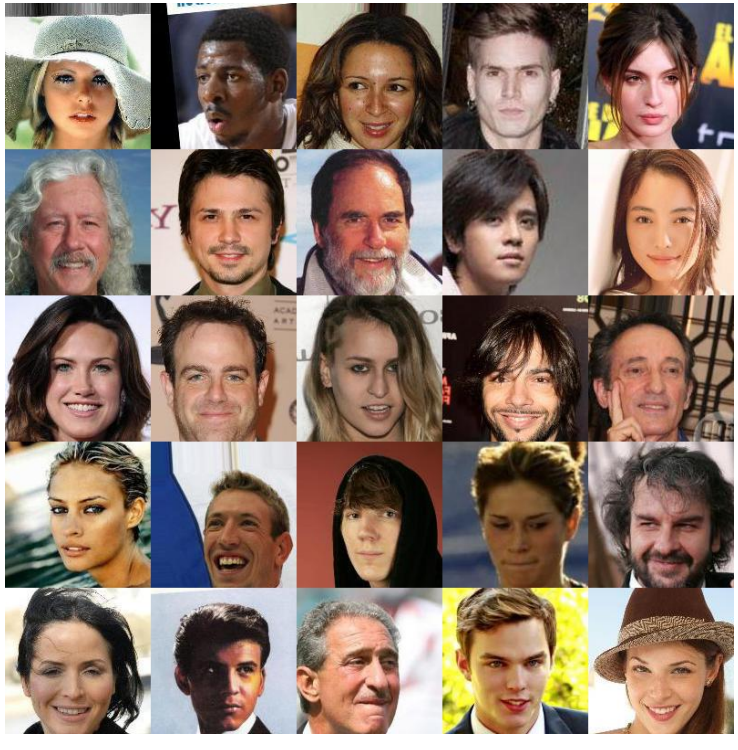
- Does give "partial credit" to nearby points
- Can directly minimize $\widehat{\mathrm{MMD}}^2(X, G_\theta(Z))$ with SGD

MNIST, mix of Gaussian kernels

$\mathbb{P}_{\text{data}}$

$\mathbb{Q}_\theta$

Celeb-A, mix of rational quadratic + linear kernels

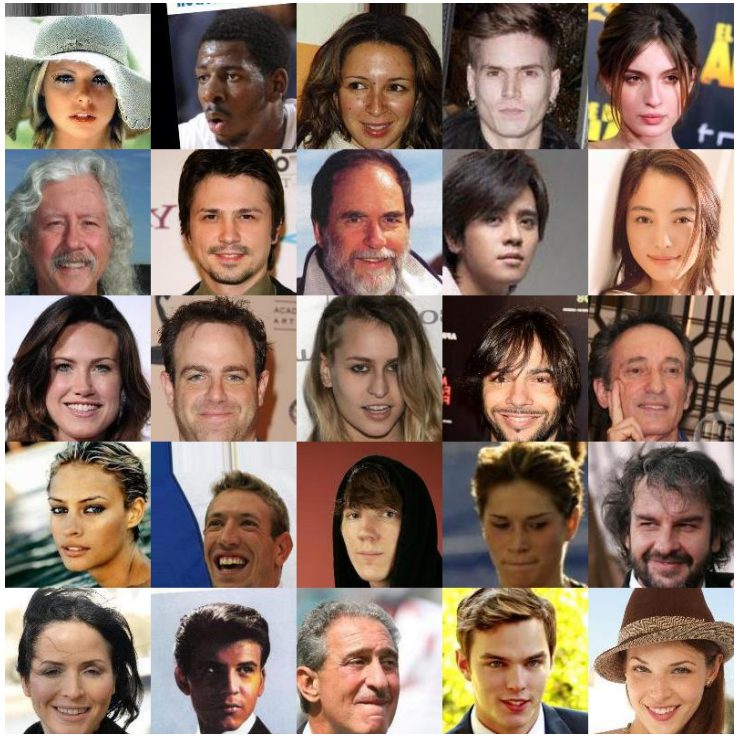


$\mathbb{P}_{\text{data}}$

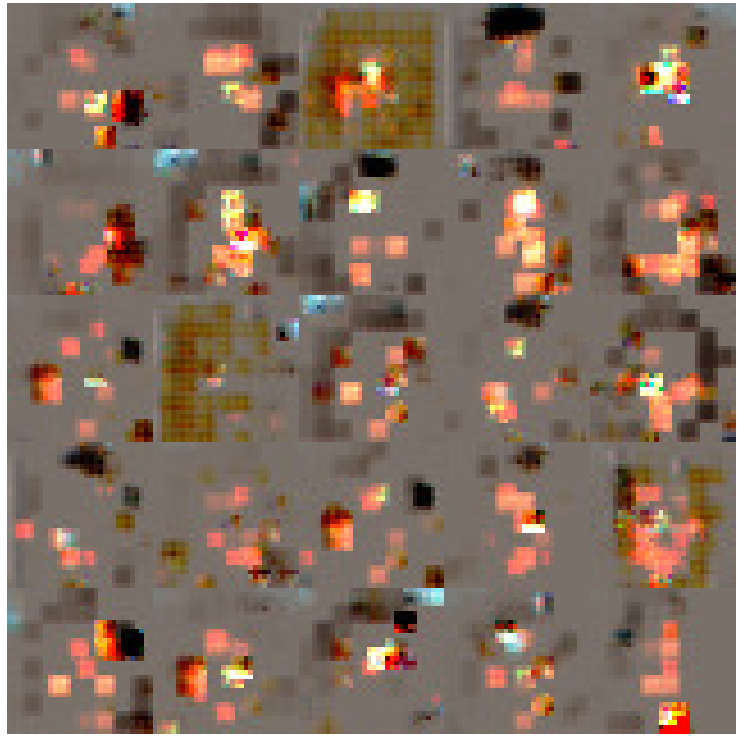
$\mathbb{P}_{\text{data}}$

Q_θ

Celeb-A, mix of rational quadratic + linear kernels



$\mathbb{P}_{\text{data}}$

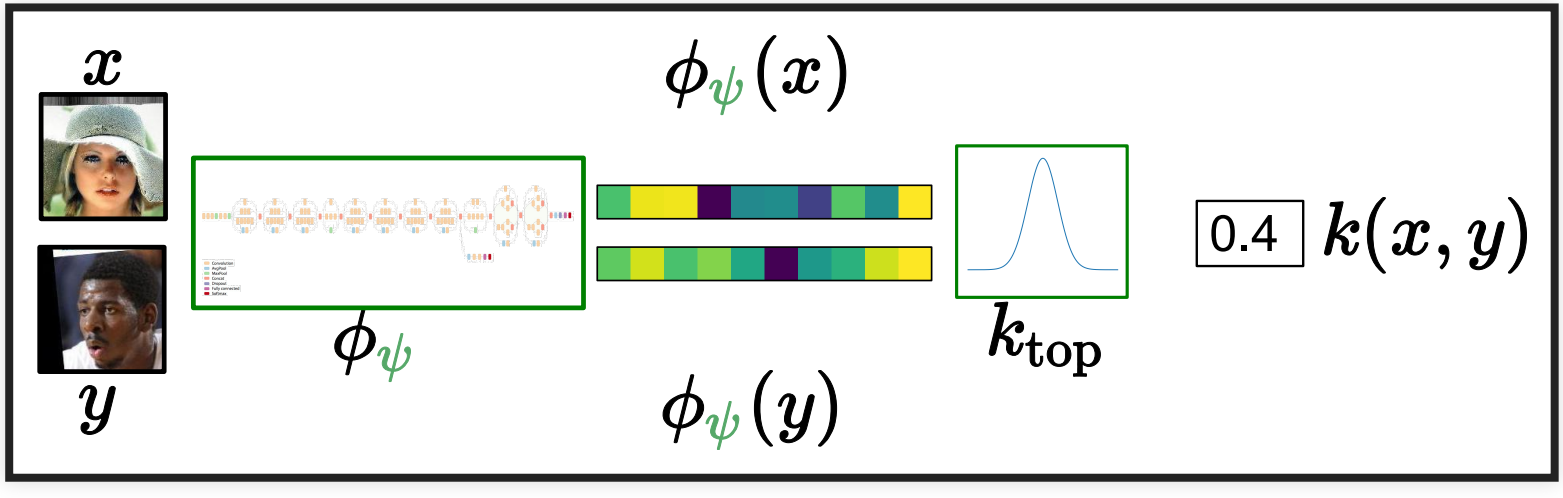


$\mathbb{Q}_\theta$

Deep kernels

$$k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi_\psi(y))$$

$$\phi_\psi : \mathcal{X} \to \mathbb{R}^D \qquad k_\psi : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$$



- k_{top} usually Gaussian, linear, ...

MMD loss with a deep kernel

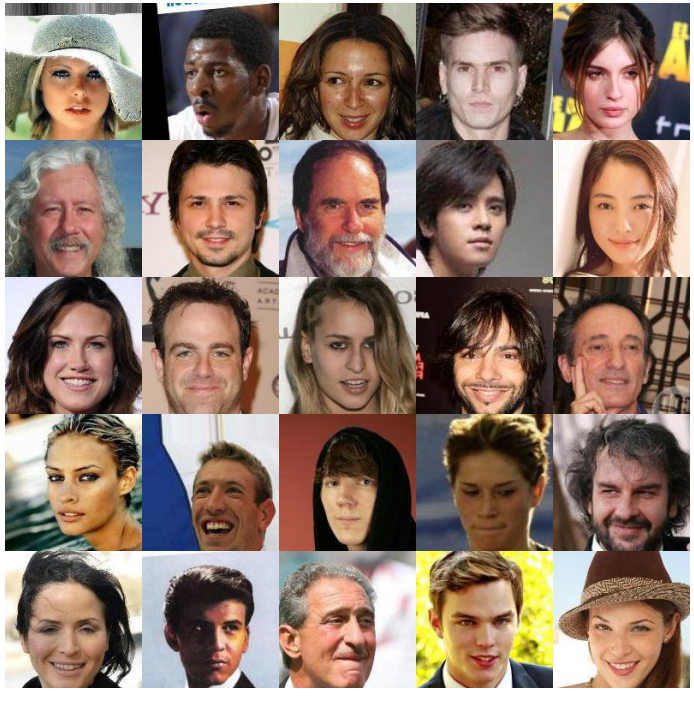
$$k(x, y) = k_{\text{top}}(\phi(x), \phi(y))$$

- $\phi : \mathcal{X} \to \mathbb{R}^{2048}$ from pretrained Inception net
- k_{top} simple: exponentiated quadratic or polynomial

MMD loss with a deep kernel

$$k(x, y) = k_{\text{top}}\left(\phi(x), \phi(y)\right)$$

- $\phi : \mathcal{X} \to \mathbb{R}^{2048}$ from pretrained Inception net
- k_{top} simple: exponentiated quadratic or polynomial

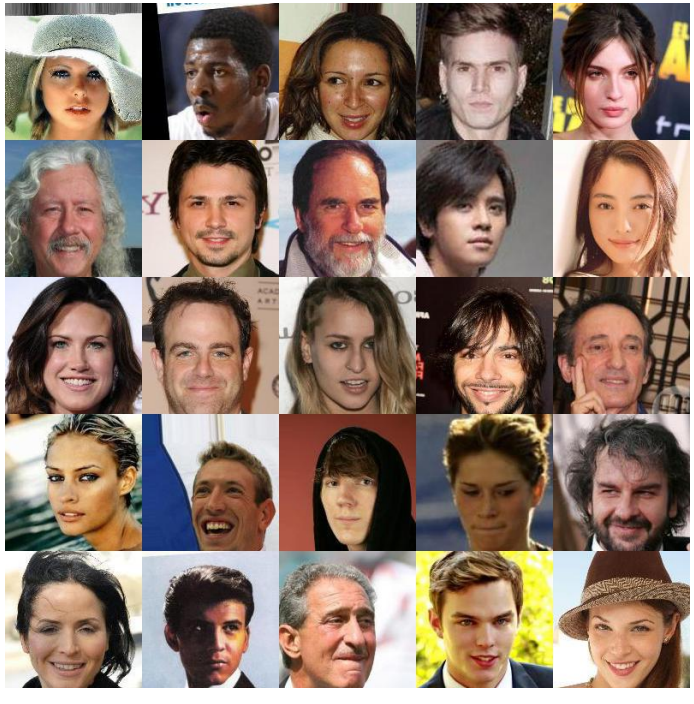


$\mathbb{P}_{\text{data}}$

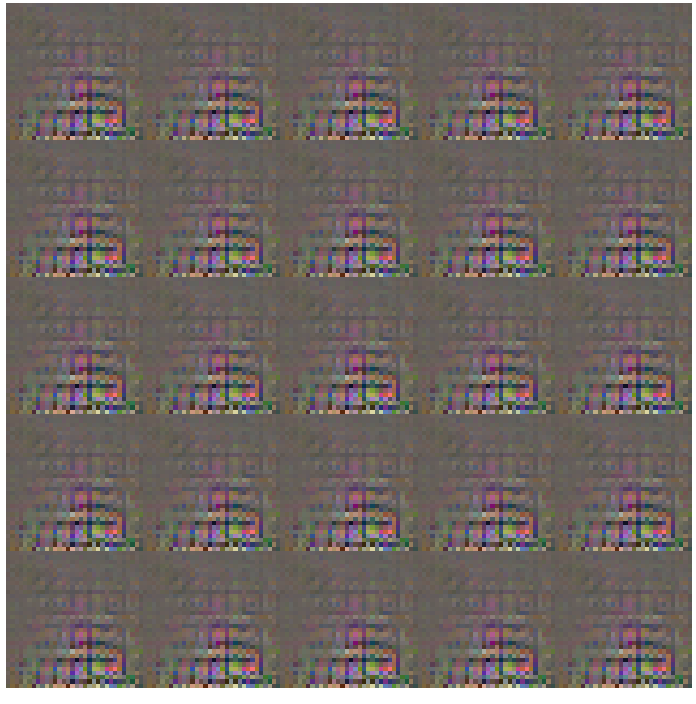
MMD loss with a deep kernel

$$k(x, y) = k_{\text{top}}\left(\phi(x), \phi(y)\right)$$

- $\phi : \mathcal{X} \to \mathbb{R}^{2048}$ from pretrained Inception net
- k_{top} simple: exponentiated quadratic or polynomial



$\mathbb{P}_{\text{data}}$



$\mathbb{Q}_\theta$

MMD loss with a deep kernel

$$k(x, y) = k_{\text{top}}\left(\phi(x), \phi(y)\right)$$

- $\phi : \mathcal{X} \to \mathbb{R}^{2048}$
- k_{top} is a characteristic kernel, e.g. RBF or inverse multiquadratic

We just got adversarial examples!

adversarial perturbation

88% tabby cat

99% guacamole

[anishathalye/obfuscated-gradients]

$\mathbb{P}_{\text{data}}$

$\mathbb{Q}_\theta$

Optimized MMD: MMD GANs [Li+ NeurIPS-17]

- Don't just use one kernel, use a *class* parameterized by ψ:

$$k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi_\psi(y))$$

Optimized MMD: MMD GANs [Li+ NeurIPS-17]

- Don't just use one kernel, use a *class* parameterized by ψ:

$$k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi_\psi(y))$$

- New distance based on *all* these kernels:

$$\mathcal{D}^{\Psi}_{\text{MMD}}(\mathbb{P}, \mathbb{Q}) = \sup_{\psi \in \Psi} \text{MMD}_\psi(\mathbb{P}, \mathbb{Q})$$

Optimized MMD: MMD GANs [Li+ NeurIPS-17]

- Don't just use one kernel, use a *class* parameterized by ψ:

$$k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi_\psi(y))$$

- New distance based on *all* these kernels:

$$\mathcal{D}^{\Psi}_{\text{MMD}}(\mathbb{P}, \mathbb{Q}) = \sup_{\psi \in \Psi} \text{MMD}_\psi(\mathbb{P}, \mathbb{Q})$$

- Minimax optimization problem

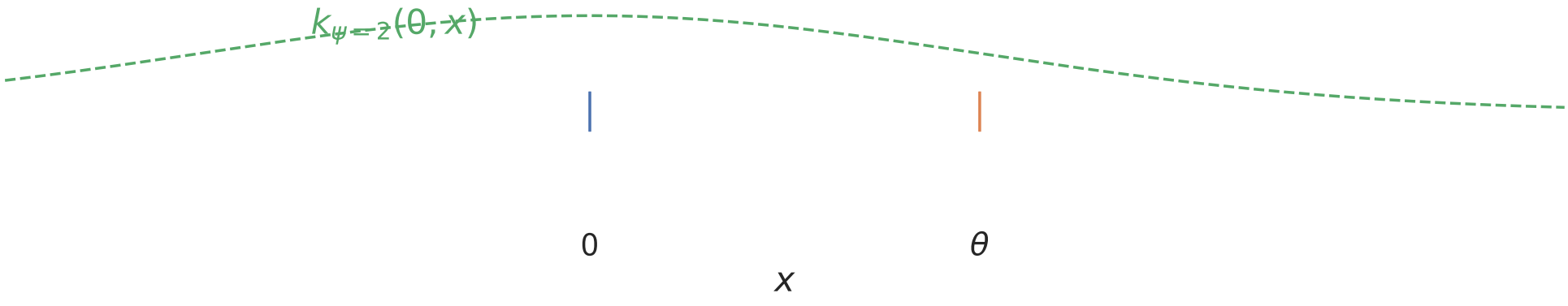
$$\inf_\theta \sup_\psi \text{MMD}_\psi(\mathbb{P}_{\text{data}}, \mathbb{Q}_\theta)$$

Non-smoothness of Optimized MMD

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

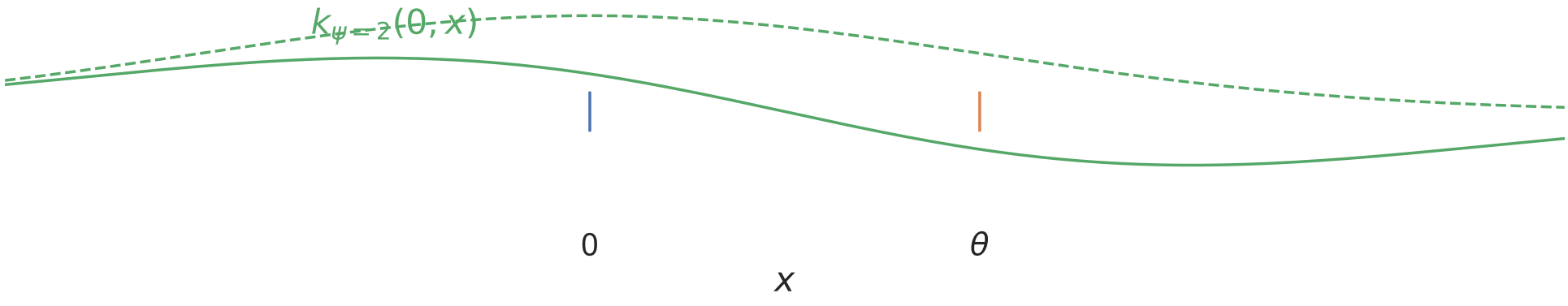
Non-smoothness of Optimized MMD

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



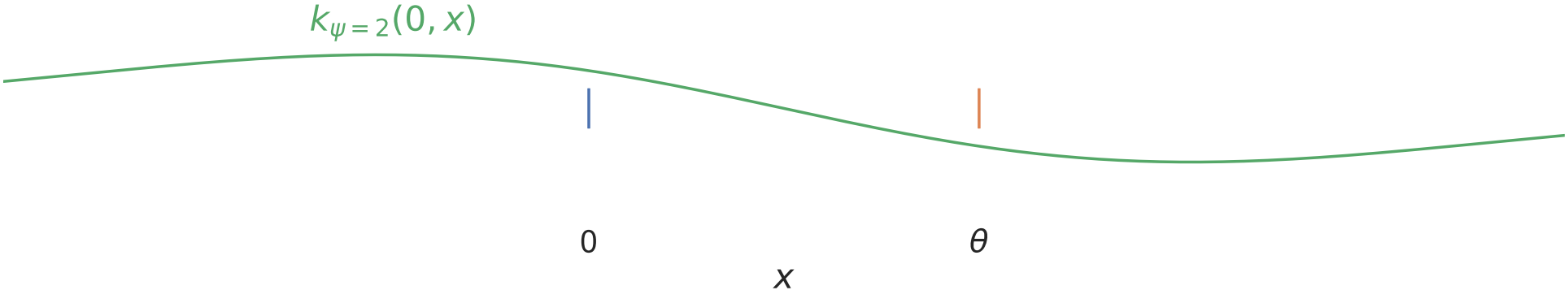
Non-smoothness of Optimized MMD

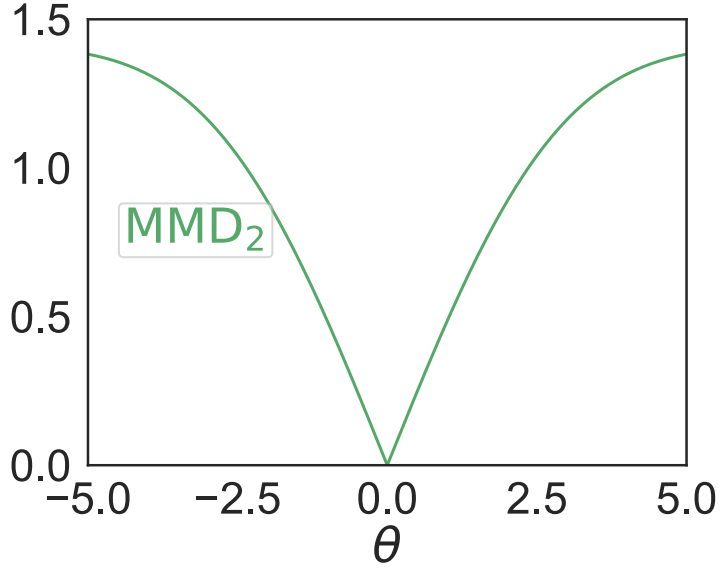
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:



Non-smoothness of Optimized MMD

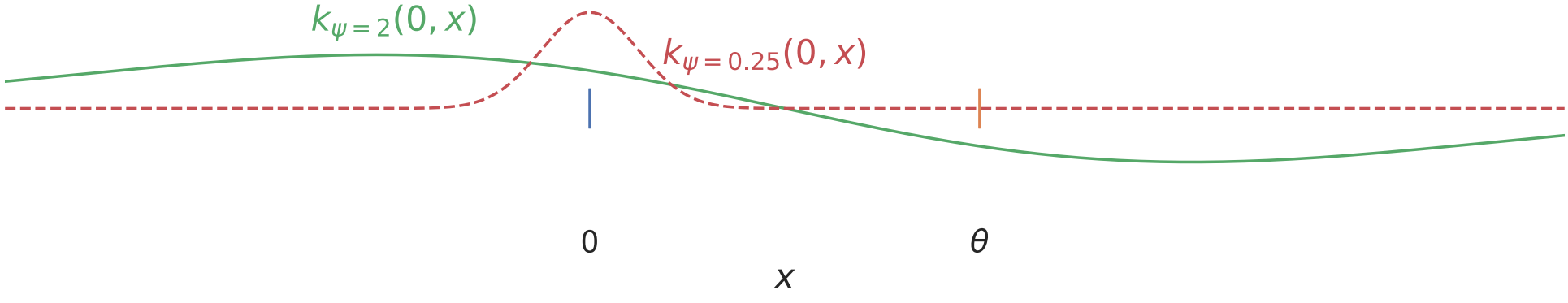
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

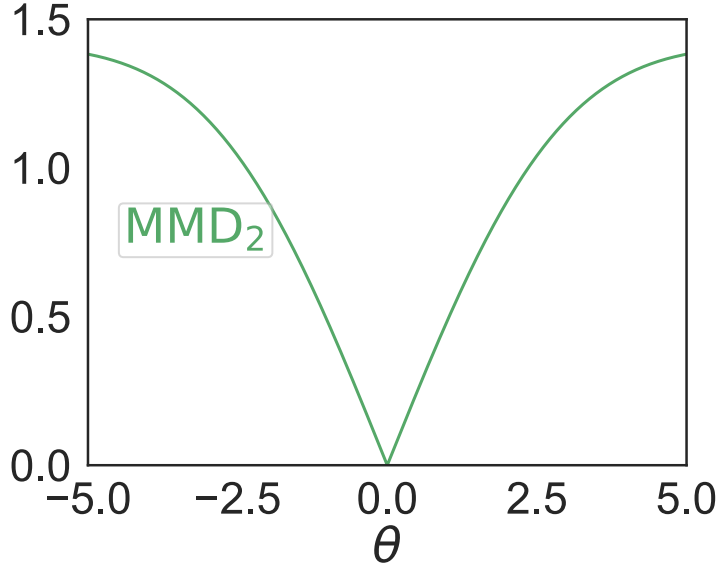




Non-smoothness of Optimized MMD

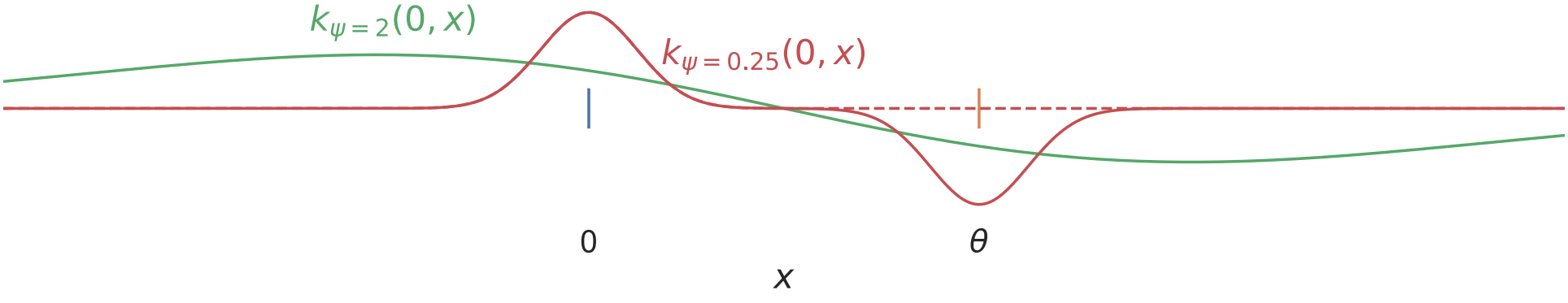
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

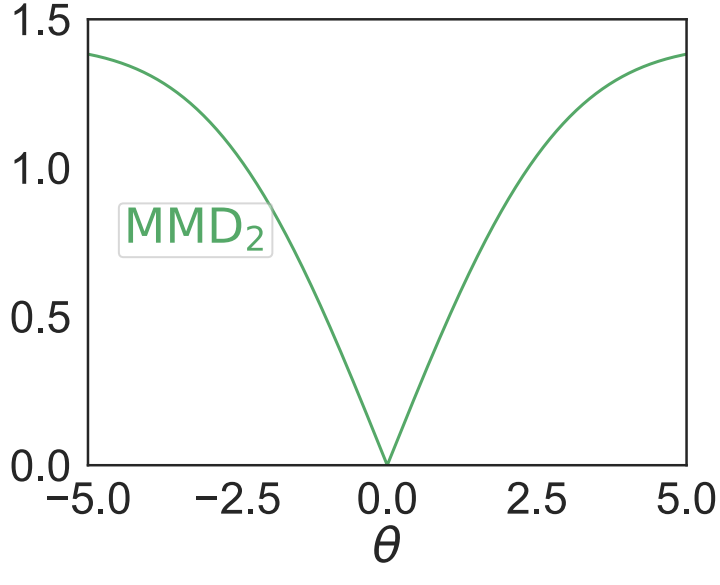




Non-smoothness of Optimized MMD

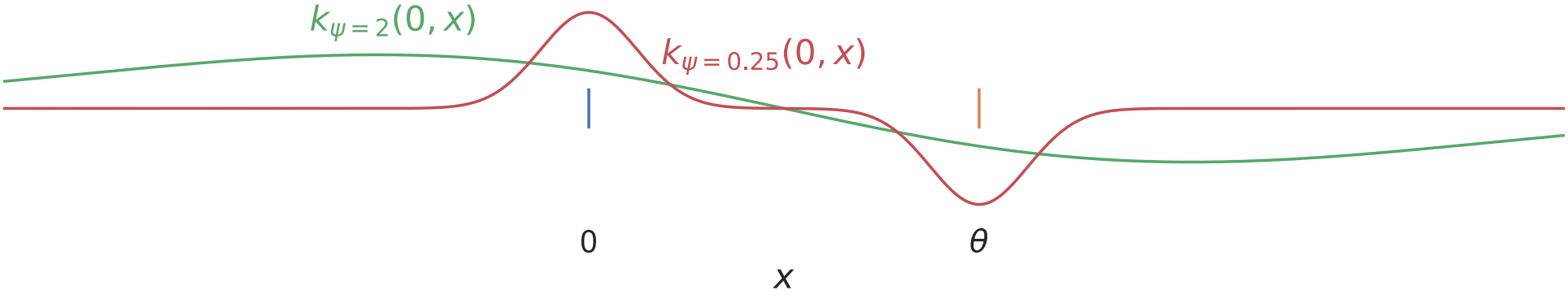
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

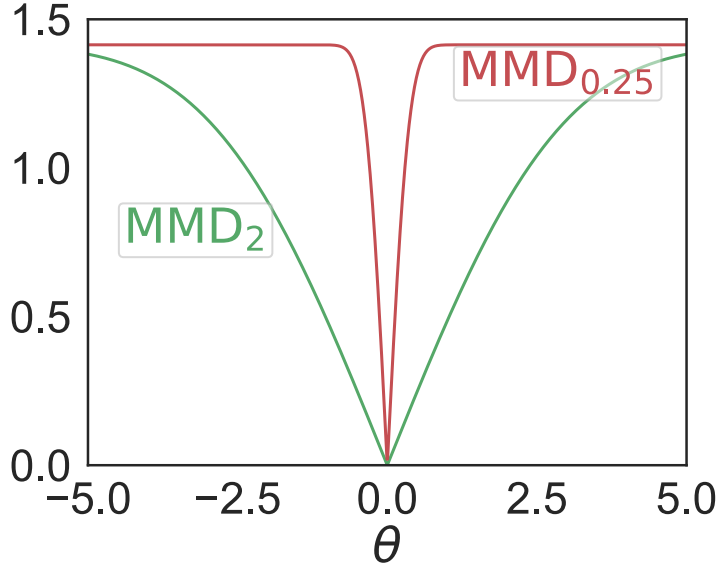




Non-smoothness of Optimized MMD

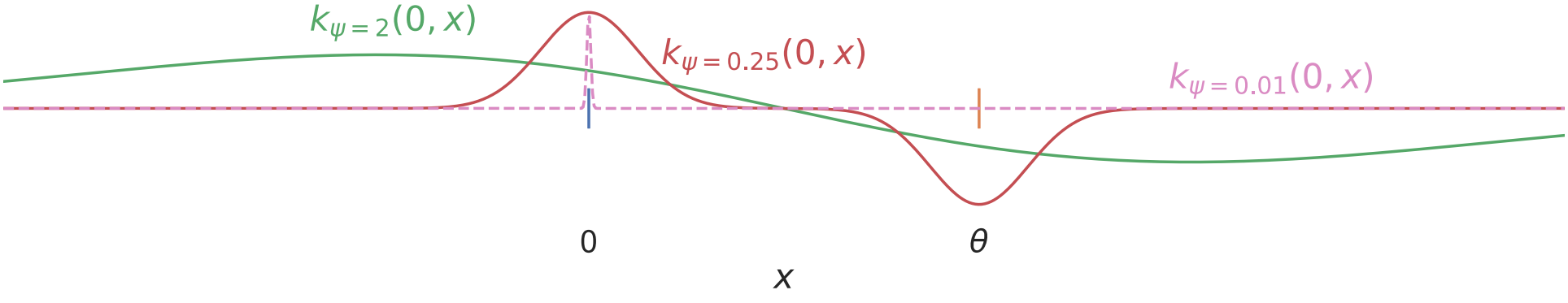
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

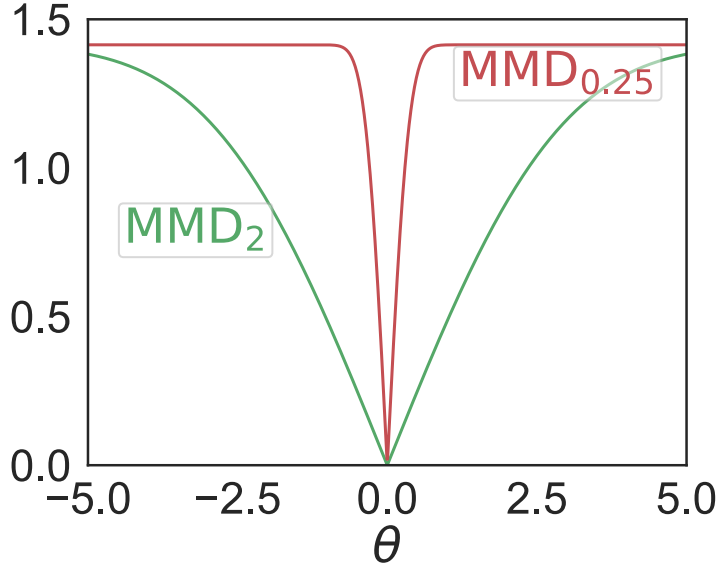




Non-smoothness of Optimized MMD

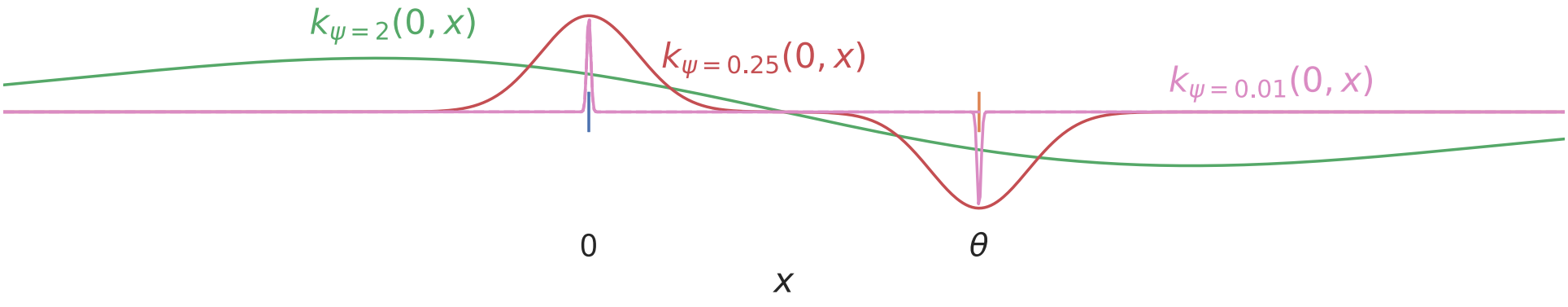
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

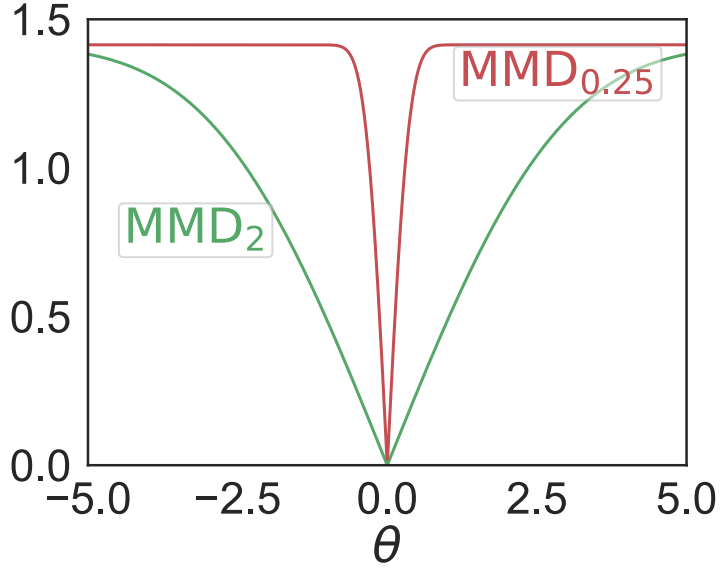




Non-smoothness of Optimized MMD

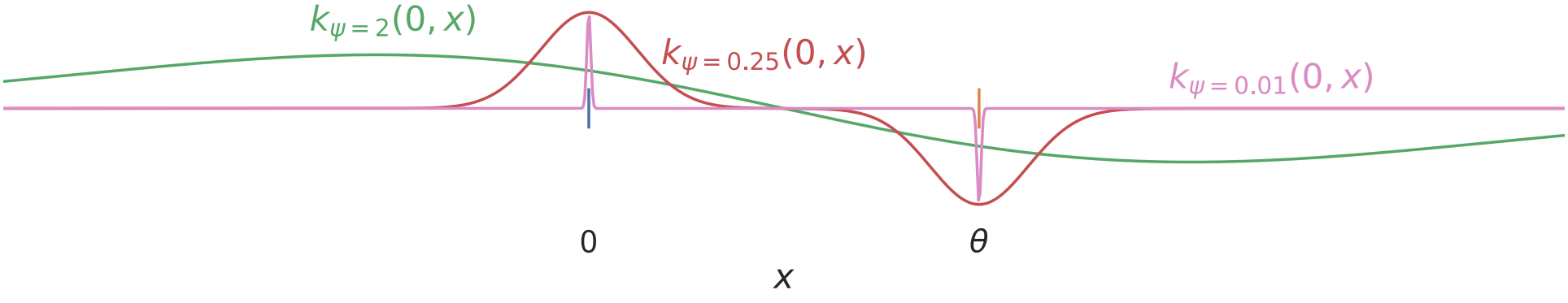
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

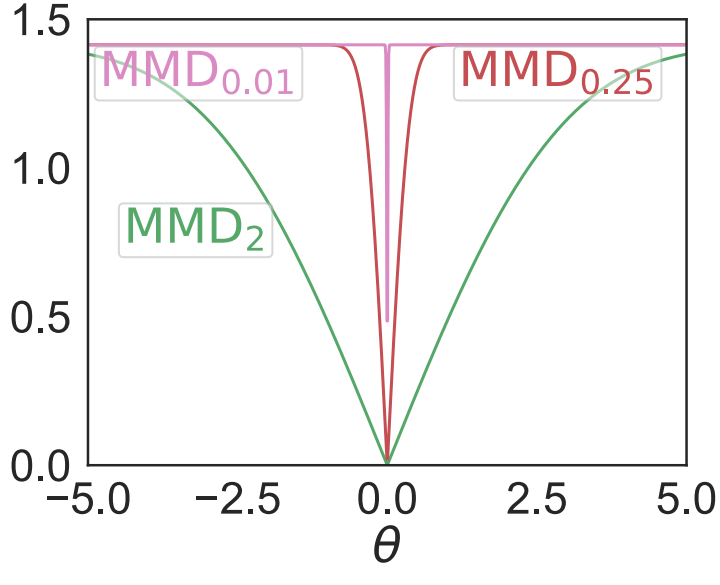




Non-smoothness of Optimized MMD

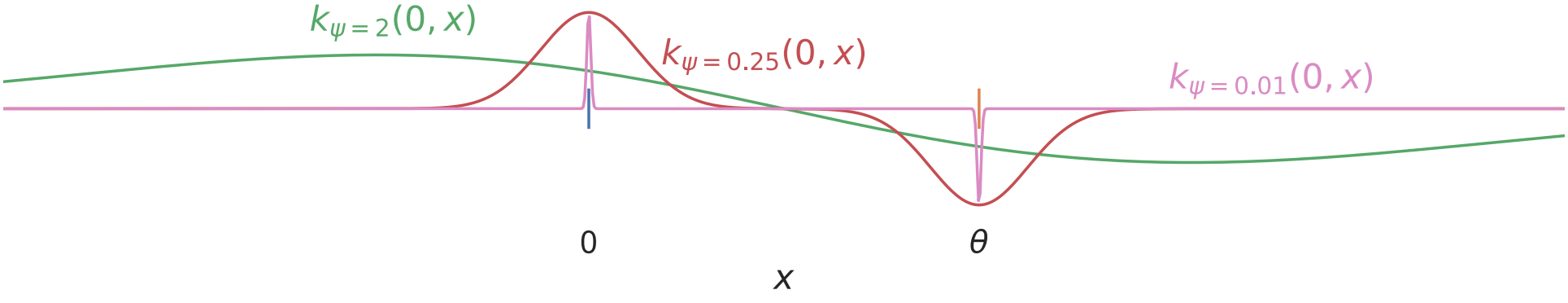
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

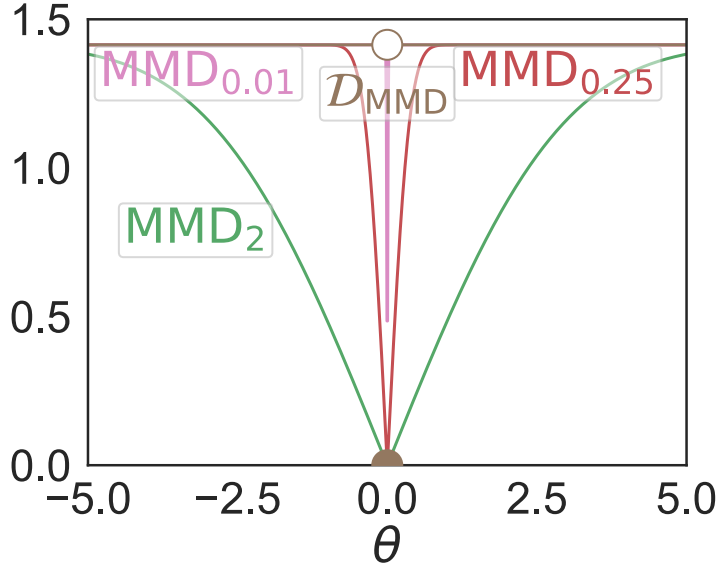




Non-smoothness of Optimized MMD

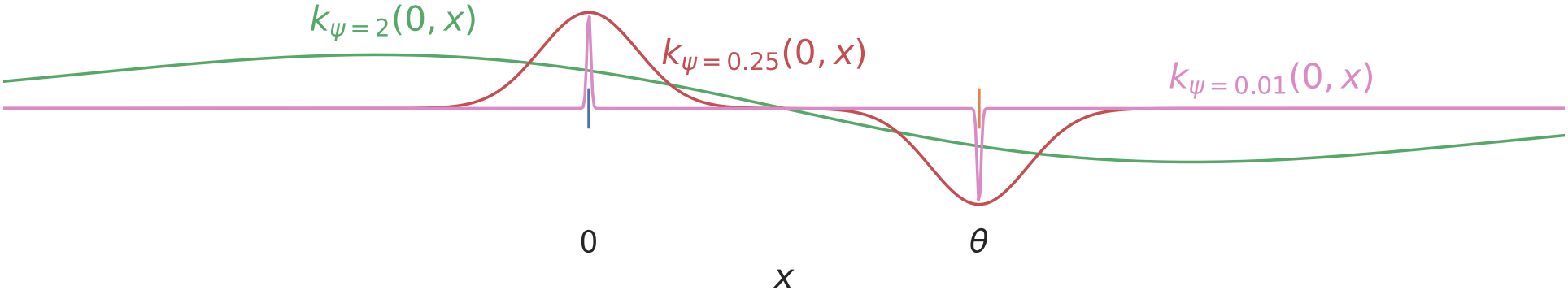
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

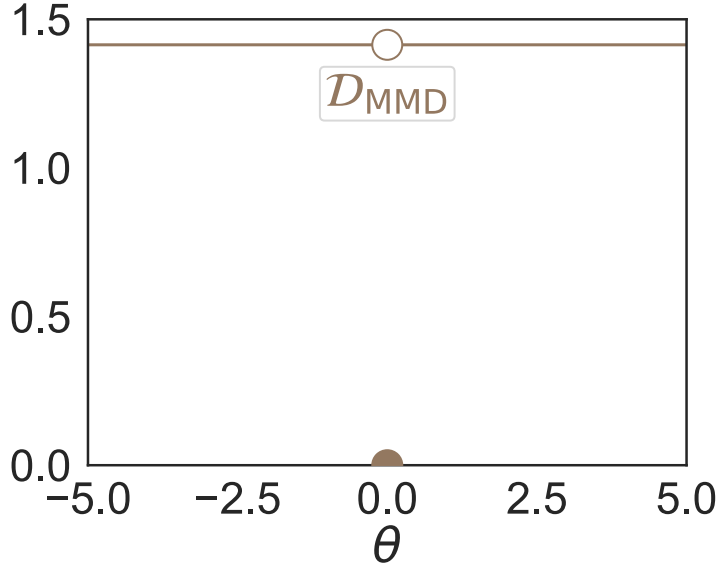




Non-smoothness of Optimized MMD

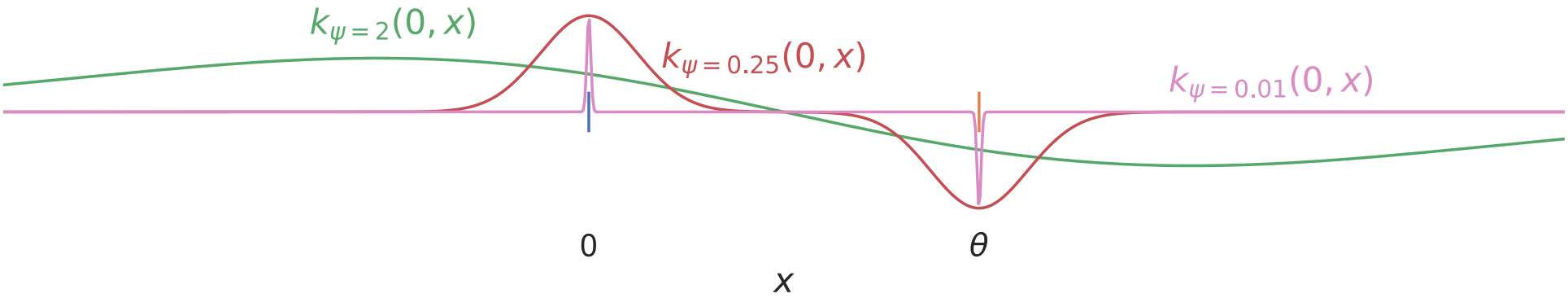
Illustrative problem in $\mathbb{R}$, DiracGAN [Mescheder+ ICML-18]:

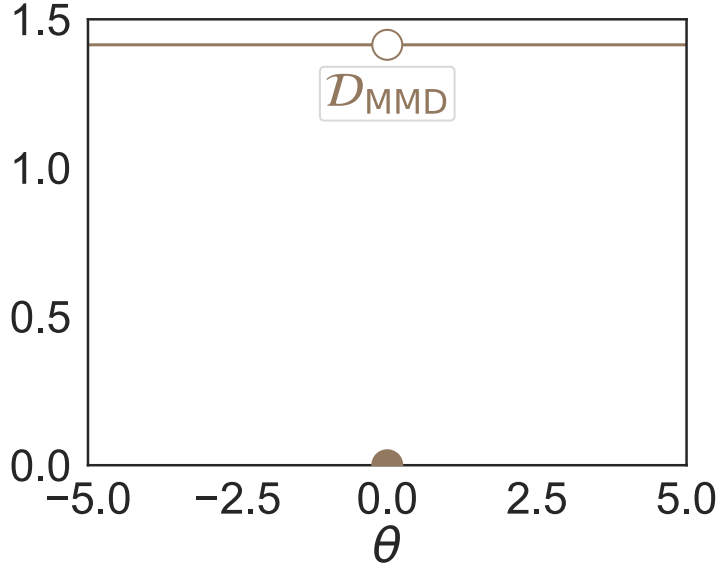


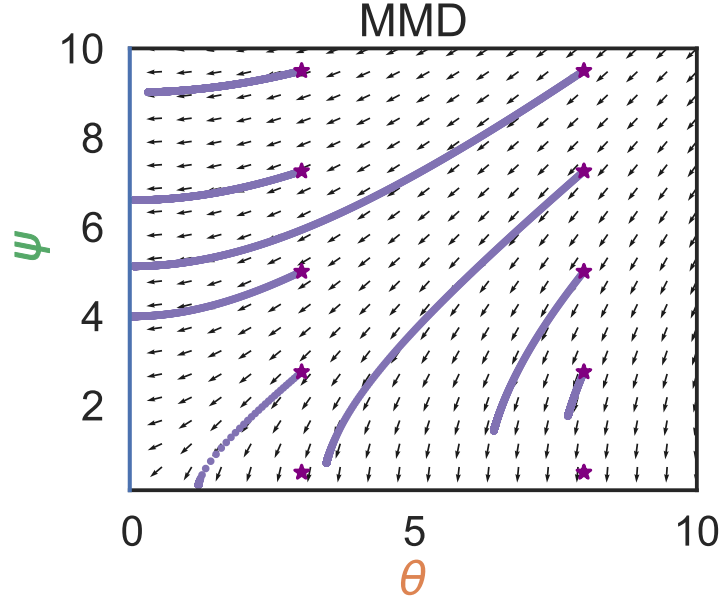


Non-smoothness of Optimized MMD

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

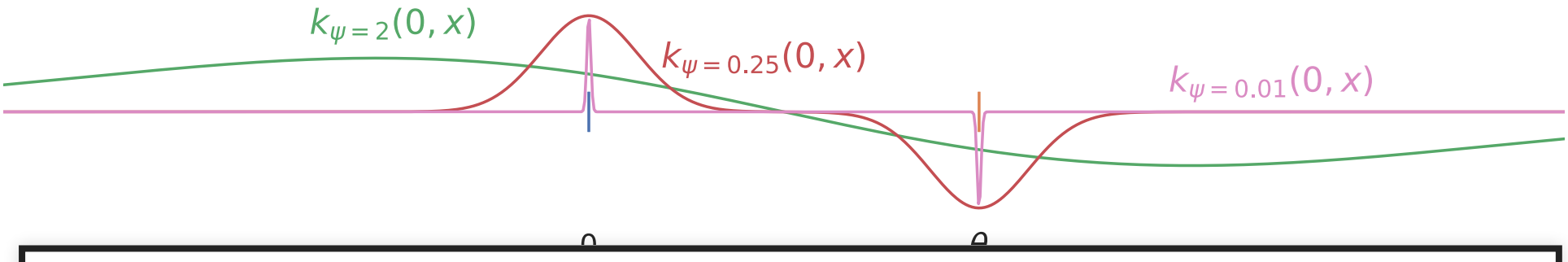




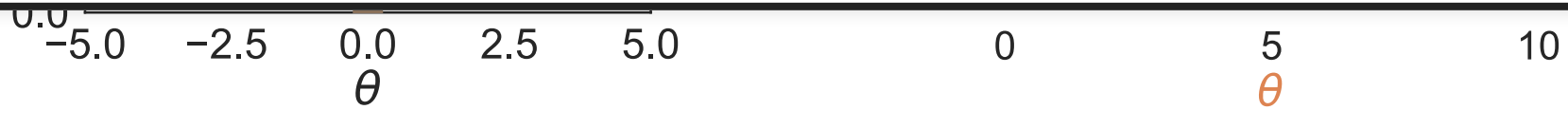


Non-smoothness of Optimized MMD

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

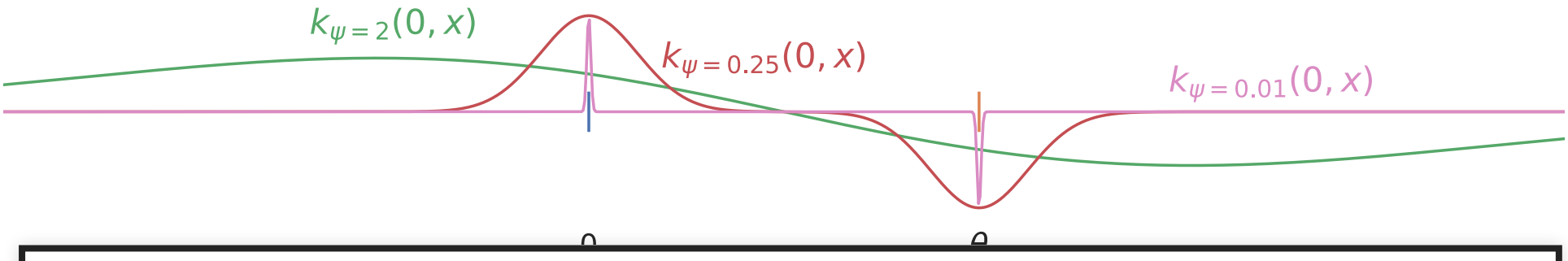


- Just need to stay away from tiny bandwidths ψ

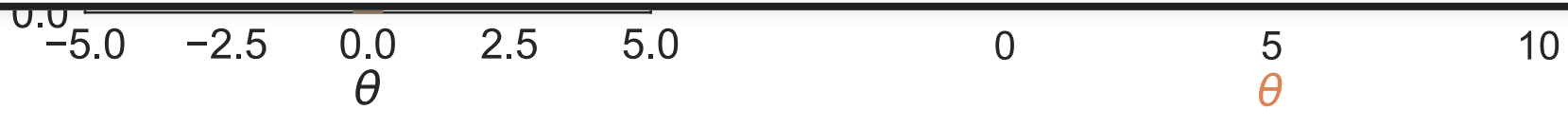


Non-smoothness of Optimized MMD

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

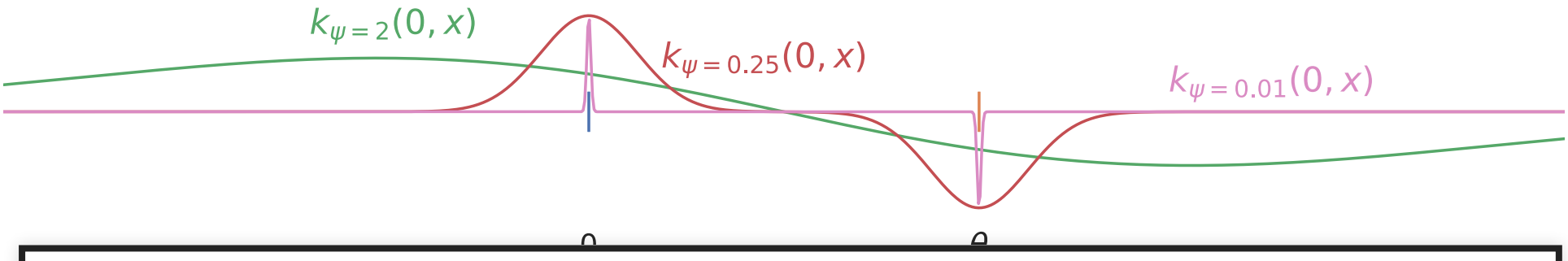


- Just need to stay away from tiny bandwidths ψ
- ...deep kernel analogue is hard.

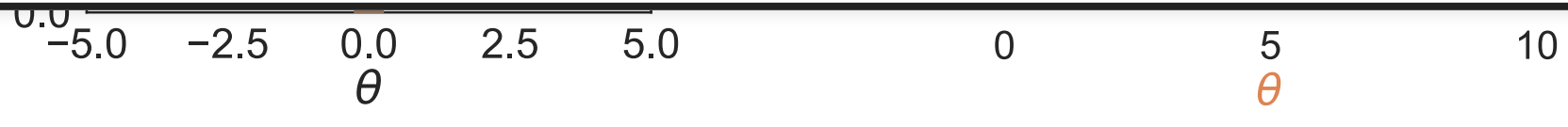


Non-smoothness of Optimized MMD

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

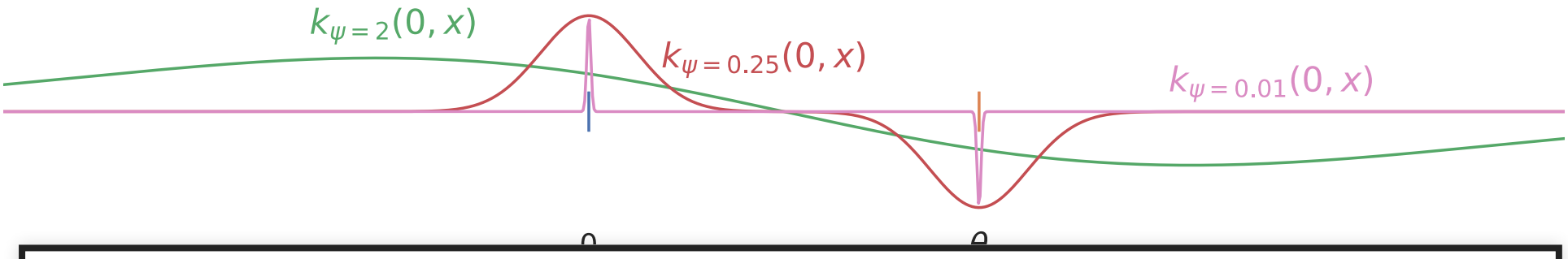


- Just need to stay away from tiny bandwidths ψ
- ...deep kernel analogue is hard.
- Instead, keep witness function from being too steep

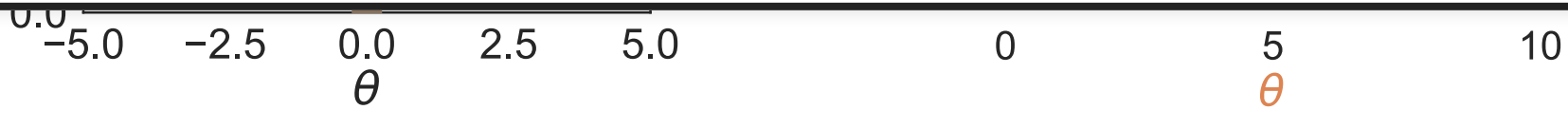


Non-smoothness of Optimized MMD

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



- Just need to stay away from tiny bandwidths ψ
- ...deep kernel analogue is hard.
- Instead, keep witness function from being too steep
- 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 Ψ gives uniformly Lipschitz critics, $\mathcal{D}_{\mathrm{MMD}}^{\Psi}$ is smooth
- Original MMD-GAN paper [Li+ NeurIPS-17]: box constraint

MMD-GAN with gradient control

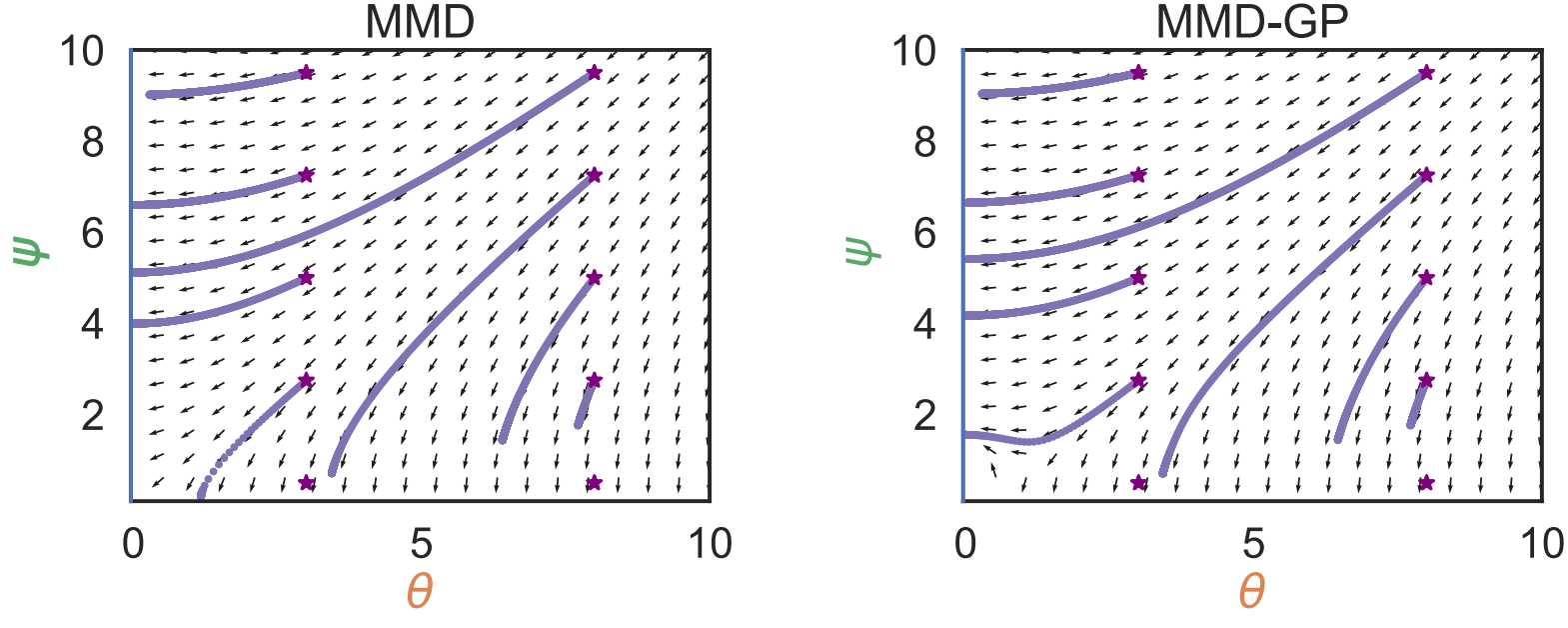
- If Ψ gives uniformly Lipschitz critics, $\mathcal{D}_{\mathrm{MMD}}^{\Psi}$ is smooth
- Original MMD-GAN paper [Li+ NeurIPS-17]: box constraint
- We [Bińkowski+ ICLR-18] used gradient penalty on critic instead

MMD-GAN with gradient control

- If Ψ gives uniformly Lipschitz critics, $\mathcal{D}_{\mathrm{MMD}}^{\Psi}$ 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 Ψ gives uniformly Lipschitz critics, $\mathcal{D}^{\Psi}_{\mathrm{MMD}}$ 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 $\mathbb{E}_{\tilde{X} \sim \mathbb{S}}[\|\nabla f(\tilde{X})\|^2] \leq 1$

New distance: Scaled MMD

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

Can do directly with kernel properties...but too expensive!

New distance: Scaled MMD

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

Can do directly with kernel properties...but too expensive!

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

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

New distance: Scaled MMD

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

Can do directly with kernel properties...but too expensive!

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

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

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

New distance: Scaled MMD

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

Can do directly with kernel properties...but too expensive!

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

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

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

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

$$\mathcal{D}^{\mathbb{S},\Psi,\lambda}_{\mathrm{SMMD}} \ \mathbf{has}\ \mathcal{F} = \bigcup_{\psi \in \Psi} \left\{ f : \|f\|_{\mathcal{H}_\psi} \leq \sigma_{\mathbb{S},k,\lambda} \right\}$$

Deriving the Scaled MMD

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

Deriving the Scaled MMD

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

Deriving the Scaled MMD

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

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[f(\tilde{X})^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\varphi(\tilde{X})\otimes\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

Deriving the Scaled MMD

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

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[f(\tilde{X})^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\varphi(\tilde{X})\otimes\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[\|\nabla f(\tilde{X})\|^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\sum_{i=1}^{d}\partial_i\varphi(\tilde{X})\otimes\partial_i\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

Deriving the Scaled MMD

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

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[f(\tilde{X})^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\varphi(\tilde{X})\otimes\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[\|\nabla f(\tilde{X})\|^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\sum_{i=1}^{d}\partial_i\varphi(\tilde{X})\otimes\partial_i\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

Constraint can be written $\langle f, D_\lambda f\rangle_{\mathcal{H}}\leq 1$

Deriving the Scaled MMD

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

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[f(\tilde{X})^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\varphi(\tilde{X})\otimes\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[\|\nabla f(\tilde{X})\|^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\sum_{i=1}^{d}\partial_i\varphi(\tilde{X})\otimes\partial_i\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

Constraint can be written $\langle f, D_\lambda f\rangle_{\mathcal{H}}\leq 1$

$\langle f, D_\lambda f\rangle_{\mathcal{H}}$

Deriving the Scaled MMD

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

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[f(\tilde{X})^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\varphi(\tilde{X})\otimes\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[\|\nabla f(\tilde{X})\|^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\sum_{i=1}^{d}\partial_i\varphi(\tilde{X})\otimes\partial_i\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

Constraint can be written $\langle f, D_\lambda f\rangle_{\mathcal{H}}\leq 1$

$$\langle f, D_\lambda f\rangle_{\mathcal{H}}\leq \|D_\lambda\|\,\|f\|_{\mathcal{H}}^2$$

Deriving the Scaled MMD

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

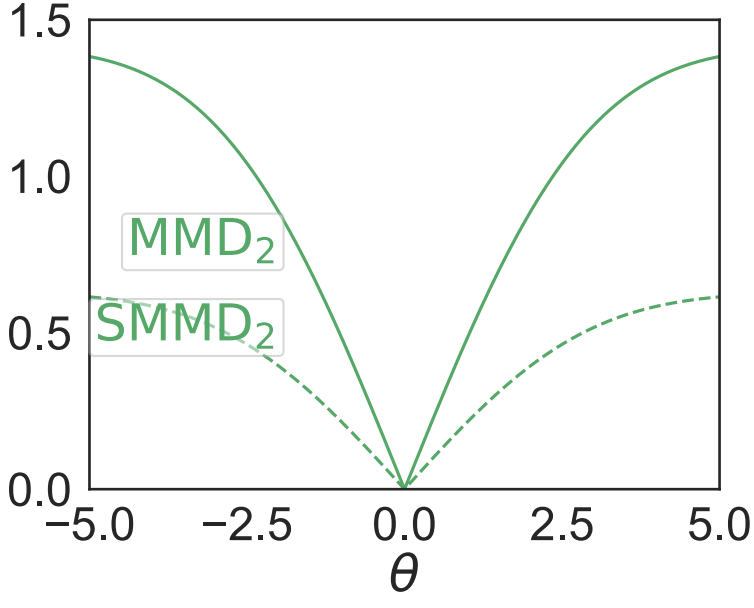
$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[f(\tilde{X})^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\varphi(\tilde{X})\otimes\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

$$\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}[\|\nabla f(\tilde{X})\|^2]=\left\langle f,\mathop{\mathbb{E}}_{\tilde{X}\sim\mathbb{S}}\left[\sum_{i=1}^{d}\partial_i\varphi(\tilde{X})\otimes\partial_i\varphi(\tilde{X})\right]f\right\rangle_{\mathcal{H}}$$

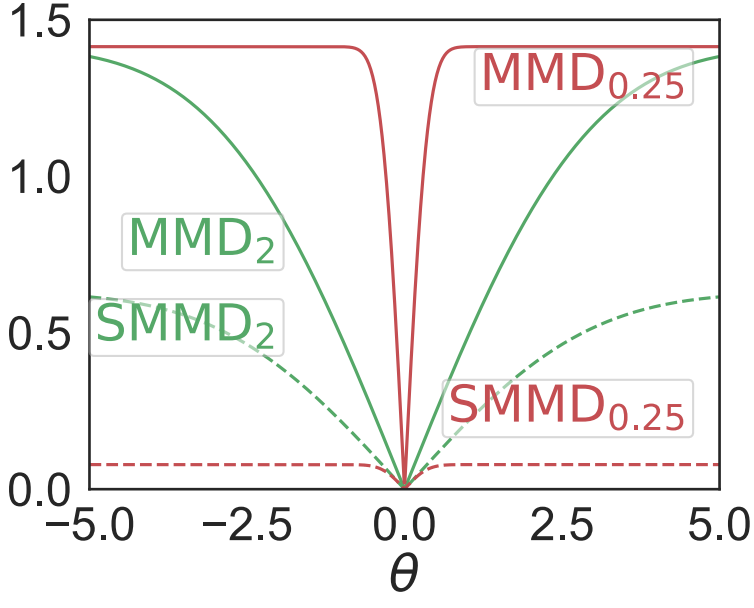
Constraint can be written $\langle f, D_\lambda f\rangle_{\mathcal{H}}\leq 1$

$$\langle f, D_\lambda f\rangle_{\mathcal{H}}\leq\|D_\lambda\|\,\|f\|_{\mathcal{H}}^2\leq\sigma_{\mathbb{S},k,\lambda}^{-2}\|f\|_{\mathcal{H}}^2$$

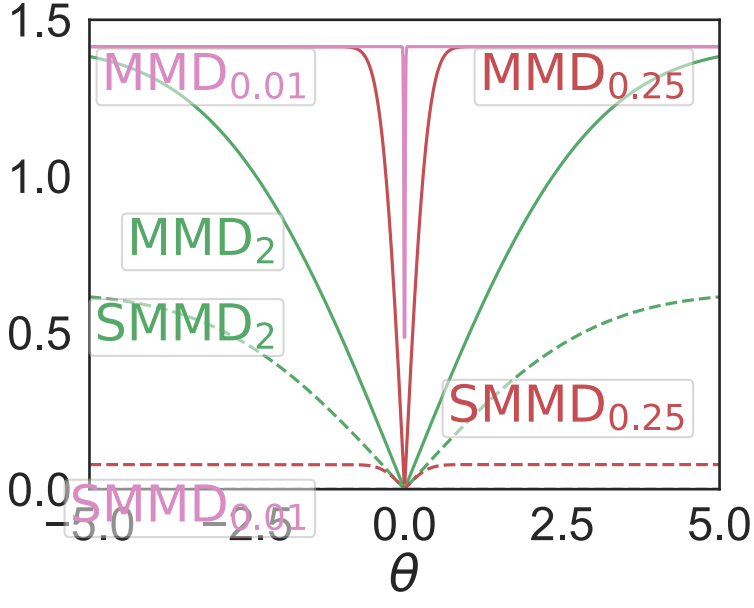
Smoothness of $\mathcal{D}_{\mathrm{SMMD}}$



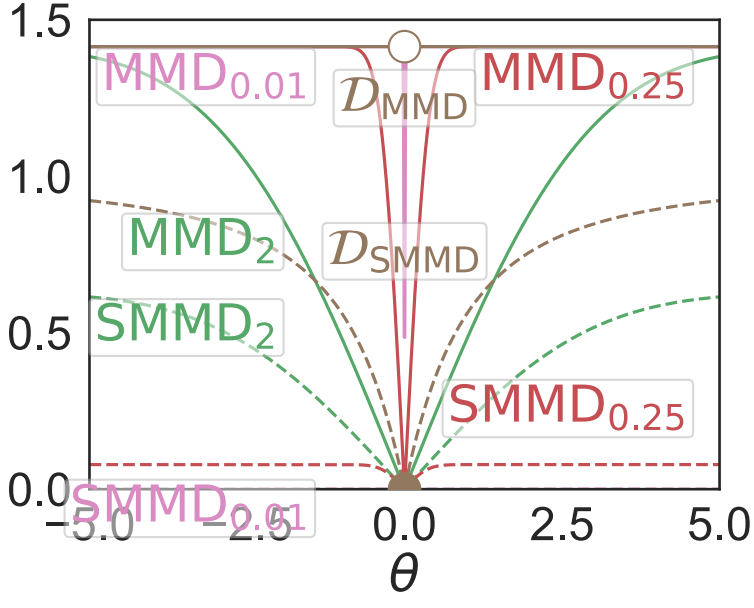
Smoothness of $\mathcal{D}_{\mathrm{SMMD}}$



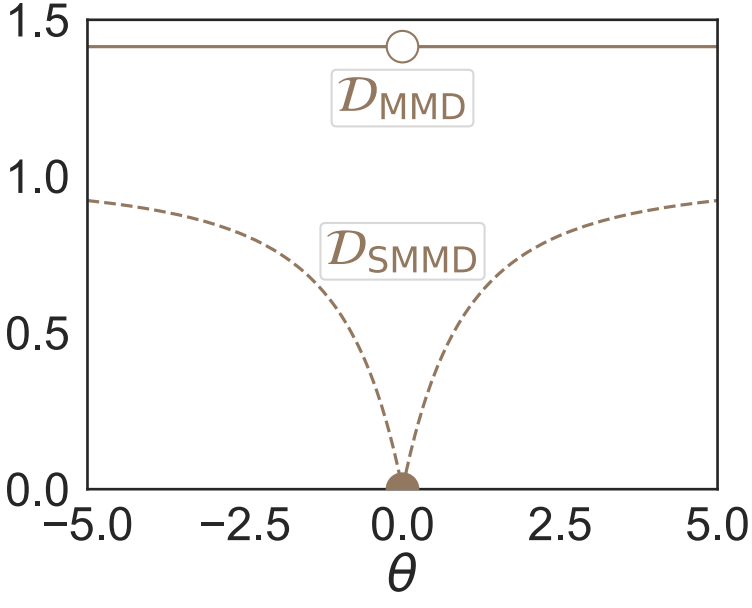
Smoothness of $\mathcal{D}_{\mathrm{SMMD}}$

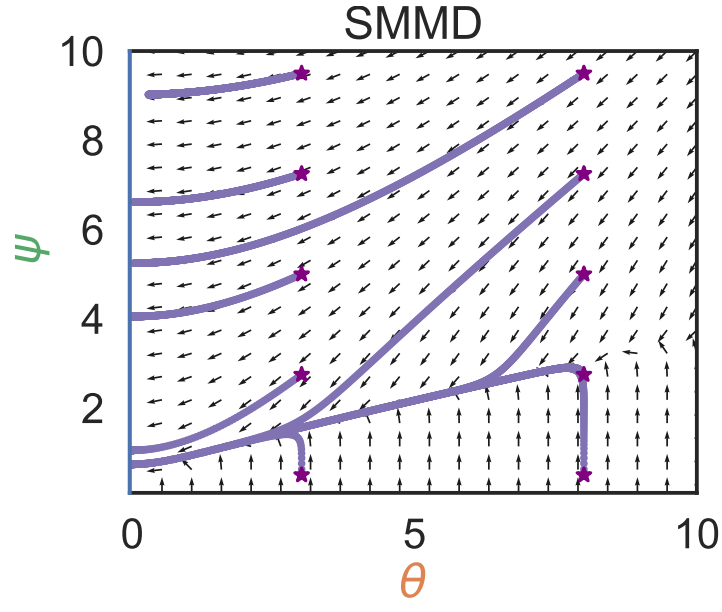


Smoothness of $\mathcal{D}_{\mathrm{SMMD}}$

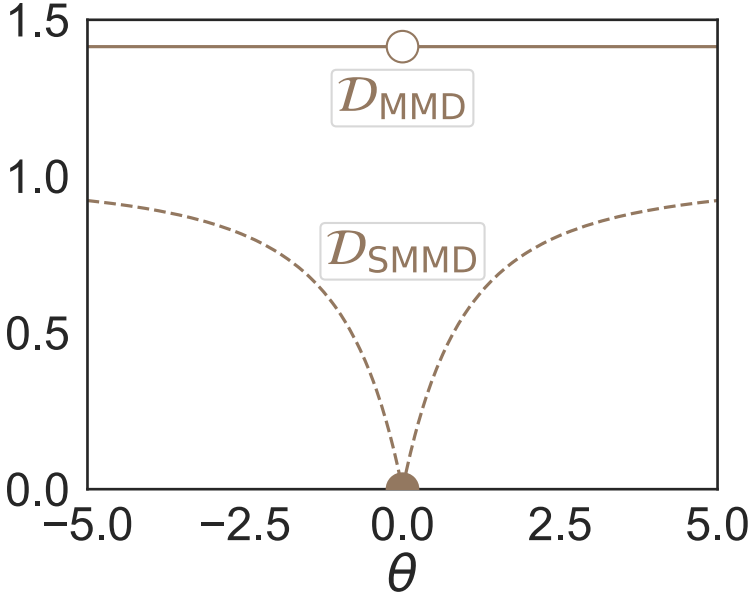


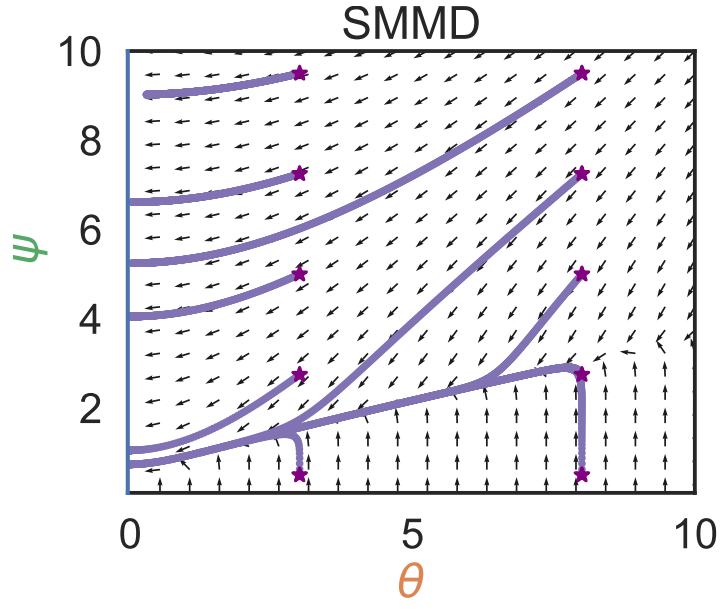
Smoothness of $\mathcal{D}_{\mathrm{SMMD}}$





Smoothness of $\mathcal{D}_{\mathrm{SMMD}}$





Theorem: $\mathcal{D}_{\mathrm{SMMD}}^{\mathbb{S},\Psi,\lambda}$ is continuous.

If $\mathbb{S}$ has a density; k_{top} is Gaussian/linear/...;

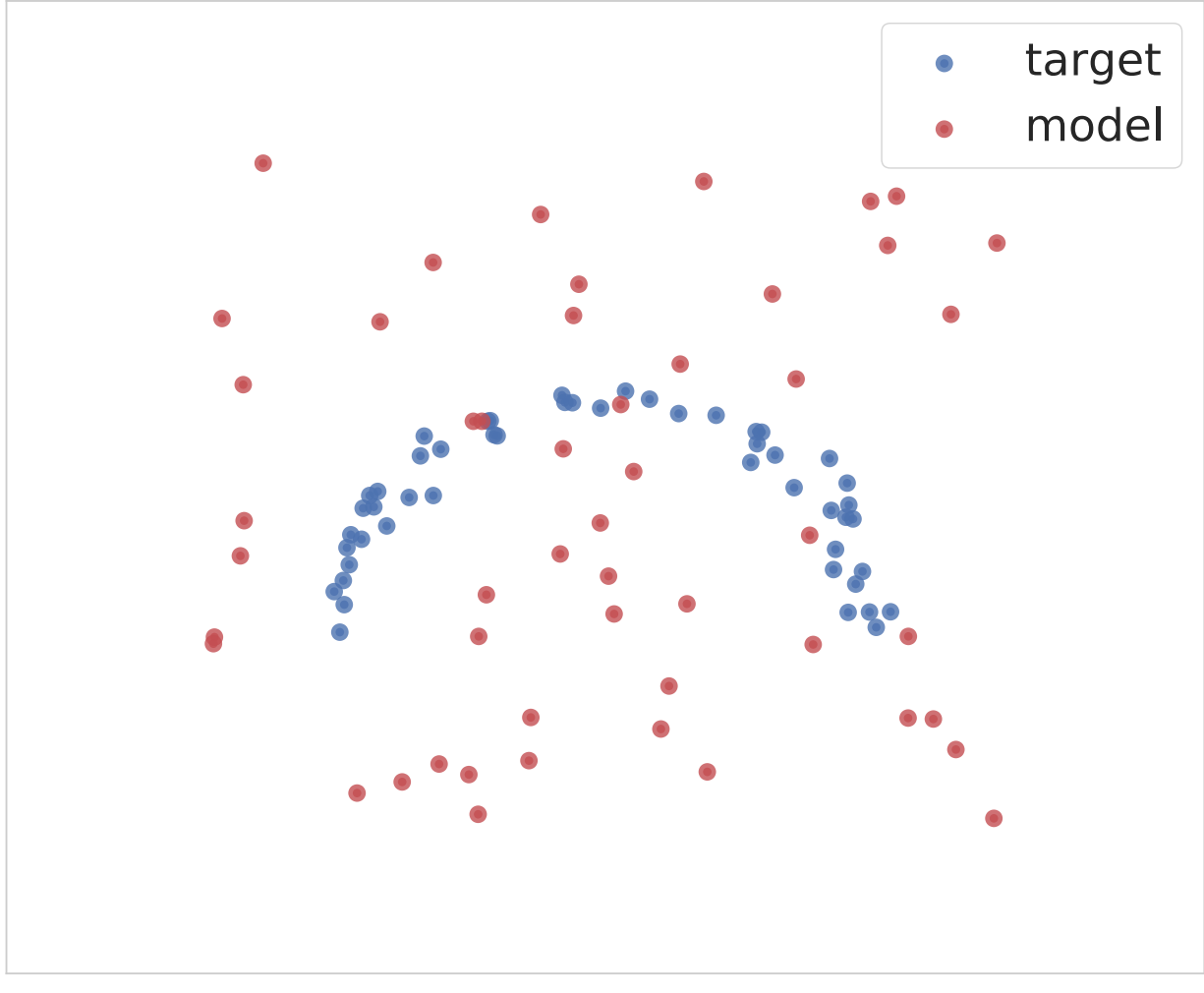
ϕ_ψ is fully-connected, Leaky-ReLU, non-increasing width;

all weights in Ψ have bounded condition number; then

$$\mathcal{W}(\mathbb{Q}_n, \mathbb{P}) \to 0 \text{ implies } \mathcal{D}_{\mathrm{SMMD}}^{\mathbb{S},\Psi,\lambda}(\mathbb{Q}_n, \mathbb{P}) \to 0.$$

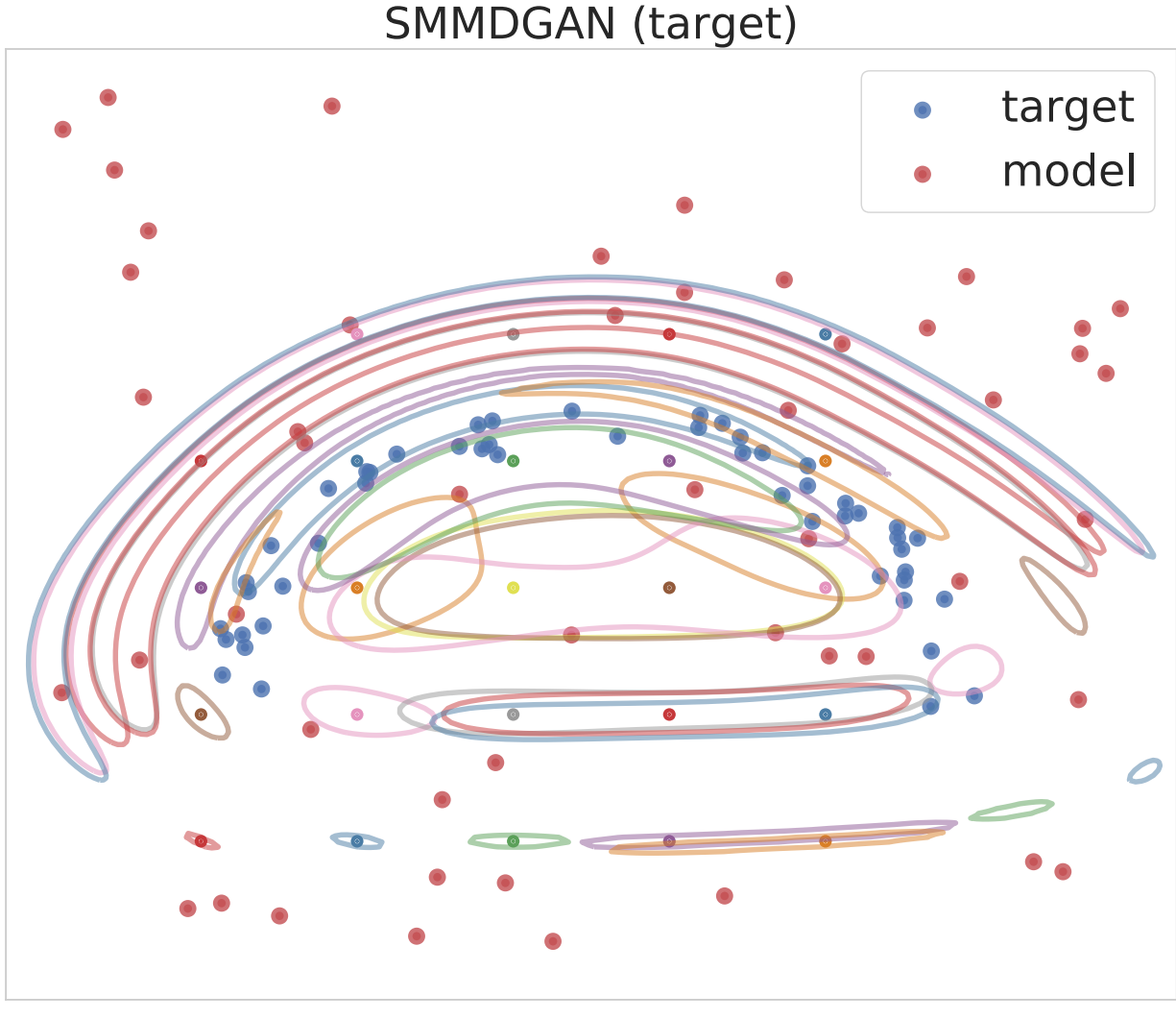
$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

Target $\mathbb{P}$ and model $\mathbb{Q}_\theta$ samples



$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

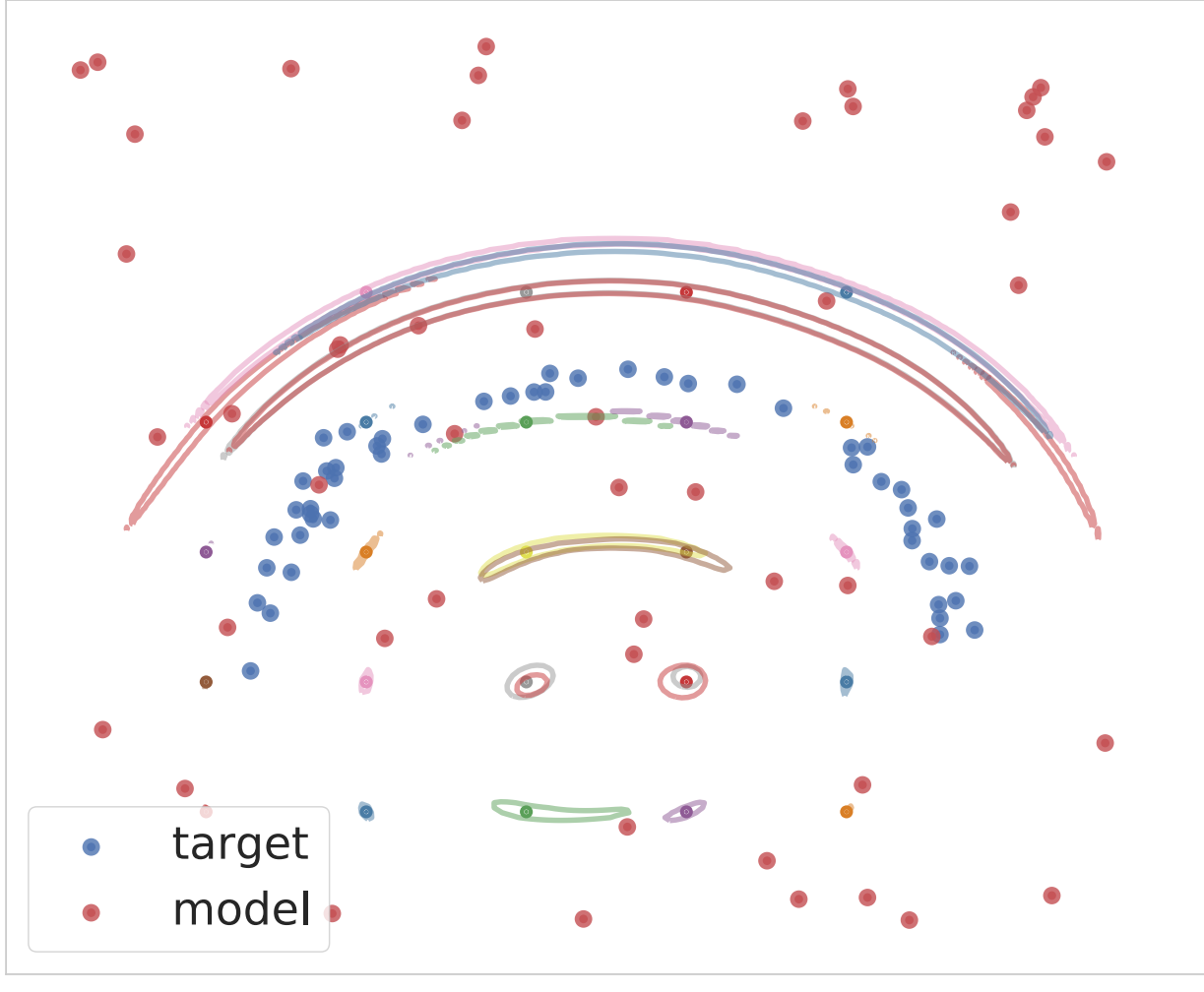
Kernels from $\mathrm{SMMD}_{\mathbb{P},k,\lambda}$, early in optimization



$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

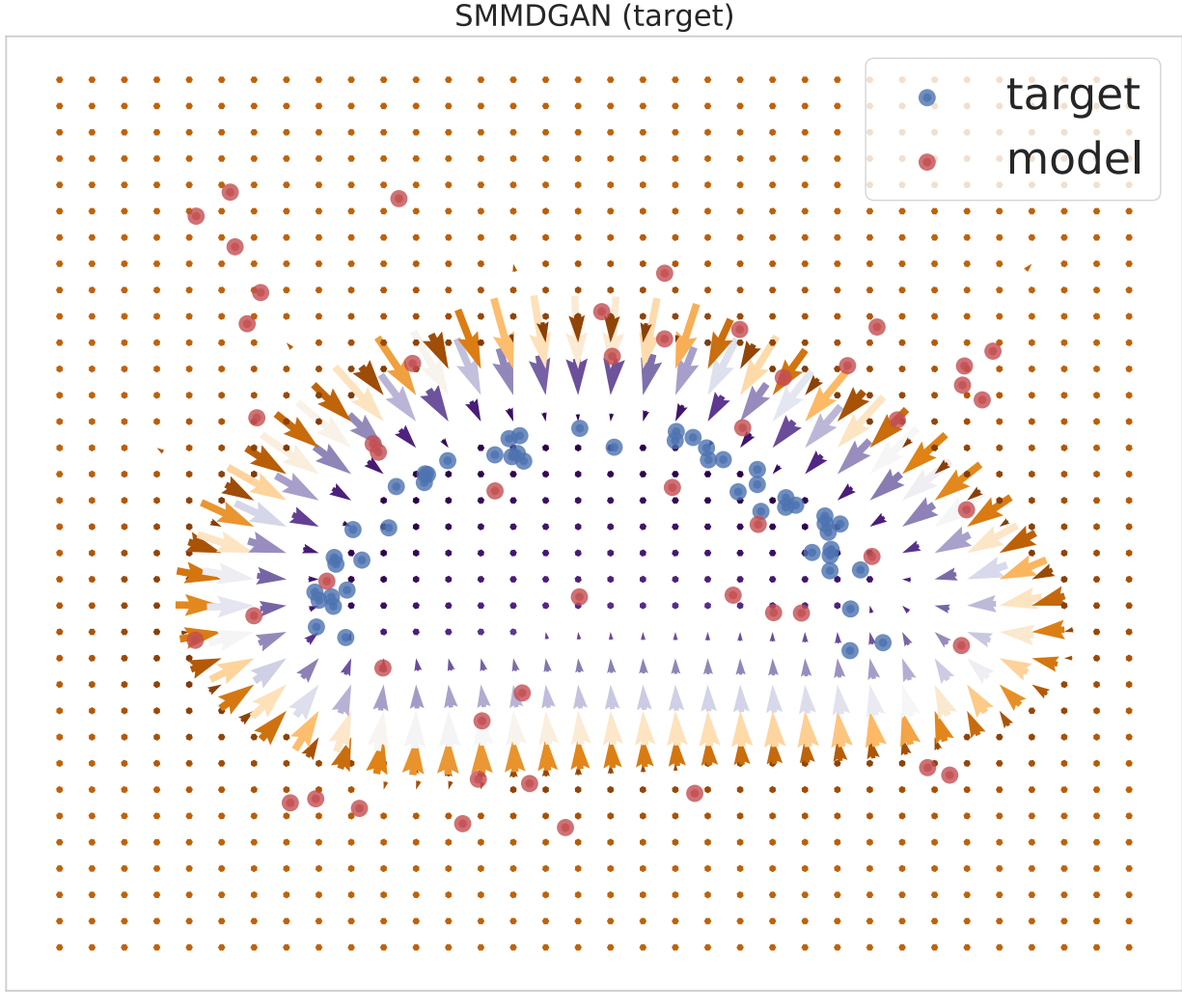
Kernels from MMD_k (early)

MMDGAN (no GP)



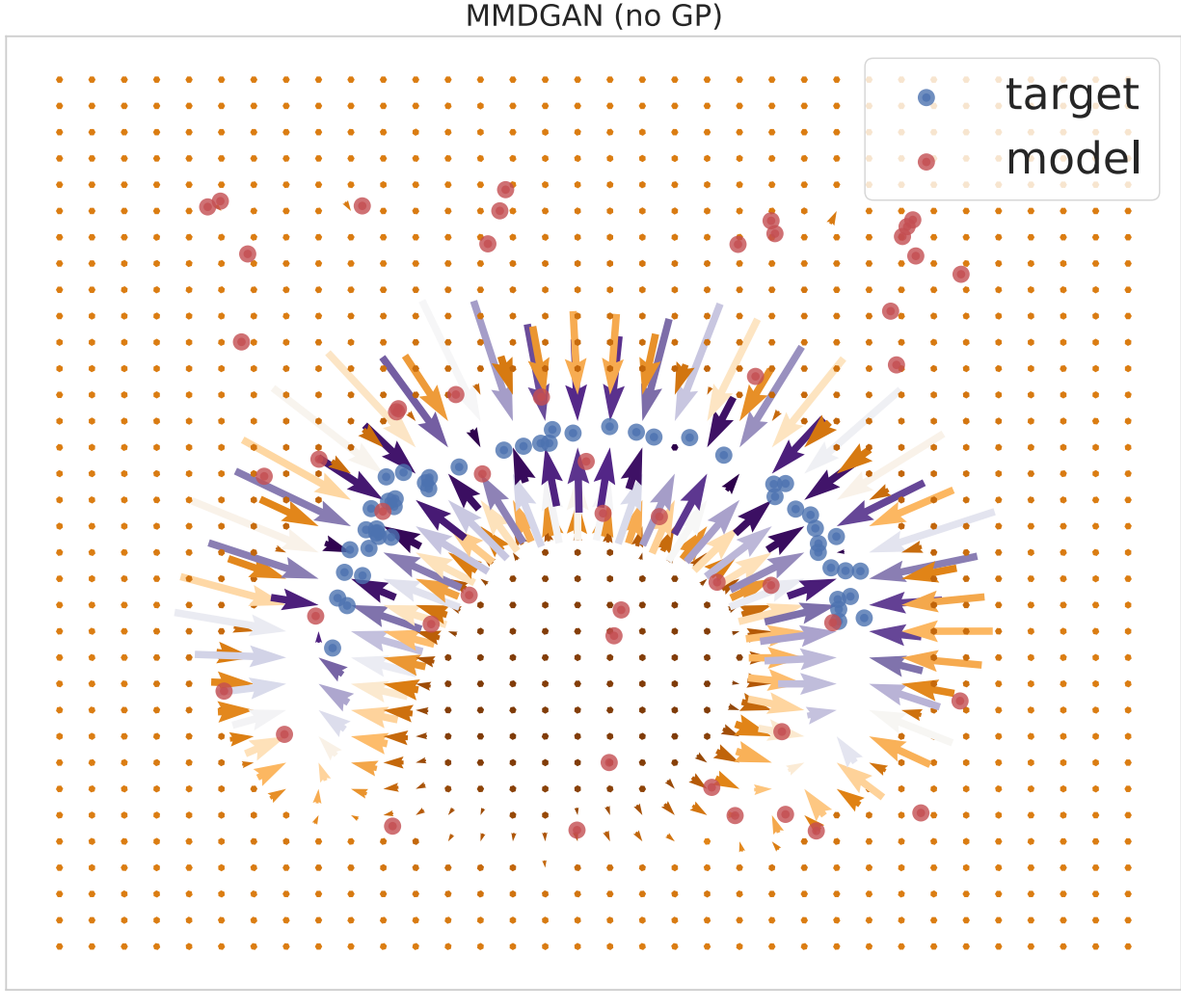
$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

Critic gradients from $\mathrm{SMMD}_{\mathbb{P},k,\lambda}$ (early)



$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

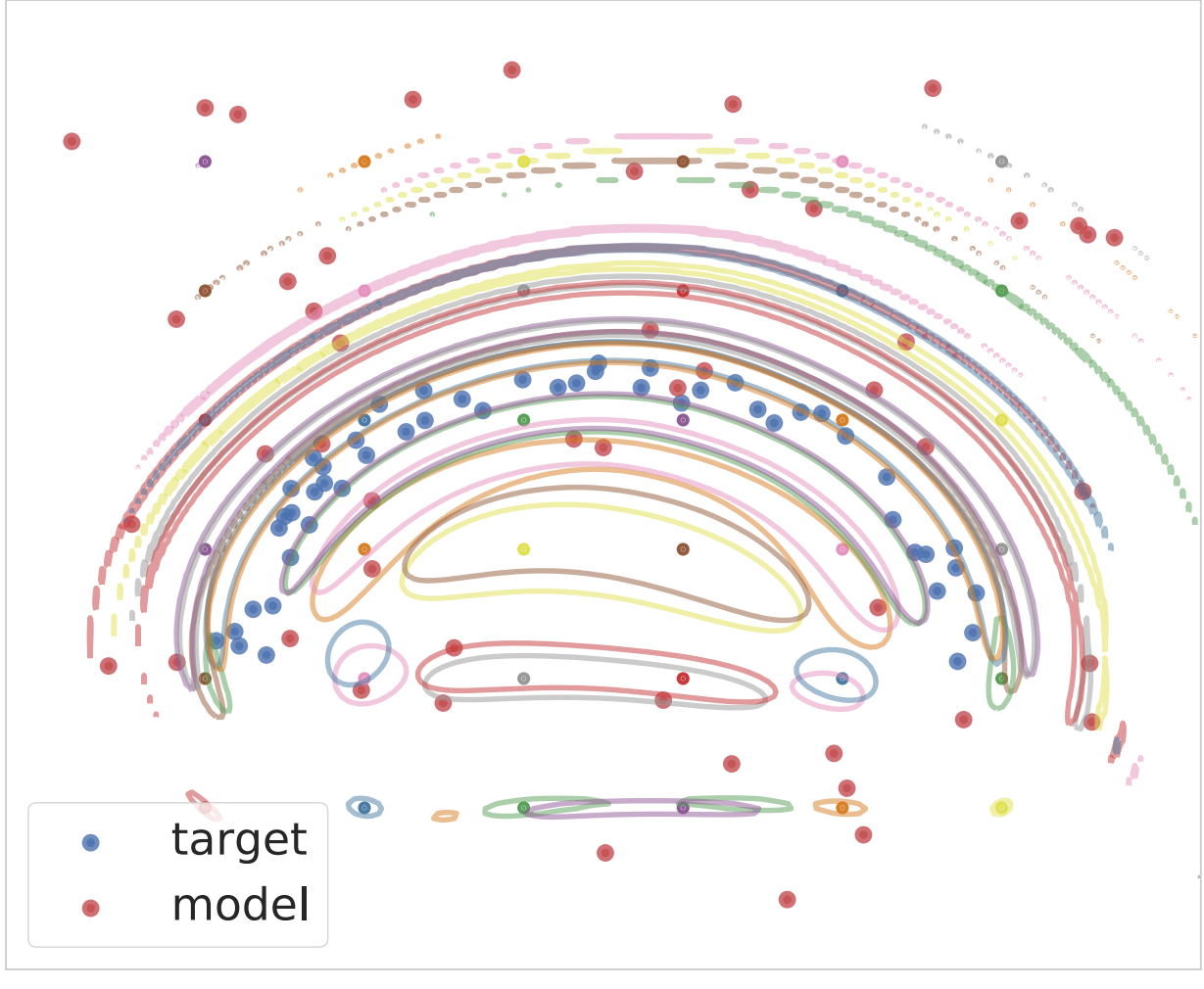
Critic gradients from MMD_k (early)



$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

Kernels from $\mathrm{SMMD}_{\mathbb{P},k,\lambda}$, **late** in optimization

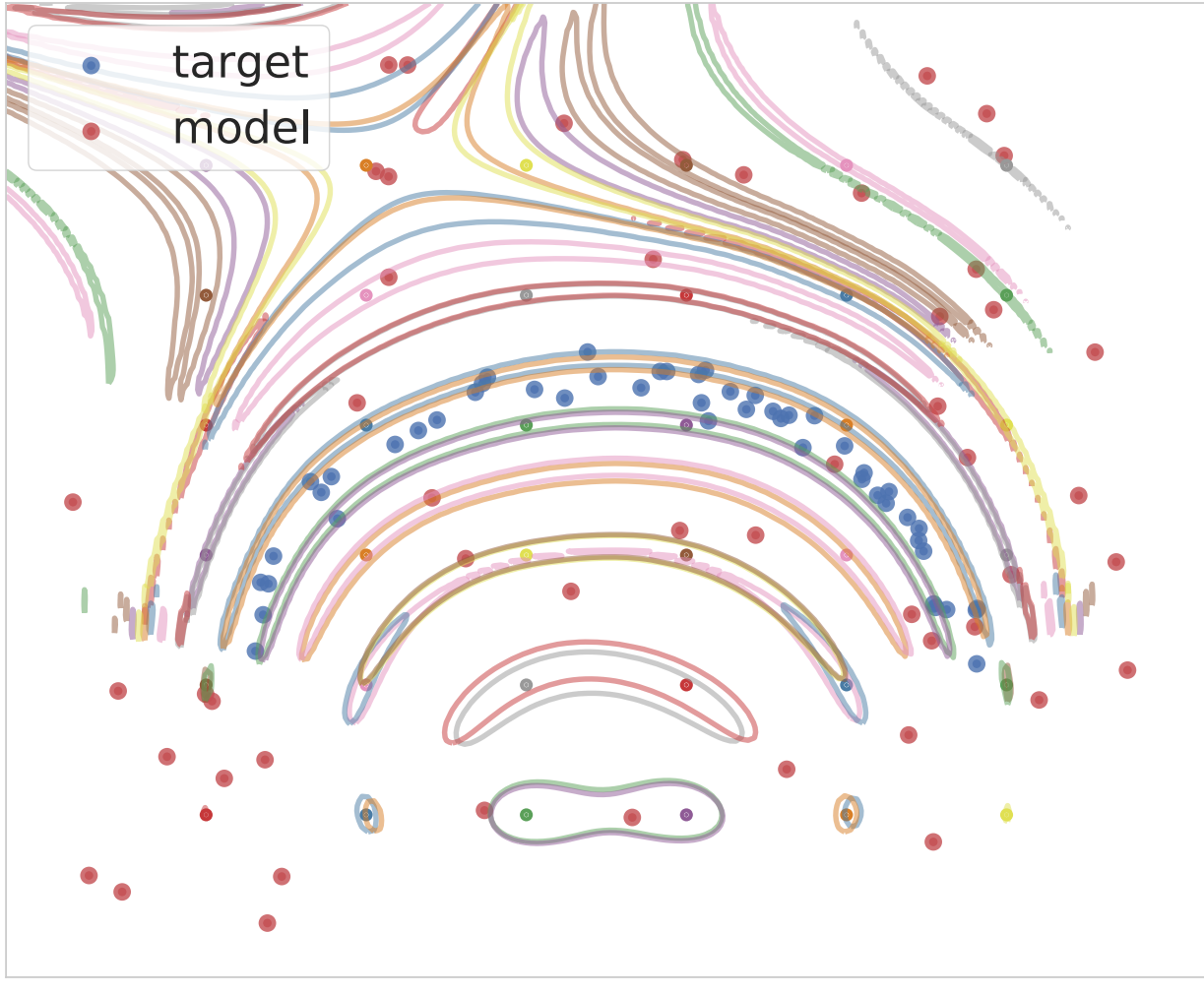
SMMDGAN (target)



$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

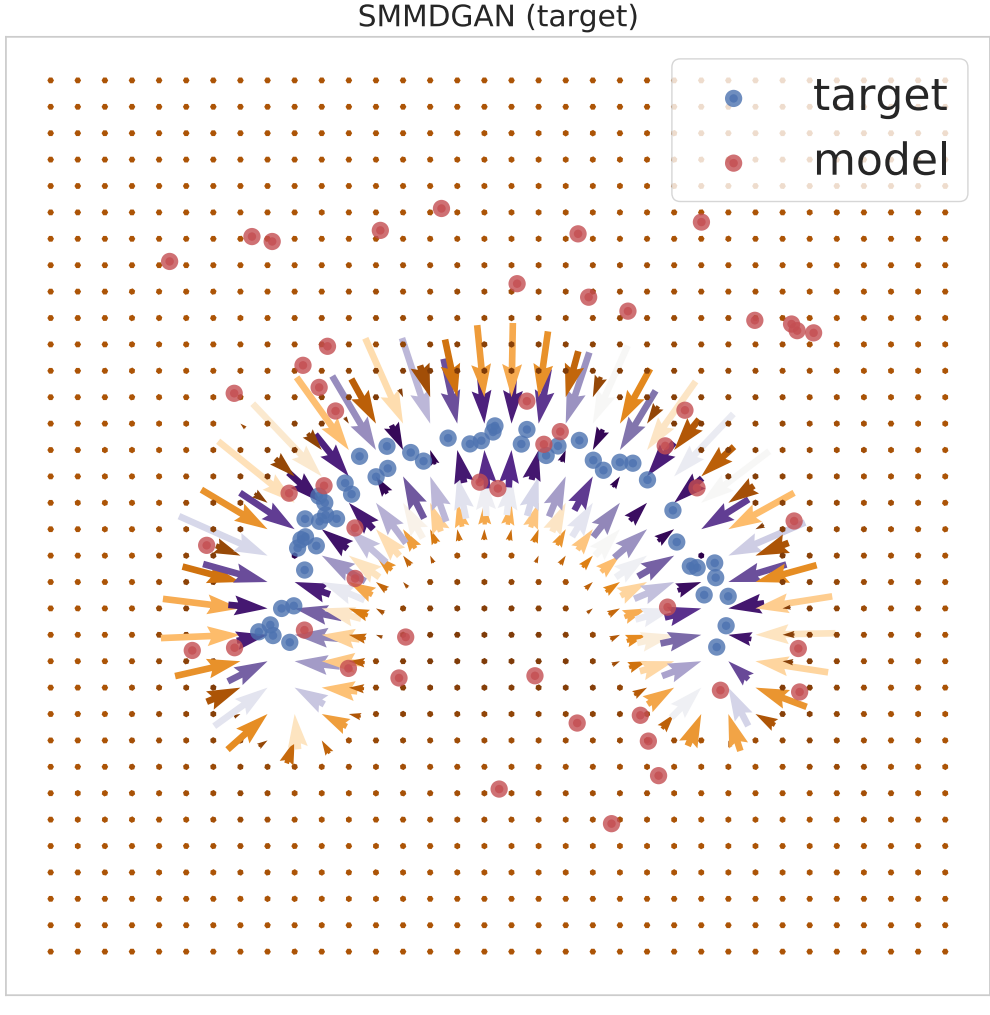
Kernels from MMD_k (late)

MMDGAN (no GP)



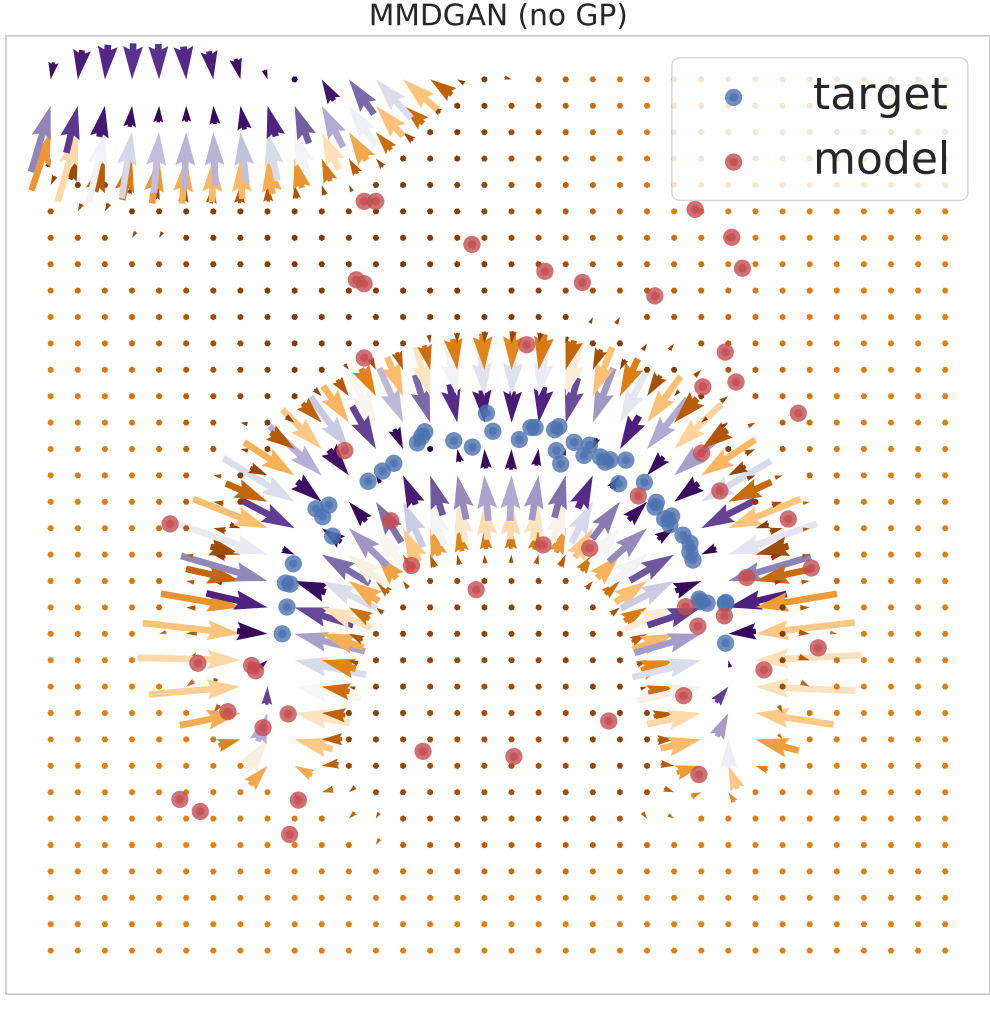
$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

Critic gradients from $\mathrm{SMMD}_{\mathbb{P},k,\lambda}$ (late)



$\mathcal{D}_{\mathrm{SMMD}}$: 2d example

Critic gradients from MMD_k (late)

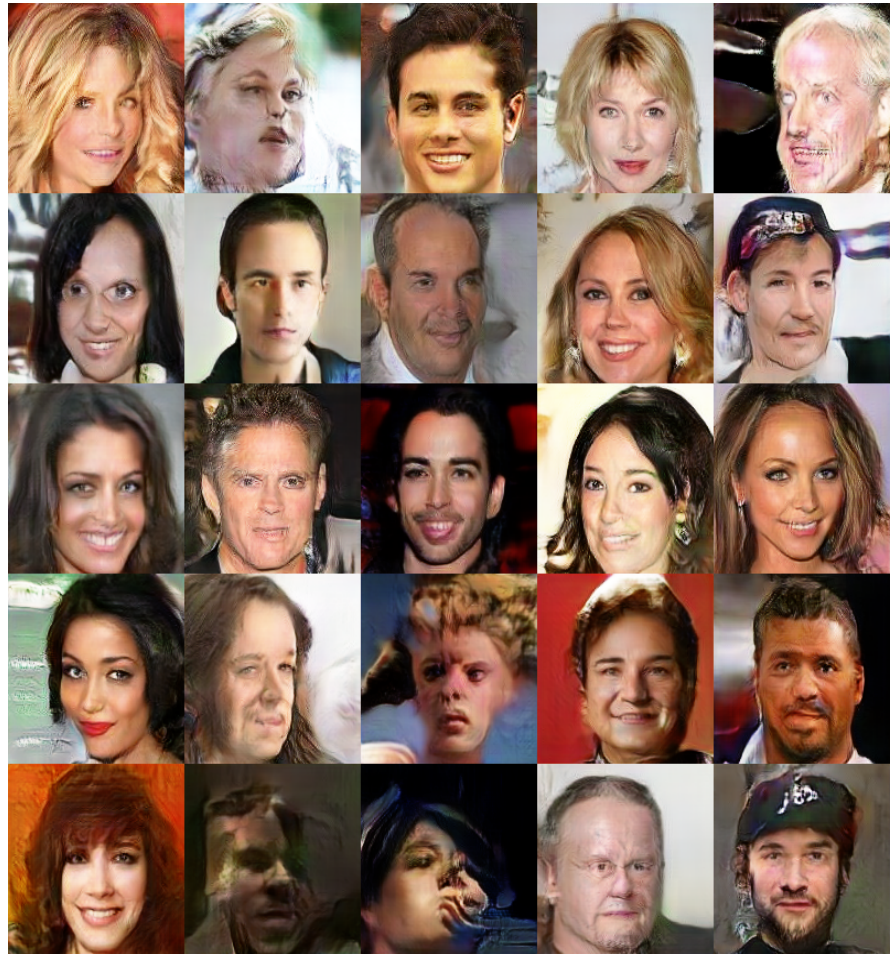


Model on 160×160 CelebA

SN-SMMD-GAN

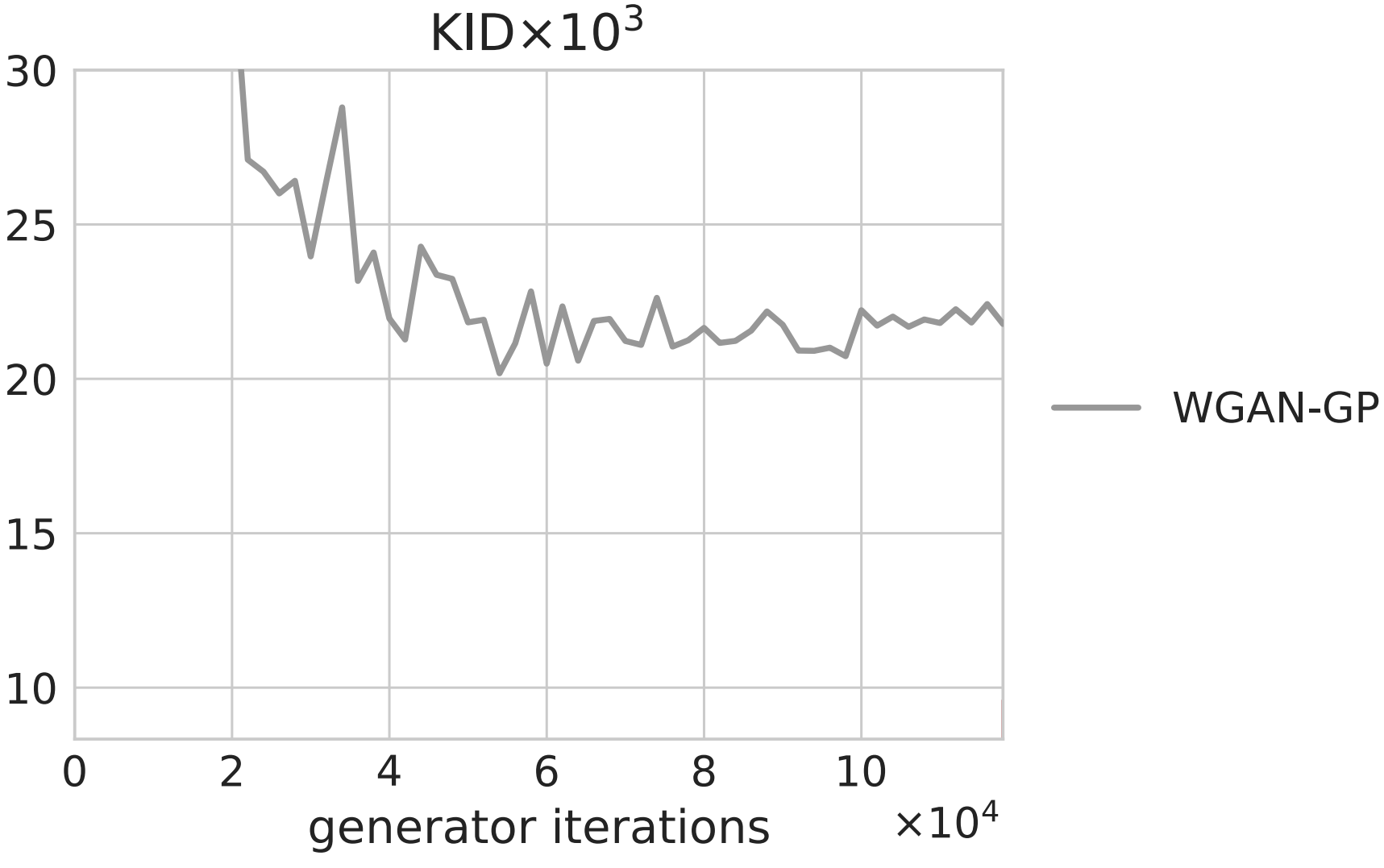
KID: 0.006

WGAN-GP

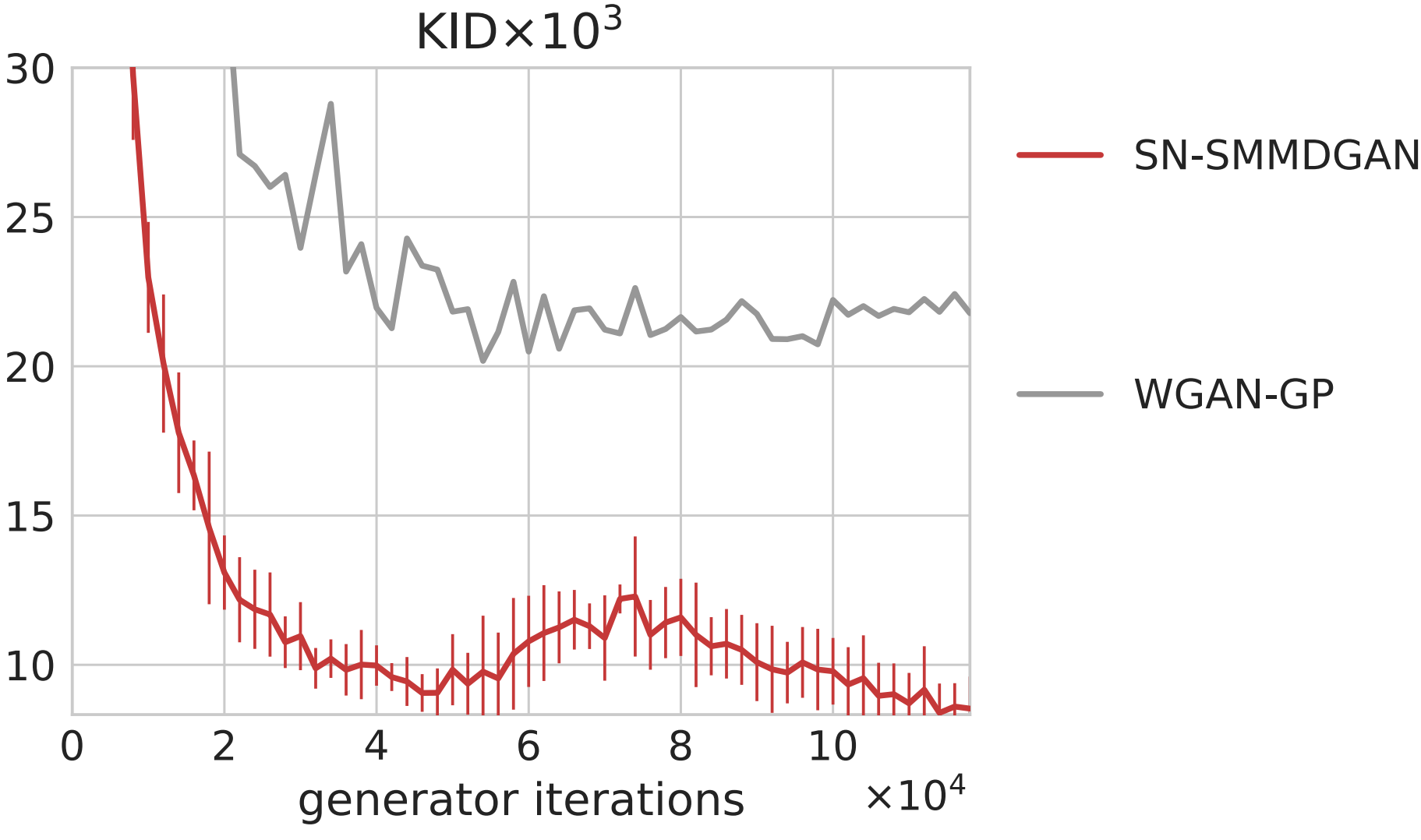


KID: 0.022

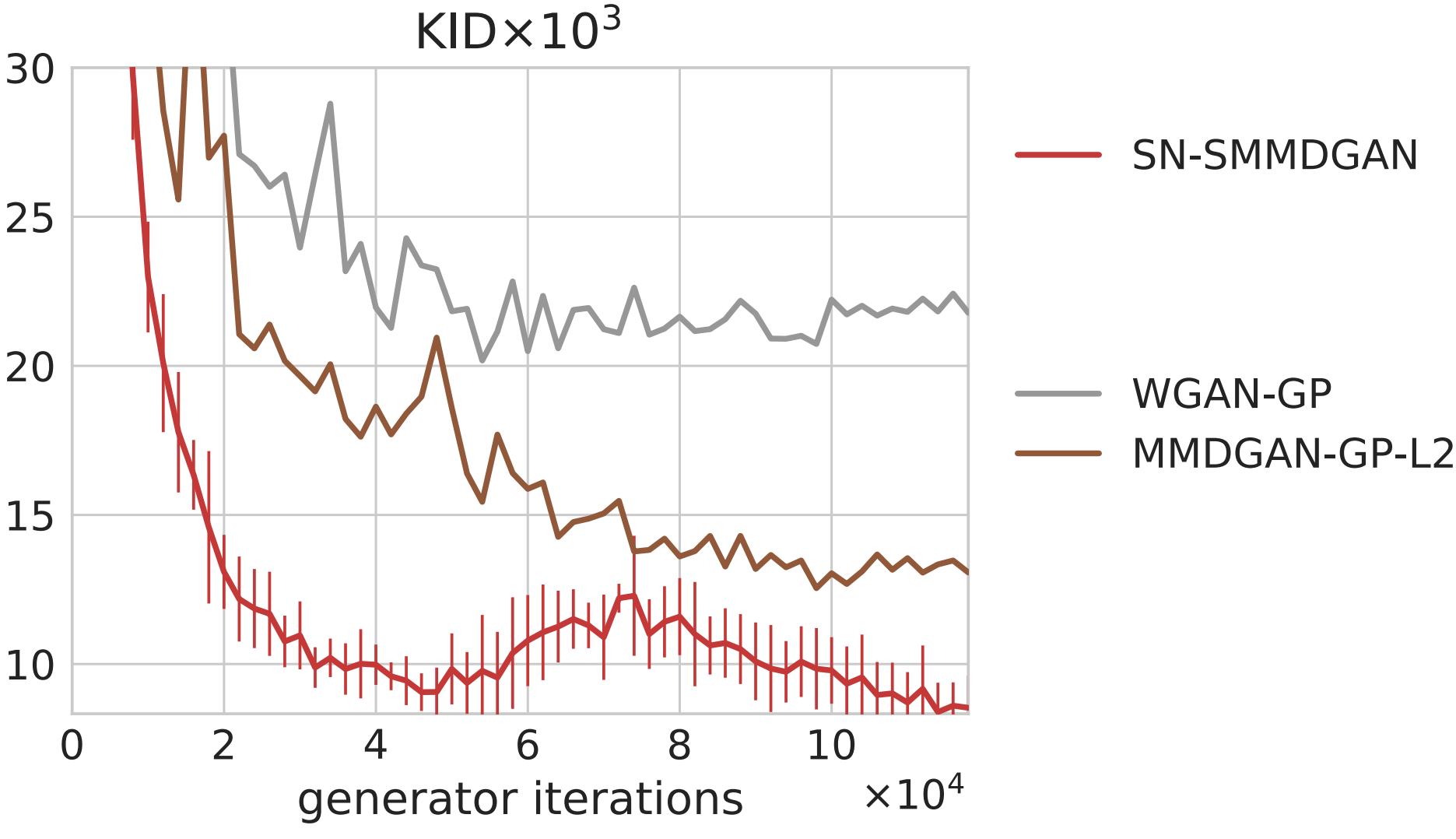
Training process on CelebA



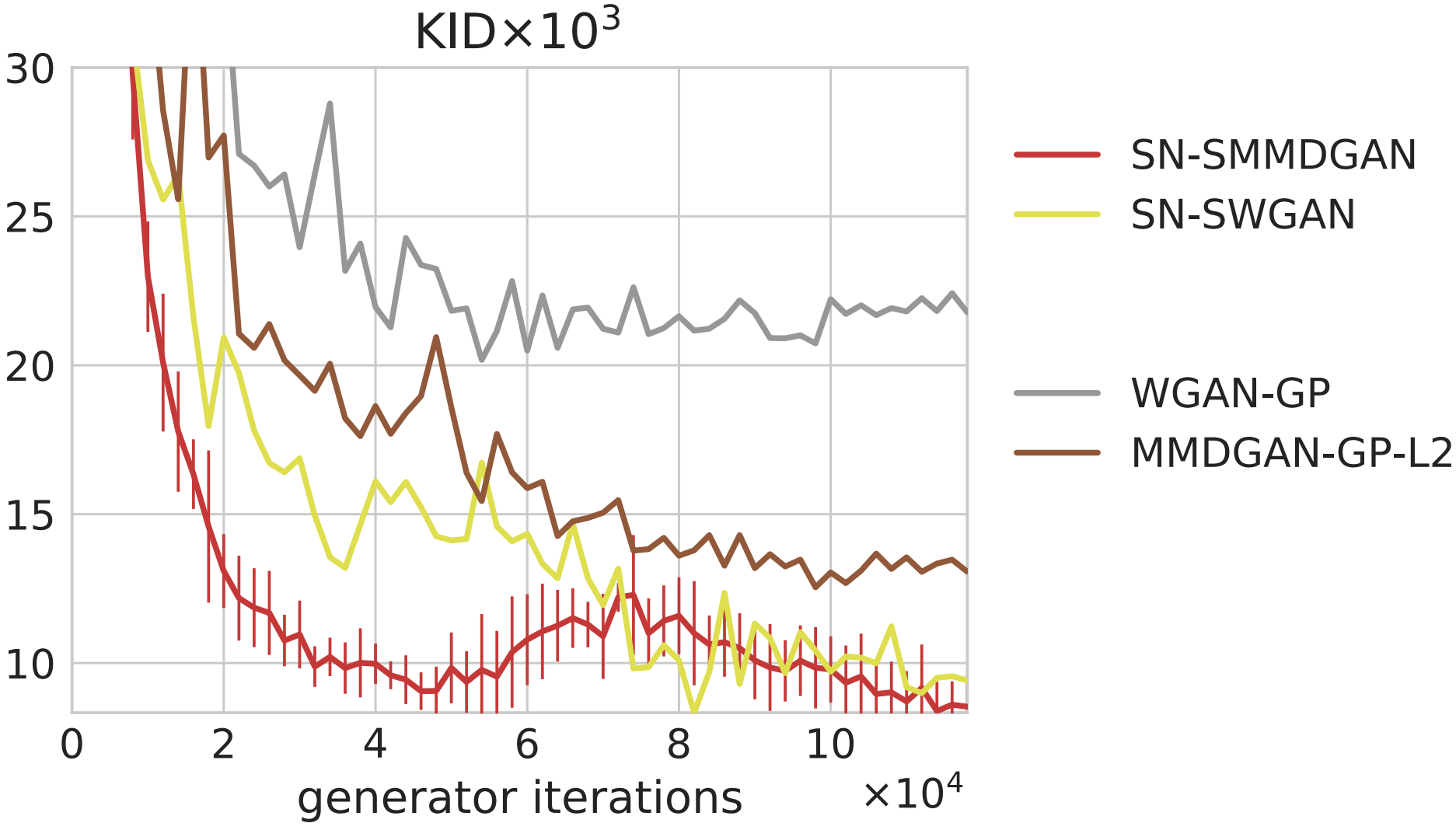
Training process on CelebA



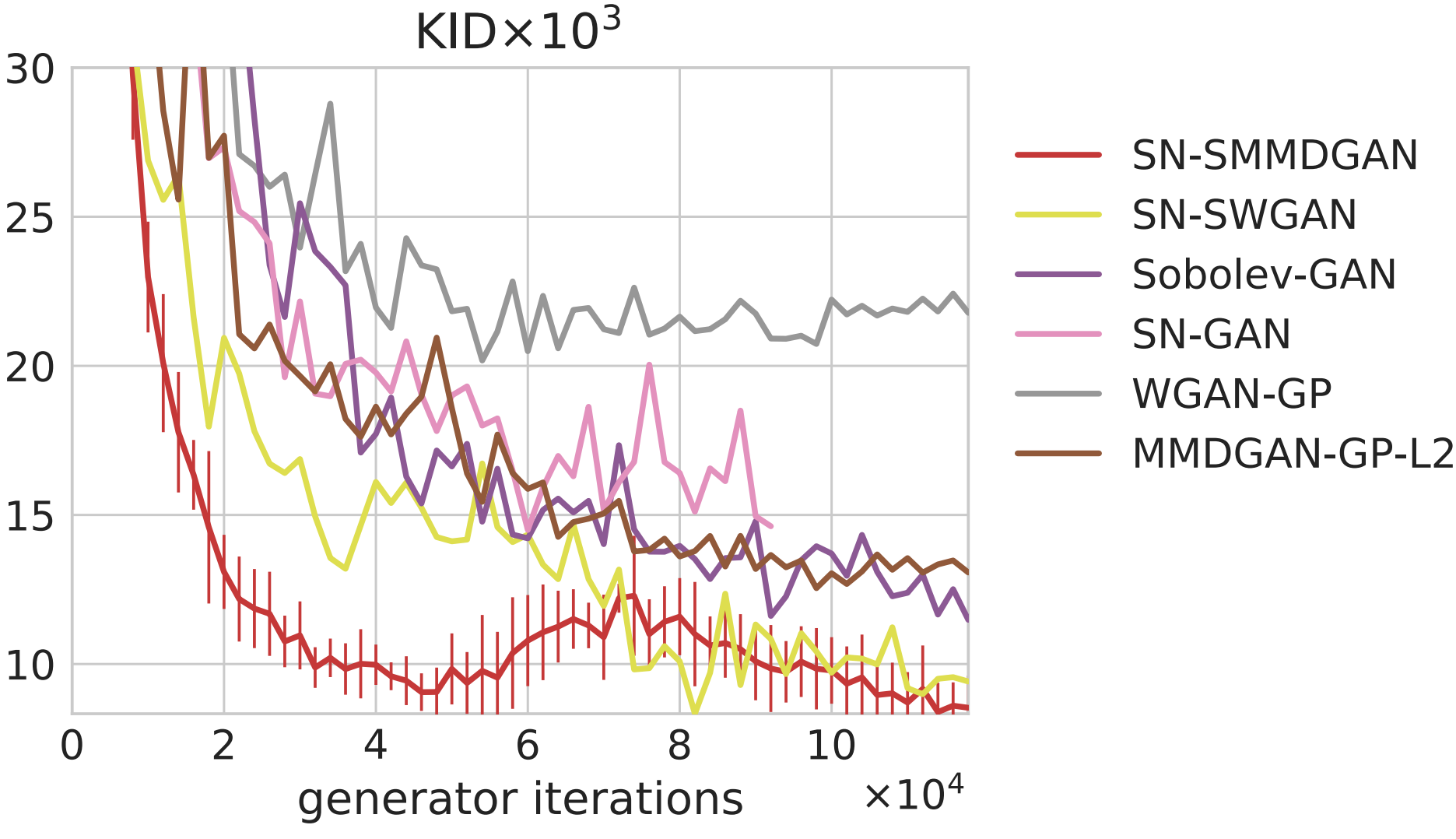
Training process on CelebA



Training process on CelebA



Training process on CelebA



Model on 64×64 ImageNet

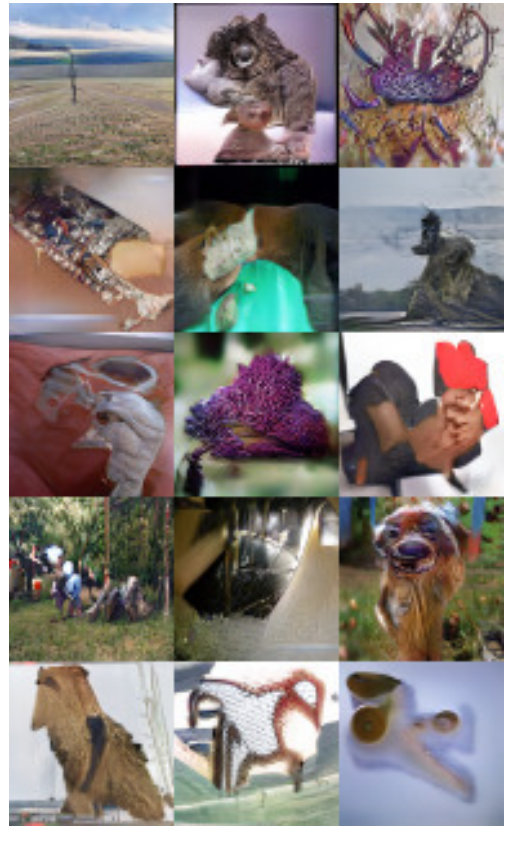
SN-SMMDGAN

KID: 0.035

SN-GAN

KID: 0.044

BGAN



KID: 0.047

Recap

- Can train generative models by minimizing a *flexible*, *smooth* distance between distributions
- Combine kernels with gradient penalties
- Strong practical results, some understanding of theory

Demystifying MMD GANs
Bińkowski*, Sutherland*, Arbel, and Gretton [ICLR 2018]

On Gradient Regularizers for MMD GANs
Arbel*, Sutherland*, Bińkowski, and Gretton [NeurIPS 2018]

Links + code: see `dougal.me`

Recap

- Can train generative models by minimizing a *flexible*, *smooth* distance between distributions
- Combine kernels with gradient penalties
- Strong practical results, some understanding of theory
- But haven't totally solved GANs yet!

Demystifying MMD GANs
Bińkowski*, Sutherland*, Arbel, and Gretton [ICLR 2018]

On Gradient Regularizers for MMD GANs
Arbel*, Sutherland*, Bińkowski, and Gretton [NeurIPS 2018]

Links + code: see `dougal.me`

Recap

- Can train generative models by minimizing a *flexible*, *smooth* distance between distributions
- Combine kernels with gradient penalties
- Strong practical results, some understanding of theory
- But haven't totally solved GANs yet!

Demystifying MMD GANs
Bińkowski*, Sutherland*, Arbel, and Gretton [ICLR 2018]

On Gradient Regularizers for MMD GANs
Arbel*, Sutherland*, Bińkowski, and Gretton [NeurIPS 2018]

Links + code: see `dougal.me`

Thanks!

Backup slides

MMD GANs versus WGANs

- Linear-kernel MMD GAN, $k(x, y) = \phi(x)\phi(y)$:

$$\text{loss} = \left| \mathop{\mathbb{E}}_{\mathbb{P}} \phi(X) - \mathop{\mathbb{E}}_{\mathbb{Q}} \phi(Y) \right|$$

$$f(t) = \text{sign}\left(\mathop{\mathbb{E}}_{\mathbb{P}} \phi(X) - \mathop{\mathbb{E}}_{\mathbb{Q}} \phi(Y) \right) \phi(t)$$

- WGAN has:

$$\text{loss} = \mathop{\mathbb{E}}_{\mathbb{P}} \phi(X) - \mathop{\mathbb{E}}_{\mathbb{Q}} \phi(Y)$$

$$f(t) = \phi(t)$$

MMD GANs versus WGANs

- Linear-kernel MMD GAN, $k(x, y) = \phi(x)\phi(y)$:

$$\text{loss} = \left|\mathop{\mathbb{E}}_{\mathbb{P}} \phi(X) - \mathop{\mathbb{E}}_{\mathbb{Q}} \phi(Y)\right|$$

$$f(t) = \text{sign}\left(\mathop{\mathbb{E}}_{\mathbb{P}} \phi(X) - \mathop{\mathbb{E}}_{\mathbb{Q}} \phi(Y)\right)\phi(t)$$

- WGAN has:

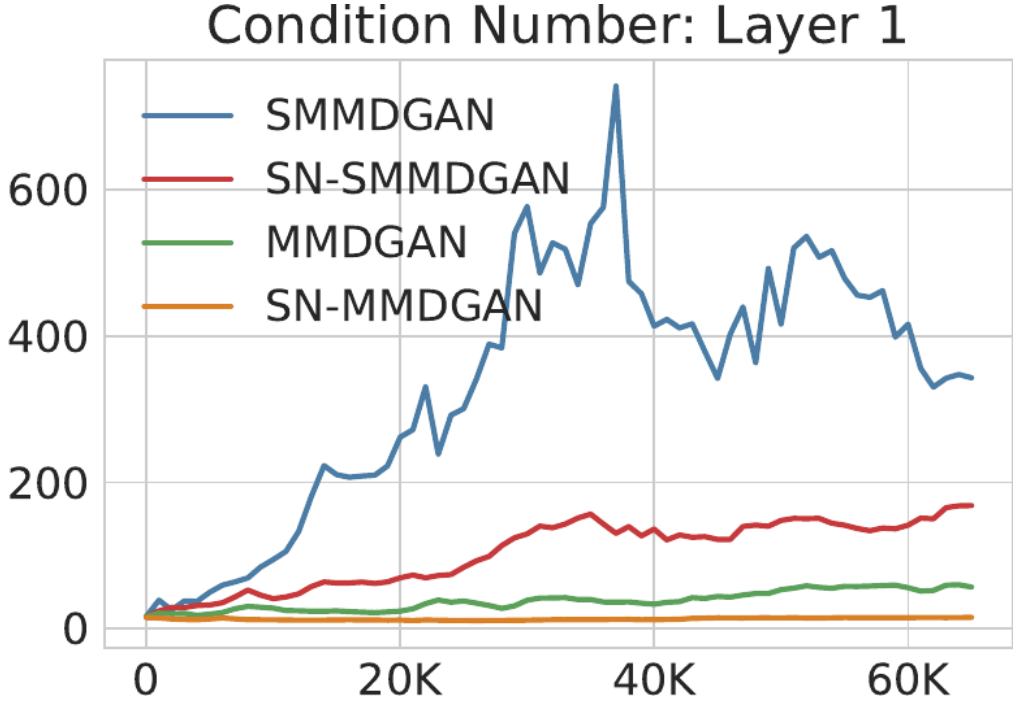
$$\text{loss} = \mathop{\mathbb{E}}_{\mathbb{P}} \phi(X) - \mathop{\mathbb{E}}_{\mathbb{Q}} \phi(Y)$$

$$f(t) = \phi(t)$$

- Linear-kernel MMD GAN-GP and WGAN-GP almost the same
- MMD GAN “offloads” some of the critic's work to closed-form optimization in the RKHS

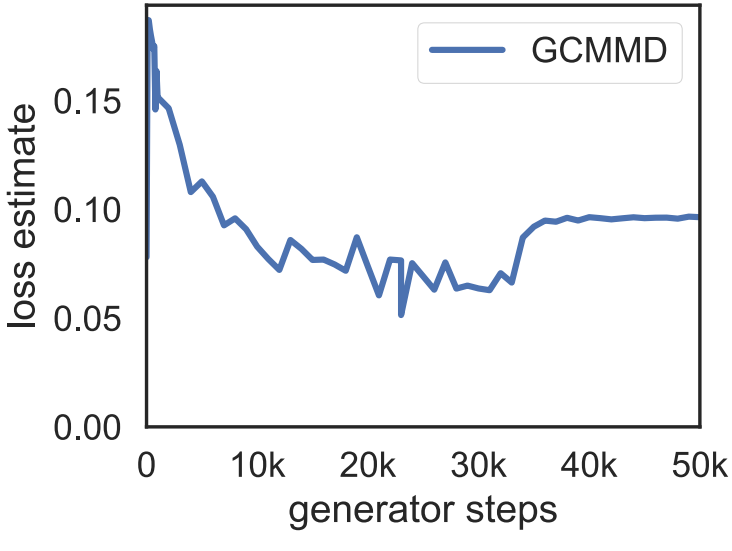
Keeping weight condition numbers bounded

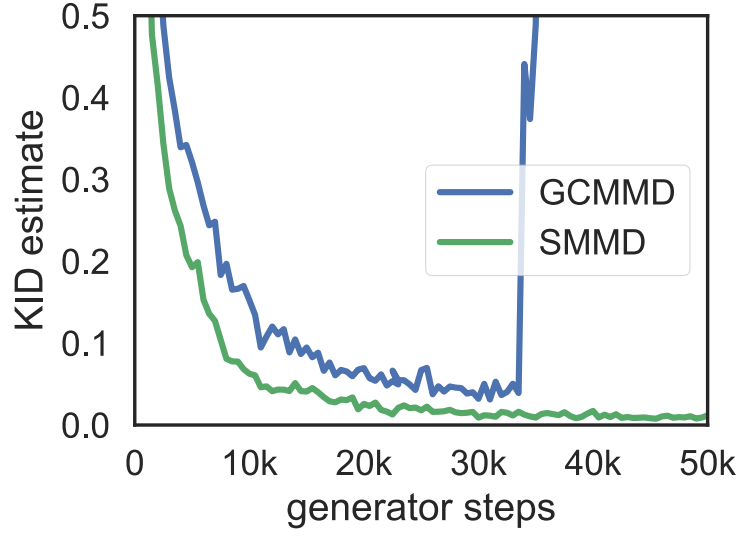
- Spectral parameterization [Miyato+ ICLR-18]:
- $W = \gamma \bar{W} / \|\bar{W}\|_{\text{op}}$; learn γ and $\bar{W}$ freely
- Encourages diversity without limiting representation



Rank collapse

- Occasional optimization failure without spectral param:
  - Generator doing reasonably well
  - Critic filters become low-rank
  - Generator corrects it by breaking everything else
  - Generator gets stuck





What if we just did spectral normalization?

- $W = \bar{W}/\|\bar{W}\|_{\text{op}}$, so that $\|W\|_{\text{op}} = 1$, $\|\phi_\psi\|_L \leq 1$
- Works well for original GANs [Miyato+ ICLR-18]
- ...but doesn't work at all as only constraint in a WGAN
- Limits representation too much
  - In DiracGAN, only allows bandwidth 1
  - $\|x \mapsto \sigma(W_n \cdots \sigma(W_1 x))\|_L \ll \|W_n\|_{\text{op}} \cdots \|W_1\|_{\text{op}}$

Continuity theorem proof

- $k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi(y))$ means
  $d_\psi(x, y) = \|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_{k_\psi}} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \|x - y\|$

Continuity theorem proof

- $k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi(y))$ means
  $d_\psi(x, y) = \|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_{k_\psi}} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \|x - y\|$
- Can show $\text{MMD}_\psi \leq \mathcal{W}_{d_\psi} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \mathcal{W}$

Continuity theorem proof

- $k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi(y))$ means
  $d_\psi(x, y) = \|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_{k_\psi}} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \|x - y\|$
- Can show $\text{MMD}_\psi \leq \mathcal{W}_{d_\psi} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \mathcal{W}$
- By assumption on k_{top}, $\sigma^{-2}_{\mathbb{S},k,\lambda} \geq \gamma^2_{k_{\text{top}}} \mathbb{E}[\|\nabla \phi_\psi(\tilde{X})\|_F^2]$

Continuity theorem proof

- $k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi(y))$ means
  $d_\psi(x, y) = \|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_{k_\psi}} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \|x - y\|$
- Can show $\text{MMD}_\psi \leq \mathcal{W}_{d_\psi} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \mathcal{W}$
- By assumption on k_{top}, $\sigma^{-2}_{\mathbb{S},k,\lambda} \geq \gamma^2_{k_{\text{top}}} \mathbb{E}[\|\nabla \phi_\psi(\tilde{X})\|_F^2]$
- $\text{SMMD}^2 \leq \dfrac{L^2_{k_{\text{top}}} \|\phi_\psi\|^2_{\text{Lip}}}{\gamma^2_{k_{\text{top}}} \mathbb{E} \|\nabla_{\tilde{X}} \phi_\psi(\tilde{X})\|_F^2} \mathcal{W}$

Continuity theorem proof

- $k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi(y))$ means
  $d_\psi(x, y) = \|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_{k_\psi}} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \|x - y\|$
- Can show $\text{MMD}_\psi \leq \mathcal{W}_{d_\psi} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \mathcal{W}$
- By assumption on k_{top}, $\sigma^{-2}_{\mathbb{S},k,\lambda} \geq \gamma^2_{k_{\text{top}}} \mathbb{E}[\|\nabla \phi_\psi(\tilde{X})\|_F^2]$
- $\text{SMMD}^2 \leq \dfrac{L^2_{k_{\text{top}}} \|\phi_\psi\|^2_{\text{Lip}}}{\gamma^2_{k_{\text{top}}} \mathbb{E} \|\nabla_{\tilde{X}} \phi_\psi(\tilde{X})\|_F^2} \mathcal{W}$
- Because Leaky-ReLU, $\phi_\psi(X) = \alpha(\psi)\phi_{\bar{\psi}}(X)$, $\|\phi_{\bar{\psi}}\|_{\text{Lip}} \leq 1$

Continuity theorem proof

- $k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi(y))$ means
  $d_\psi(x, y) = \|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_{k_\psi}} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \|x - y\|$
- Can show $\text{MMD}_\psi \leq \mathcal{W}_{d_\psi} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \mathcal{W}$
- By assumption on k_{top}, $\sigma^{-2}_{\mathbb{S},k,\lambda} \geq \gamma^2_{k_{\text{top}}} \mathbb{E}[\|\nabla \phi_\psi(\tilde{X})\|_F^2]$
- $\text{SMMD}^2 \leq \dfrac{L^2_{k_{\text{top}}} \|\phi_\psi\|^2_{\text{Lip}}}{\gamma^2_{k_{\text{top}}} \mathbb{E} \|\nabla_{\tilde{X}} \phi_\psi(\tilde{X})\|_F^2} \mathcal{W}$
- Because Leaky-ReLU, $\phi_\psi(X) = \alpha(\psi)\phi_{\bar{\psi}}(X)$, $\|\phi_{\bar{\psi}}\|_{\text{Lip}} \leq 1$
- For Lebesgue-almost all $\tilde{X}$, $\|\nabla_{\tilde{X}} \phi_{\bar{\psi}}(\tilde{X})\|_F^2 \geq \dfrac{d_{\text{top}} \alpha^L}{\kappa^L}$

Continuity theorem proof

- $k_\psi(x, y) = k_{\text{top}}(\phi_\psi(x), \phi(y))$ means
  $d_\psi(x, y) = \|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_{k_\psi}} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \|x - y\|$
- Can show $\text{MMD}_\psi \leq \mathcal{W}_{d_\psi} \leq L_{k_{\text{top}}} \|\phi_\psi\|_{\text{Lip}} \mathcal{W}$
- By assumption on k_{top}, $\sigma^{-2}_{\mathbb{S},k,\lambda} \geq \gamma^2_{k_{\text{top}}} \mathbb{E}[\|\nabla \phi_\psi(\tilde{X})\|_F^2]$
- $\text{SMMD}^2 \leq \frac{L^2_{k_{\text{top}}} \|\phi_\psi\|^2_{\text{Lip}}}{\gamma^2_{k_{\text{top}}} \mathbb{E} \|\nabla_{\tilde{X}} \phi_\psi(\tilde{X})\|_F^2} \mathcal{W} \leq \frac{L^2_{k_{\text{top}}} \kappa^L}{\gamma^2_{k_{\text{top}}} d_{\text{top}} \alpha^L} \mathcal{W}$
- Because Leaky-ReLU, $\phi_\psi(X) = \alpha(\psi)\phi_{\bar{\psi}}(X)$, $\|\phi_{\bar{\psi}}\|_{\text{Lip}} \leq 1$
- For Lebesgue-almost all $\tilde{X}$, $\|\nabla_{\tilde{X}} \phi_{\bar{\psi}}(\tilde{X})\|_F^2 \geq \frac{d_{\text{top}} \alpha^L}{\kappa^L}$

Implicit generative model evaluation

- No likelihoods, so...how to compare models?

Implicit generative model evaluation

- No likelihoods, so...how to compare models?
- Main approach:
  look at a bunch of pictures and see if they're pretty or not

Implicit generative model evaluation

- No likelihoods, so...how to compare models?
- Main approach:
  look at a bunch of pictures and see if they're pretty or not
  - Easy to find (really) bad samples

Implicit generative model evaluation

- No likelihoods, so...how to compare models?
- Main approach:
  look at a bunch of pictures and see if they're pretty or not
  - Easy to find (really) bad samples
  - Hard to see if modes are missing / have wrong probabilities

Implicit generative model evaluation

- No likelihoods, so...how to compare models?
- Main approach:
  look at a bunch of pictures and see if they're pretty or not
  - Easy to find (really) bad samples
  - Hard to see if modes are missing / have wrong probabilities
  - Hard to compare models beyond certain threshold

Implicit generative model evaluation

- No likelihoods, so...how to compare models?
- Main approach:
  look at a bunch of pictures and see if they're pretty or not
  - Easy to find (really) bad samples
  - Hard to see if modes are missing / have wrong probabilities
  - Hard to compare models beyond certain threshold
- Need better, quantitative methods

Implicit generative model evaluation

- No likelihoods, so...how to compare models?
- Main approach:
  look at a bunch of pictures and see if they're pretty or not
  - Easy to find (really) bad samples
  - Hard to see if modes are missing / have wrong probabilities
  - Hard to compare models beyond certain threshold
- Need better, quantitative methods
- Our method: Kernel Inception Distance (KID)
  - $\mathrm{MMD}_k(\mathbb{P}_{\mathrm{data}}, \mathbb{Q}_\theta)^2$, k cubic on pretrained Inception rep

Implicit generative model evaluation

- No likelihoods, so...how to compare models?
- Main approach:
  look at a bunch of pictures and see if they're pretty or not
  - Easy to find (really) bad samples
  - Hard to see if modes are missing / have wrong probabilities
  - Hard to compare models beyond certain threshold
- Need better, quantitative methods
- Our method: Kernel Inception Distance (KID)
  - $\mathrm{MMD}_k(\mathbb{P}_{\mathrm{data}}, \mathbb{Q}_\theta)^2$, k cubic on pretrained Inception rep
  - `tf.contrib.gan.eval.kernel_inception_distance`

Inception score [Salimans+ NIPS-16]

- Previously standard quantitative method
- Based on ImageNet classifier label predictions
  - Classifier should be confident on individual images
  - Predicted labels should be diverse across sample

Inception score [Salimans+ NIPS-16]

- Previously standard quantitative method
- Based on ImageNet classifier label predictions
  - Classifier should be confident on individual images
  - Predicted labels should be diverse across sample
- No notion of target distribution $\mathbb{P}_{\mathbf{data}}$

Inception score [Salimans+ NIPS-16]

- Previously standard quantitative method
- Based on ImageNet classifier label predictions
  - Classifier should be confident on individual images
  - Predicted labels should be diverse across sample
- No notion of target distribution $\mathbb{P}_{\mathbf{data}}$
- Scores completely meaningless on LSUN, Celeb-A, SVHN, ...

Inception score [Salimans+ NIPS-16]

- Previously standard quantitative method
- Based on ImageNet classifier label predictions
  - Classifier should be confident on individual images
  - Predicted labels should be diverse across sample
- No notion of target distribution $\mathbb{P}_{\text{data}}$
- Scores completely meaningless on LSUN, Celeb-A, SVHN, ...
- Not great on CIFAR-10 either

Inception score [Salimans+ NIPS-16]

- Previously standard quantitative method
- Base
  - Cl
  - Pr
- No r
- Scor
- Not

s

VHN, ...

A Note on the Inception Score

Shane Barratt, Rishi Sharma

(Submitted on 6 Jan 2018)

Deep generative models are powerful tools that have produced impressive results in recent years. These advances have been for the most part empirically driven, making it essential that we use high quality evaluation metrics. In this paper, we provide new insights into the Inception Score, a recently proposed and widely used evaluation metric for generative models, and demonstrate that it fails to provide useful guidance when comparing models. We discuss both suboptimalities of the metric itself and issues with its application. Finally, we call for researchers to be more systematic and careful when evaluating and comparing generative models, as the advancement of the field depends upon it.

Fréchet Inception Distance (FID) [Heusel+ NIPS-17]

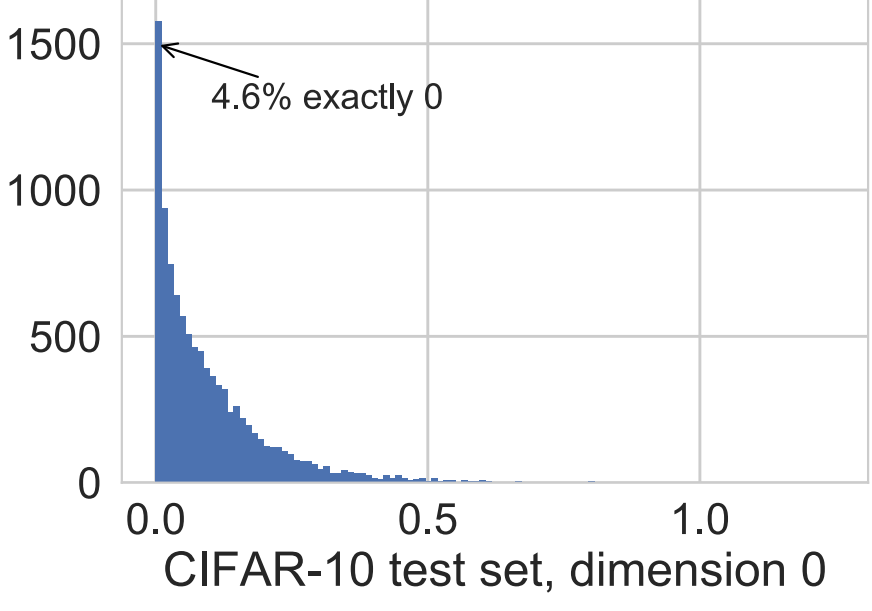
- Fit normals to Inception hidden layer activations of $\mathbb{P}$ and $\mathbb{Q}$
- Compute Fréchet (Wasserstein-2) distance between fits

Fréchet Inception Distance (FID) [Heusel+ NIPS-17]

- Fit normals to Inception hidden layer activations of $\mathbb{P}$ and $\mathbb{Q}$
- Compute Fréchet (Wasserstein-2) distance between fits
- Meaningful on not-ImageNet datasets

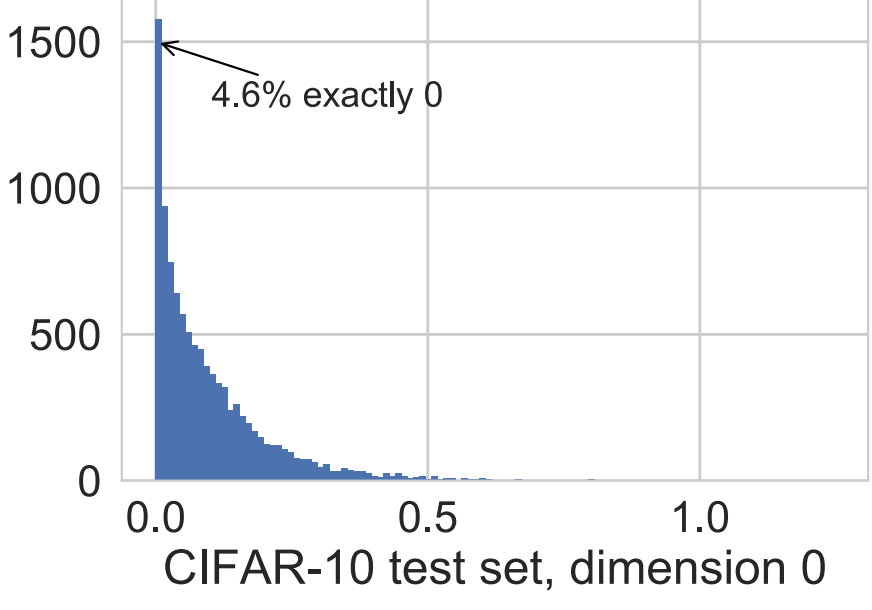
Fréchet Inception Distance (FID) [Heusel+ NIPS-17]

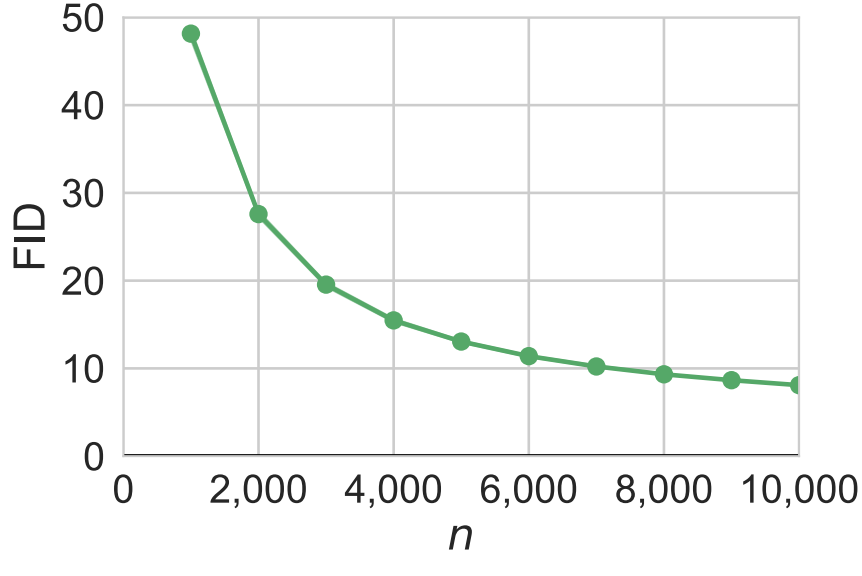
- Fit normals to Inception hidden layer activations of $\mathbb{P}$ and $\mathbb{Q}$
- Compute Fréchet (Wasserstein-2) distance between fits
- Meaningful on not-ImageNet datasets



Fréchet Inception Distance (FID) [Heusel+ NIPS-17]

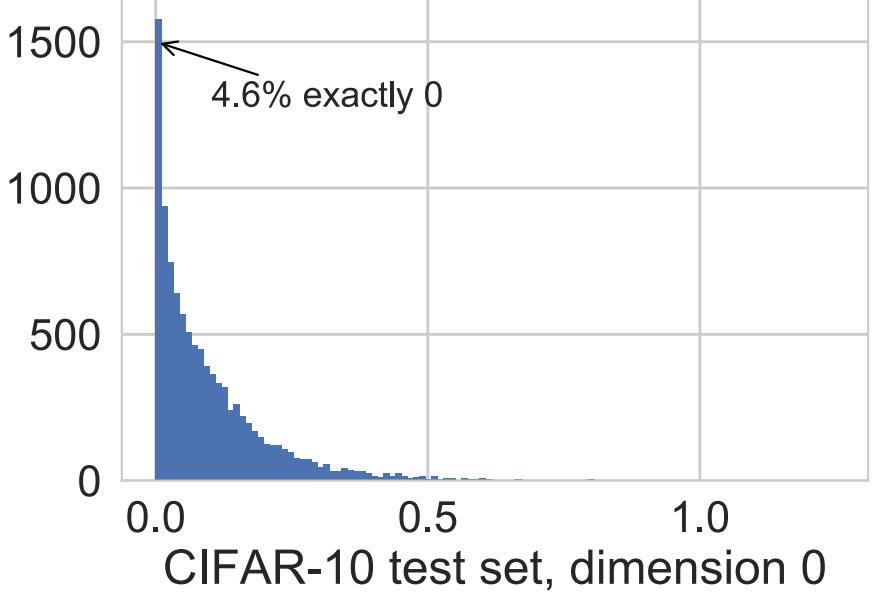
- Fit normals to Inception hidden layer activations of $\mathbb{P}$ and $\mathbb{Q}$
- Compute Fréchet (Wasserstein-2) distance between fits
- Meaningful on not-ImageNet datasets
- Estimator extremely biased, tiny variance

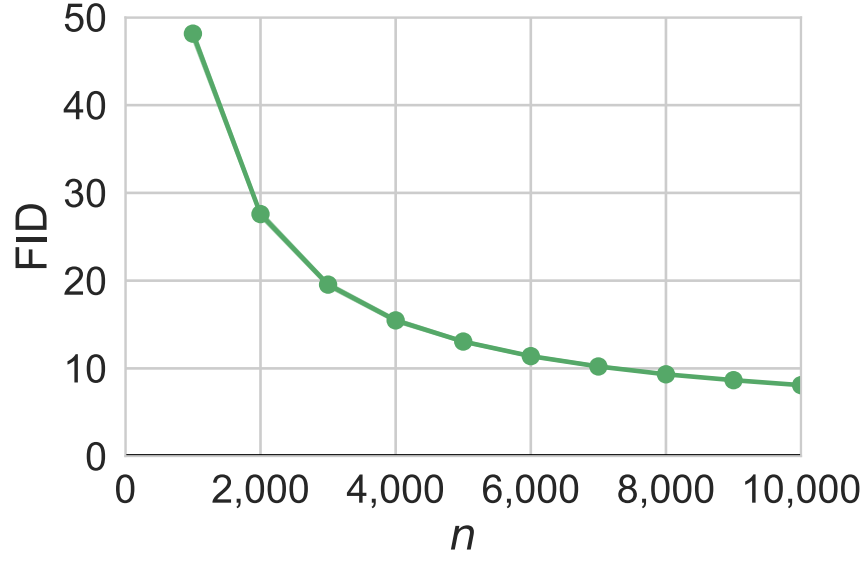




Fréchet Inception Distance (FID) [Heusel+ NIPS-17]

- Fit normals to Inception hidden layer activations of $\mathbb{P}$ and $\mathbb{Q}$
- Compute Fréchet (Wasserstein-2) distance between fits
- Meaningful on not-ImageNet datasets
- Estimator extremely biased, tiny variance
- $\mathbf{FID}(\mathbb{P}_1, \mathbb{Q}) < \mathbf{FID}(\mathbb{P}_2, \mathbb{Q}), \mathbb{E}\,\mathbf{FID}(\hat{\mathbb{P}}_1, \mathbb{Q}) > \mathbb{E}\,\mathbf{FID}(\hat{\mathbb{P}}_2, \mathbb{Q})$





Fréchet Inception Distance (FID) [Heusel+ NIPS-17]

- Fit normals to Inception hidden layer activations of $\mathbb{P}$ and $\mathbb{Q}$

takerum commented 11 days ago • edited

Member

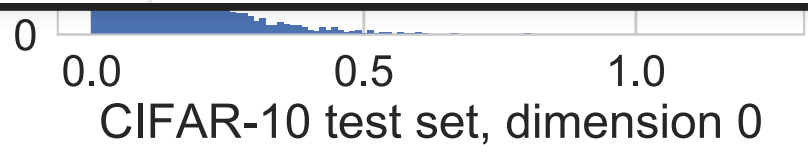
> Table 2 in SN-GAN paper, How do you calculate the FID score of real data? (7.8)

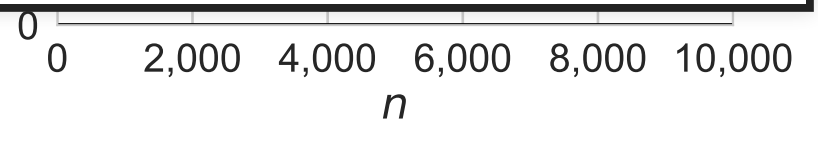
We sample 10000 images on test set and 5000 images on training set and calculate FID on the two sets of the images.

> Table 2 How many generated samples are used for getting FID or Inception scores? 5000 or 50000?

> Your paper used 5000 samples to compute FID and Inception score. But in improved-gan paper, they use 50000 samples to get Inceptions scores.

Both of the original paper and our paper use 50,000 samples for calculating the "mean" and "std" of the inception scores.
The original paper and our paper calculate inception score with 5,000 samples and repeated 10 times to estimate the mean and variance of inception scores on each independently generated set of images.
For FID, we calculate it with 5,000 samples and report the value, because we found that the variation of FID within independent sets is very small compared to the value of FID.



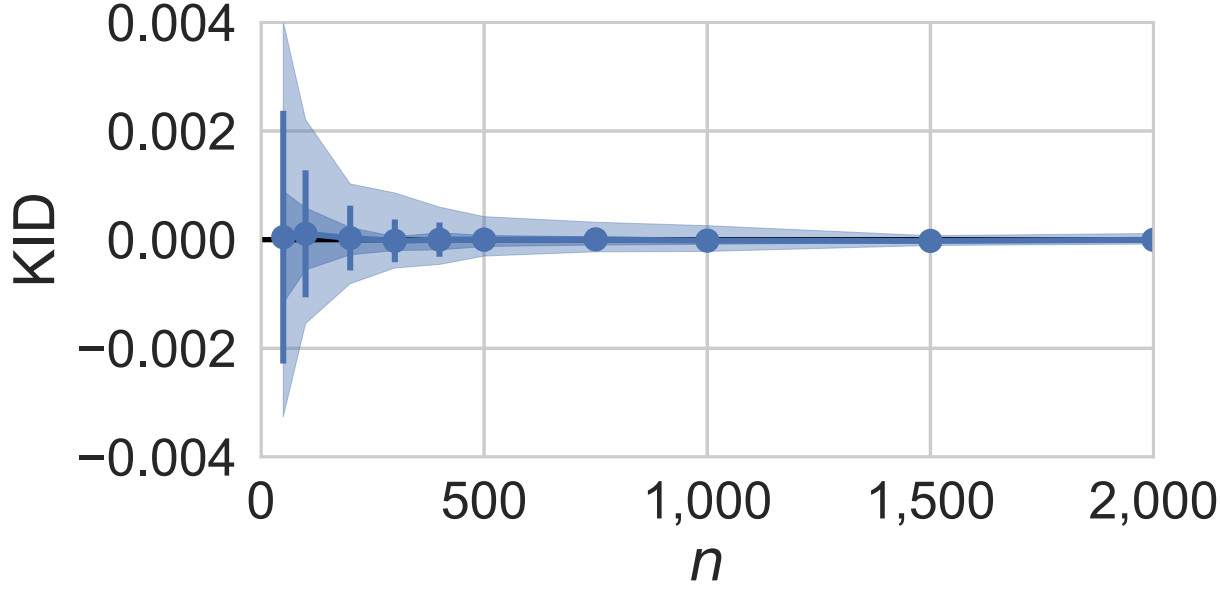


New method: Kernel Inception Distance (KID)

- $\widehat{\mathrm{MMD}}^2$ between Inception hidden layer activations
- Use default polynomial kernel: $k(x, y) = \left(\frac{1}{d}\langle x, y\rangle + 1\right)^3$

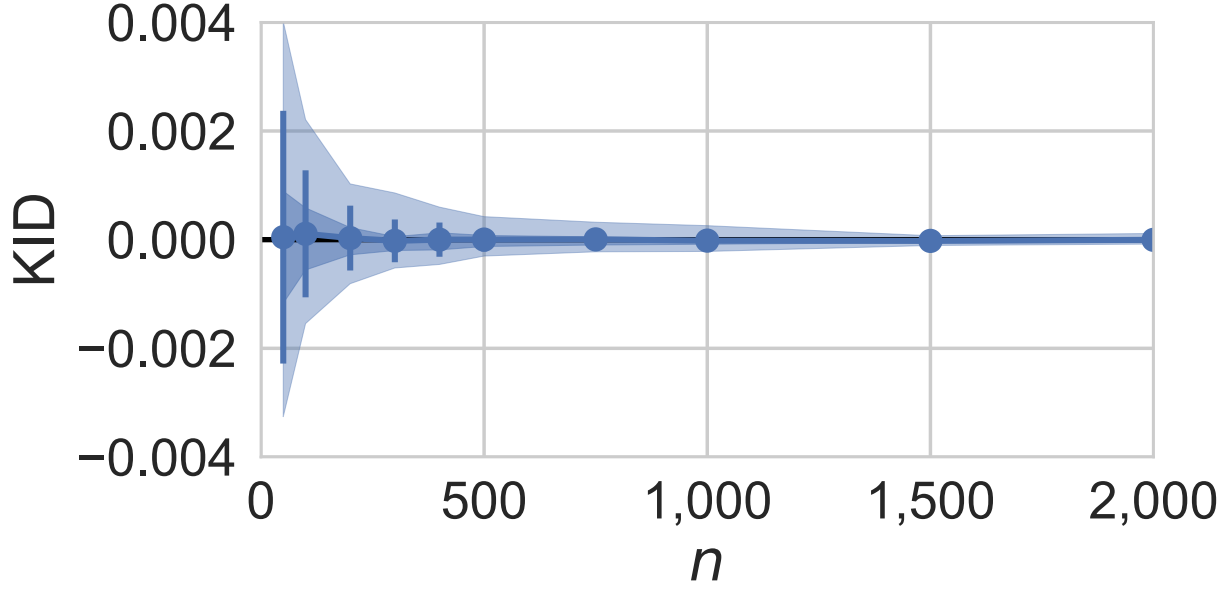
New method: Kernel Inception Distance (KID)

- $\widehat{\mathrm{MMD}}^2$ between Inception hidden layer activations
- Use default polynomial kernel: $k(x, y) = \left(\frac{1}{d}\langle x, y\rangle + 1\right)^3$
- Unbiased estimator, reasonable with few samples



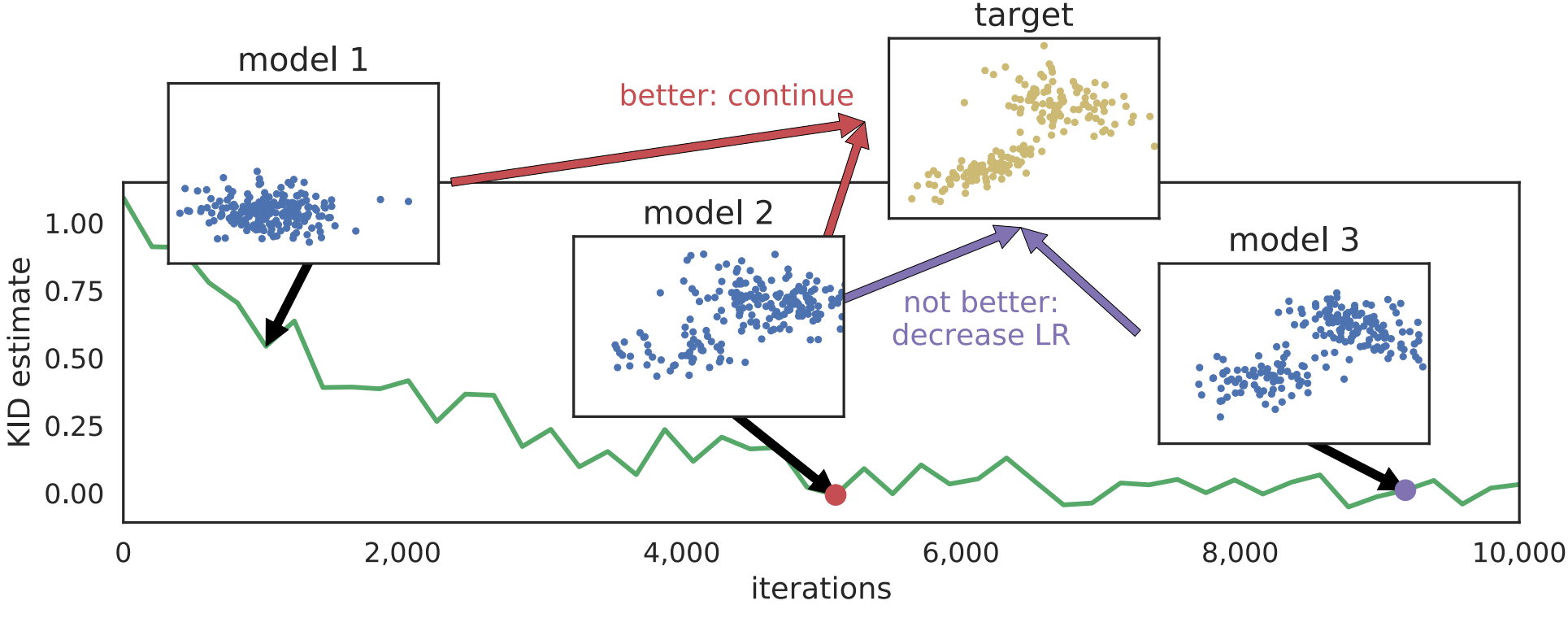
New method: Kernel Inception Distance (KID)

- $\widehat{\mathrm{MMD}}^2$ between Inception hidden layer activations
- Use default polynomial kernel: $k(x, y) = \left(\frac{1}{d}\langle x, y\rangle + 1\right)^3$
- Unbiased estimator, reasonable with few samples
- In `tensorflow.contrib.gan.eval` (tensorflow#21066)



Automatic learning rate adaptation with KID

- Models need appropriate learning rate schedule to work well
- Automate with three-sample MMD test [Bounliphone+ ICLR-16]:



Controlling critic complexity