Kernel Distances for Deep Generative Models

Michael Arbel*[1] Dougal J. Sutherland*[1]
Mikołaj Bińkowski[2] Arthur Gretton[1]

[1]Gatsby Computational Neuroscience Unit
University College London

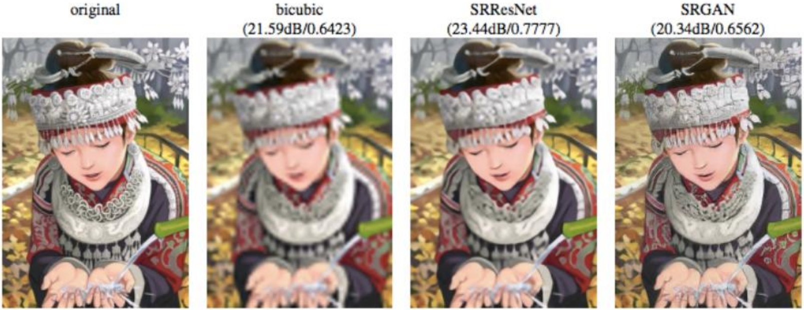
[2]Department of Mathematics
Imperial College London

March 28, 2019

Generative Adversarial Networks

Many successful applications:

- Single-image super-resolution

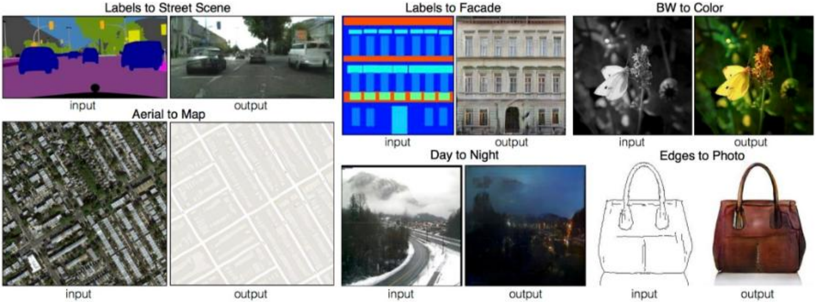


Ledig et al 2015

Generative Adversarial Networks

Many successful applications:

- Image generation tasks: Image to image translation



Isola et al 2016

Generative Adversarial Networks

Many successful applications:

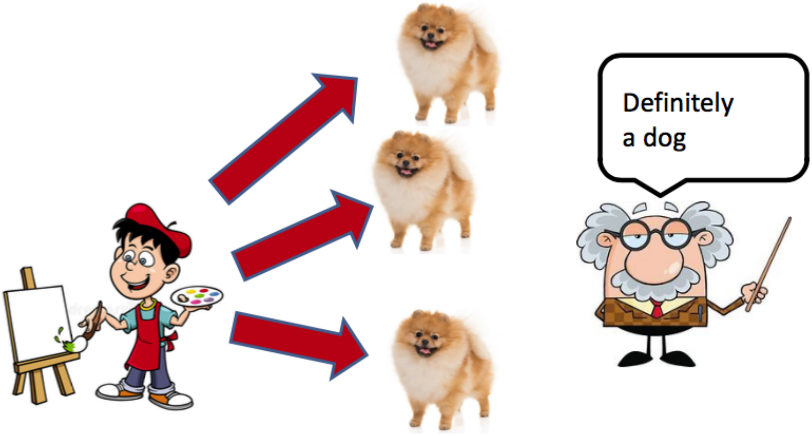
- Text to image generation

Zhang et al 2016

GAN's are hard to train!

Several failure cases

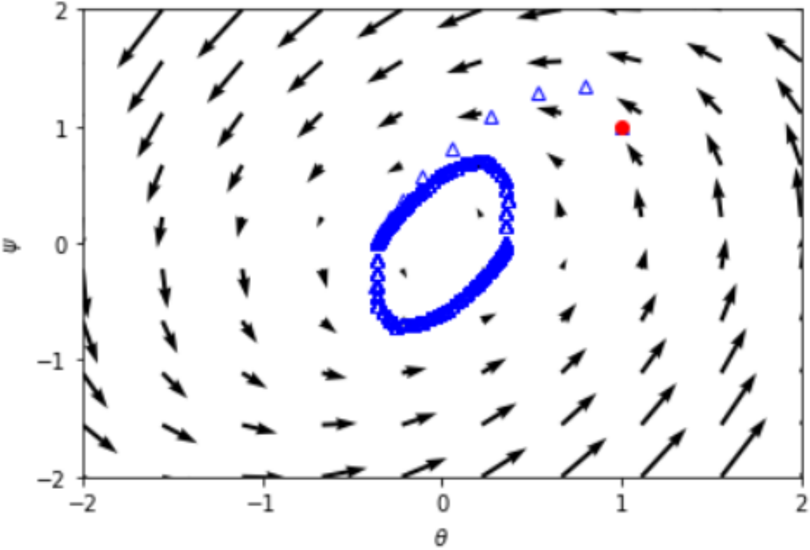
- Mode collapse:



GAN's are hard to train!

Several failure cases

- Oscillations: [Mescheder et al., 2018, Balduzzi et al., 2018]



GAN's are hard to train!

Different angles:

- Optimization: [Roth et al., 2017, Mescheder et al., 2018]
- Game theory: [Heusel et al., 2017, Balduzzi et al., 2018]
- Metric :
  [Arjovsky et al., 2017, Lin et al., 2018, Petzka et al., 2017]

GAN's are hard to train!

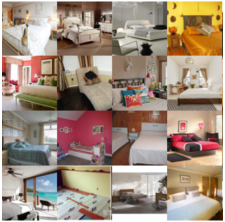
Different angles:

- Optimization: [Roth et al., 2017, Mescheder et al., 2018]
- Game theory: [Heusel et al., 2017, Balduzzi et al., 2018]
- Metric :
  [Arjovsky et al., 2017, Lin et al., 2018, Petzka et al., 2017]

- What losses for training GAN's?
- How to construct such losses?

Implicit generative models (IGM)

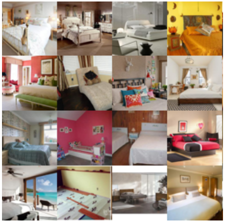
Given samples from a distribution $\mathbb{P}$ over $\mathcal{X}$, want a model that can produce new samples from $\mathbb{Q}_\theta \approx \mathbb{P}$



$X \sim \mathbb{P}$

Implicit generative models (IGM)

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

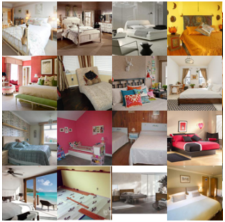


$X \sim \mathbb{P}$

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

Implicit generative models (IGM)

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



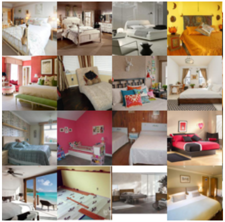
$X \sim \mathbb{P}$

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

- EGM: $_\theta$ has density $q_\theta(Y)$, no samples from Q_θ required.

Implicit generative models (IGM)

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



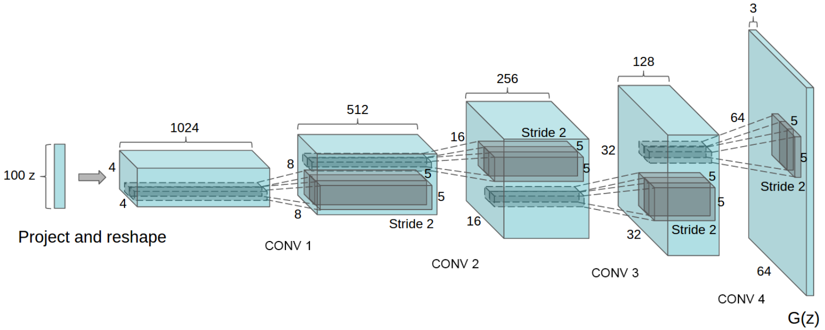
$X \sim \mathbb{P}$

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

- EGM: ${}_\theta$ has density $q_\theta(Y)$, no samples from Q_θ required.
- IGM: $Y = G_\theta(Z)$ with known distribution for Z. Training by sampling form ${}_\theta$.

Implicit generative models (IGM)

Deep network (params θ) mapping from noise $\mathbb{Z}$ to image $\mathcal{X}$



DCGAN generator [Radford et al., 2015]

$\mathbb{Z}$ is uniform on $[-1, 1]^{100}$

Choose θ by minimizing some cost

Generative Adversarial Networks [Goodfellow et al., 2014]

- Loss function:

$$L_{\mathcal{F}}(\theta) = \sup_{\phi \in \mathcal{F}} \mathbb{E}_{x \sim \mathbb{P}}[\log(\phi(x))] + \mathbb{E}_{x \sim \mathbb{Q}_\theta}[\log(1 - \phi(x))] \quad (1)$$

Generative Adversarial Networks [Goodfellow et al., 2014]

- Loss function:

$$L_{\mathcal{F}}(\theta) = \sup_{\phi \in \mathcal{F}} \mathbb{E}_{x \sim \mathbb{P}}[\log(\phi(x))] + \mathbb{E}_{x \sim \mathbb{Q}_\theta}[\log(1 - \phi(x))] \quad (1)$$

- Optimal classifer (over all possible classifiers):

$$L^*(\theta) = -\log(4) + 2JSD(\mathbb{P}, \mathbb{Q}_\theta) \quad (2)$$

Generative Adversarial Networks [Goodfellow et al., 2014]

- Loss function:

$$L_{\mathcal{F}}(\theta) = \sup_{\phi \in \mathcal{F}} \mathbb{E}_{x \sim \mathbb{P}}[\log(\phi(x))] + \mathbb{E}_{x \sim \mathbb{Q}_\theta}[\log(1 - \phi(x))] \quad (1)$$

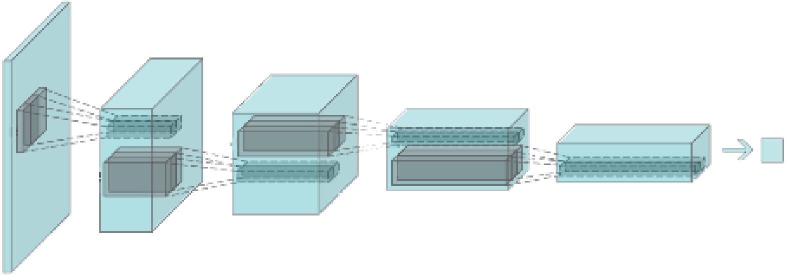
- Optimal classifer (over all possible classifiers):

$$L^*(\theta) = -\log(4) + 2JSD(\mathbb{P}, \mathbb{Q}_\theta) \quad (2)$$

- Lower-bound: $L_{\mathcal{F}}(\theta) \leq L^*(\theta)$.

Generative Adversarial Networks [Goodfellow et al., 2014]

Deep network (params ψ) mapping from image space $\mathcal{X}$ to some value



DCGAN critic [Radford et al., 2015]

Generative Adversarial Networks [Goodfellow et al., 2014]

- Min-max problem: $\min_\theta \max_\psi \mathcal{L}(\theta, \psi)$

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{X \sim \mathbb{P}}[\log \phi_\psi(X)] + \mathbb{E}_{Z \sim \mathbb{Z}}[\log(1 - \phi_\psi(G_\theta(Z)))]$$

- Solved approximately by alternating:
  - k SGD steps on ϕ
  - 1 SGD step on θ

Connection with Game Theory

Two agents:

-  (Generator G_θ): minimize $\mathcal{L}(\theta, \psi)$ in θ.
-  (Critic ϕ_ψ): maximize $\mathcal{L}(\theta, \psi)$ in ψ.

- Generator (student)

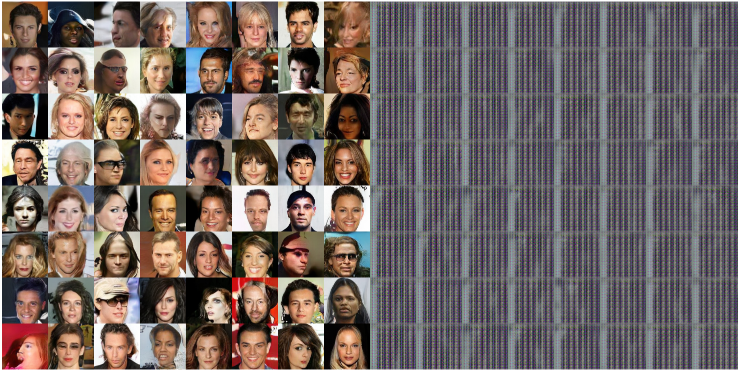
- Task: critic must teach generator to draw images (here dogs)

- Critic (teacher)

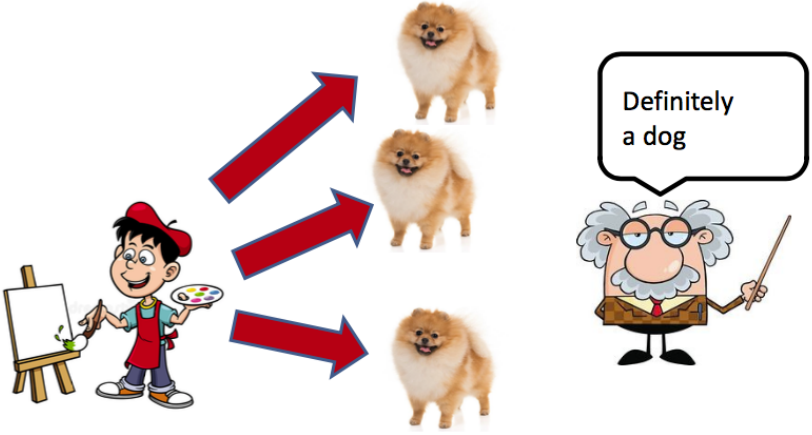
Connection with Game Theory

Not all Nash-Equilibria are of interest!!

$$\min_{\theta} \max_{\psi} L(\theta, \psi) \neq \max_{\psi} \min_{\theta} L(\theta, \psi)$$

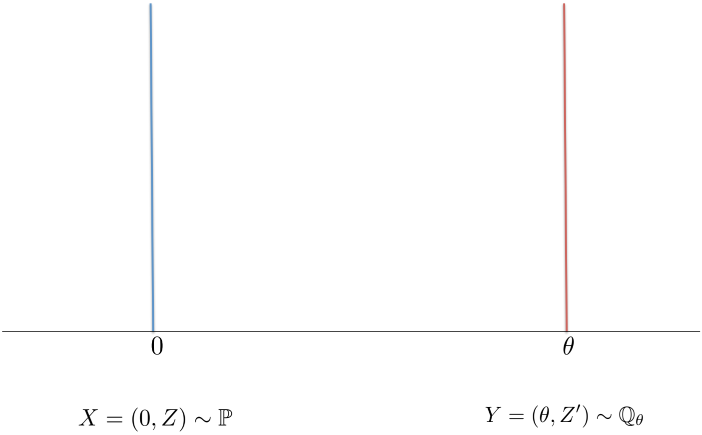


Mode collapse

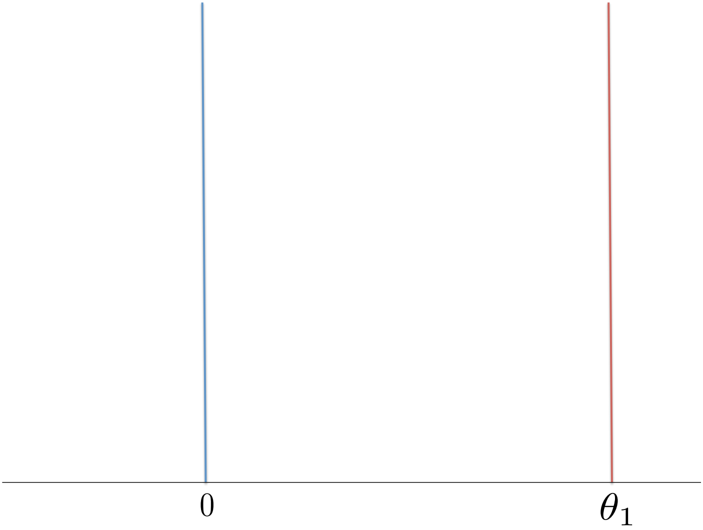


Classification not enough!
Need to compare sets

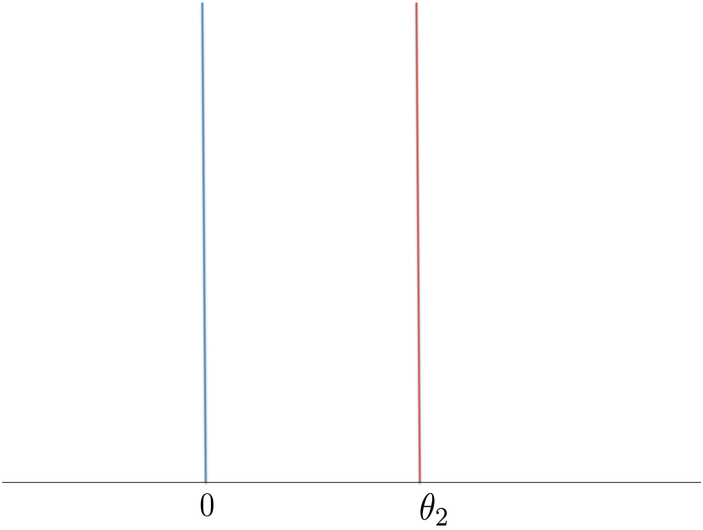
$$L^*(\theta) = -\log(4) + 2JSD(\mathbb{P}, \mathbb{Q}_\theta) \tag{3}$$



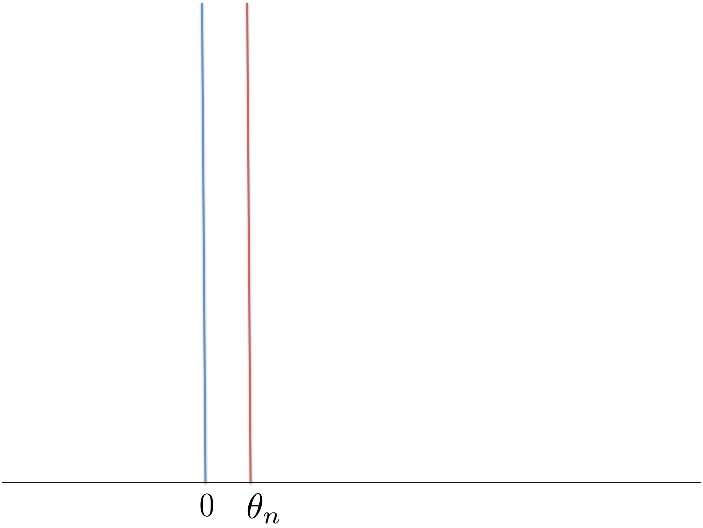
$$JSD(\mathbb{P}, \mathbb{Q}_{\theta_1}) = \log(2) \quad (3)$$



$$JSD(\mathbb{P}, \mathbb{Q}_{\theta_2}) = \log(2) \tag{3}$$



$$JSD(\mathbb{P}, \mathbb{Q}_{\theta_n}) = \log(2) \quad (3)$$



Weak continuity

Definition

A sequence $(\mathbb{Q}_n)_n$ converges weakly to $\mathbb{Q}$ if for all bounded continuous functions f:

$$\mathbb{E}_{\mathbb{Q}_n}[f(X)] \to \mathbb{E}_{\mathbb{Q}}[f(X)] \tag{4}$$

Weak continuity

Definition

A sequence $(\mathbb{Q}_n)_n$ converges weakly to $\mathbb{Q}$ if for all bounded continuous functions f:

$$\mathbb{E}_{\mathbb{Q}_n}[f(X)] \to \mathbb{E}_{\mathbb{Q}}[f(X)] \tag{4}$$

Definition

A functional $\mathbb{Q} \mapsto F(\mathbb{Q})$ is continuous under the weak topology if for all $\mathbb{Q}_n \rightharpoonup \mathbb{Q}$:

$$F(\mathbb{Q}_n) \to F(\mathbb{Q}) \tag{5}$$

Weak continuity

Definition

A sequence $(\mathbb{Q}_n)_n$ converges weakly to $\mathbb{Q}$ if for all bounded continuous functions f:

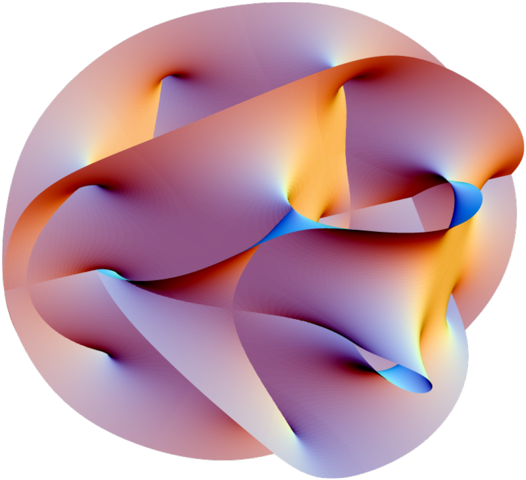
$$\mathbb{E}_{\mathbb{Q}_n}[f(X)] \to \mathbb{E}_{\mathbb{Q}}[f(X)] \quad (4)$$

Definition

A functional $\mathbb{Q} \mapsto F(\mathbb{Q})$ is continuous under the weak topology if for all $\mathbb{Q}_n \rightharpoonup \mathbb{Q}$:

$$F(\mathbb{Q}_n) \to F(\mathbb{Q}) \quad (5)$$

$\mathbb{Q} \mapsto JSD(\mathbb{P}, \mathbb{Q})$ is not continuous under the weak topology!



$$Y = G_\theta(Z) \qquad Z \sim [0,1]^q \tag{6}$$

Training IGMs

Criteria for choosing the loss $L(\mathbb{P}, \mathbb{Q})$:

- (C) Weak continuity: if $Q_n \rightharpoonup Q$ then $L(\mathbb{P}, \mathbb{Q}_n) \to L(\mathbb{P}, \mathbb{Q})$.
  ($Q_n \rightharpoonup Q$ means $\mathbb{Q}_{\ltimes}[f(x)] \to \mathbb{Q}[f(x)]$ for all bounded continuous f.)

Training IGMs

Criteria for choosing the loss $L(\mathbb{P}, \mathbb{Q})$:

- (C) Weak continuity: if $Q_n \rightharpoonup Q$ then $L(\mathbb{P}, \mathbb{Q}_n) \to L(\mathbb{P}, \mathbb{Q})$. ($Q_n \rightharpoonup Q$ means $\mathbb{Q}_\ltimes[f(x)] \to \mathbb{Q}[f(x)]$ for all bounded continuous f.)
- (M) Metrization of weak convergence: $L(\mathbb{Q}, \mathbb{Q}_n) \to 0$ if and only if $Q_n \rightharpoonup Q$.

Training IGMs

Criteria for choosing the loss $L(\mathbb{P}, \mathbb{Q})$:

- (C) Weak continuity: if $Q_n \rightharpoonup Q$ then $L(\mathbb{P}, \mathbb{Q}_n) \to L(\mathbb{P}, \mathbb{Q})$. ($Q_n \rightharpoonup Q$ means $\mathbb{Q}_{\ltimes}[f(x)] \to \mathbb{Q}[f(x)]$ for all bounded continuous f.)
- (M) Metrization of weak convergence: $L(\mathbb{Q}, \mathbb{Q}_n) \to 0$ if and only if $Q_n \rightharpoonup Q$.
- (T) Tractability: $L(\mathbb{P}, \mathbb{Q})$ can be "easily" estimated by sampling from $\mathbb{P}$ and $\mathbb{Q}$.

Training IGMs

Loss	Expression	(C)	(M)	(T)
$JSD(\mathbb{P}\|\mathbb{Q})$	$\frac{1}{2}(KL(\mathbb{P}\|\mu) + KL(\mathbb{Q}\|\mu))$ $\mu = \frac{\mathbb{P}+\mathbb{Q}}{2}$	✗	✗	✗
$W_1(\mathbb{P}, \mathbb{Q})$	$\sup_{\|f\|_{Lip}\leq 1} \mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]$	✓	✓	✗
$MMD(\mathbb{P}, \mathbb{Q})$	$\sup_{\|f\|_{\mathcal{H}}\leq 1} \mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]$	✓	✓	✓

Wasserstein GAN [Arjovsky et al., 2017]

1-Wasserstein distance:

$$W_1(\mathbb{P}, \mathbb{Q}) := \sup_{\|f\|_{Lip} \leq 1} \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[ff(X)]$$

$$\|f\|_{Lip} = \sup_{X, X'} \frac{|f(X) - f(X')|}{\|X - X'\|}$$

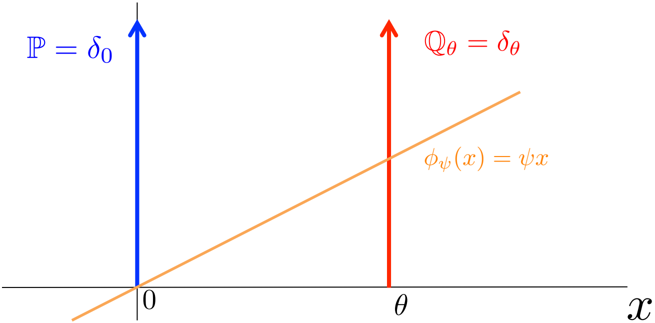
WGAN: replace f by ϕ_ψ and optimize over ψ:

$$\min_\theta \underbrace{\max_\psi \mathbb{E}_{X \sim \mathbb{P}}[\phi_\psi(X)] - \mathbb{E}_{Z \sim \mathbb{Z}}[\phi_\psi(G_\theta(Z))]}_{\hat{W}_1(\mathbb{P}, \mathbb{Q}_\theta)}$$

Non-convergence in WGAN

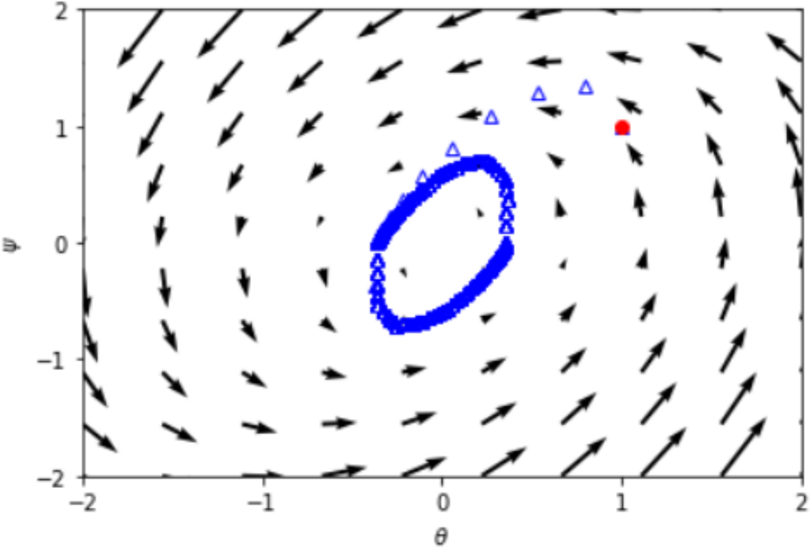
Toy problem in $\mathbb{R}$, DiracGAN [Mescheder et al., 2018]

- Point mass target $\mathbb{P} = \delta_0$, model $\mathbb{Q}_\theta = \delta_\theta$
- Test functions : $\phi_\psi(x) = \psi x$, $|\psi| \leq 1$.



Non-convergence in WGAN

- WGAN-GP reduces mode collapse but... oscillations can still happen [Mescheder et al., 2018]



Maximum Mean Discrepancy [Gretton et al., 2012]

Maximum mean discrepancy:

$$MMD(\mathbb{P}, \mathbb{Q}) = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]$$

Functions are linear combinations of features:

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} f_i \varphi_i(x)$$

Infinitely many features using kernels

- Feature map $\varphi(x) = [...\varphi_i(x)...]$
- For positive definite k

$$k(x,x') = \sum_i \varphi_i(x)\varphi_i(x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$$

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

Infinitely many features using kernels

- Feature map $\varphi(x) = [...\varphi_i(x)...]$
- For positive definite k

$$k(x, x') = \sum_i \varphi_i(x)\varphi_i(x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$$

- Infinitely many features $\varphi(x)$, but dot product in closed form
- $\mathcal{H}$: all possible linear combinations of features:

$$f = \sum_i^{\infty} f_i \varphi_i$$

Infinitely many features using kernels

- Feature map $\varphi(x) = [...\varphi_i(x)...]$
- For positive definite k

$$k(x, x') = \sum_i \varphi_i(x)\varphi_i(x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$$

- Infinitely many features $\varphi(x)$, but dot product in closed form
- $\mathcal{H}$: all possible linear combinations of features:

$$f = \sum_i^{\infty} f_i \varphi_i$$

$$f(x) = \sum_i^{\infty} f_i \varphi_i(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}}$$

Maximum Mean Discrepancy [Gretton et al., 2012]

A simple expression for maximum mean discrepancy:

$$MMD^2(\mathbb{P}, \mathbb{Q}) = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]$$

$$= \underbrace{\mathbb{E}_{\mathbb{P}}[k(X, X')]}_{(a)} + \underbrace{\mathbb{E}_{\mathbb{Q}}[k(X, X')]}_{(a)} - 2 \underbrace{\mathbb{E}_{\mathbb{P},\mathbb{Q}}[k(X, X')]}_{(b)}$$

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

Illustration of the MMD

$$\sim P$$

$$\sim Q$$

Illustration of the MMD

- *dog*$(= \mathbb{P})$ and *fish*$(= \mathbb{Q})$
- Each entry is one of $k(dog_i, dog_j)$, $k(dog_i, fish_j)$ or $k(fish_i, fish_j)$

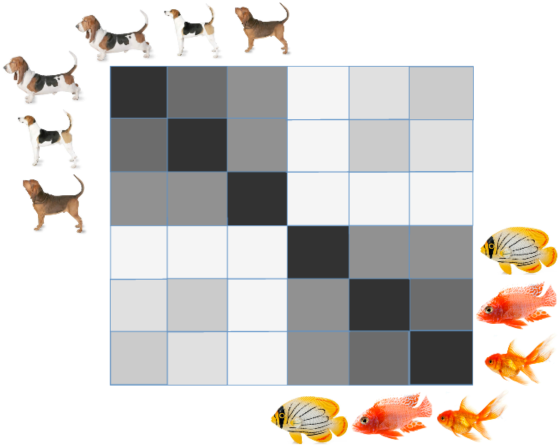
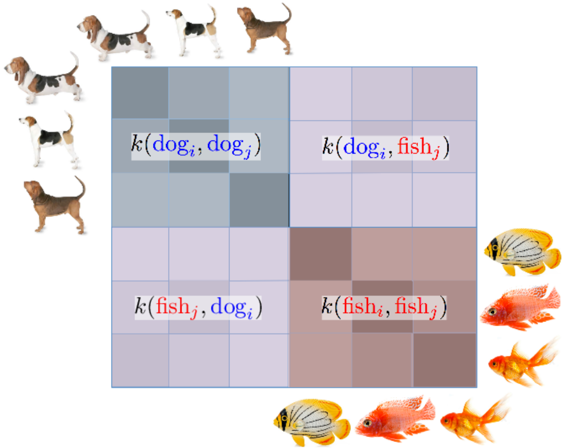
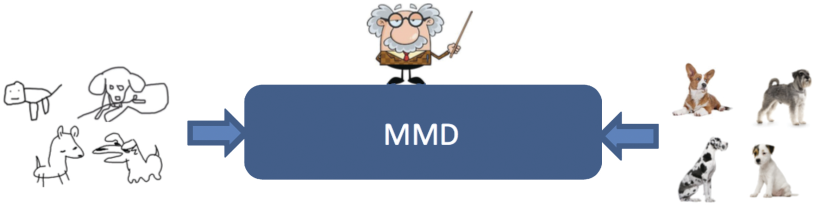


Illustration of the MMD

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(dog_i, dog_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(fish_i, fish_j) - \frac{2}{n^2} \sum_{i,j} k(dog_i, fish_j)$$



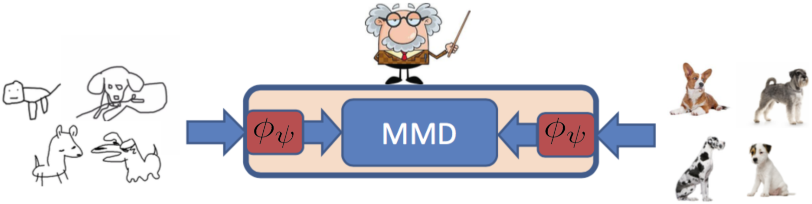
MMD as a loss [Dziugaite et al., 2015, Li et al., 2015]



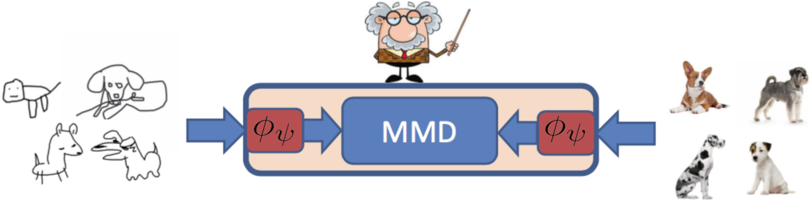
MMD as a loss [Dziugaite et al., 2015, Li et al., 2015]

Hard to pick a good kernel for images

MMD GANs: Deep kernels [Li et al., 2017]

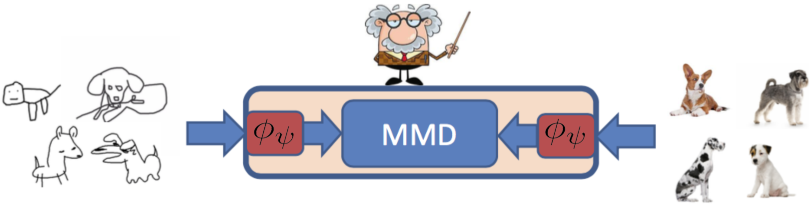


MMD GANs: Deep kernels [Li et al., 2017]



$$\min_{\theta} \underbrace{\max_{\psi} MMD^2{}_{k_\psi}(\mathbb{P}, \mathbb{Q}_\theta)}_{\mathcal{D}_{MMD}(\mathbb{P},\mathbb{Q}_\theta)}$$

MMD GANs: Deep kernels [Li et al., 2017]



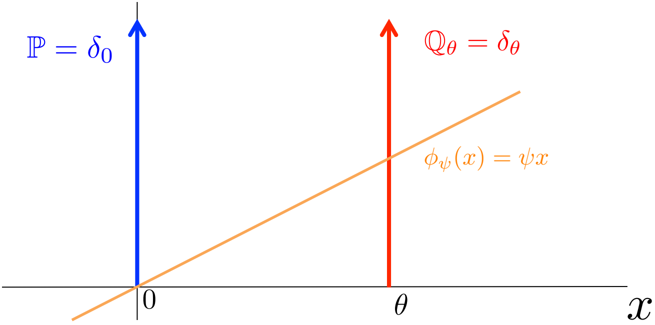
$$\underbrace{\min_\theta \max_\psi MMD^2_{k_\psi}(\mathbb{P}, \mathbb{Q}_\theta)}_{\mathcal{D}_{MMD}(\mathbb{P},\mathbb{Q}_\theta)}$$

$$k_\psi(X, Y) = K_{top}(\phi_\psi(X), \phi_\psi(Y))$$

Smoothness of $\mathcal{D}_{MMD}$

Toy problem in $\mathbb{R}$, DiracGAN [Mescheder et al., 2018]

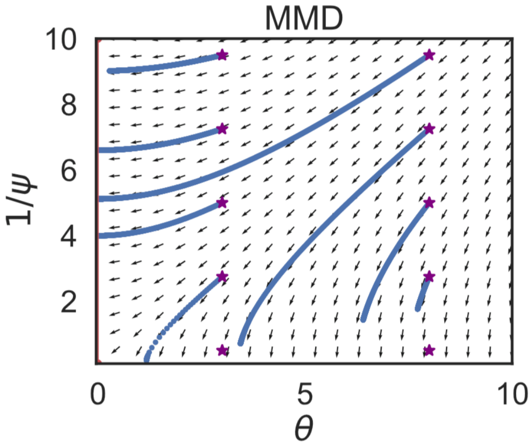
- Point mass target $\mathbb{P} = \delta_0$, model $\mathbb{Q}_\theta = \delta_\theta$
- Representation $\phi_\psi(x) = \psi x$, $\psi \in \mathbb{R}$
- kernel $K_{top}(a, b) = \exp(-\frac{1}{2}(a - b)^2)$



Smoothness of $\mathcal{D}_{MMD}$

Toy problem in $\mathbb{R}$, DiracGAN [Mescheder et al., 2018]

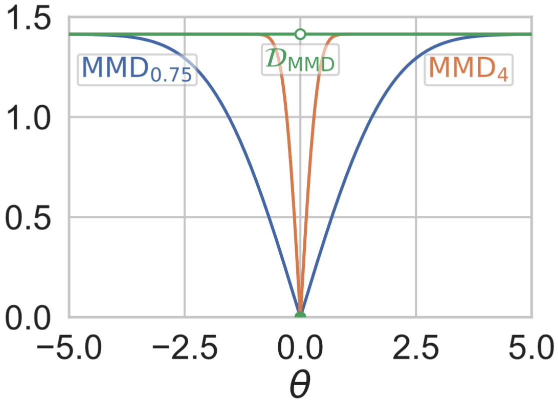
- Point mass target $\mathbb{P} = \delta_0$, model $\mathbb{Q}_\theta = \delta_\theta$
- Representation $\phi_\psi(x) = \psi x$, $\psi \in \mathbb{R}$
- kernel $k_{top}(a, b) = \exp(-\frac{1}{2}(a - b)^2)$



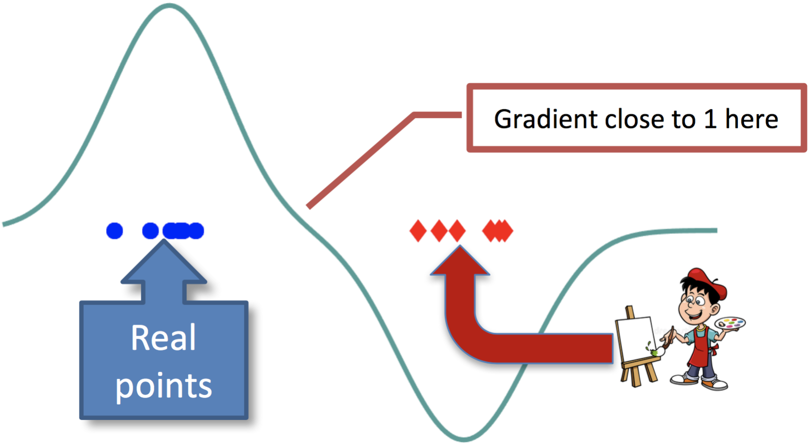
Smoothness of $\mathcal{D}_{MMD}$

Toy problem in $\mathbb{R}$, DiracGAN [Mescheder et al., 2018]

- $\mathcal{D}_{MMD} = \sup_{\psi} MMD(\phi_{\psi}(\mathbb{P}), \phi_{\psi}(\mathbb{Q}_{\theta})) = \sqrt{2}$.



Smoothness of $\mathcal{D}_{MMD}$ [Bińkowski et al., 2018]



Smoothness of $\mathcal{D}_{MMD}$ [Bińkowski et al., 2018]

Train MMD critic features with the witness function gradient penalty

$$\max_{\psi} MMD^2(\phi_\psi(X), \phi_\psi(G_\theta(Z))) - \lambda \mathbb{E}_{\widetilde{X}}[(\|\nabla_{\widetilde{X}} f_\psi(\widetilde{X})\|^2 - 1)^2]$$

where

$$\begin{aligned} \widetilde{X} =& \gamma X_i + (1-\gamma) G_\theta(Z_j) \\ & \gamma \sim \mathcal{U}([0,1]); \quad X_i \sim \mathbb{P}; \quad Z_j \sim \mathbb{Z} \end{aligned}$$

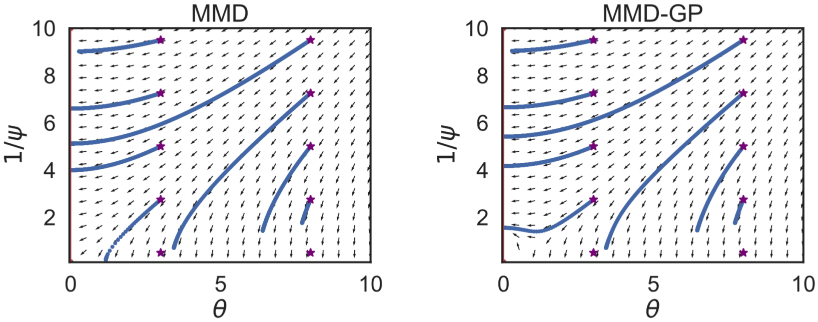
and

$$f_\psi(t) \propto \frac{1}{n}\sum_{i=1}^{n} K(\phi_\psi(X_i), t) - \frac{1}{n}\sum_{i=1}^{n} K(\phi_\psi(G_\theta(Z_j)), t)$$

Smoothness of $\mathcal{D}_{MMD}$

Toy problem in $\mathbb{R}$, DiracGAN [Mescheder et al., 2018]

- Point mass target $\mathbb{P} = \delta_0$, model $\mathbb{Q}_\theta = \delta_\theta$
- Representation $\phi_\psi(x) = \psi x$, $\psi \in \mathbb{R}$
- kernel $k_{top}(a, b) = \exp(-\frac{1}{2}(a - b)^2)$

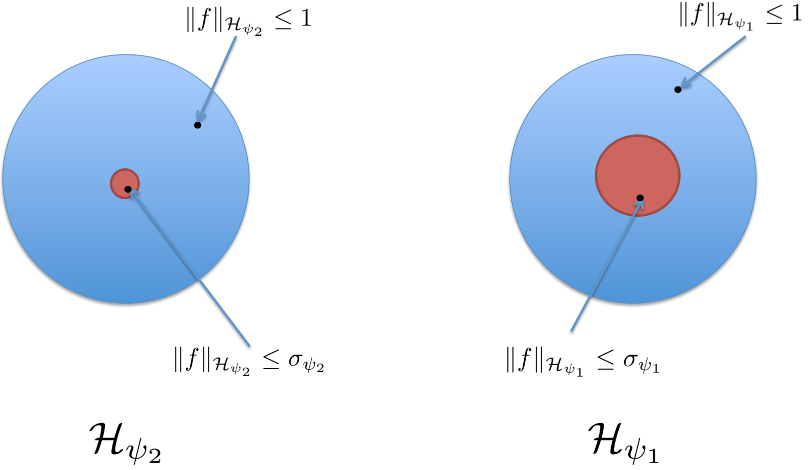


Scaled MMD [Arbel et al., 2018]

$$MMD_{k_\psi}(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}_\psi} \leq 1} \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{P}}[f(X)]$$

Scaled MMD [Arbel et al., 2018]

$$SMMD_{k_\psi}(\mathbb{P},\mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}_\psi} \leq \sigma_\psi} \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{P}}[f(X)] = \sigma_\psi MMD_{k_\psi}(\mathbb{P},\mathbb{Q})$$



Scaled MMD [Arbel et al., 2018]

Define a different norm:

$$\|f\|_{S_\psi}^2 = \mathbb{E}_\mu[\|f(X)\|^2] + \mathbb{E}_\mu[\|\nabla f(X)\|^2] + \|f\|_{\mathcal{H}_\psi}^2$$

We would like to have:

$$\|f\|_{S_\psi}^2 \leq 1$$

Scaled MMD [Arbel et al., 2018]

Define a different norm:

$$\|f\|_{S_\psi}^2 = \mathbb{E}_\mu[\|f(X)\|^2] + \mathbb{E}_\mu[\|\nabla f(X)\|^2] + \|f\|_{\mathcal{H}_\psi}^2$$

We would like to have:

$$\langle f, Cf \rangle_{\mathcal{H}_\psi} \leq 1$$

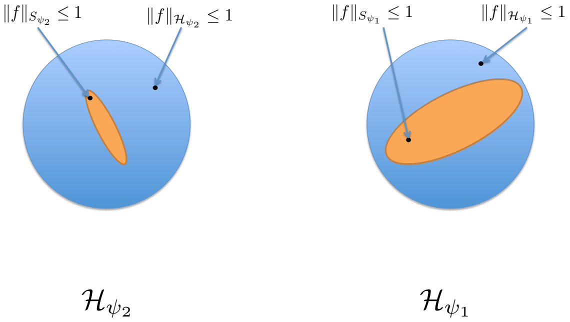
Scaled MMD [Arbel et al., 2018]

Define a different norm:

$$\|f\|^2_{S_\psi} = \mathbb{E}_\mu[\|f(X)\|^2] + \mathbb{E}_\mu[\|\nabla f(X)\|^2] + \|f\|^2_{\mathcal{H}_\psi}$$

We would like to have:

$$\langle f, Cf \rangle_{\mathcal{H}_\psi} \leq 1$$



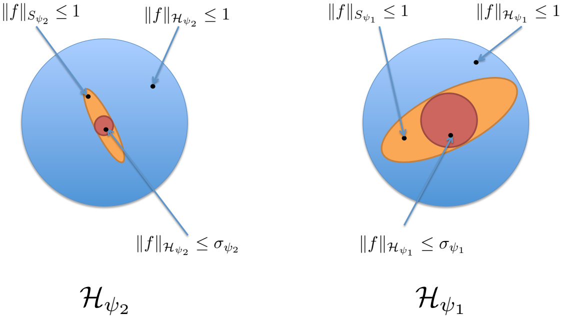
Scaled MMD [Arbel et al., 2018]

Define a different norm:

$$\|f\|_{S_\psi}^2 = \mathbb{E}_\mu[\|f(X)\|^2] + \mathbb{E}_\mu[\|\nabla f(X)\|^2] + \|f\|_{\mathcal{H}_\psi}^2$$

We only need:

$$\|f\|_{\mathcal{H}_\psi}^2 \leq \|C\|_{op}^{-1}$$



Scaled MMD [Arbel et al., 2018]

$$SMMD_\psi(\mathbb{P},\mathbb{Q}) := \sigma_\psi MMD(\phi_\psi(\mathbb{P}),\phi_\psi(\mathbb{Q}))$$

where:

$$\sigma_\psi = (\lambda + \mathbb{E}_\mu[K(\phi_\psi(X),\phi_\psi(X))] + \mathbb{E}_\mu[\sum_{i=1}^{d} \partial_i \partial_{i+d} K(\phi_\psi(X),\phi_\psi(X))])^{-\frac{1}{2}}$$

Scaled MMD [Arbel et al., 2018]

$$SMMD_{\psi}(\mathbb{P},\mathbb{Q}) := \sigma_{\psi} MMD(\phi_{\psi}(\mathbb{P}),\phi_{\psi}(\mathbb{Q}))$$

where:

$$\sigma_{\psi} = (\lambda + \mathbb{E}_{\mu}[K(\phi_{\psi}(X),\phi_{\psi}(X))] + \mathbb{E}_{\mu}[\sum_{i=1}^{d} \partial_i \partial_{i+d} K(\phi_{\psi}(X),\phi_{\psi}(X))])^{-\frac{1}{2}}$$

when K is of the form $K(a,b) = g(-\|a-b\|^2)$

$$\sigma_{\psi} = (\lambda + g(0) + 2|g'(0)|\,\mathbb{E}_{\mu}[\|\nabla \phi_{\psi}(X)\|^2])^{-\frac{1}{2}}$$

Scaled MMD GAN

Adversarial distance:

$$\mathcal{D}_{SMMD}(\mathbb{P}, G_\theta(\mathbb{Z})) := \max_{\psi} \sigma_{\psi} MMD(\phi_{\psi}(\mathbb{P}), \phi_{\psi}(G_\theta(\mathbb{Z})))$$

Generator's objective:

$$\min_{\theta} \mathcal{D}_{SMMD}(\mathbb{P}, G_\theta(\mathbb{Z}))$$

SMMD GAN

- Use a class of features ϕ_ψ
- Chose the most discriminative one:

$$\mathcal{D}_{SMMD}(\mathbb{P}, \mathbb{Q}) = \sup_{\psi} \sigma_{\psi,\mathbb{P},\lambda} MMD(\phi_\psi(\mathbb{P}), \phi_\psi(\mathbb{Q}))$$

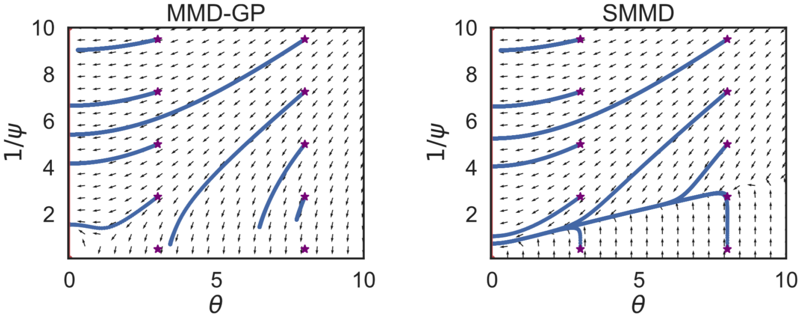
SMMD GAN

- Use a class of features ϕ_ψ
- Chose the most discriminative one:

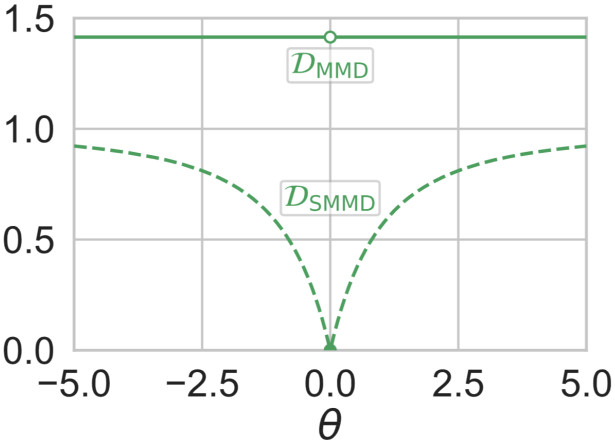
$$\mathcal{D}_{\mathit{SMMD}}(\mathbb{P}, \mathbb{Q}) = \sup_{\psi} \sigma_{\psi,\mathbb{P},\lambda} \mathit{MMD}(\phi_\psi(\mathbb{P}), \phi_\psi(\mathbb{Q}))$$

- Initialize random generator G_θ and feature ϕ_ψ
- Repeat:
  - k SGD steps in ψ to maximize $\widehat{\sigma^2}_\psi \widehat{\mathit{MMD}^2}(\phi_\psi(\mathbb{P}), \phi_\psi(\mathbb{Q}))$
  - One SGD step in θ to minimize $\widehat{\sigma^2}_\psi \widehat{\mathit{MMD}^2}(\phi_\psi(\mathbb{P}), \phi_\psi(\mathbb{Q}))$

$\mathcal{D}_{SMMD}$ vs $\mathcal{D}_{MMD}$



$\mathcal{D}_{SMMD}$ vs $\mathcal{D}_{MMD}$



Weak continuity of $\mathcal{D}_{SMMD}$

- $\|\phi_\psi\|_{Lip} \leq 1$ implies weak continuity of $\mathcal{D}_{SMMD}$...
- but $\mathbb{E}_\mu[\|\nabla_X \phi_\psi(X)\|^2] \leq 1$ generally doesn't!

Weak continuity of $\mathcal{D}_{SMMD}$

- $\|\phi_\psi\|_{Lip} \leq 1$ implies weak continuity of $\mathcal{D}_{SMMD}$...
- but $\mathbb{E}_\mu[\|\nabla_X \phi_\psi(X)\|^2] \leq 1$ generally doesn't!
- Luckily

$$\nabla_X \phi_\psi(X) = \prod_{l=1}^{L} W_l \circ M_l(X)$$

Weak continuity of $\mathcal{D}_{SMMD}$

- $\|\phi_\psi\|_{Lip} \leq 1$ implies weak continuity of $\mathcal{D}_{SMMD}$...
- but $\mathbb{E}_\mu[\|\nabla_X \phi_\psi(X)\|^2] \leq 1$ generally doesn't!
- Luckily

$$\nabla_X \phi_\psi(X) = \prod_{l=1}^{L} W_l \circ M_l(X)$$

- If W_l have full rank, decreasing dimensions + leaky-ReLu:

$$\|\nabla \phi_\psi(X)\| \geq \|\phi_\psi\|_{Lip} \frac{\alpha^L}{\kappa^L}$$

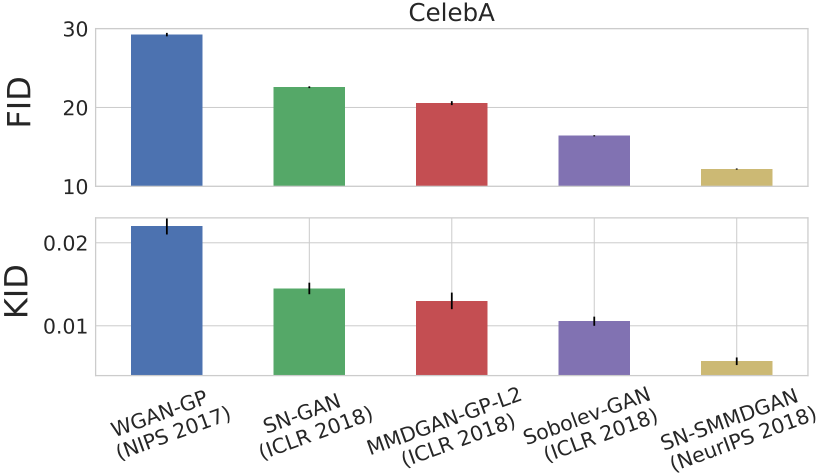
Weak continuity of $\mathcal{D}_{SMMD}$

Theorem: $\mathcal{D}_{SMMD}(\mathbb{P}, \mathbb{Q})$ is continuous wrt. the weak topology if:

- μ has a density w.r.t Lebesgue measure.
- ϕ_ψ is fully connected with Leaky-ReLU and non-increasing width.
- The condition number of the weights per-layer in ϕ_ψ is bounded.

Experimental results: celebA 160×160

202 599 face images, resized and cropped to 160×160.



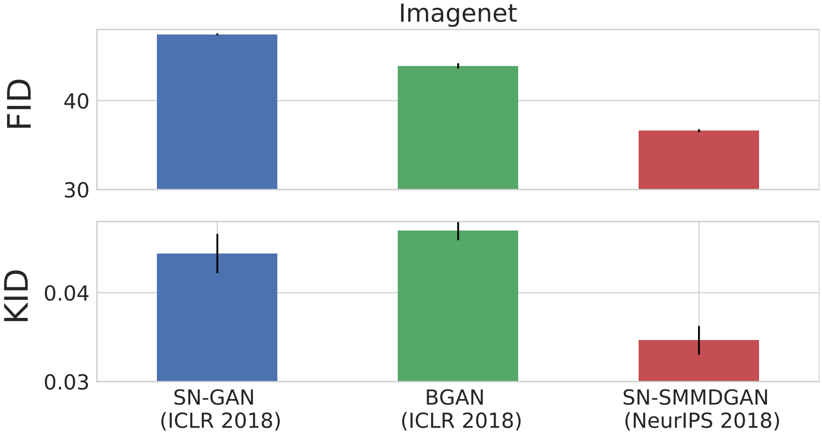
Experimental results: celebA 160×160

WGAN-GP (NIPS 2017)

SN-SMMDGAN (ours)

Experimental results: Imagenet 64×64

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



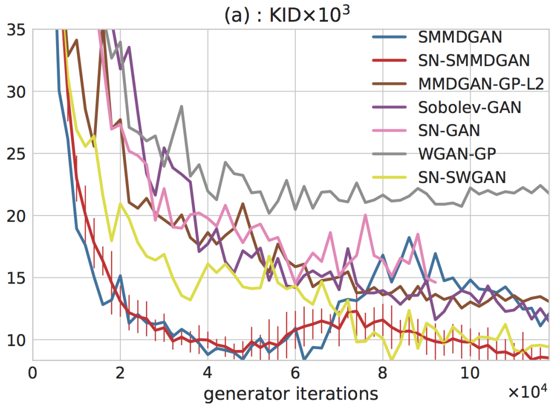
Experimental results: Imagenet 64×64

SN-GAN (ICLR 2018)

SN-SMMDGAN (ours)

Experimental results

Faster training: performance scores vs generator iterations on CelebA



- Spectral parametrization improves training ! (SMMDGAN vs SN-SMMDGAN)

Conclusion

- Weak continuity of the loss functional is crucial for successful training of IGMs.
- Adapting the amplitude of the MMD to the smoothness of the kernel provides a simple way to achieve weak continuity.
- Some insights on the choice of the critic's architecture.
- State of the art results on challenging datasets.

Future directions:

- How do adversarial distances relate to other well-known distances? (Not generally equivalent in the strict metric sense.)
- The choice of the distributions for the regularizing factor.

Thank you !

- Arbel, M., Sutherland, D. J., Bińkowski, M., and Gretton, A. (2018).
  On gradient regularizers for MMD GANs.
  arXiv:1805.11565 [cs, stat].
  arXiv: 1805.11565.
- Arjovsky, M., Chintala, S., and Bottou, L. (2017).
  Wasserstein gan.
- Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. (2018).
  The Mechanics of n-Player Differentiable Games.
  arXiv:1802.05642 [cs].
  arXiv: 1802.05642.
- Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018).
  Demystifying MMD GANs.
- Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015).
  Training generative neural networks via Maximum Mean Discrepancy optimization.