On Gradient Regularizers for MMD-GANs

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

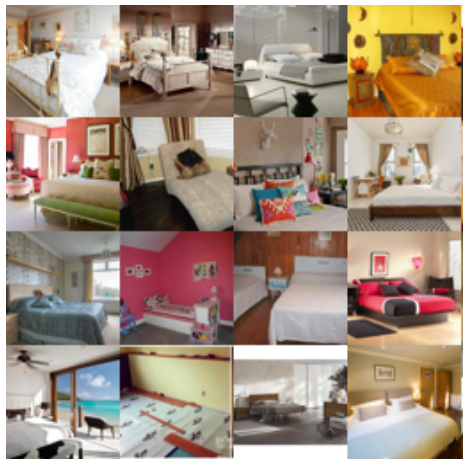
[1]Gatsby Computational Neuroscience Unit
University College London

[2]Department of Mathematics
Imperial College London

December 19, 2018

Implicit generative models

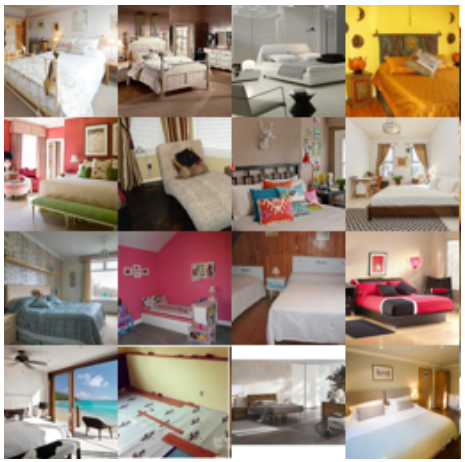
Given samples from a distributions $\mathbb{P}$ over $\mathcal{X}$, want a model that can produce new samples from $\mathbb{Q}_\theta \approx \mathbb{P}$



$X \sim \mathbb{P}$

Implicit generative models

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

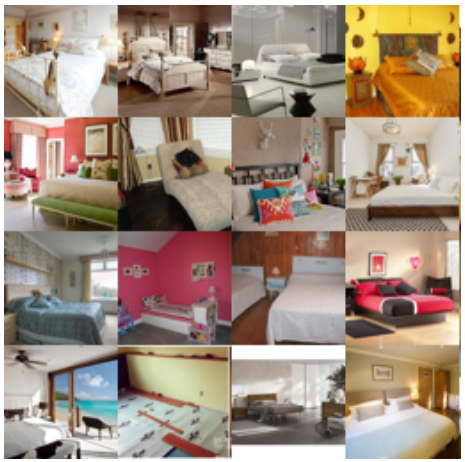


$X \sim \mathbb{P}$

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

Implicit generative models

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



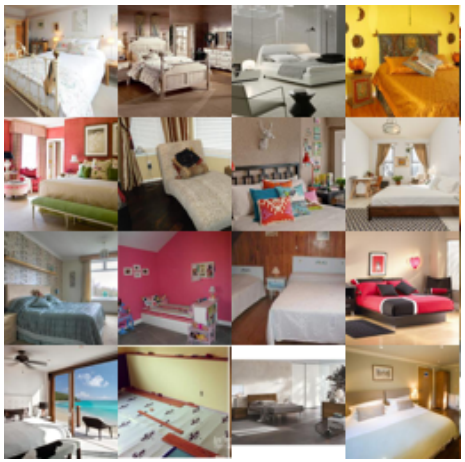
$X \sim \mathbb{P}$

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

$Y = G_\theta(Z)$ where Z is uniform on $[-1, 1]^{100}$

Implicit generative models

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



$X \sim \mathbb{P}$

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

$Y = G_\theta(Z)$ where Z is uniform on $[-1, 1]^{100}$

Don't necessarily care about likelihoods, interpretability, ...

What is a Generative Adversarial Network (GAN)?

Two networks:

-  (Generator G_θ): creates (fake) samples $Y = G_\theta(Z)$ from random noise Z:
-  (Critic ϕ_ψ): determine whether samples are fake or real

- Generator (student)

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

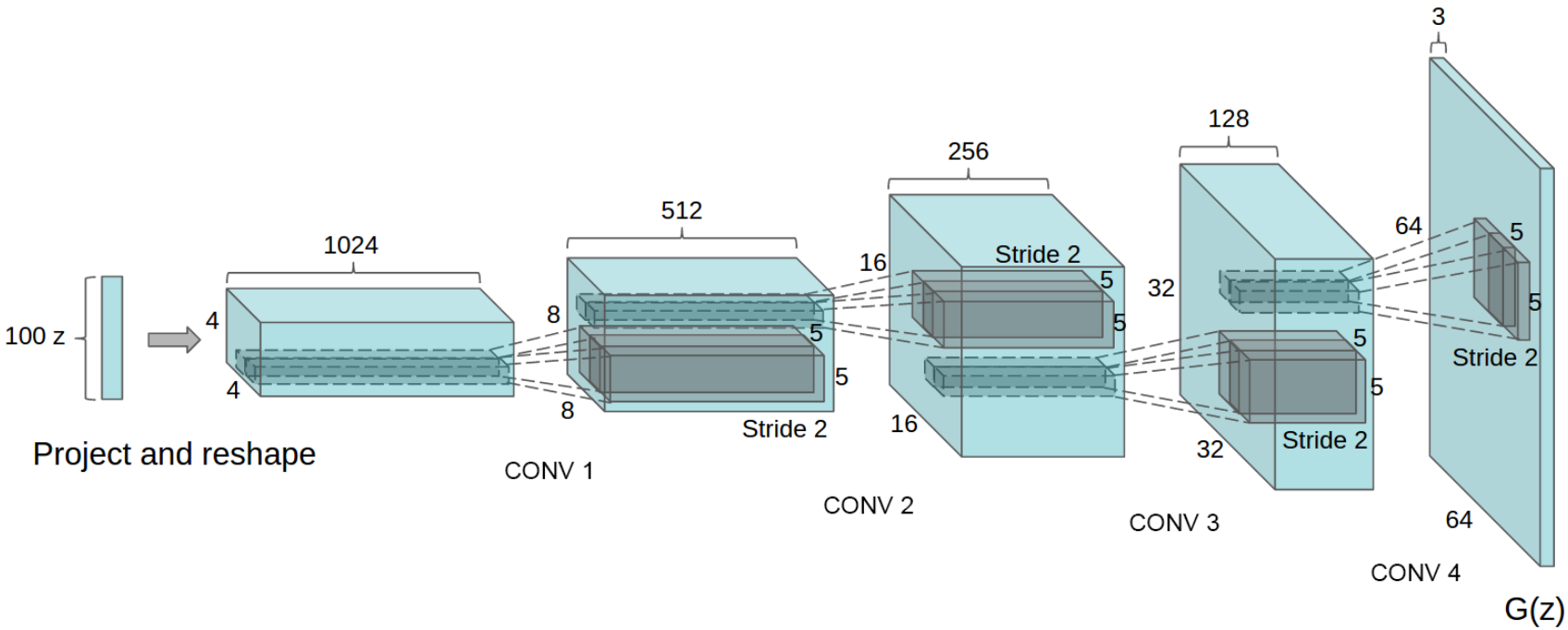




- Critic (teacher)

What is a Generative Adversarial Network (GAN)?

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



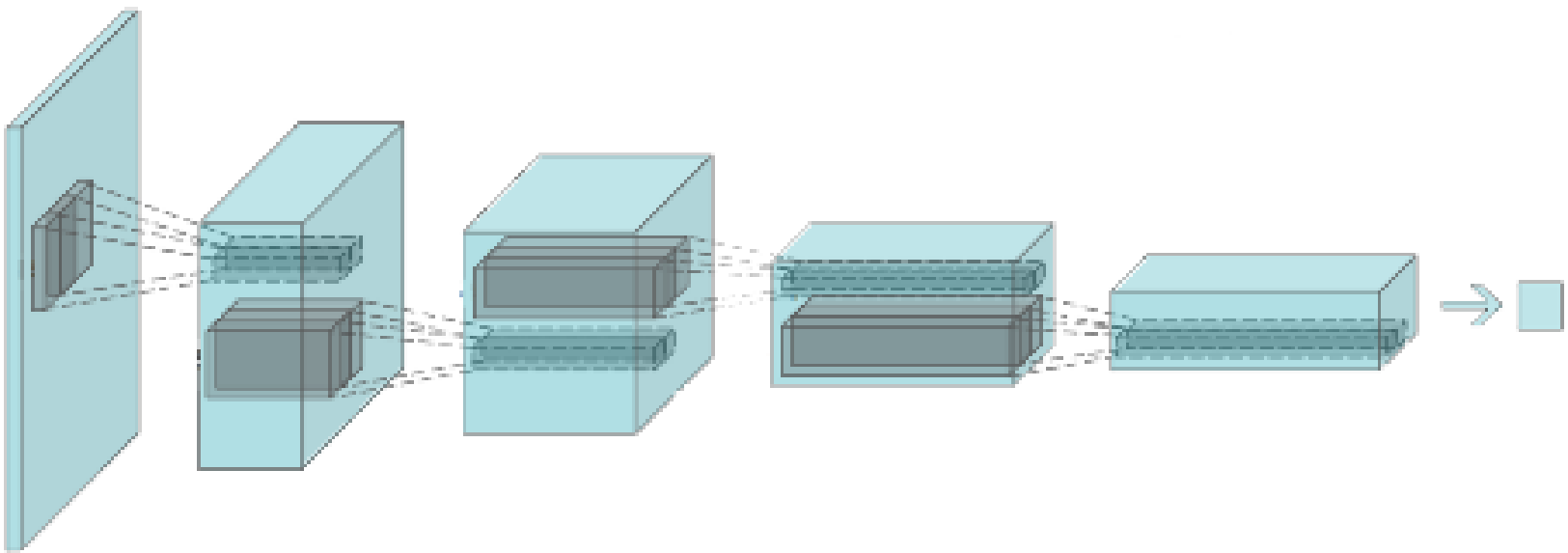
DCGAN generator [Radford et al., 2015]

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

Choose θ by minimizing a cost

What is a Generative Adversarial Network (GAN)?

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

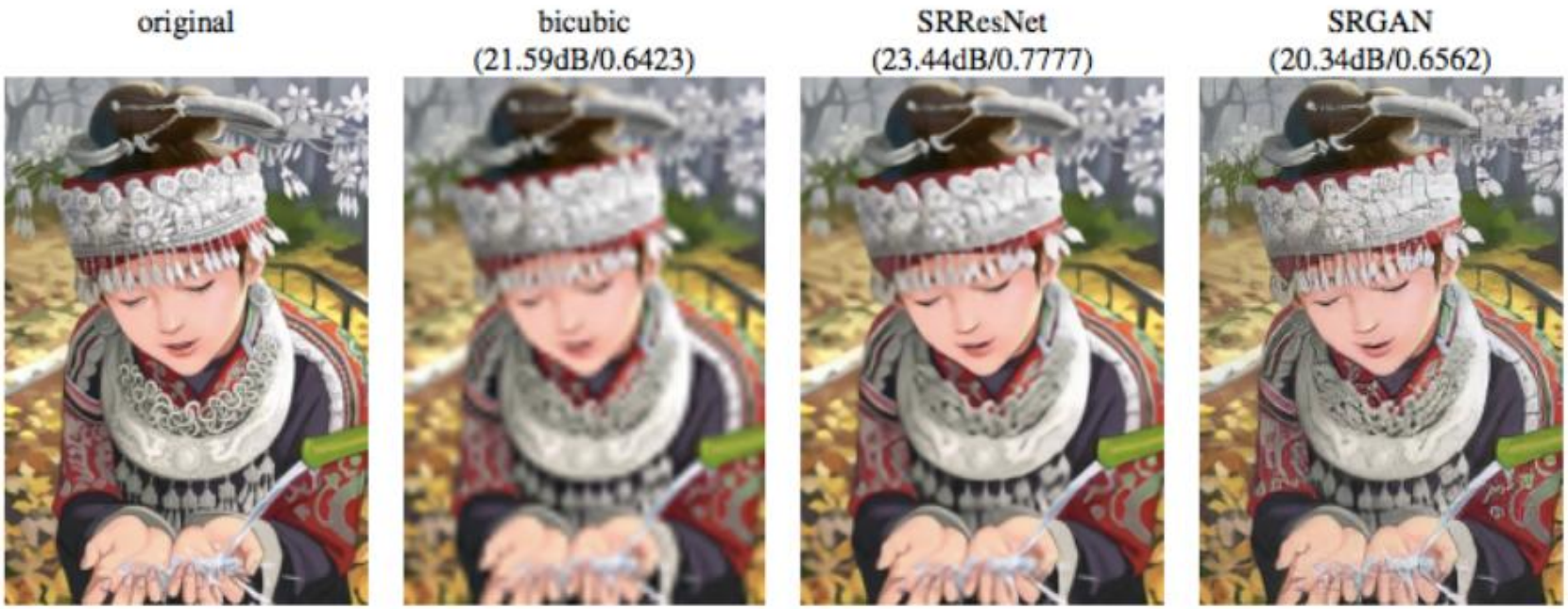


DCGAN critic [Radford et al., 2015]

Choose ψ by minimizing a cost

Why using GANs?

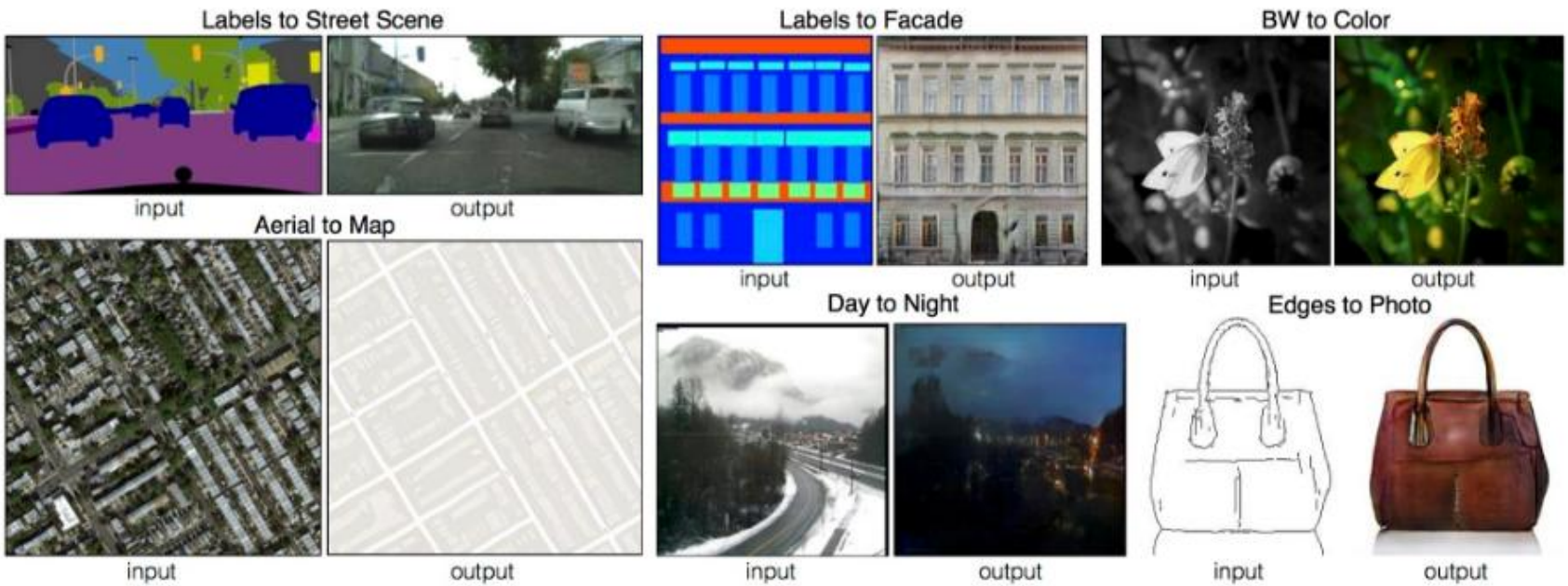
- Image generation tasks
  - Single-image super-resolution



Ledig et al 2015

Why using GANs?

- Image generation tasks
  - Image to image translation



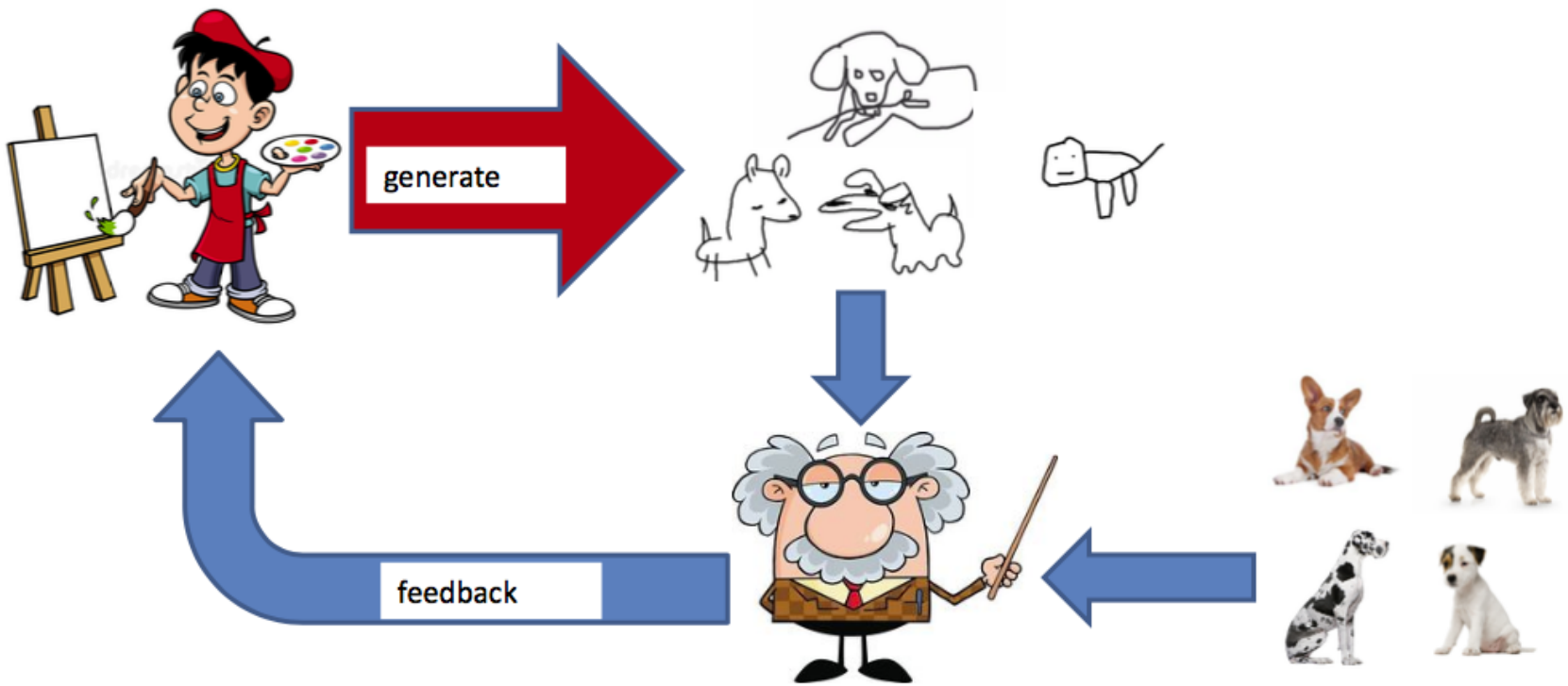
Isola et al 2016

Why using GANs?

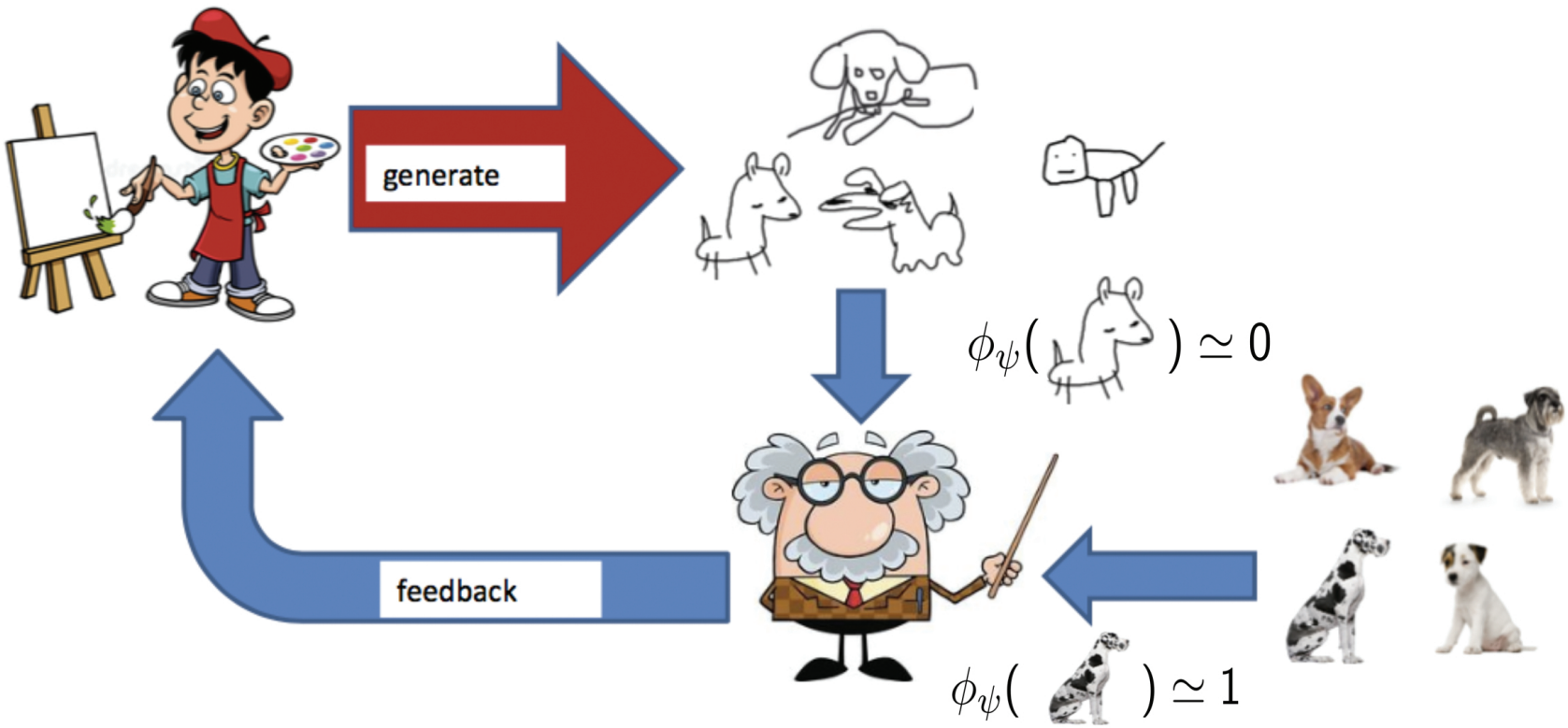
- Image generation tasks
  - Text to image generation

Zhang et al 2016

What is a Generative Adversarial Network (GAN)?



What is a Generative Adversarial Network (GAN)?



Loss function

- The critic's objective:

$$\mathcal{J}^{(C)}(\theta, \psi) = \mathbb{E}_{\text{  }\sim\mathbb{P}}[\log(\phi_\psi(\text{  })] + \mathbb{E}_{\text{  }\sim\text{  }}[\log(1 - \phi_\psi(\text{  }))]$$

- The generator cost $\mathcal{J}^{(G)}$:

$$\mathcal{J}^{(G)} = -\mathcal{J}^{(C)}$$

- min-max problem:

$$\min_\theta \max_\psi \mathbb{E}_{X\sim\mathbb{P}}[\log\phi_\psi(X)] + \mathbb{E}_{Z\sim\mathbb{Z}}[\log(1 - \phi_\psi(G_\theta(Z)))]$$

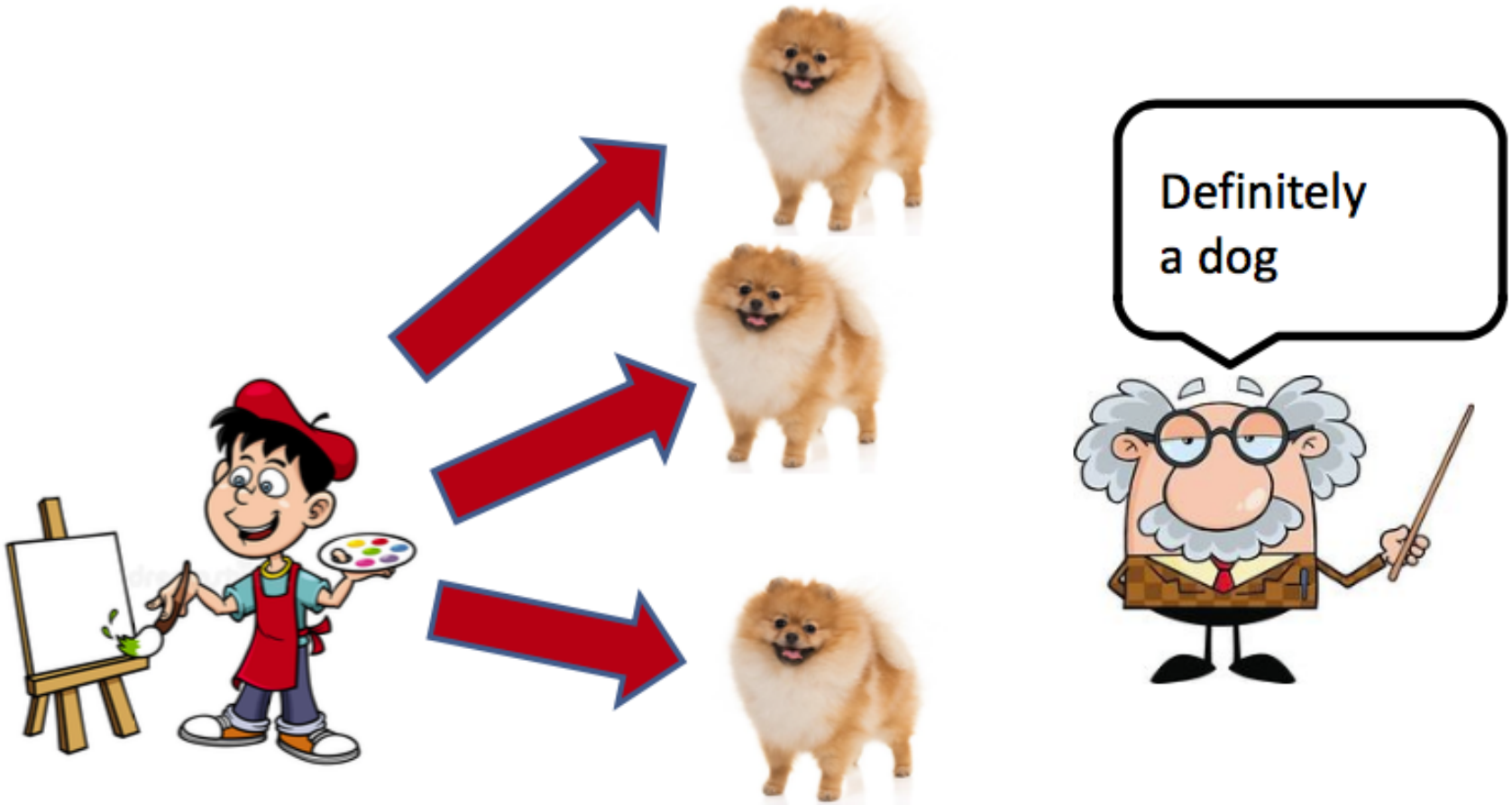
Optimization

- min-max problem:

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

- alternate:
  - 5 SGD steps in ψ to maximize $\mathcal{J}^{\mathcal{C}}(\theta, \psi)$
  - 1 SGD step in θ to minimize $\mathcal{J}^{\mathcal{C}}(\theta, \psi)$

Mode collapse



Classification not enough!
Need to compare sets

Wasserstein GAN [Arjovsky et al., 2017]

Remove log:

$$\min_{\theta} \max_{\psi} \mathbb{E}_{X\sim\mathbb{P}}[\log\phi_\psi(X)] + \mathbb{E}_{Z\sim\mathbb{Z}}[\log(1-\phi_\psi(G_\theta(Z)))]$$

Wasserstein GAN [Arjovsky et al., 2017]

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

Wasserstein GAN [Arjovsky et al., 2017]

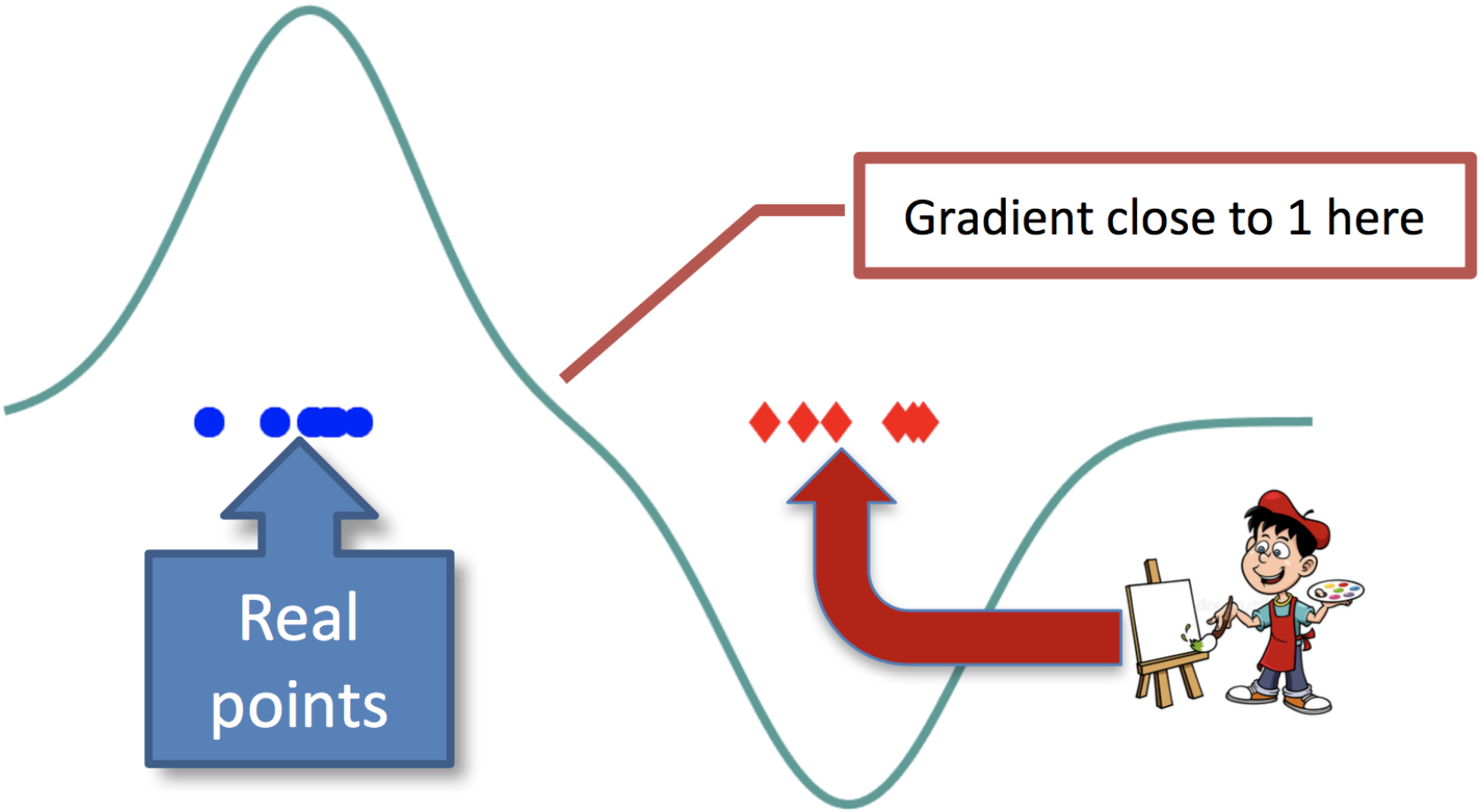
1-Wasserstein distance:

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

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

WGAN: replace f by ϕ_ψ and optimize over ψ.

Wasserstein GAN [Arjovsky et al., 2017]



WGAN-GP [Gulrajani et al., 2016]

Gradient penalty: penalizes non-Lipschitzness:

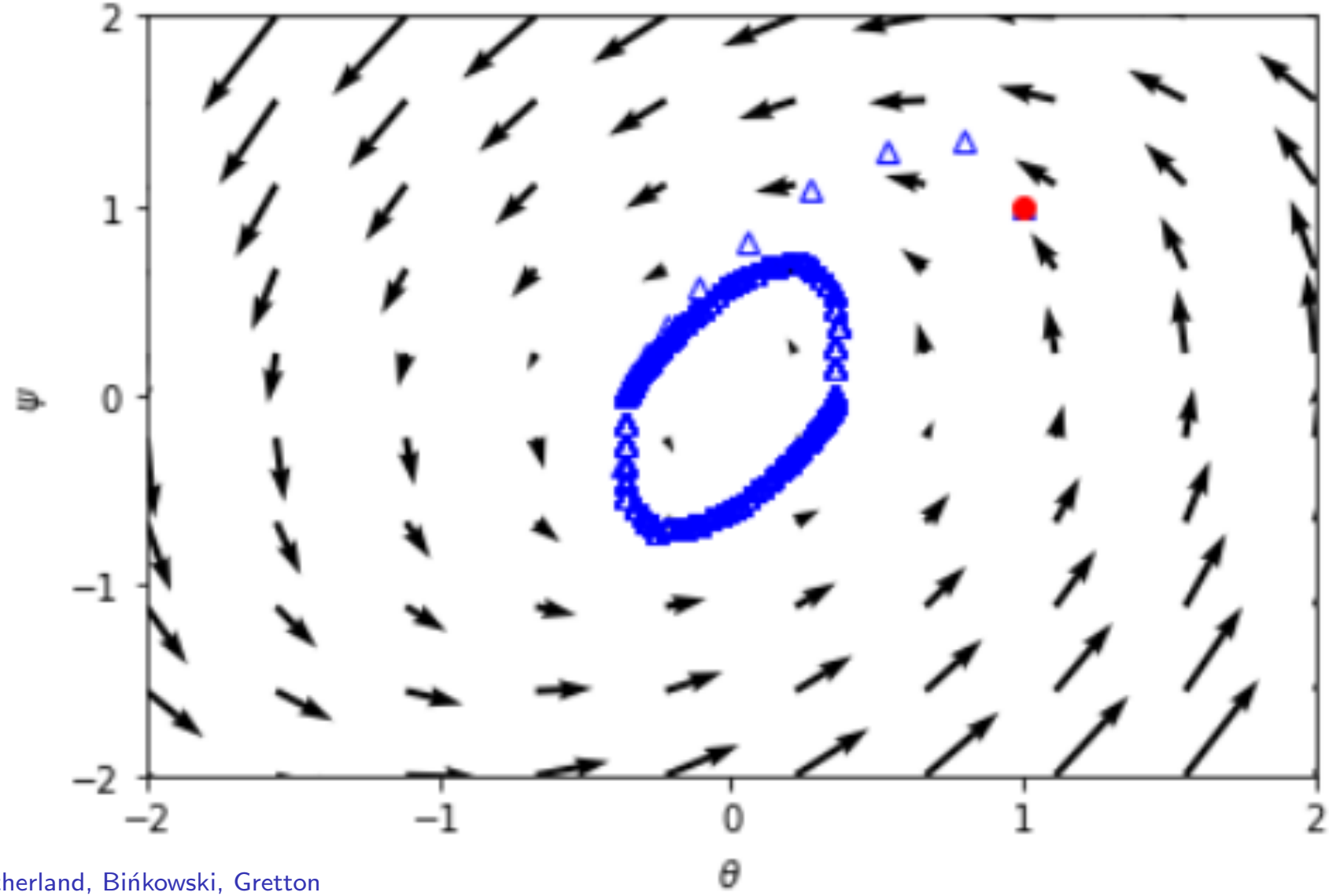
$$\max_{\psi} \mathbb{E}_{X\sim\mathbb{P}}[\phi_\psi(X)] - \mathbb{E}_{Z\sim\mathbb{Z}}[\phi_\psi(G_\theta(Z))] - \lambda\, \mathbb{E}_{\widetilde{X}}[(\|\nabla_{\widetilde{X}}\phi_\psi(\widetilde{X})\| - 1)^2]$$

where

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

Non-convergence in GANs

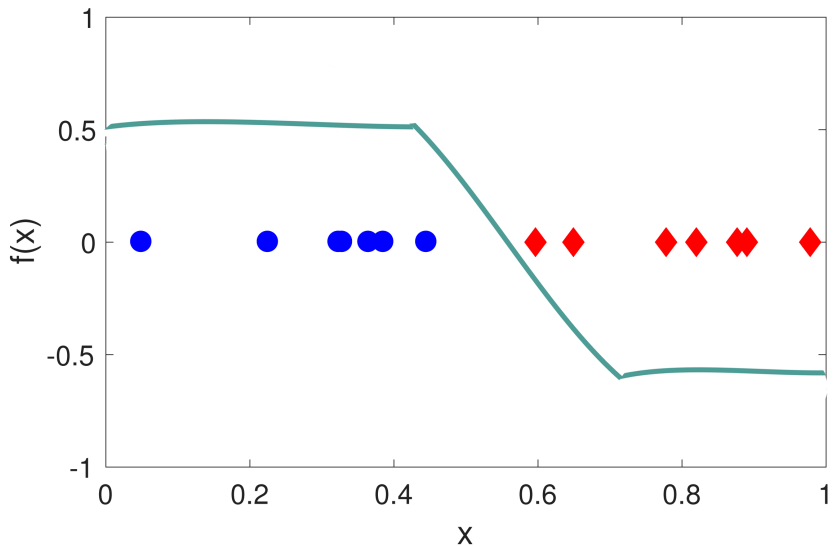
- WGAN-GP reduces mode collapse but... oscillations can still happen [Mescheder et al., 2018]



Integral probability metric

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

$$\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]$$

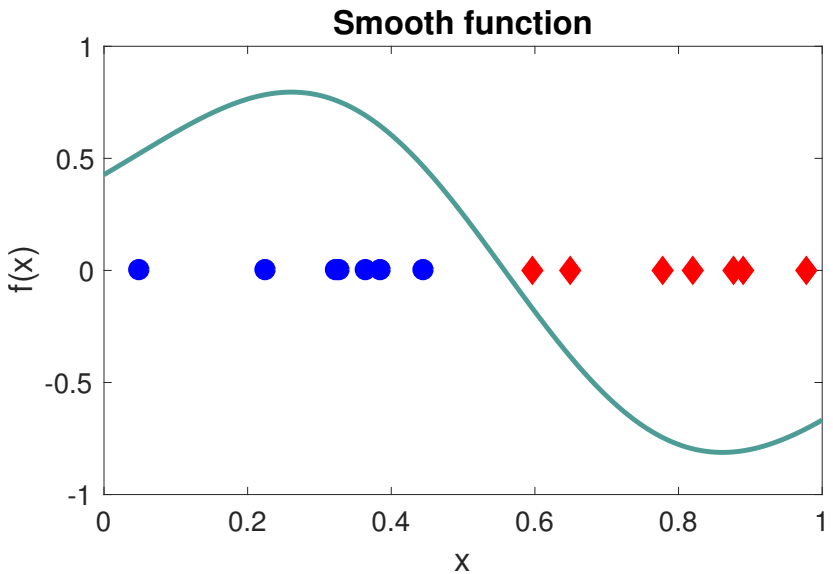


For $W_1(\mathbb{P}, \mathbb{Q})$, " well behaved" f means $\|f\|_{Lip} \leq 1$.

Maximum Mean Discrepancy [Gretton et al., 2012]

Maximum mean discrepancy: find smooth function for $\mathbb{P}$ vs $\mathbb{Q}$ to maximize

$$\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]$$



" smooth function" f means $f \in \mathcal{H}$ and $\|f\|_{\mathcal{H}} \leq 1$.
Optimal f called " witness function"

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:

$$f = \sum_{i}^{\infty} f_i \varphi_i$$

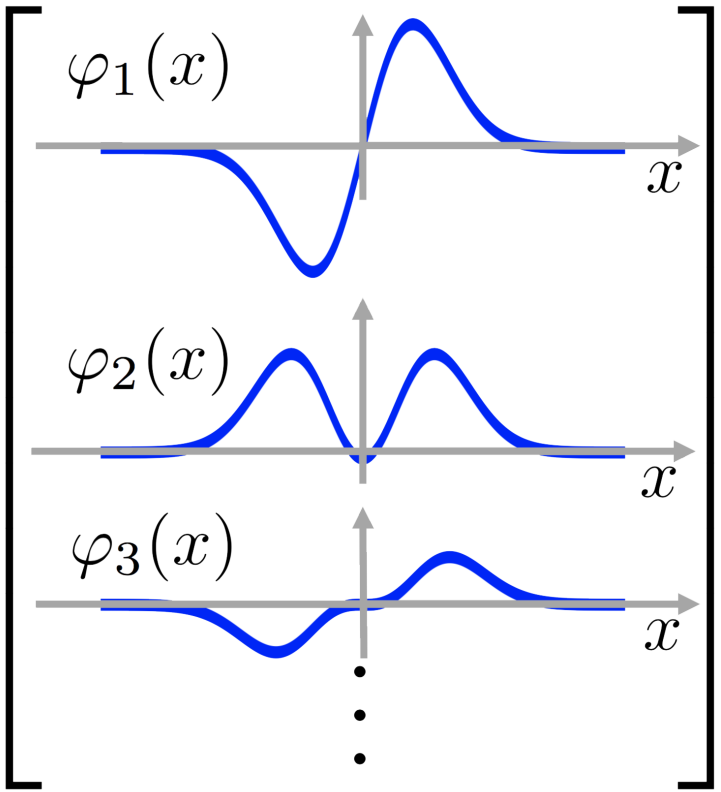
$$f(x) = \sum_{i}^{\infty} f_i \varphi_i(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}}$$

Infinitely many features using kernels

Exponentiated quadratic kernel

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

$$\varphi(x) = \left[\begin{array}{c} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \\ \vdots \end{array} \right]$$



Features: Gaussian Processes for Machine learning, Rasmussen and Williams, Ch. 4.

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 $k(fish_i, fish_j)$

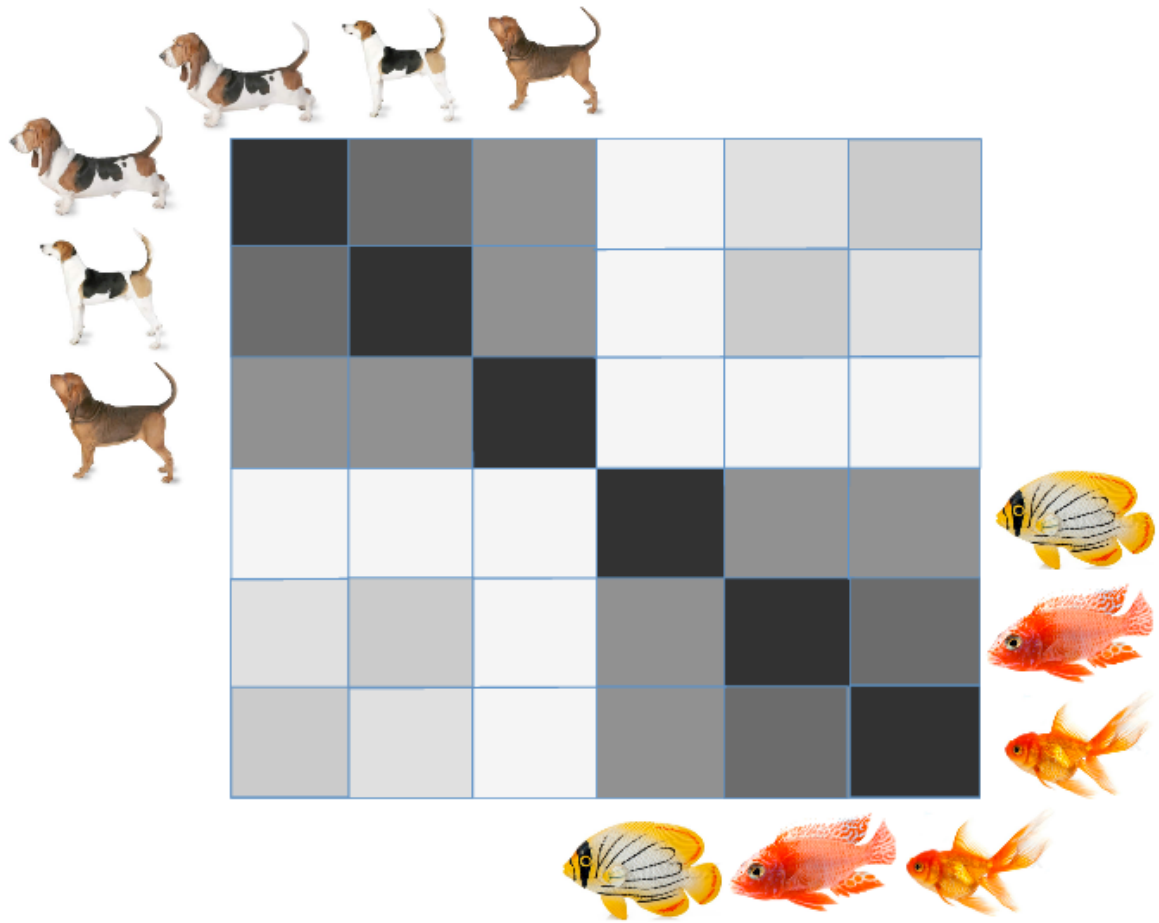
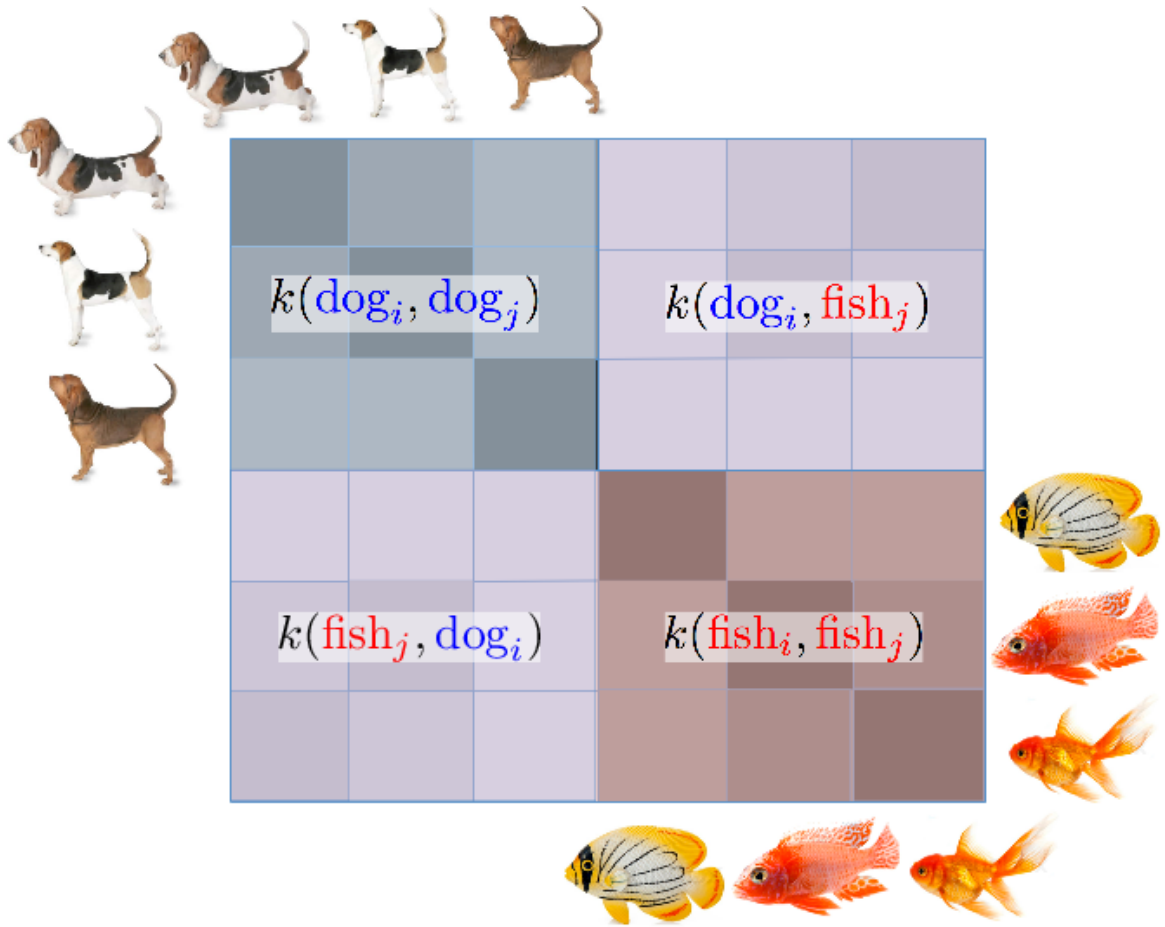
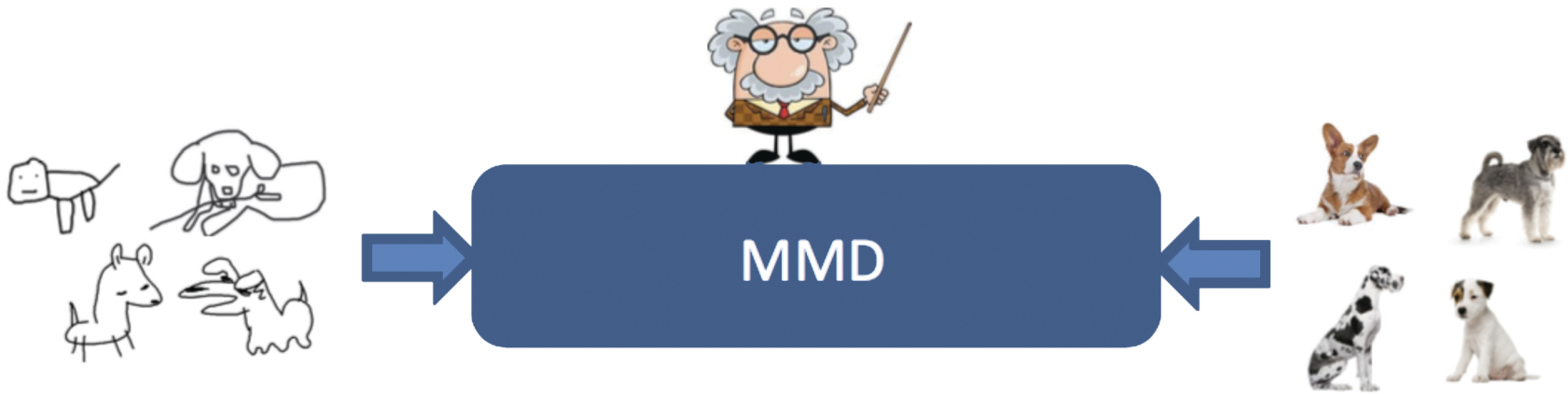


Illustration of the MMD

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



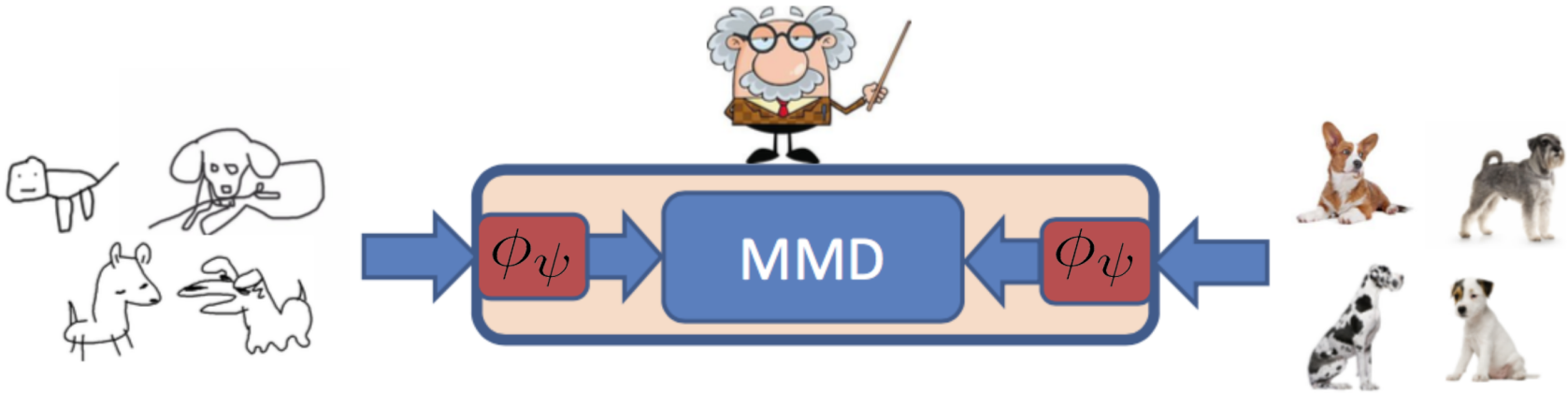
MMD as a loss [Dziugaite et al., 2015, Li et al., 2015]



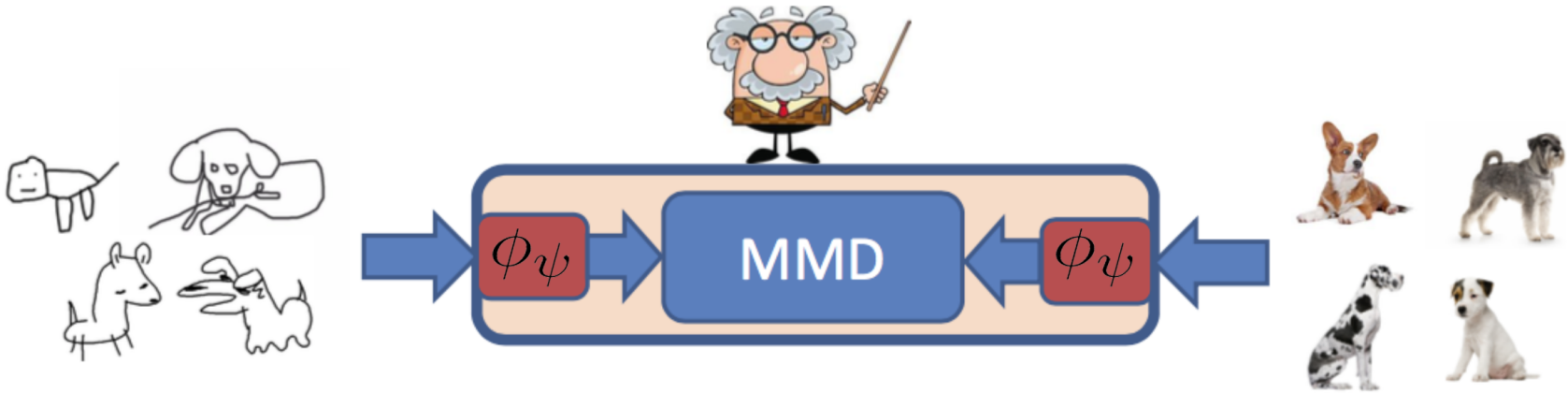
MMD as a loss [Dziugaite et al., 2015, Li et al., 2015]

Hard to pick a good kernel for images

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



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

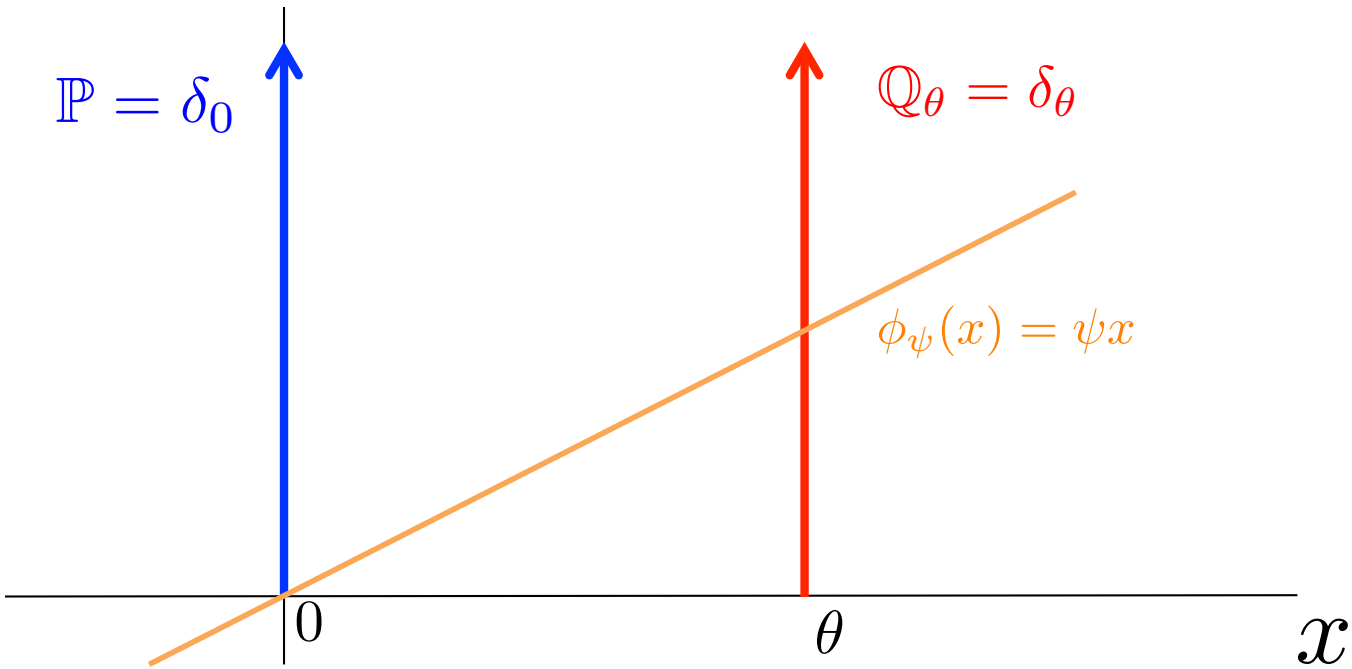


$$\min_\theta \underbrace{\max_\psi MMD^2(\phi_\psi(X), \phi_\psi(G_\theta(Z)))}_{\mathcal{D}_{MMD}(\mathbb{P},\mathbb{Q}_\theta)}$$

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

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

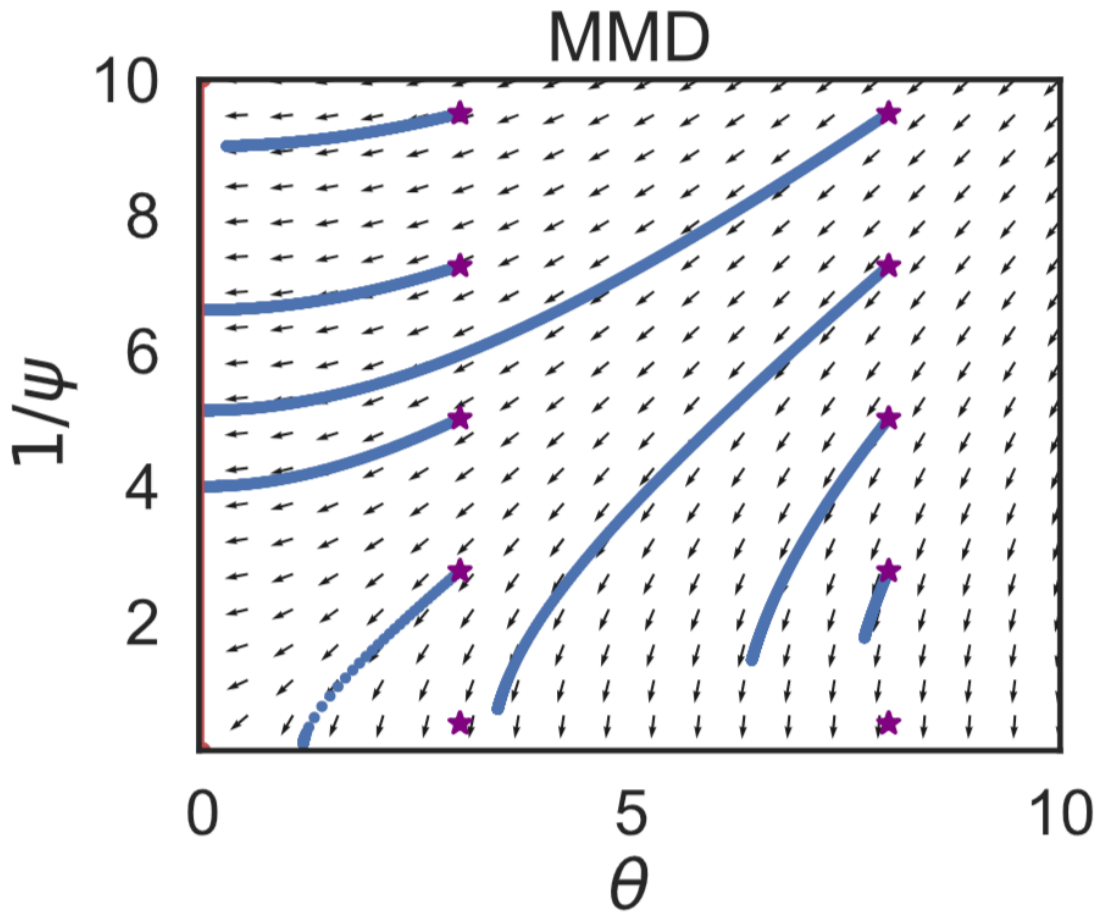
- Point mass target $\mathbb{P} = \delta_0$, model $\mathbb{Q}_\theta = \delta_\theta$
- Representation $\phi_\psi(x) = \psi x$, $\psi \in \mathbb{R}$
- kernel $K(a, b) = \exp(-\frac{1}{2}(a-b)^2)$



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

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

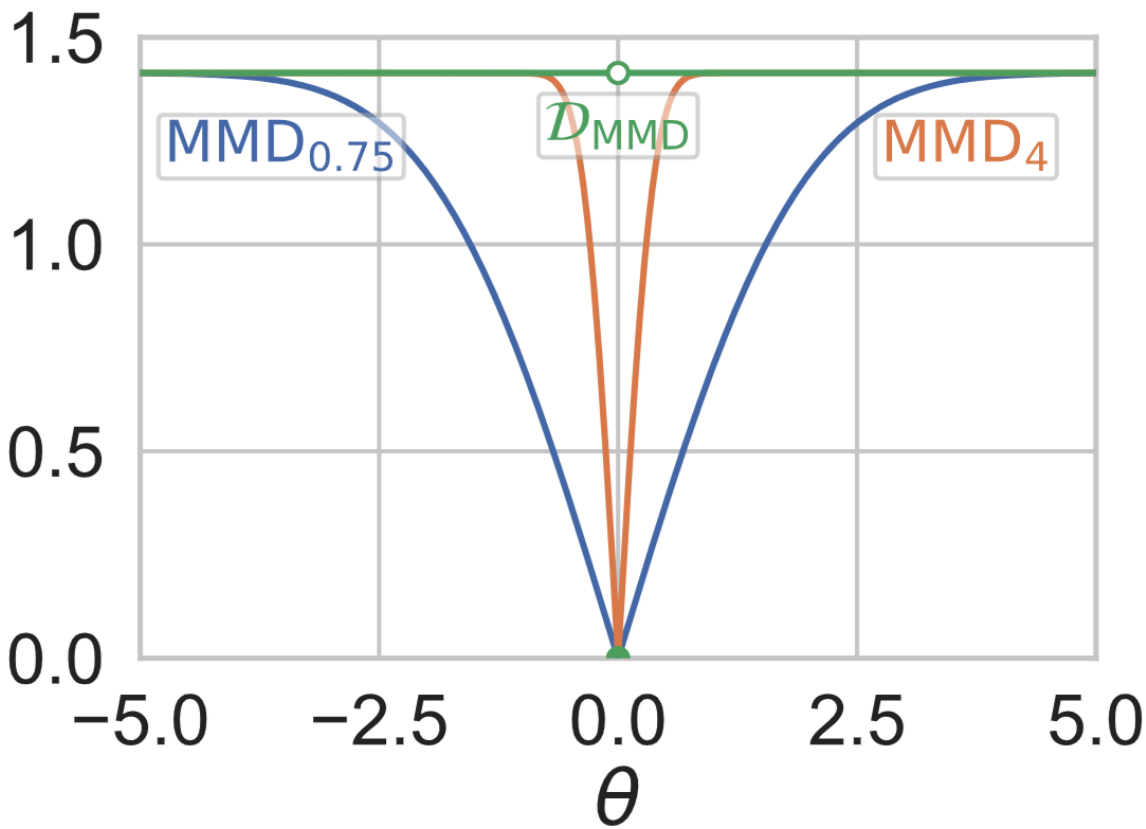
- Point mass target $\mathbb{P} = \delta_0$, model $\mathbb{Q}_\theta = \delta_\theta$
- Representation $\phi_\psi(x) = \psi x$, $\psi \in \mathbb{R}$
- kernel $k_{top}(a, b) = \exp(-\frac{1}{2}(a-b)^2)$



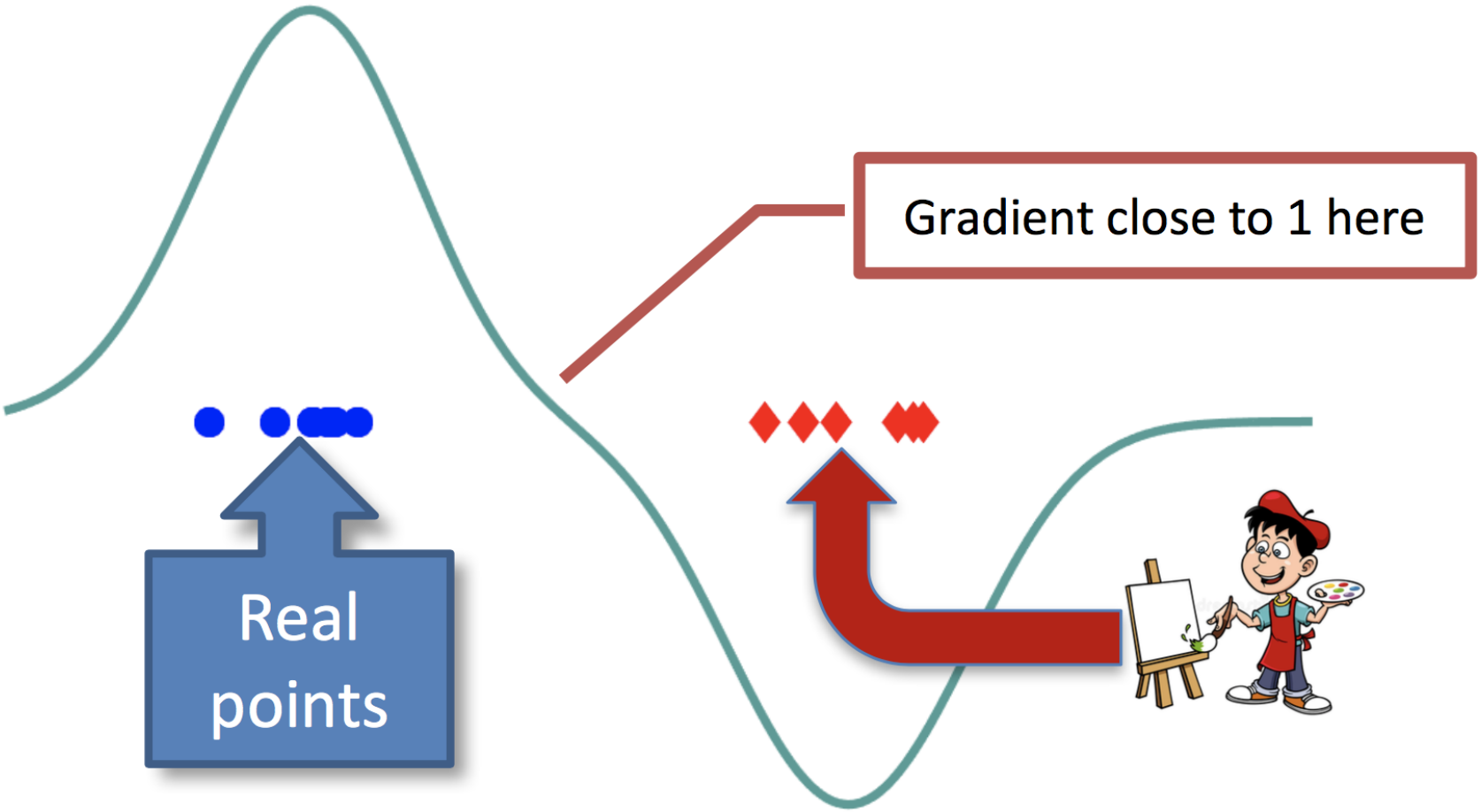
Smoothness of $\mathcal{D}_{MMD}$

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

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

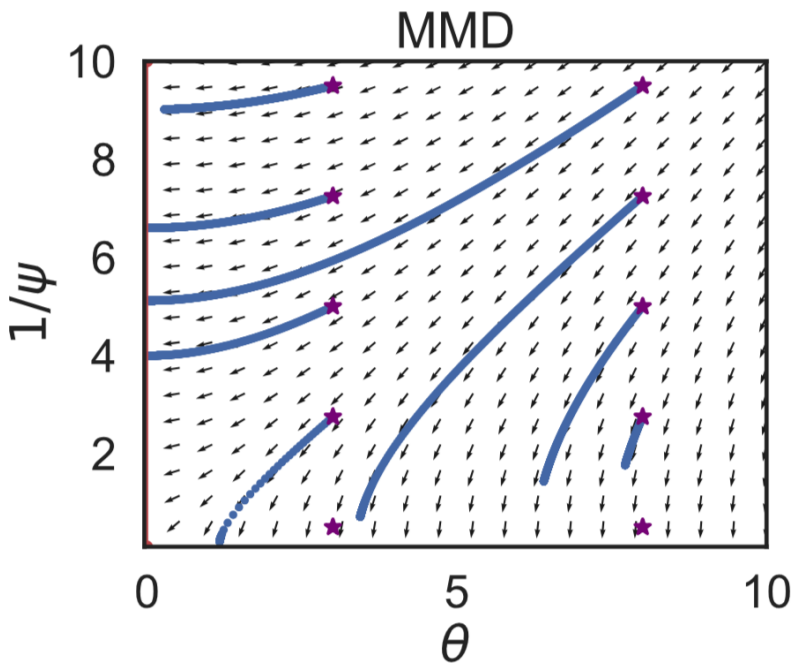
and

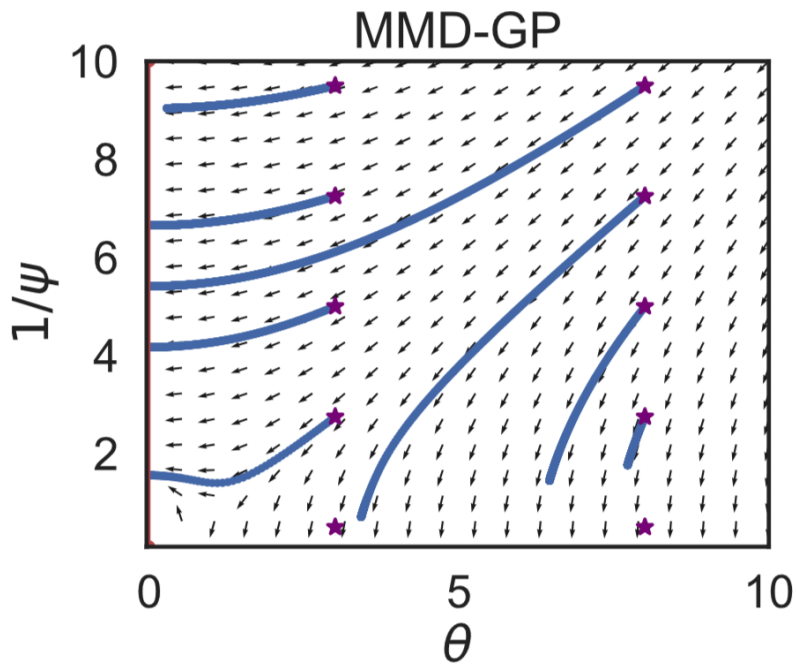
$$f_\psi(t) \propto \frac{1}{n}\sum_{i=1}^{n} K(\phi_\psi(X_i), t) - \frac{1}{n}\sum_{i=1}^{n} K(\phi_\psi(G_\theta(Z_j)), t)$$

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

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

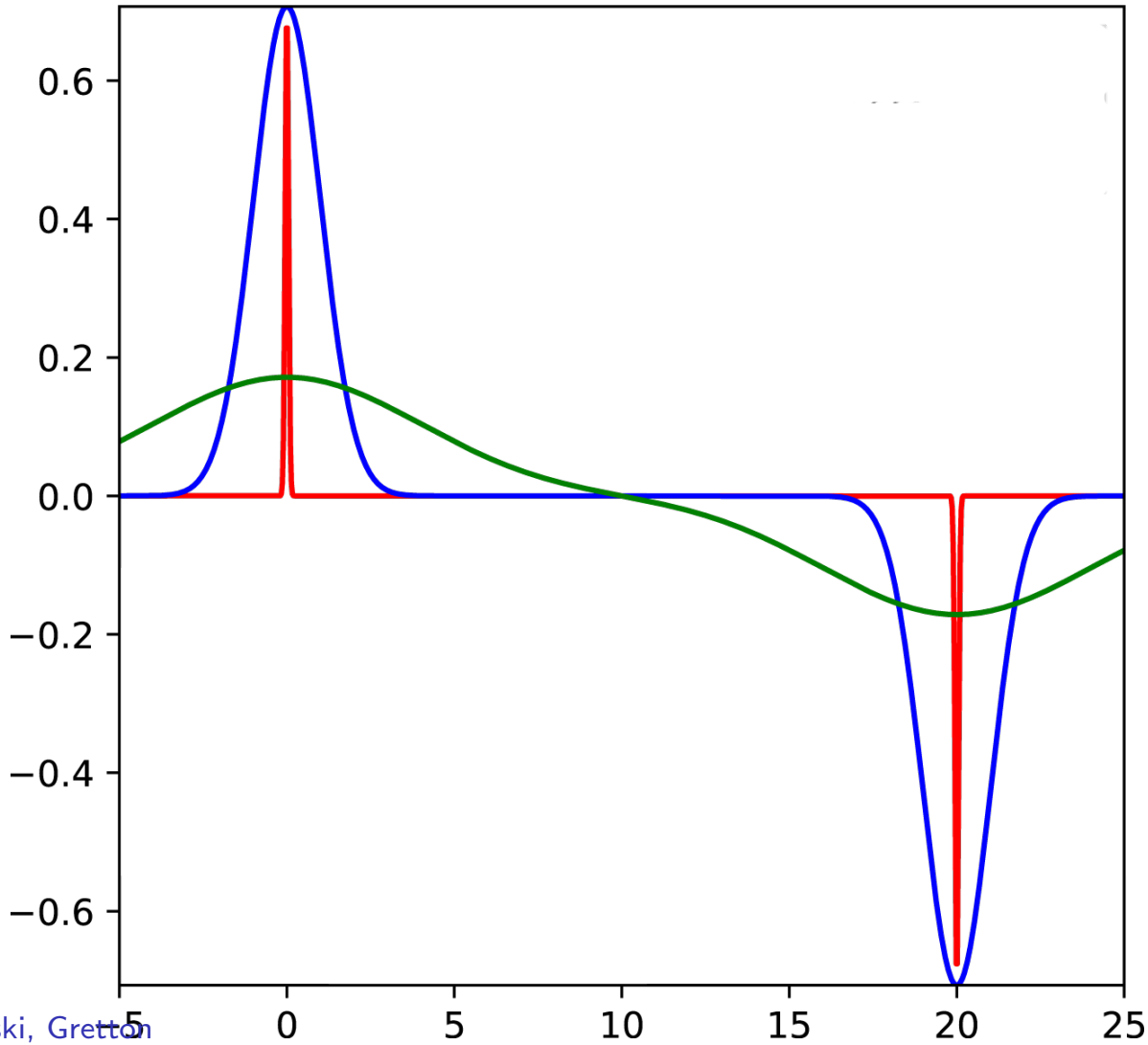
- Point mass target $\mathbb{P} = \delta_0$, model $\mathbb{Q}_\theta = \delta_\theta$
- Representation $\phi_\psi(x) = \psi x$, $\psi \in \mathbb{R}$
- kernel $k_{top}(a, b) = \exp(-\frac{1}{2}(a-b)^2)$





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

- $\mathcal{D}_{MMD}$ is not even continuous in θ !
- Additive gradient penalty doesn't help in practice



A better gradient constraint

New MMD GAN witness regularizer (NeurIPS 2018)

- Constrain the gradient in closed form!

Gradient controlled MMD:

$$GCMMD := \sup_{\|f\|_S \leq 1} \mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]$$

where:

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

A better gradient constraint

New MMD GAN witness regularizer (NeurIPS 2018)

- Constrain the gradient in closed form!

Gradient controlled MMD:

$$GCMMD := \sup_{\|f\|_S \leq 1} \mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]$$

where:

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

problem: not computationally feasible: $O(n^3)$ per iteration.

A better gradient constraint

New MMD GAN witness regularizer (NeurIPS 2018)

- Constrain the gradient in closed form!

The scaled MMD:

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

where:

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

A better gradient constraint

New MMD GAN witness regularizer (NeurIPS 2018)

- Constrain the gradient in closed form!

The scaled MMD:

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

where:

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

- Guarantees $\mathbb{E}_\mu[\|\nabla f(X)\|^2] \leq 1$

A better gradient constraint

New MMD GAN witness regularizer (NeurIPS 2018)

- Constrain the gradient in closed form!

The scaled MMD:

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

where:

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

- Guarantees $\mathbb{E}_\mu[\|\nabla f(X)\|^2] \leq 1$
- Replace expensive constraint with cheap upper bound

$$\mathbb{E}_\mu[\|\nabla f(X)\|^2] \leq \|f\|_S^2 \leq \sigma_\psi^{-2} \|f\|_{\mathcal{H}}^2 \leq 1$$

Scaled MMD GAN

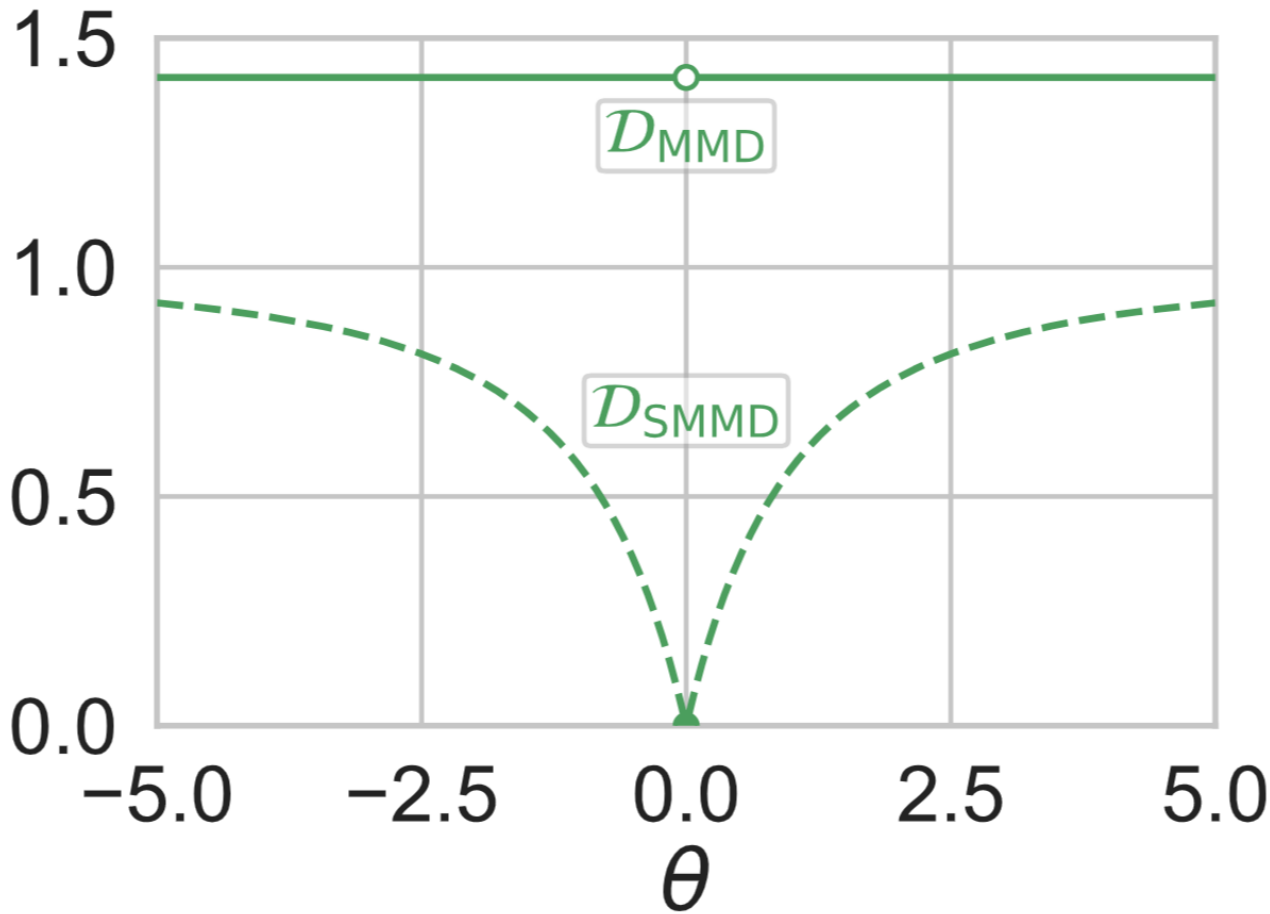
Adversarial distance:

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

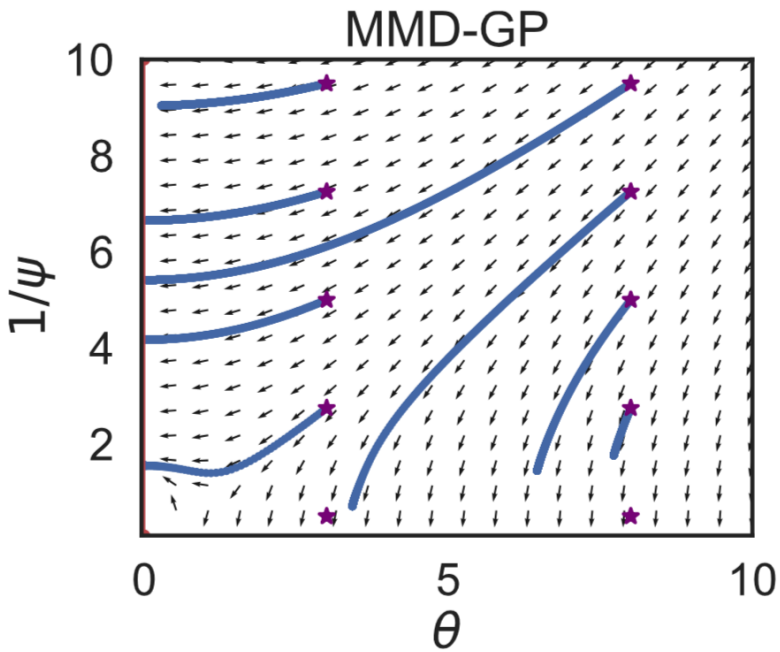
Generator's objective:

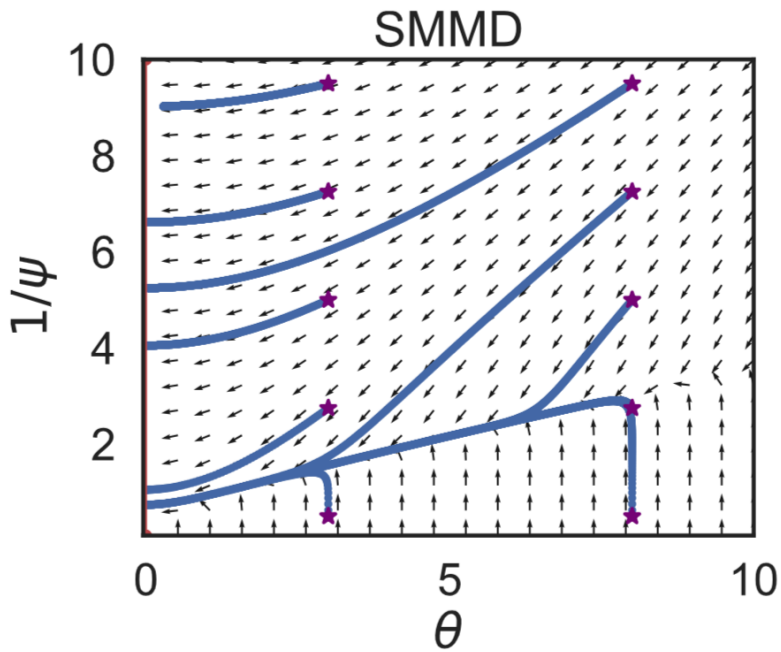
$$\min_{\theta} \mathcal{D}_{SMMD}(\mathbb{P}, G_\theta(\mathbb{Z}))$$

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



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





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:

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

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

Evaluation of GANs

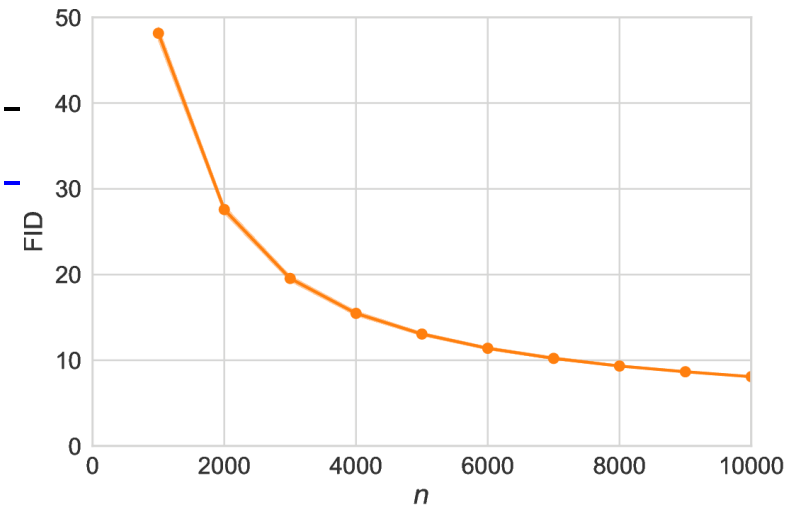
The Frechet inception distance [Heusel et al., 2017] Fits Gaussians to features in a pre-trained inception network:

$$FID(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2 + \|\Sigma_{\mathbb{P}}^{\frac{1}{2}} - \Sigma_{\mathbb{Q}}^{\frac{1}{2}}\|^2$$

where $\mu_{\mathbb{P}}$ and $\Sigma_{\mathbb{P}}$ are the mean and covariance of features of samples from $\mathbb{P}$

Problem: bias. For finite samples can consistently give incorrect answer.

- Bias demo, CIFAR-10 train vs test



Evaluation of GANs

The FID can give the wrong answer in practice. Let $d = 2048$, and define

$$\mathbb{P}_1 = relu(\mathcal{N}(0, I_d)) \quad \mathbb{P}_2 = relu(\mathcal{N}(1, .8\Sigma + .2I_d)) \quad \mathbb{Q} = relu(\mathcal{N}(1, I_d))$$

where $\Sigma = \frac{4}{d} CC^T$, with C a $d \times d$ matrix with iid standard normal entries.

Evaluation of GANs

The FID can give the wrong answer in practice. Let $d = 2048$, and define

$$\mathbb{P}_1 = relu(\mathcal{N}(0, I_d)) \quad \mathbb{P}_2 = relu(\mathcal{N}(1, .8\Sigma + .2I_d)) \quad \mathbb{Q} = relu(\mathcal{N}(1, I_d))$$

where $\Sigma = \frac{4}{d} CC^T$, with C a $d \times d$ matrix with iid standard normal entries. From a random draw of C:

$$FID(\mathbb{P}_1, \mathbb{Q}) \approx 1123.0 > 1114.8 \approx FID(\mathbb{P}_2, \mathbb{Q})$$

With $m = 50000$ samples,

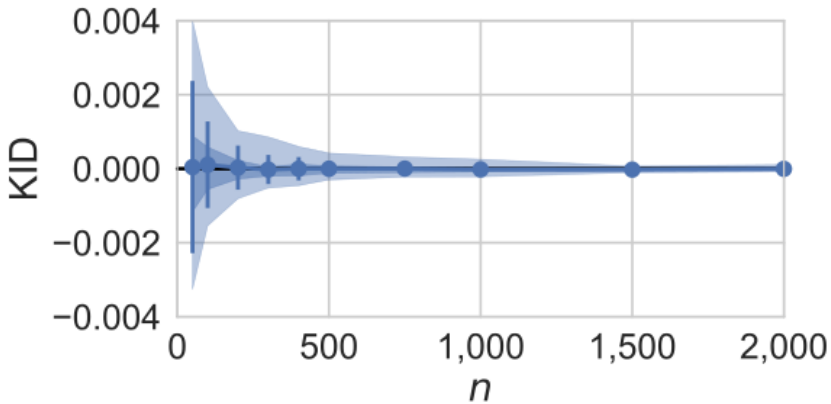
$$FID(\mathbb{P}_1, \mathbb{Q}) \approx 1133.7 < 1136.2 \approx FID(\mathbb{P}_2, \mathbb{Q})$$

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

The kernel inception distance (KID)

The Kernel inception distance [Bińkowski et al., 2018] Measures similarity of the samples' representations in the inception architecture
MMD with kernel

$$k(x, y) = (\frac{1}{d} x^T y + 1)^3$$

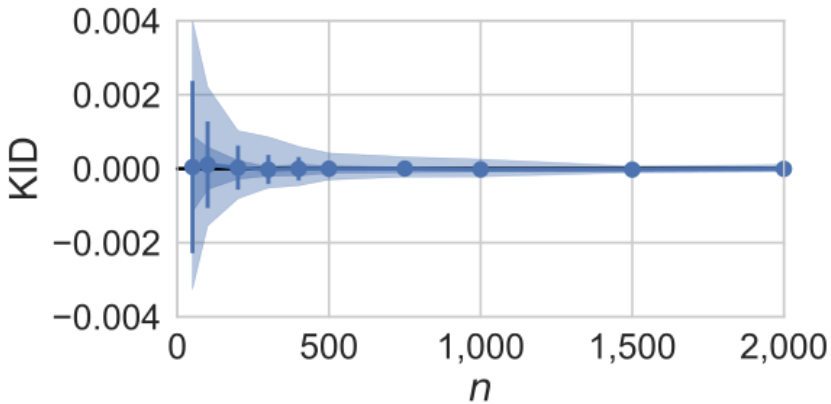


- Checks match for feature means, variances, skewness
- Unbiased: eg CIFAR-10 train/test

The kernel inception distance (KID)

The Kernel inception distance [Bińkowski et al., 2018] Measures similarity of the samples' representations in the inception architecture
MMD with kernel

$$k(x, y) = (\frac{1}{d} x^T y + 1)^3$$

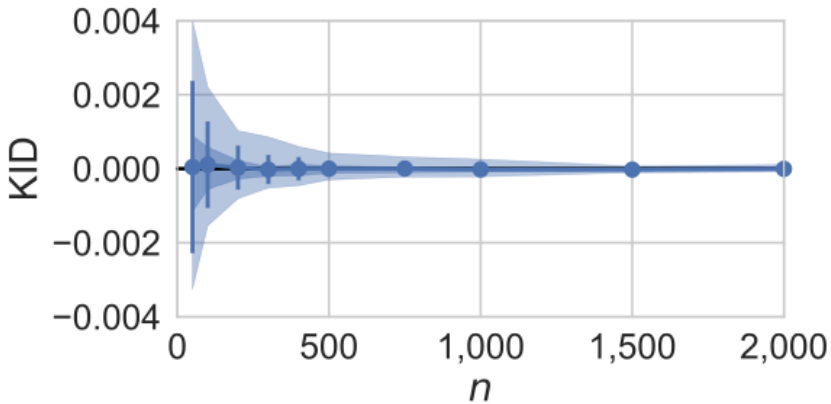


- Checks match for feature means, variances, skewness
- Unbiased: eg CIFAR-10 train/test
- ... "but isn't KID computationally costly? "

The kernel inception distance (KID)

The Kernel inception distance [Bińkowski et al., 2018] Measures similarity of the samples' representations in the inception architecture
MMD with kernel

$$k(x, y) = (\frac{1}{d} x^T y + 1)^3$$



- Checks match for feature means, variances, skewness
- Unbiased: eg CIFAR-10 train/test
- ... "but isn't KID computationally costly? "
- "Block" KID implementation is cheaper than FID: see paper

Benchmarks for comparison (ICLR 2018)

SPECTRAL NORMALIZATION FOR GENERATIVE ADVERSARIAL NETWORKS

Takeru Miyato[1], Toshiki Kataoka[1], Masanori Koyama[2], Yuichi Yoshida[3]
{miyato, kataoka}@preferred.jp
koyama.masanori@gmail.com
yoshi...i.ac.jp
...works, Inc. [2]Ritsumeikan University [3]National Institute of Informatics

We combine with scaled MMD

SOBOLEV GAN

Youssef Mroueh[†], Chun-Liang Li[○,⋆], Tom Sercu[†,⋆], Anant Raj[◊,⋆] & Yu Cheng[†]
† IBM Research AI
○ Carnegie Mellon University
◊ Max Planck Institute for Intelligent Systems
⋆ denotes Equal Contribution
{mroueh,chengyu}@us.ibm.com, chunlial@cs.cmu.edu,
tom.sercul@ibm.com,anant.raj@tuebingen.mpg.de

DEMYSTIFYING MMD GANS

Mikołaj Bińkowski*
Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

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

Our ICLR 2018 paper

BOUNDARY-SEEKING GENERATIVE ADVERSARIAL NETWORKS

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

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

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

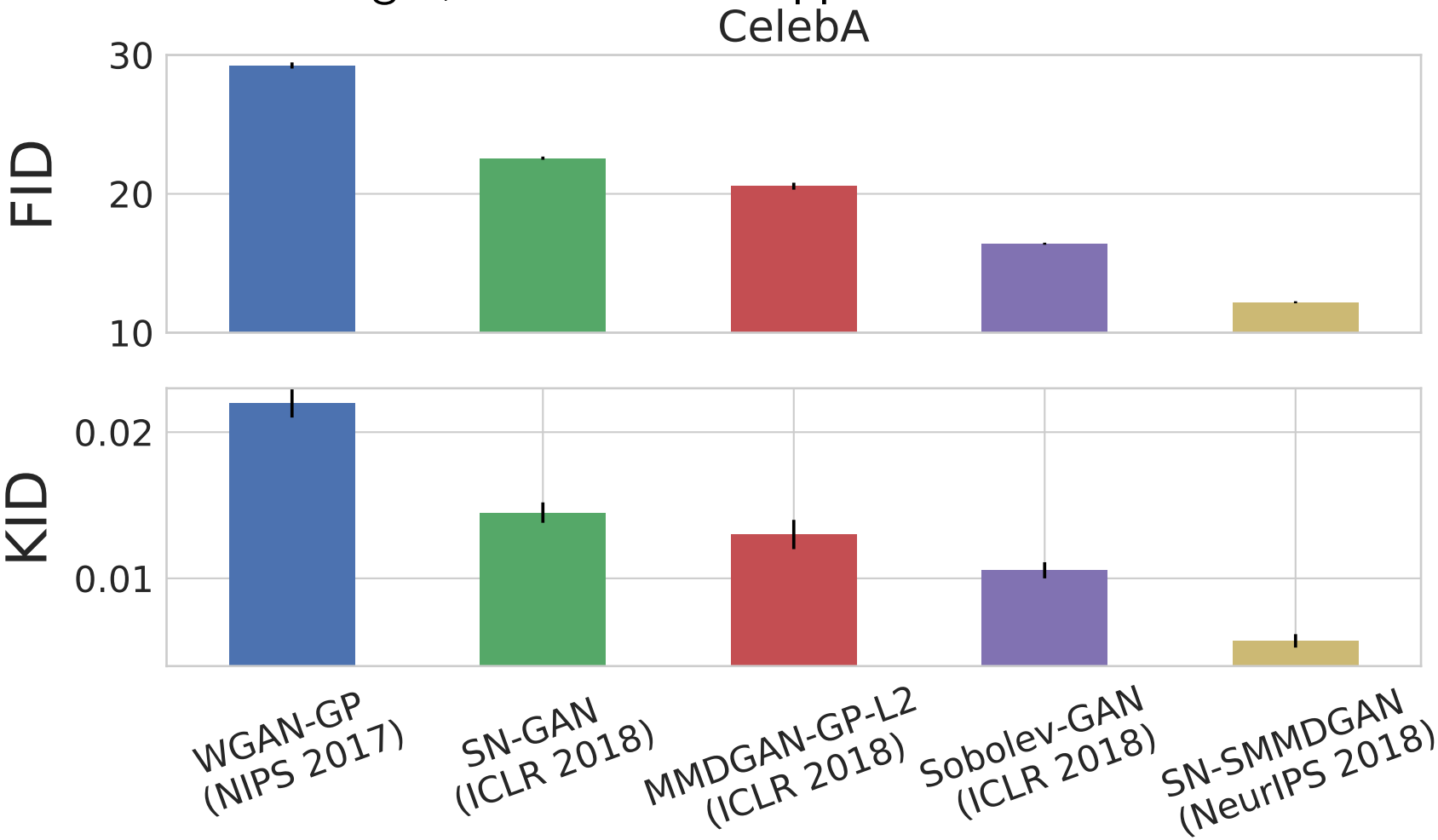
Adam Trischler
MSR
adam.trischler@microsoft.com

Kyunghyun Cho
New York University,
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

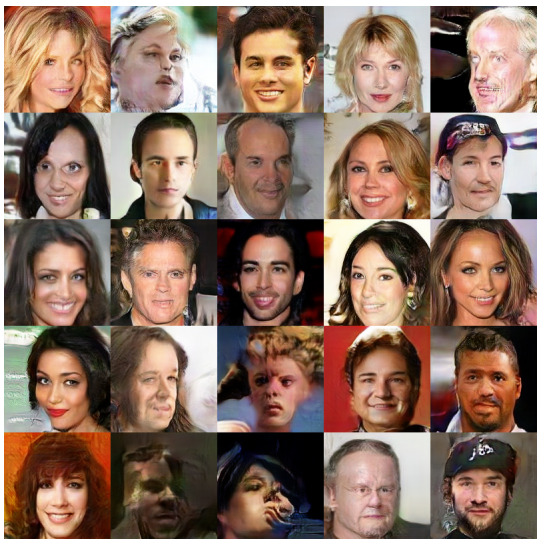
Yoshua Bengio
MILA, University of Montréal, CIFAR, IVADO
yoshua.bengio@umontreal.ca

Experimental results: celebA 160×160

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



Experimental results: celebA 160×160

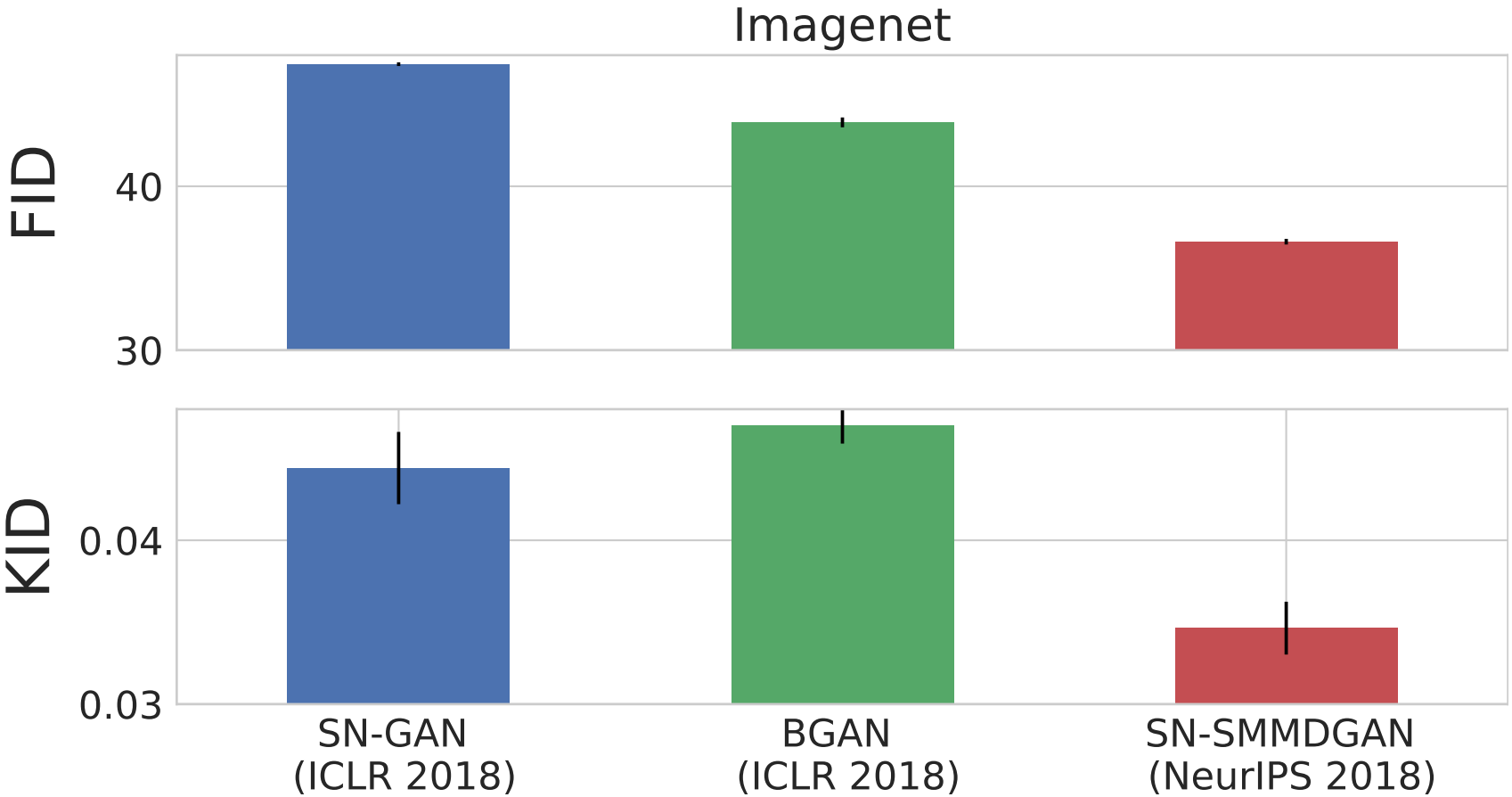


WGAN-GP (NIPS 2017)

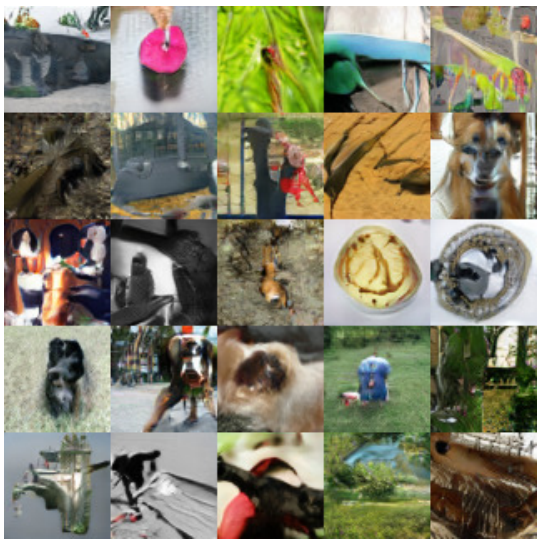
SN-SMMDGAN (ours)

Experimental results: Imagenet 64×64

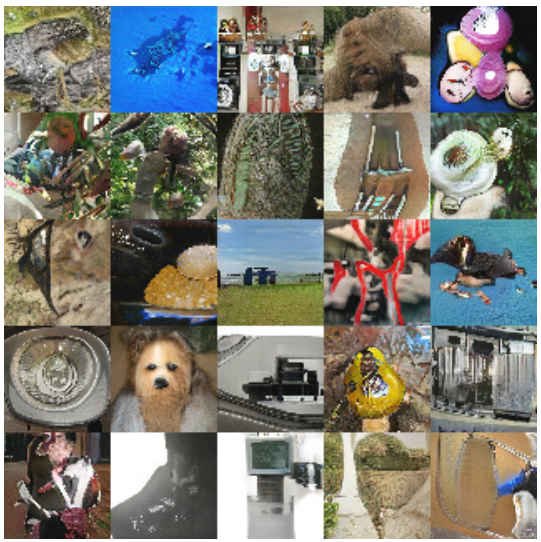
ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64×64. Around 20 000 classes.



Experimental results: Imagenet 64×64



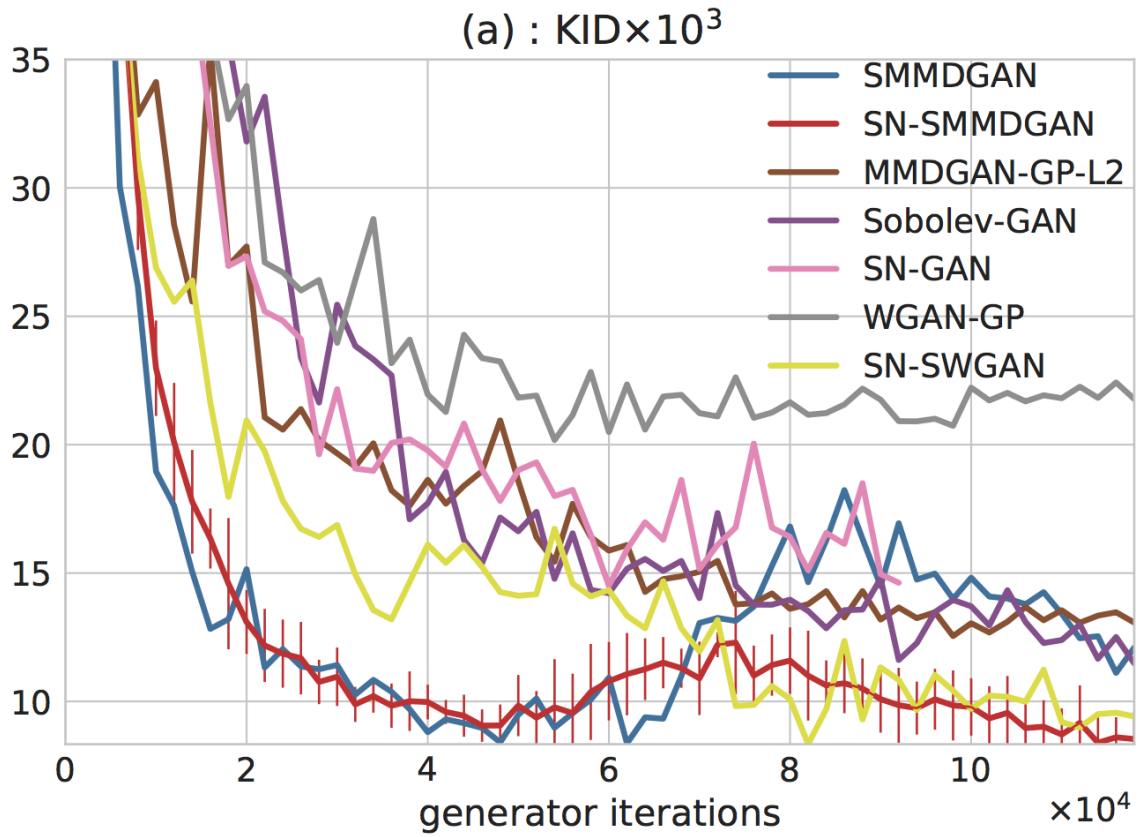
SN-GAN (ICLR 2018)



SN-SMMDGAN (ours)

Experimental results

Faster training: performance scores vs generator iterations on CelebA

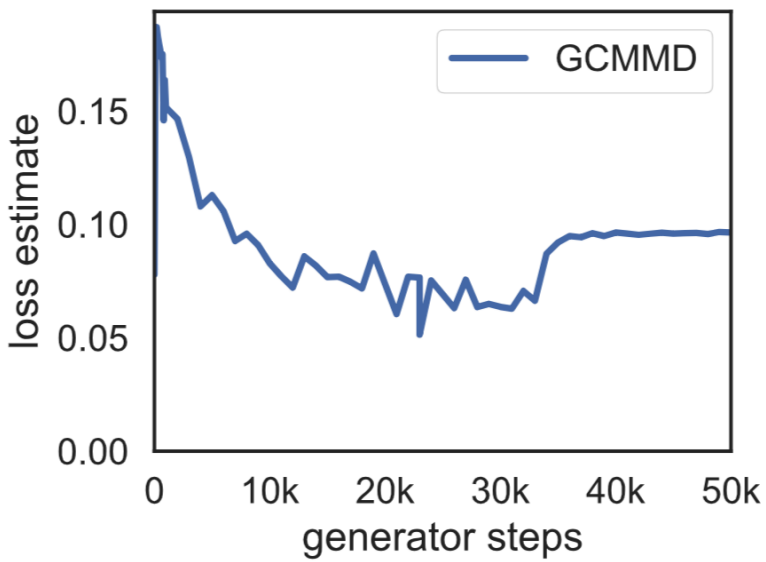


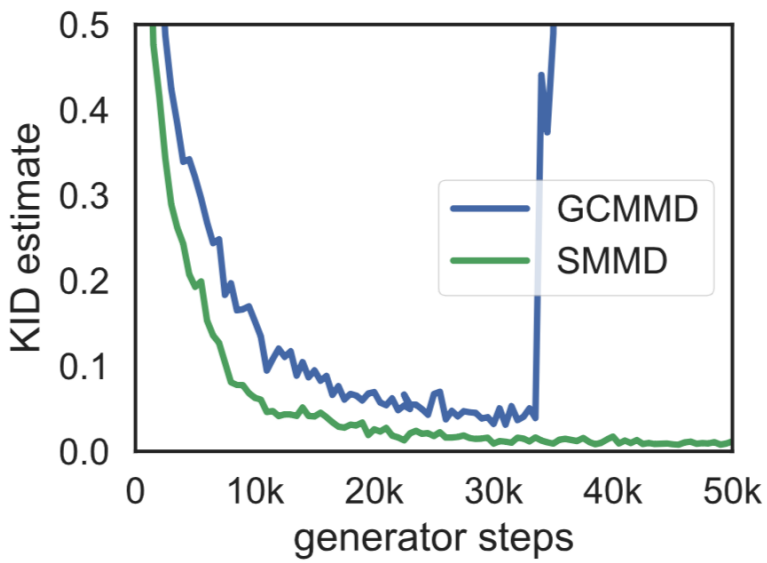
- Spectral parametrization improves training ! (SMMDGAN vs SN-SMMDGAN)

Rank collapse

Optimization failure we sometimes see with SMMD and GCMMD:

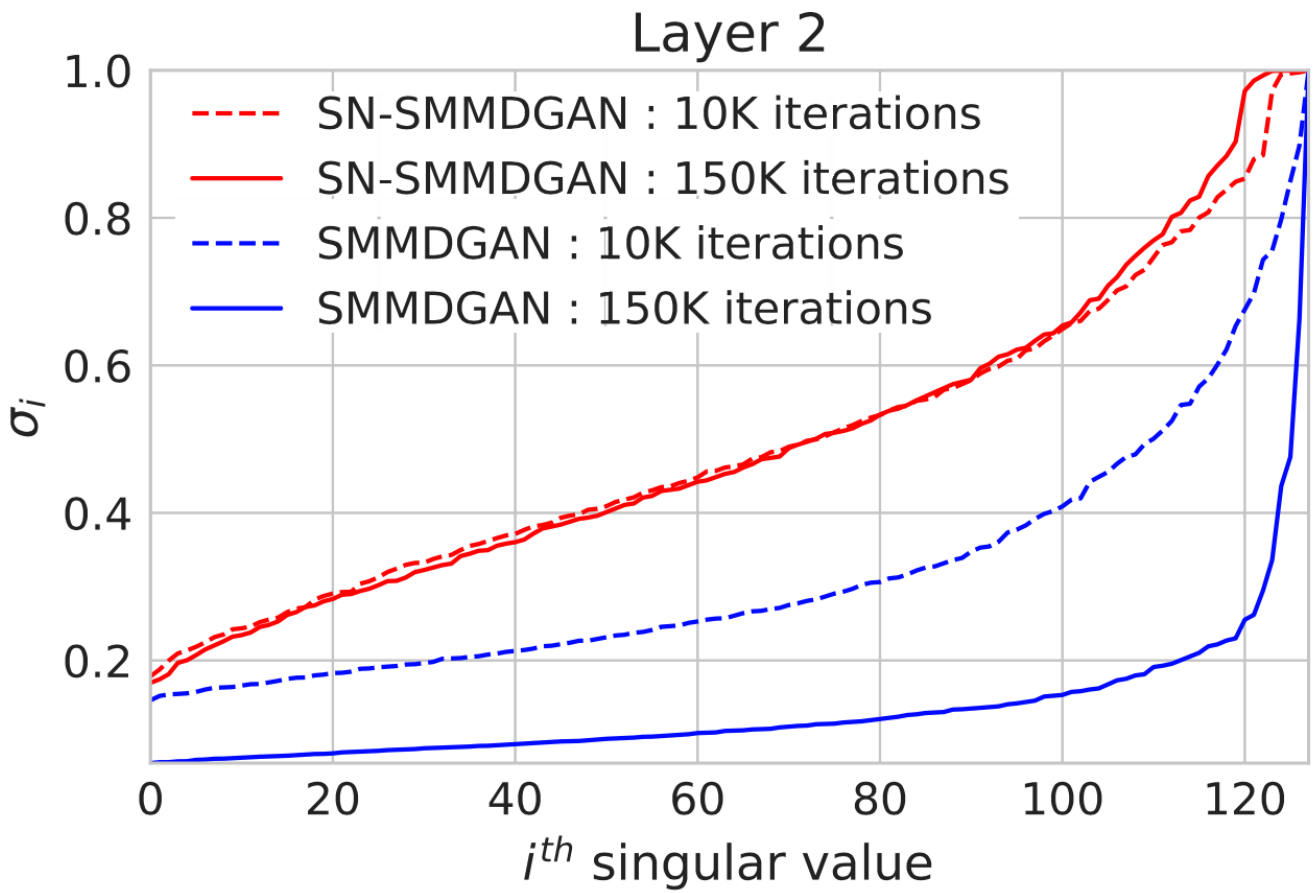
- Generator doing reasonably well
- Critic filters become low-rank
- Generator corrects by breaking everything else



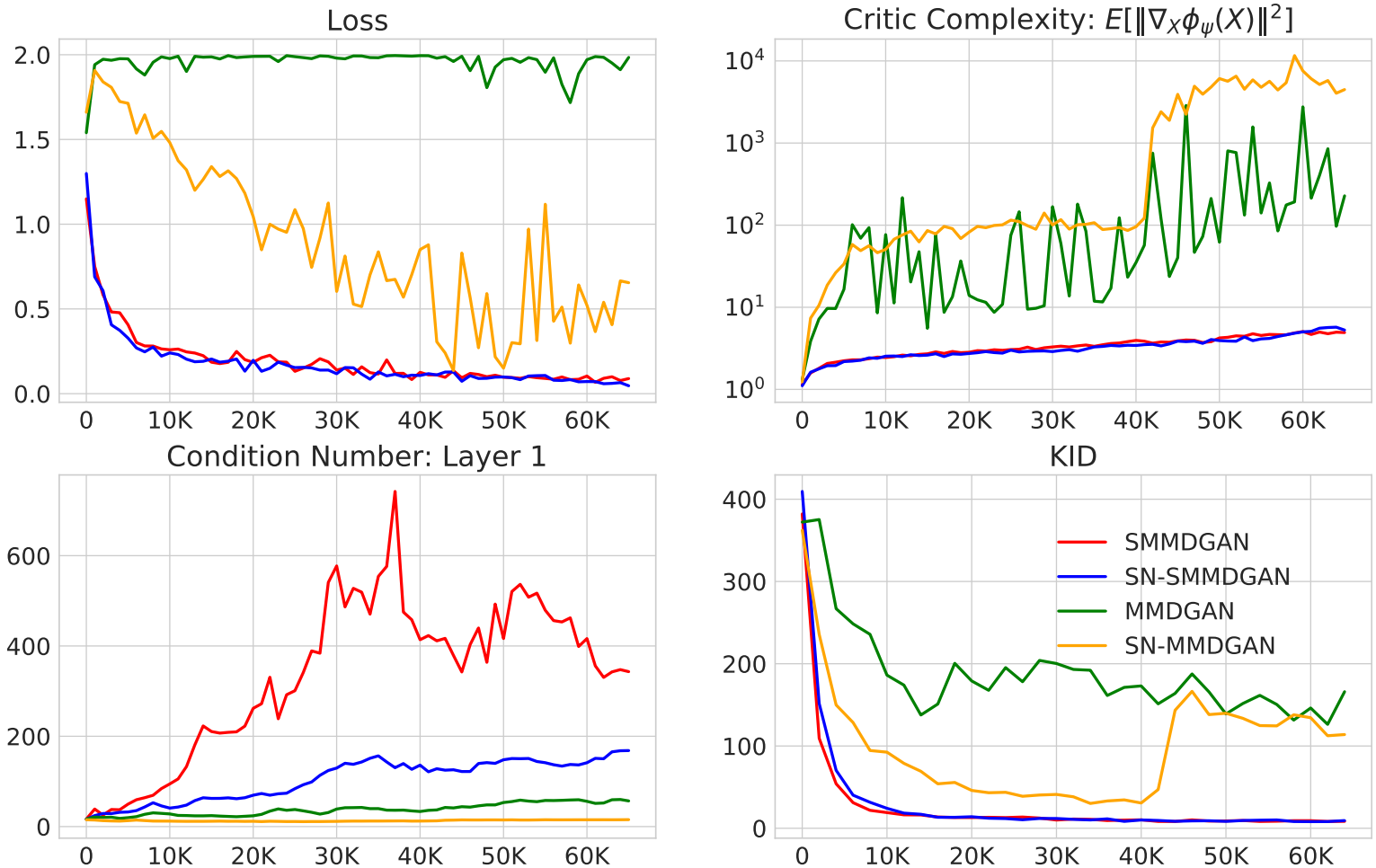


Spectral parametrization [Miyato et al., 2018]

- Use filters: $W = \gamma \bar{W} / \|\bar{W}\|_{op}$, learn γ and $\bar{W}$ freely.
- Encourages diversity without limiting representation
- In practice, controls the condition number of W



Experimental results: ablation studies

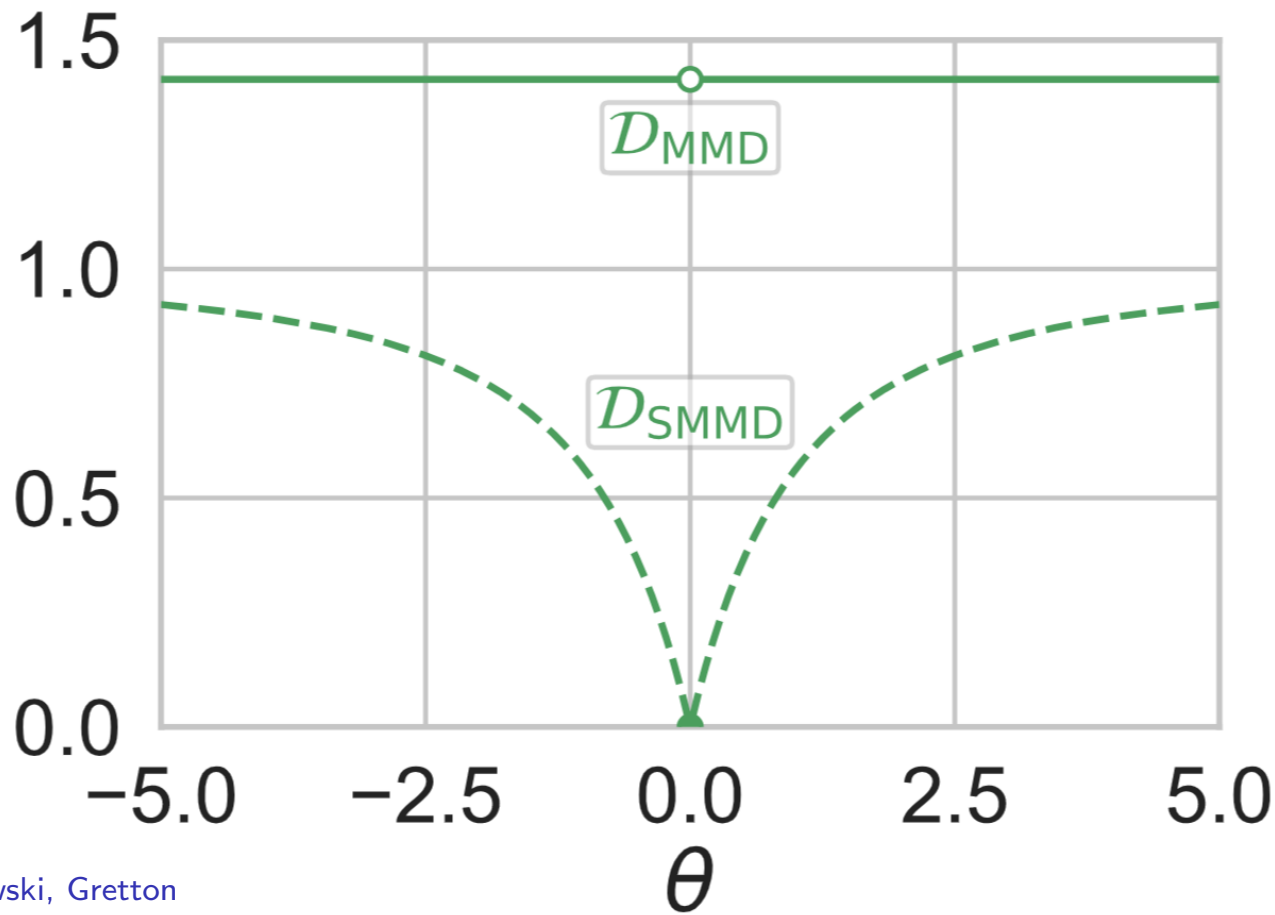


- SN: controls the condition number of the weights W
- SMMD: controls the critic complexity
- SN + MMD: unstable training!
- SN + SMMD: optimal performance (Why?)

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

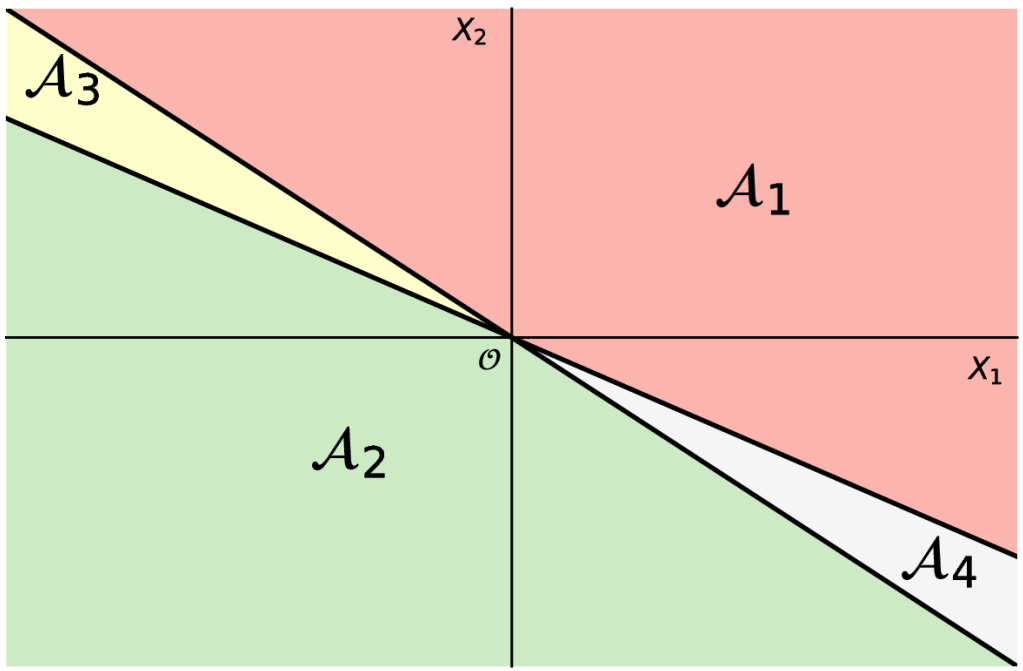
$$\mathcal{D}_{SMMD}(\mathbb{P}, G_\theta(\mathbb{Z})) := \max_\psi \sigma_\psi MMD(\phi_\psi(\mathbb{P}), \phi_\psi(G_\theta(\mathbb{Z})))$$

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



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

- $\mathcal{D}_{SMMD}$ is smooth if $\|\phi_\psi\|_{Lip} \leq 1$ for all X.
- $\mathcal{D}_{SMMD}$ only constrains $\mathbb{E}_\mu[\|\nabla\phi_\psi(X)\|^2] \leq 1$.
- In general: $\mathbb{E}_\mu[\|\nabla\phi_\psi(X)\|^2] \leq 1$ does not imply $\|\nabla\phi_\psi(X)\|^2 \leq 1$ for all X ...



But it holds true under some conditions

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

Adversarial distance:

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

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

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

- μ has positive density.
- ϕ_ψ is fully connected with Leaky-ReLU and non-increasing width.
- Condition number of the weights per-layer in ϕ_ψ is bounded.

Theorem: proof outline

$$
\begin{aligned}
\mathit{SMMD}(\mathbb{P}, \mathbb{Q}) &\leq C \frac{\|\phi_\psi\|_{\mathit{Lip}}}{\mathbb{E}_\mu[\|\nabla \phi_\psi(X)\|]} W_1(\mathbb{P}, \mathbb{Q}) \\
&\leq C \frac{\beta_\psi \|\phi_{\bar{\psi}}\|_{\mathit{Lip}}}{\beta_\psi \, \mathbb{E}_\mu[\|\nabla \phi_{\bar{\psi}}(X)\|]} W_1(\mathbb{P}, \mathbb{Q})
\end{aligned}
$$

Theorem: proof outline

$$SMMD(\mathbb{P},\mathbb{Q}) \leq C \frac{\|\phi_\psi\|_{Lip}}{\mathbb{E}_\mu[\|\nabla\phi_\psi(X)\|]} W_1(\mathbb{P},\mathbb{Q})$$
$$\leq C \frac{\beta_\psi \|\phi_{\bar{\psi}}\|_{Lip}}{\beta_\psi\, \mathbb{E}_\mu[\|\nabla\phi_{\bar{\psi}}(X)\|]} W_1(\mathbb{P},\mathbb{Q})$$

For fully connected network with α-leaky ReLU and condition number bounded by κ:

$$\|\nabla\phi_\psi(X)\|^2 \geq \frac{d_L \alpha^L}{\kappa^L}$$

for almost all $X \in \mathbb{R}^d$

Conclusion

- MMD-based losses are effective and principled for training implicit generative models.
- Regularizing the parametric kernels for MMD GANs is crucial to get well behaved losses.
- Adapting the amplitude of the MMD to the smoothness of the kernel provides a simple way to achieve regularization.
- A new insight on desired properties for the critic network.
- State of the art results on challenging datasets

Future directions:

- How to relate the generator and critic architecture?
- Why adversarial distances suffer less from the curse of dimensionality?

Thank you !


Arjovsky, M., Chintala, S., and Bottou, L. (2017).
Wasserstein gan.


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.
arXiv:1505.03906 [cs, stat].
arXiv: 1505.03906.


Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012).
Optimal kernel choice for large-scale two-sample tests.
In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1205–1213. Curran Associates, Inc.


Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., and Courville, A. (2016).