

GANs with integral probability metrics: some results and conjectures

Arthur Gretton



Gatsby Computational Neuroscience Unit,
University College London

MILA, Montreal, 2019

A motivation: comparing two samples

- Given: Samples from unknown distributions P and Q .
- Goal: do P and Q differ?



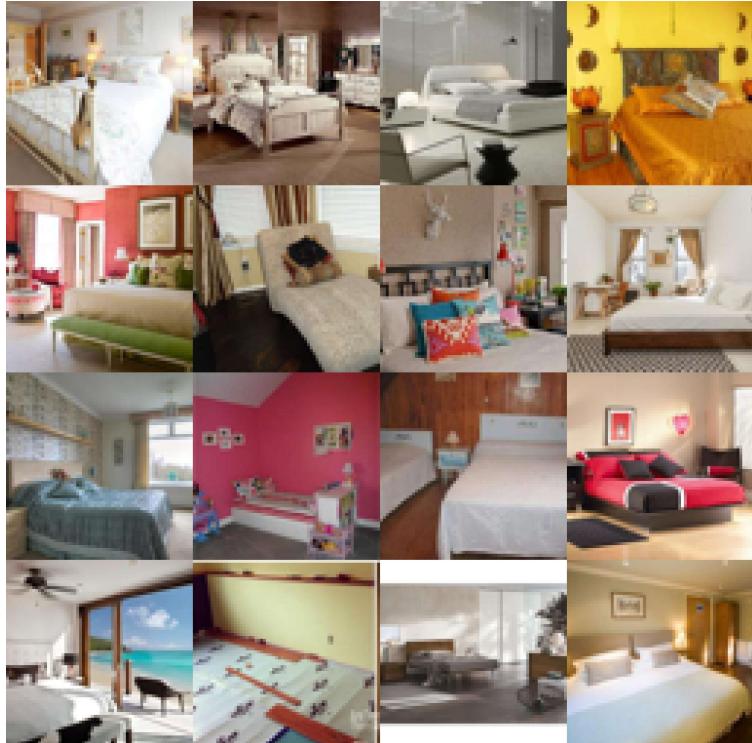
$\sim P$



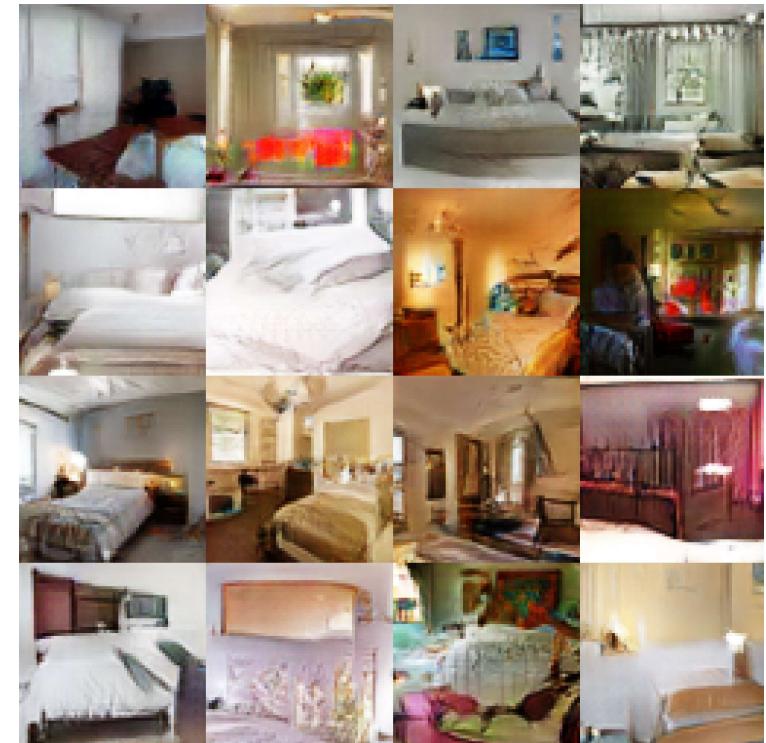
$\sim Q$

Training implicit generative models

- Have: One collection of samples X from unknown distribution P .
- Goal: generate samples Q that look like P



LSUN bedroom samples P



Generated Q , MMD GAN

Using a critic $D(P, Q)$ to train a GAN

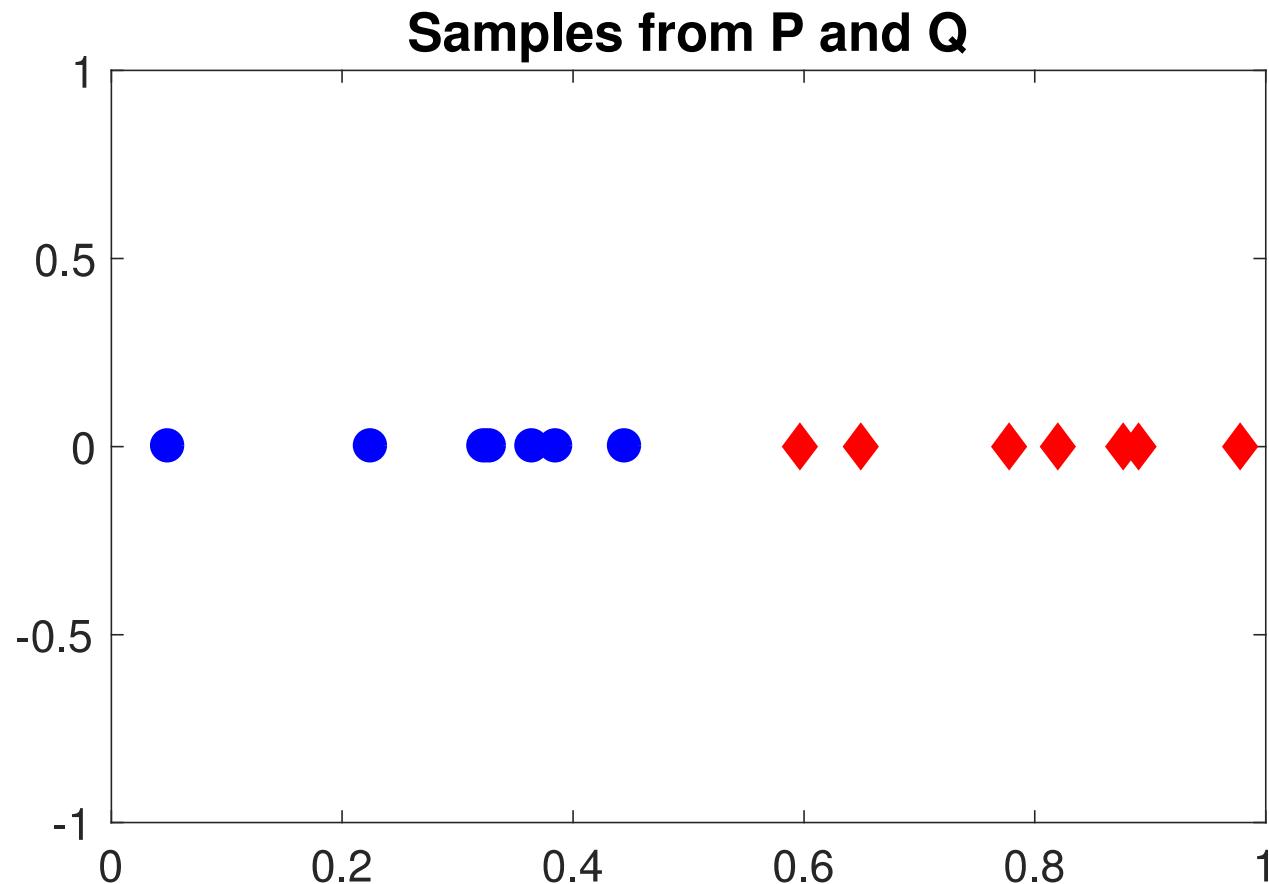
Outline

- Measures of distance between distributions
 - The MMD: an integral probability metric
 - f-divergences vs integral probability metrics
- Gradient penalties for GAN critics
 - The optimisation viewpoint
 - The regularisation viewpoint
- Theory
 - Relation of MMD critic and Wasserstein
 - Gradient bias
- Evaluating GAN performance, experiments

The Maximum Mean Discrepancy: An Integral Probability Metric

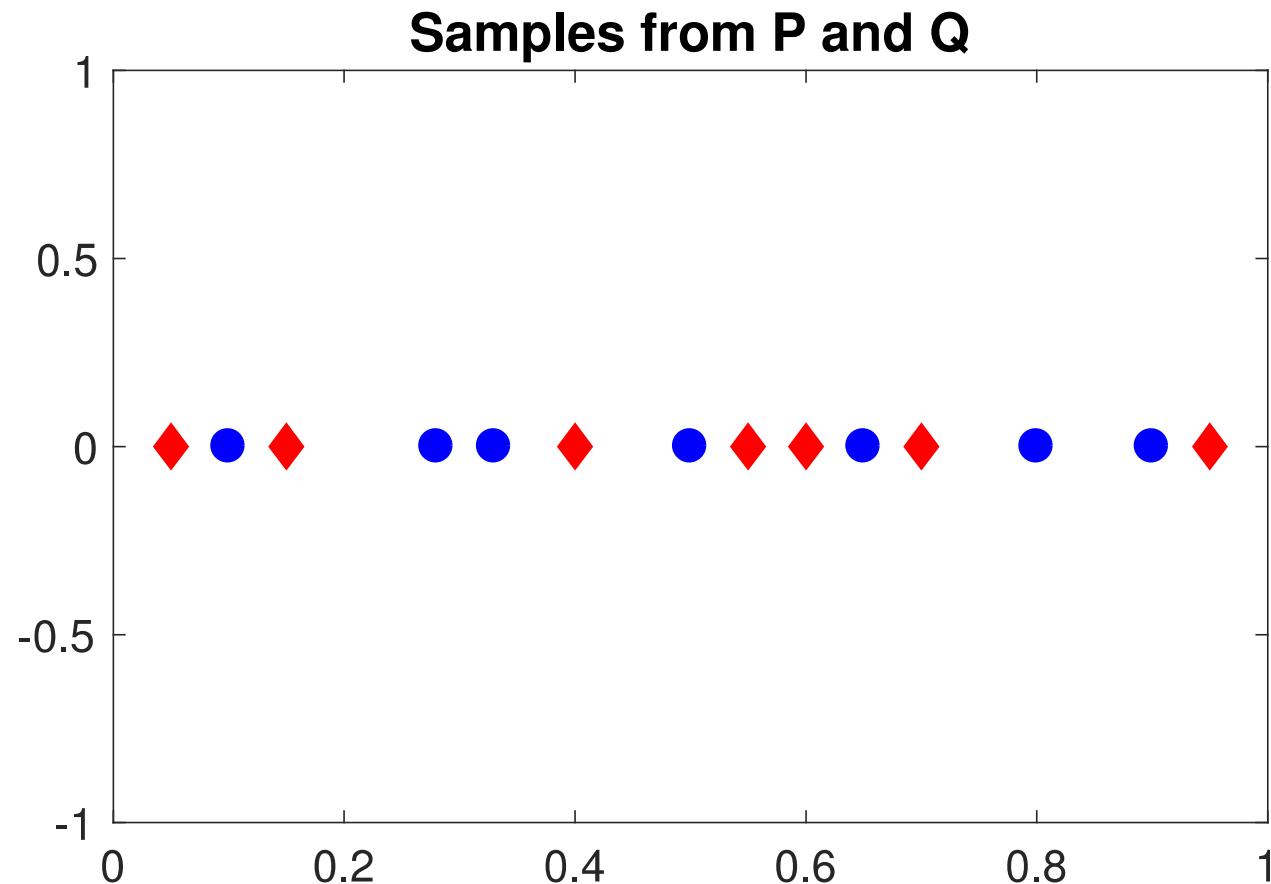
Integral probability metrics

Are P and Q different?



Integral probability metrics

Are P and Q different?

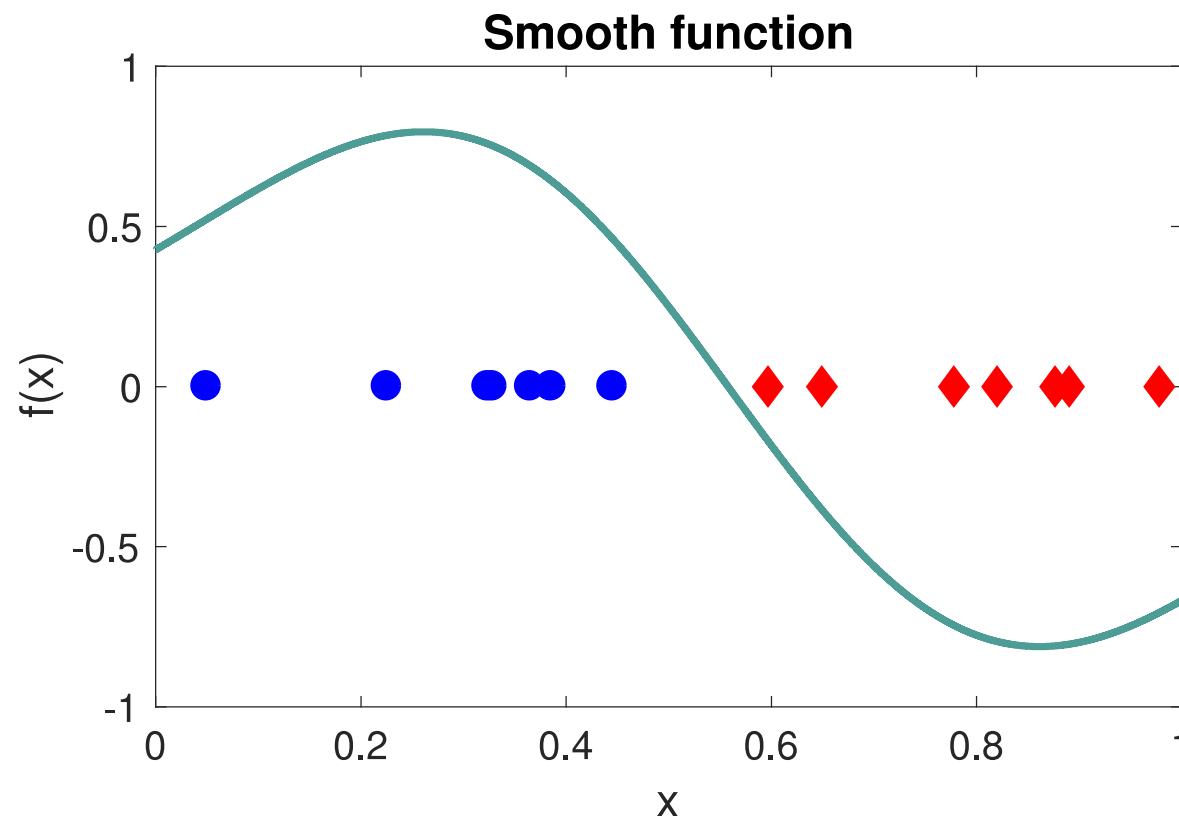


Integral probability metrics

Integral probability metric:

Find a "well behaved function" $f(x)$ to maximize

$$\mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)$$

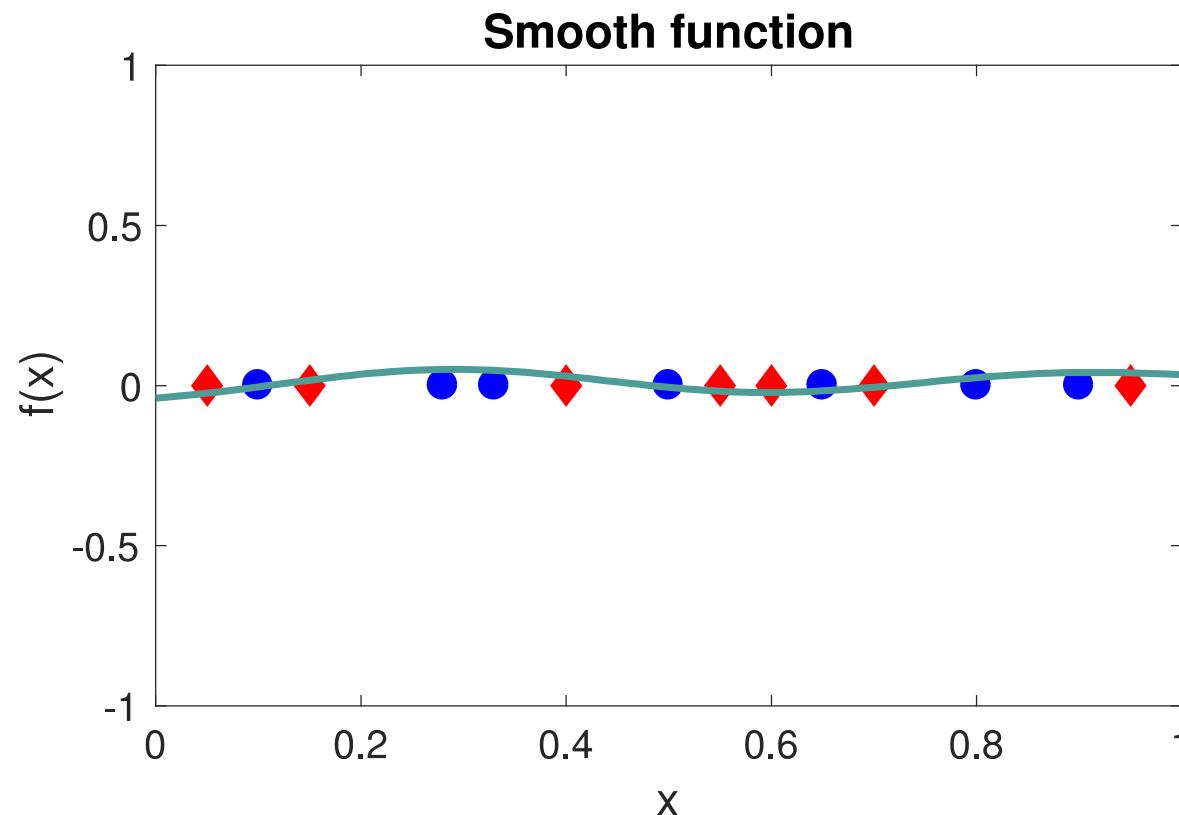


Integral probability metrics

Integral probability metric:

Find a "well behaved function" $f(x)$ to maximize

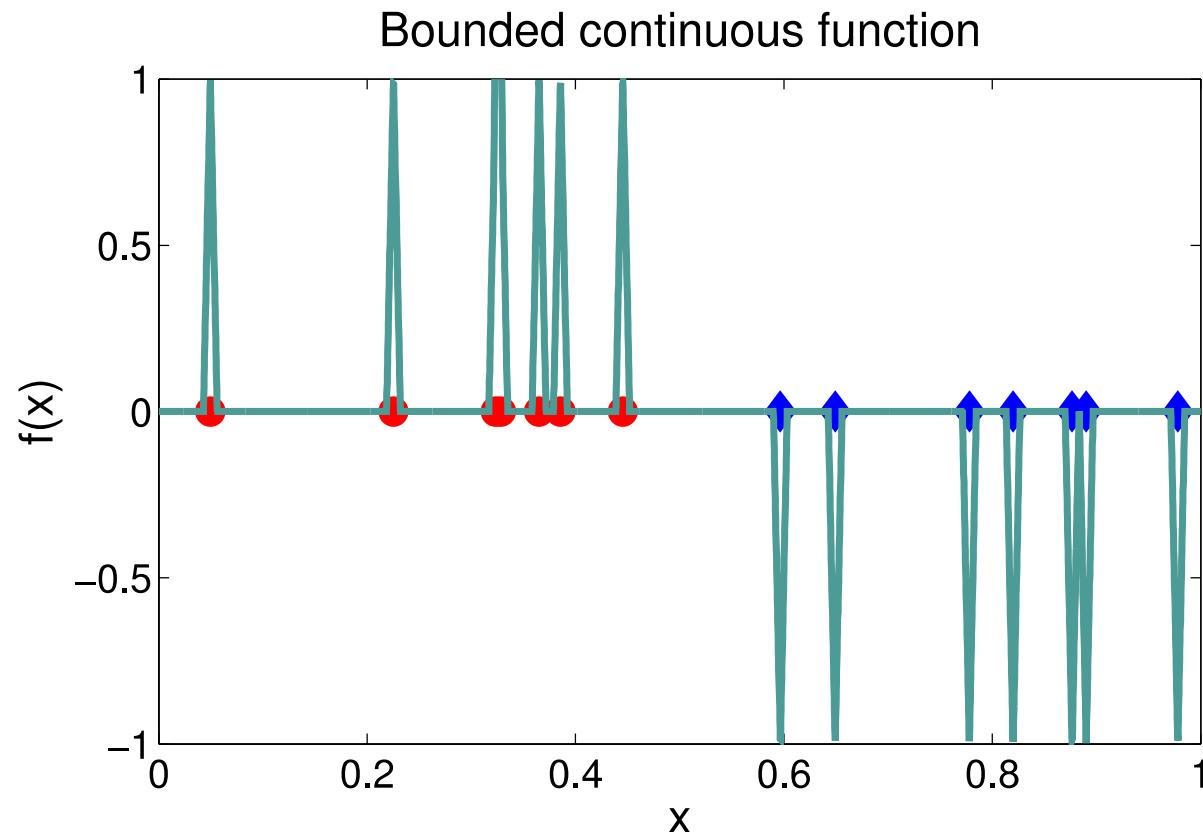
$$\mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)$$



Integral probability metrics

What if the function is not well behaved?

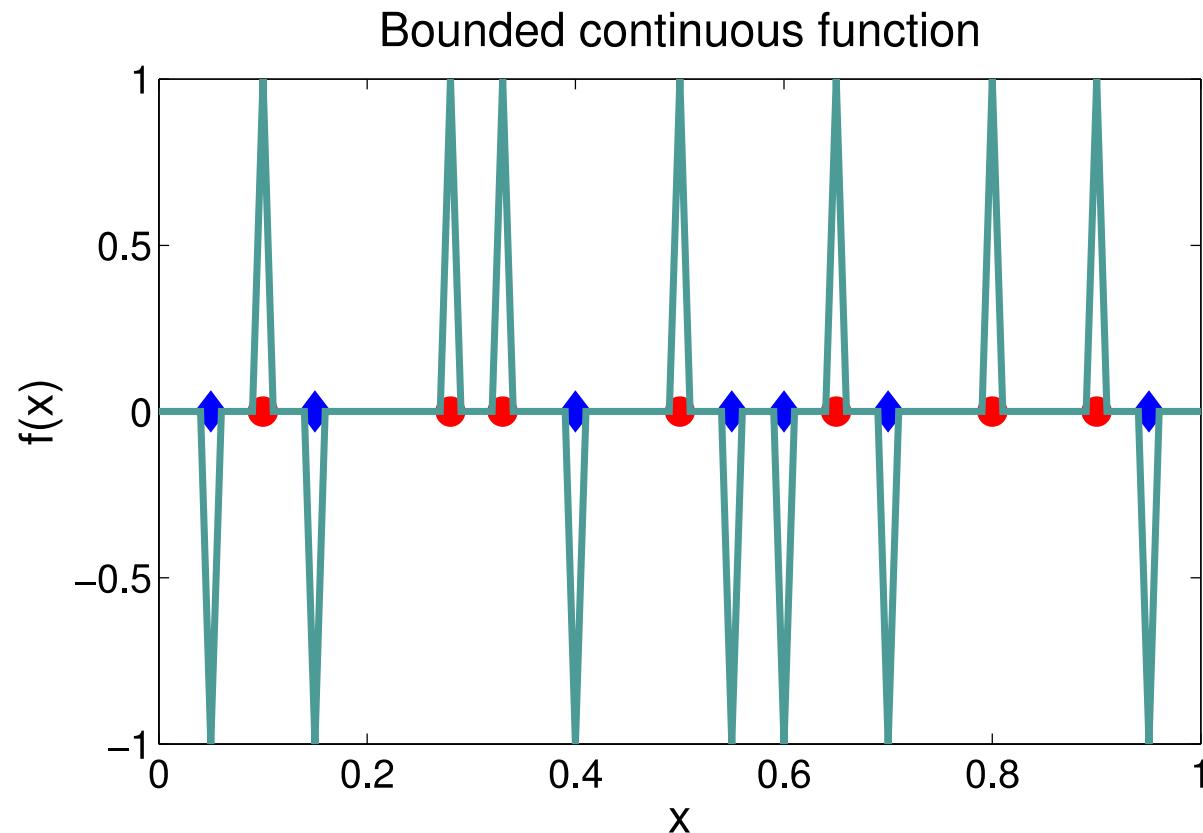
$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$



Integral probability metrics

What if the function is not well behaved?

$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)$$



The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

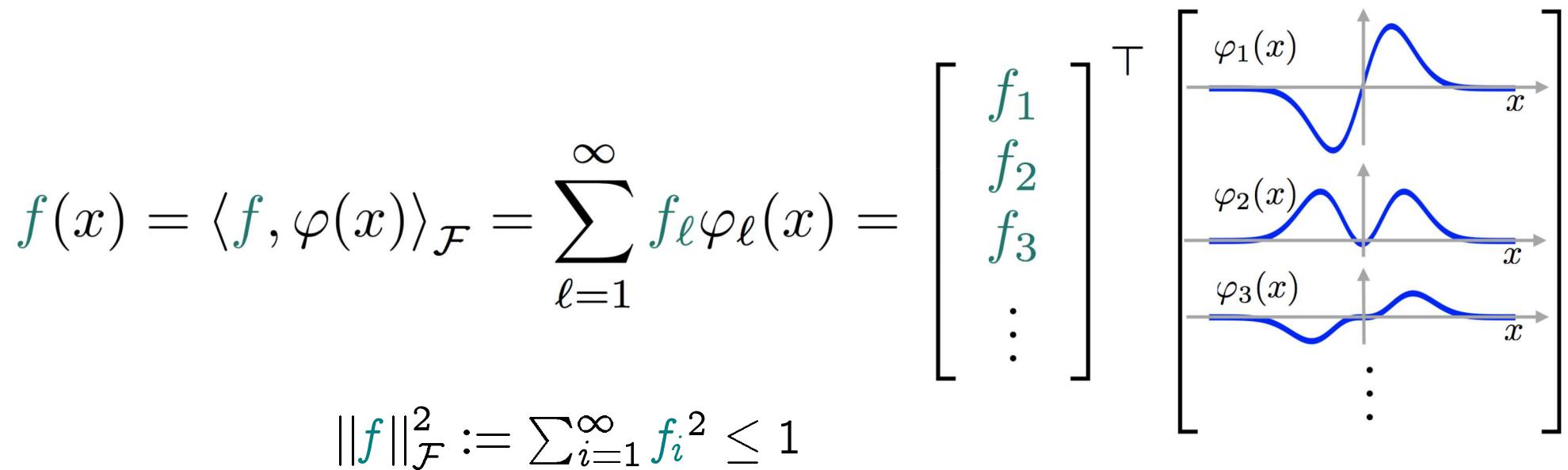
The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

$$f(x) = \langle \mathbf{f}, \varphi(x) \rangle_{\mathcal{F}} = \sum_{\ell=1}^{\infty} f_{\ell} \varphi_{\ell}(x) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}^T \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \\ \vdots \end{bmatrix}$$


$$\|\mathbf{f}\|_{\mathcal{F}}^2 := \sum_{i=1}^{\infty} f_i^2 \leq 1$$

The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

For characteristic RKHS \mathcal{F} , $MMD(P, Q; \mathcal{F}) = 0$ iff $P = Q$

Other choices for witness function class:

- Bounded continuous [Dudley, 2002]
- Bounded variation 1 (Kolmogorov metric) [Müller, 1997]
- Lipschitz (Wasserstein distances) [Dudley, 2002]

The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

Expectations of functions are linear combinations
of expected features

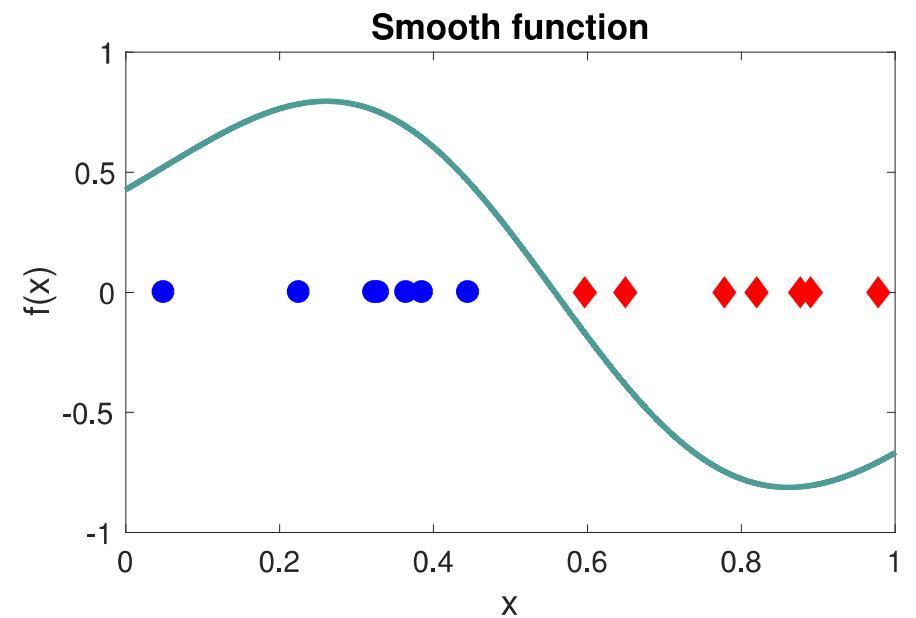
$$\mathbf{E}_P(f(X)) = \langle f, \mathbf{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

Integral prob. metric vs feature mean difference

The MMD:

$$\begin{aligned} MMD(P, Q; F) \\ = \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)] \end{aligned}$$



Integral prob. metric vs feature mean difference

The MMD:

use

$$\begin{aligned} MMD(P, Q; F) &= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)] \\ &= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \end{aligned}$$
$$\mathbf{E}_P f(X) = \langle \mu_P, f \rangle_{\mathcal{F}}$$

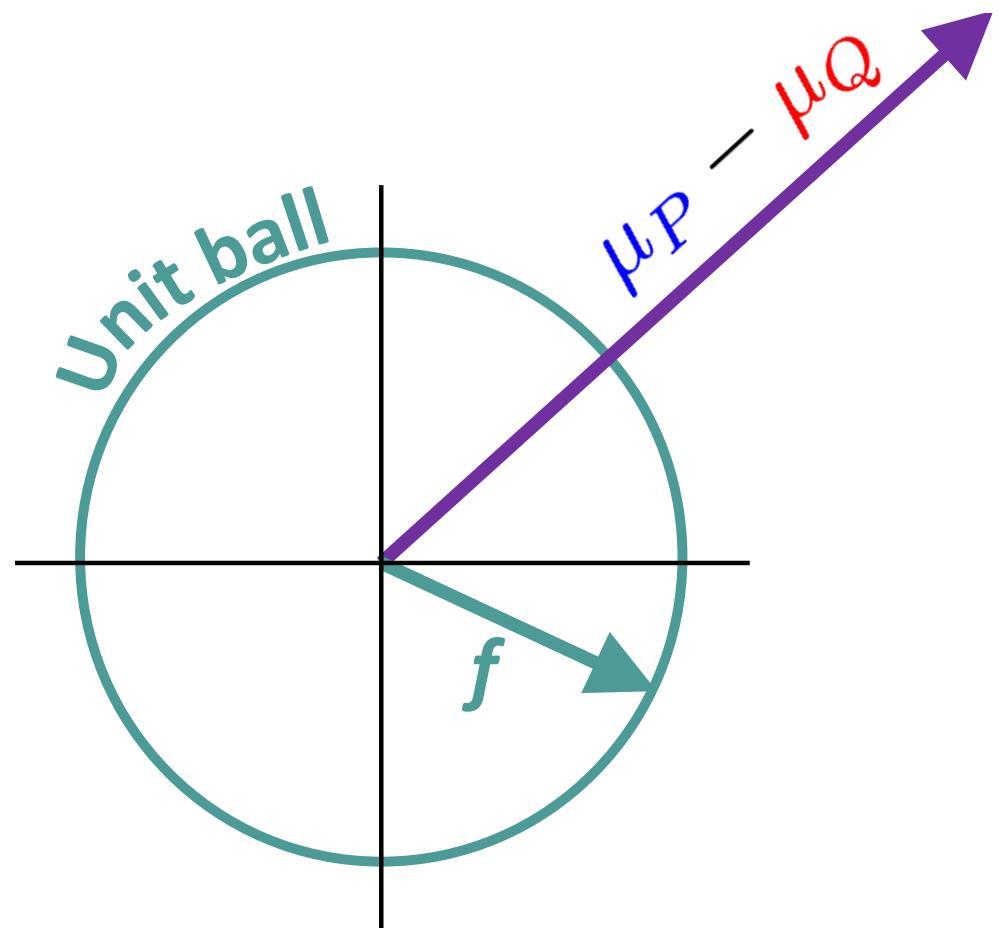
Integral prob. metric vs feature mean difference

The MMD:

$$MMD(P, Q; F)$$

$$= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$

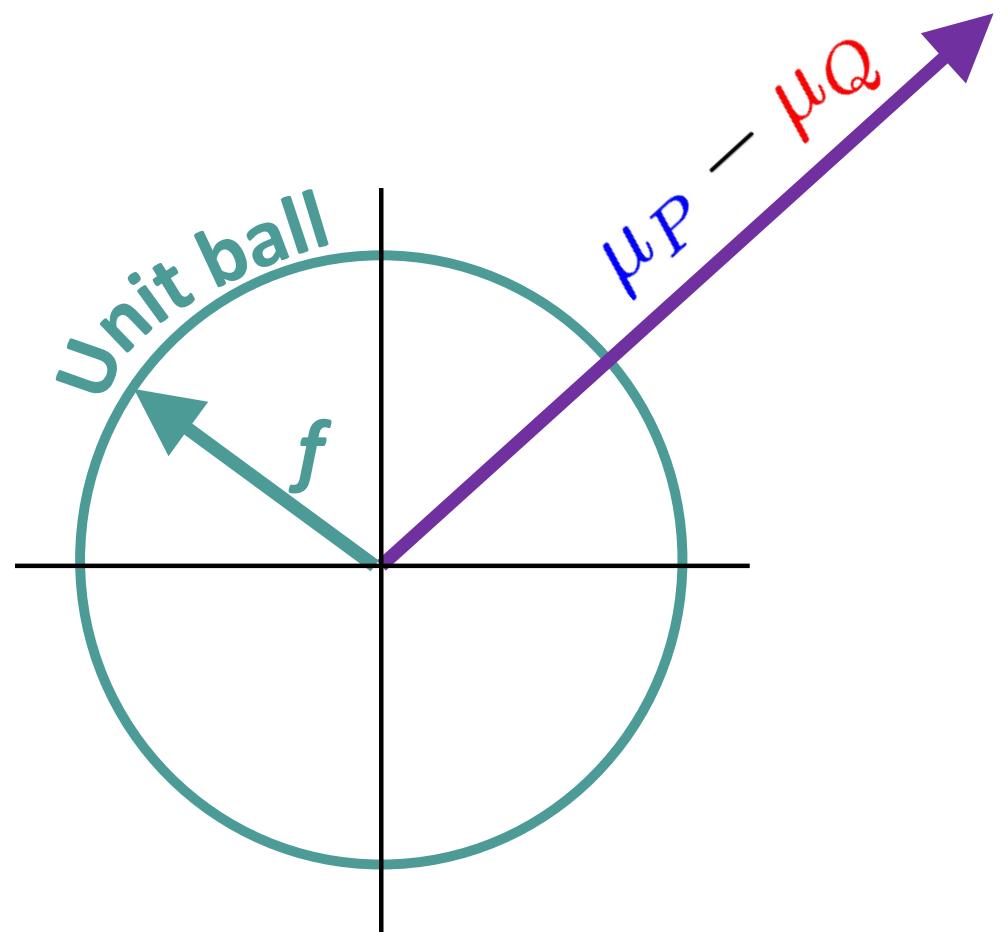
$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature mean difference

The MMD:

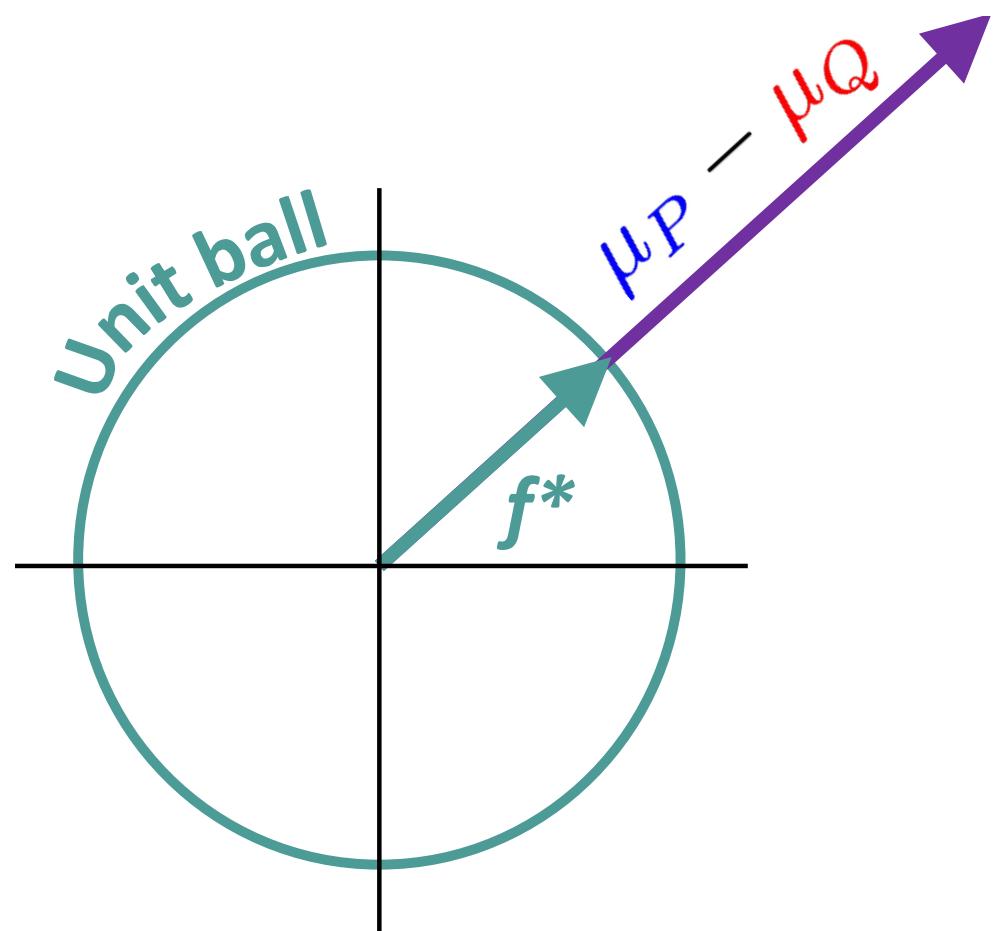
$$\begin{aligned} MMD(P, Q; \mathcal{F}) &= \sup_{f \in \mathcal{F}} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)] \\ &= \sup_{f \in \mathcal{F}} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \end{aligned}$$



Integral prob. metric vs feature mean difference

The MMD:

$$\begin{aligned} MMD(P, Q; \mathcal{F}) &= \sup_{f \in \mathcal{F}} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)] \\ &= \sup_{f \in \mathcal{F}} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \end{aligned}$$



$$f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$$

Integral prob. metric vs feature mean difference

The MMD:

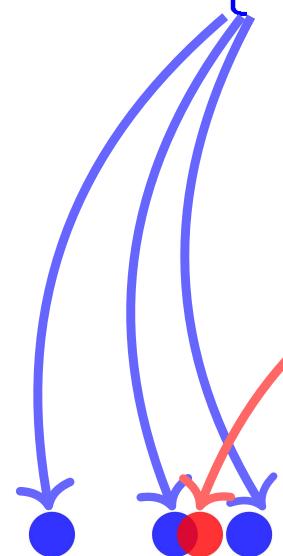
$$\begin{aligned} MMD(P, Q; F) &= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)] \\ &= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\ &= \|\mu_P - \mu_Q\| \end{aligned}$$

IPM view equivalent to feature mean difference (kernel case only)

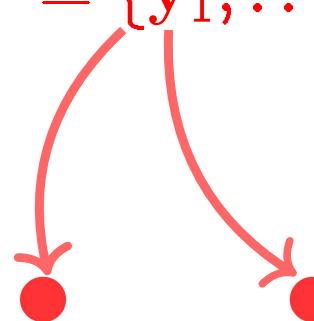
Construction of MMD witness

Construction of empirical **witness function** (proof: next slide!)

Observe $X = \{x_1, \dots, x_n\} \sim P$

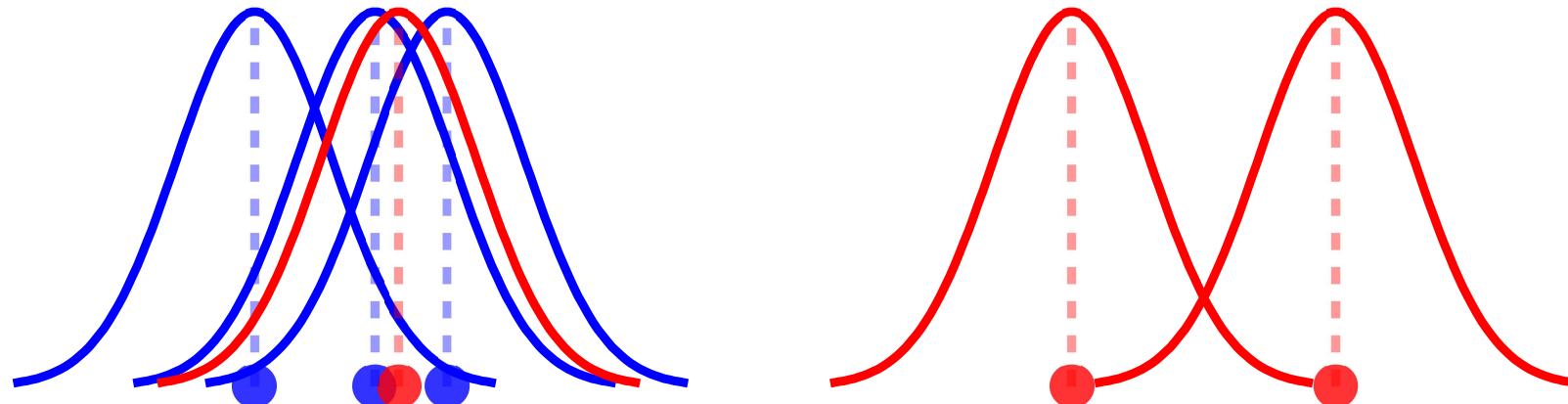


Observe $Y = \{y_1, \dots, y_n\} \sim Q$



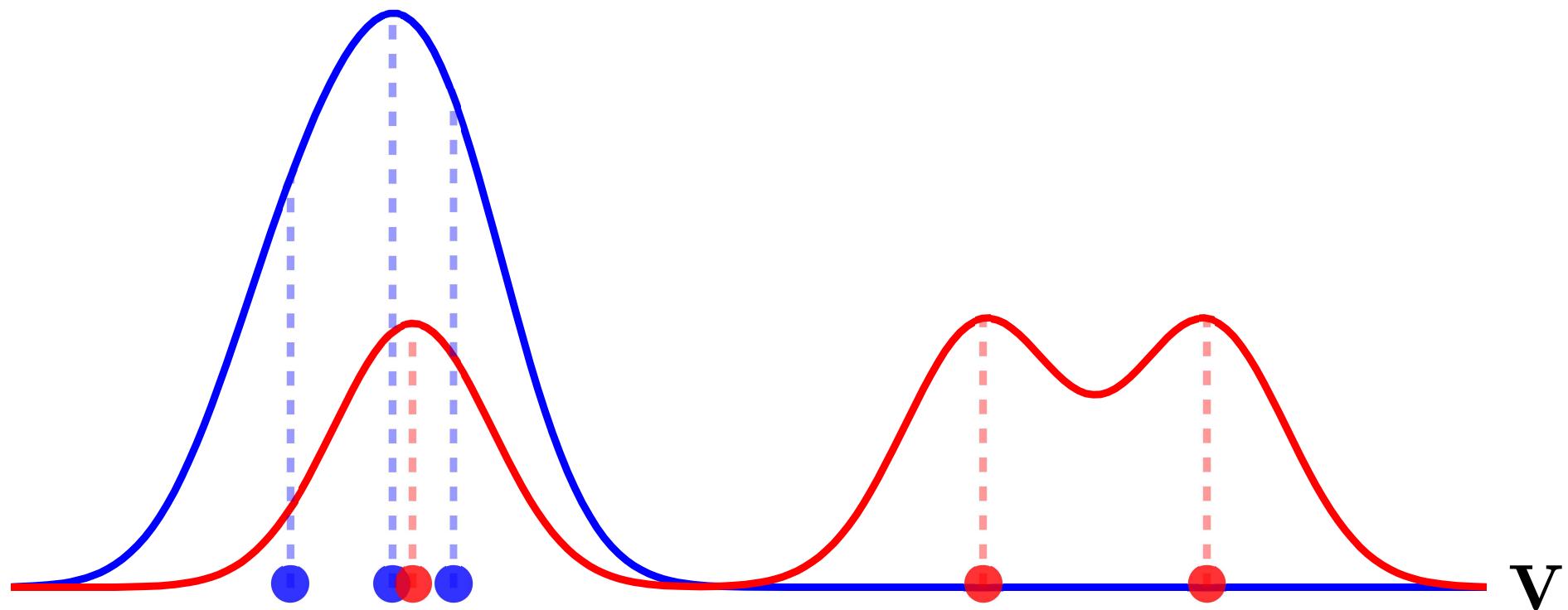
Construction of MMD witness

Construction of empirical **witness function** (proof: next slide!)



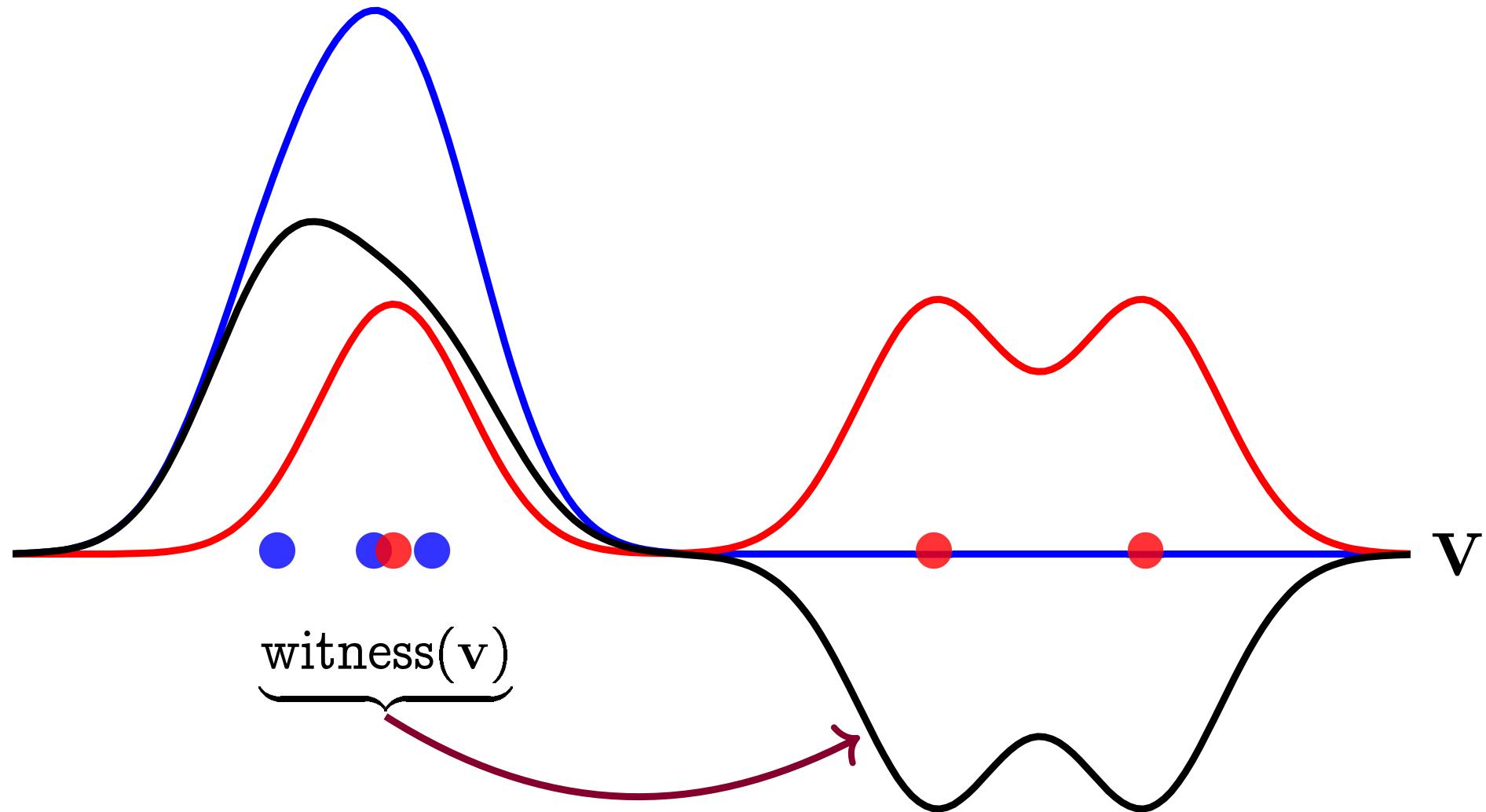
Construction of MMD witness

Construction of empirical **witness function** (proof: next slide!)



Construction of MMD witness

Construction of empirical **witness function** (proof: next slide!)



Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

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

Derivation of empirical witness function

Recall the **witness function** expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

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

The empirical witness function at v

$$f^*(v) = \langle f^*, \varphi(v) \rangle_{\mathcal{F}}$$

Derivation of empirical witness function

Recall the **witness function** expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

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

The empirical witness function at v

$$\begin{aligned} f^*(v) &= \langle f^*, \varphi(v) \rangle_{\mathcal{F}} \\ &\propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \end{aligned}$$

Derivation of empirical witness function

Recall the witness function expression

$$f^* \propto \mu_P - \mu_Q$$

The empirical feature mean for P

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

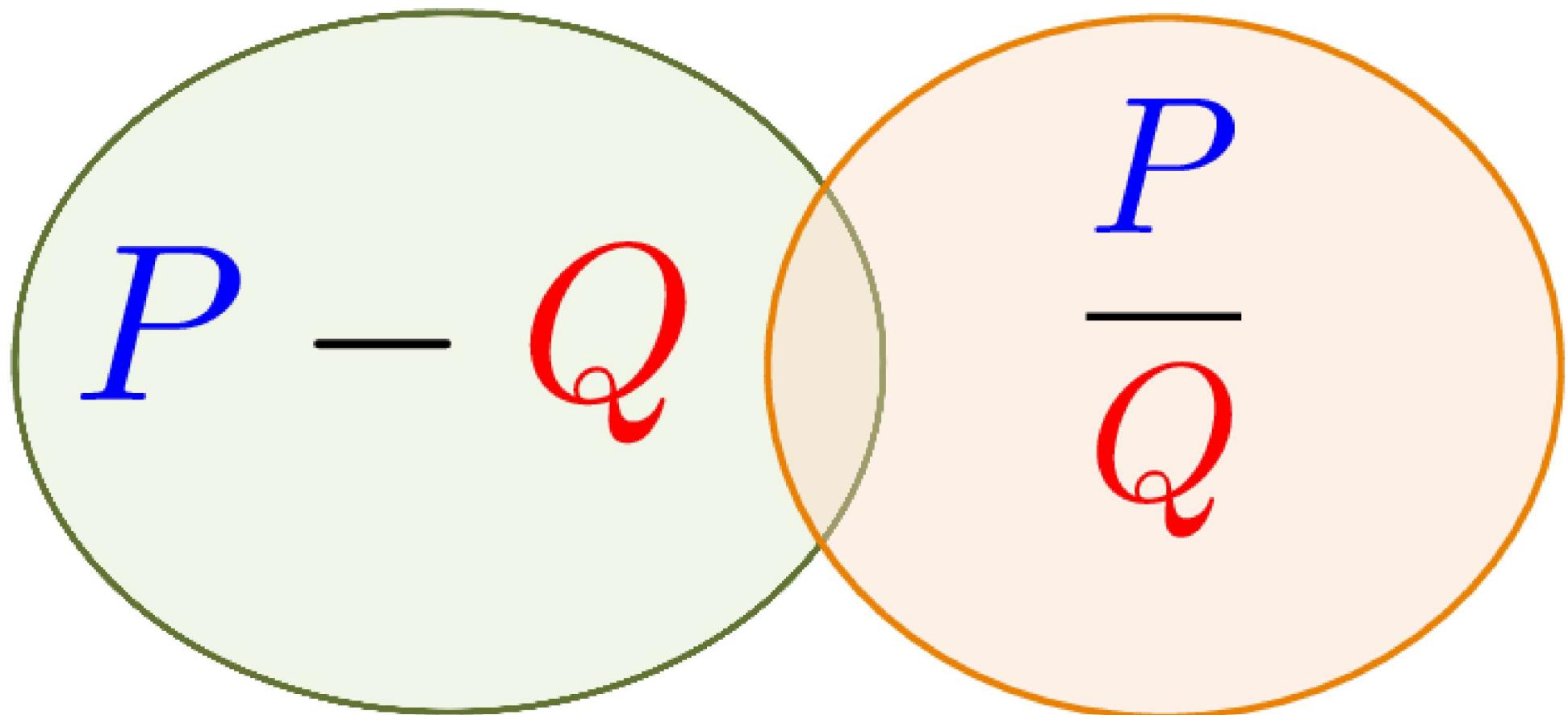
The empirical witness function at v

$$\begin{aligned} f^*(v) &= \langle \textcolor{teal}{f}^*, \varphi(v) \rangle_{\mathcal{F}} \\ &\propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \\ &= \frac{1}{n} \sum_{i=1}^n k(\textcolor{blue}{x}_i, v) - \frac{1}{n} \sum_{i=1}^n k(\textcolor{red}{y}_i, v) \end{aligned}$$

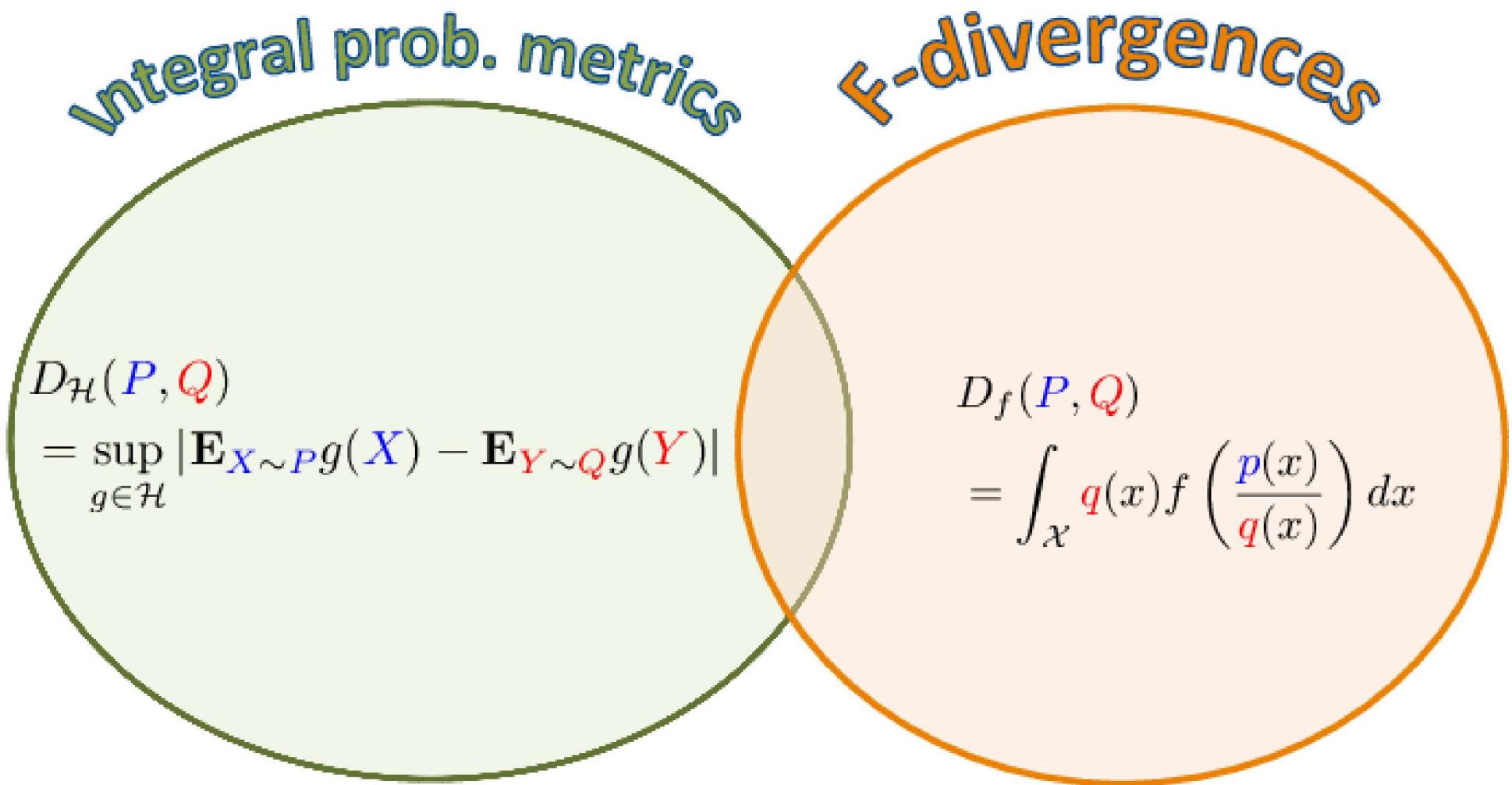
Don't need explicit feature coefficients $f^* := \begin{bmatrix} f_1^* & f_2^* & \dots \end{bmatrix}$

Interlude: divergence measures

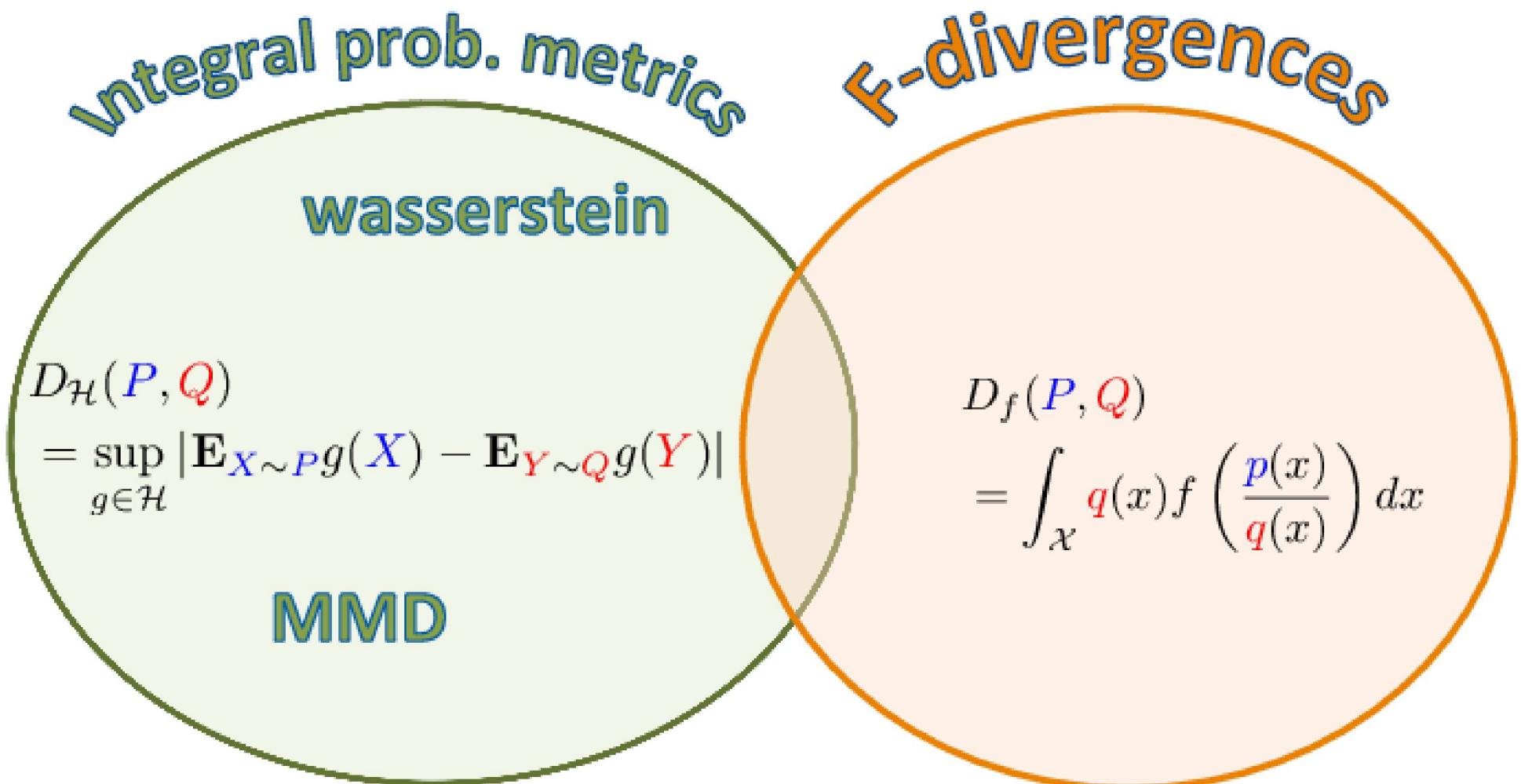
Divergences



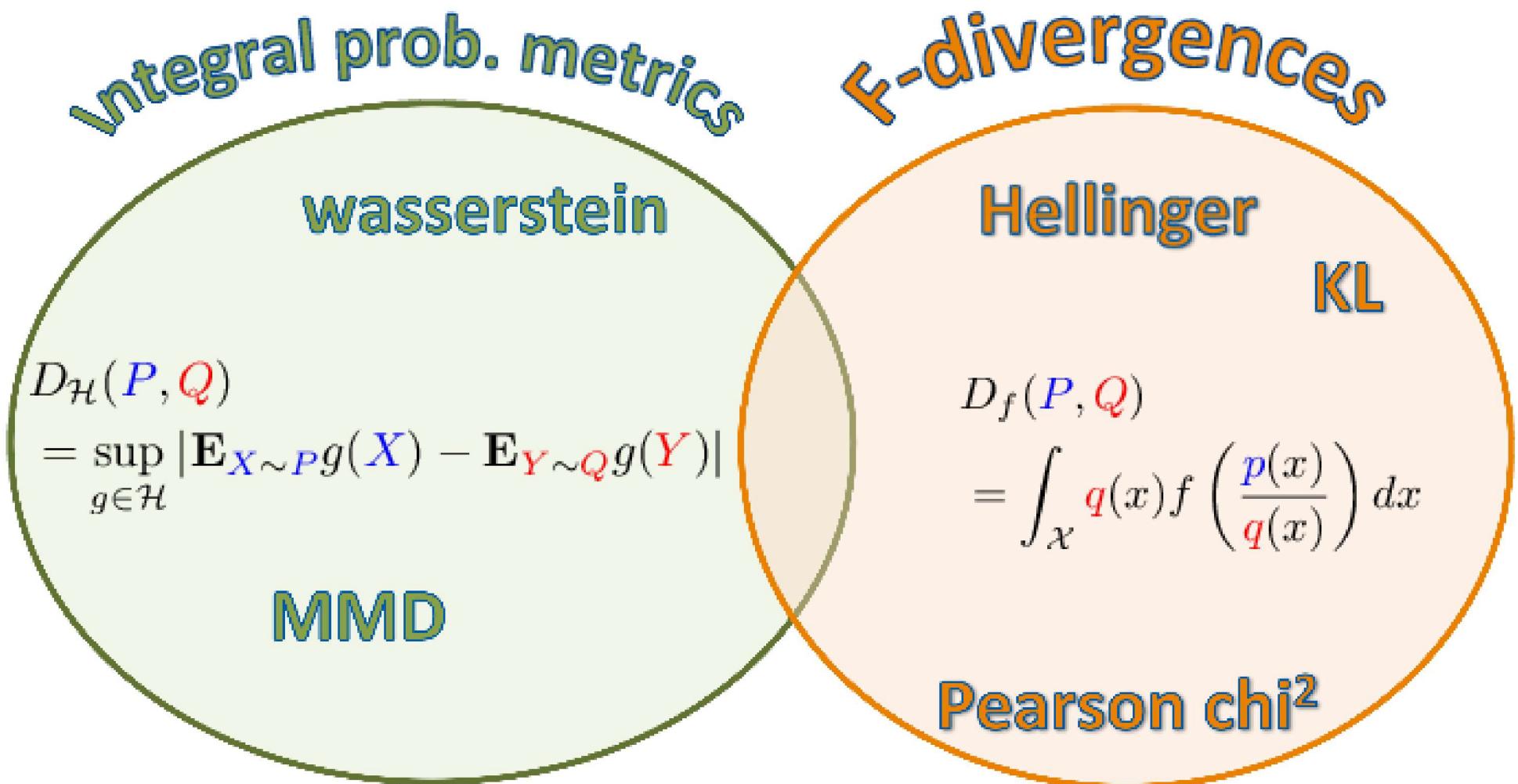
Divergences



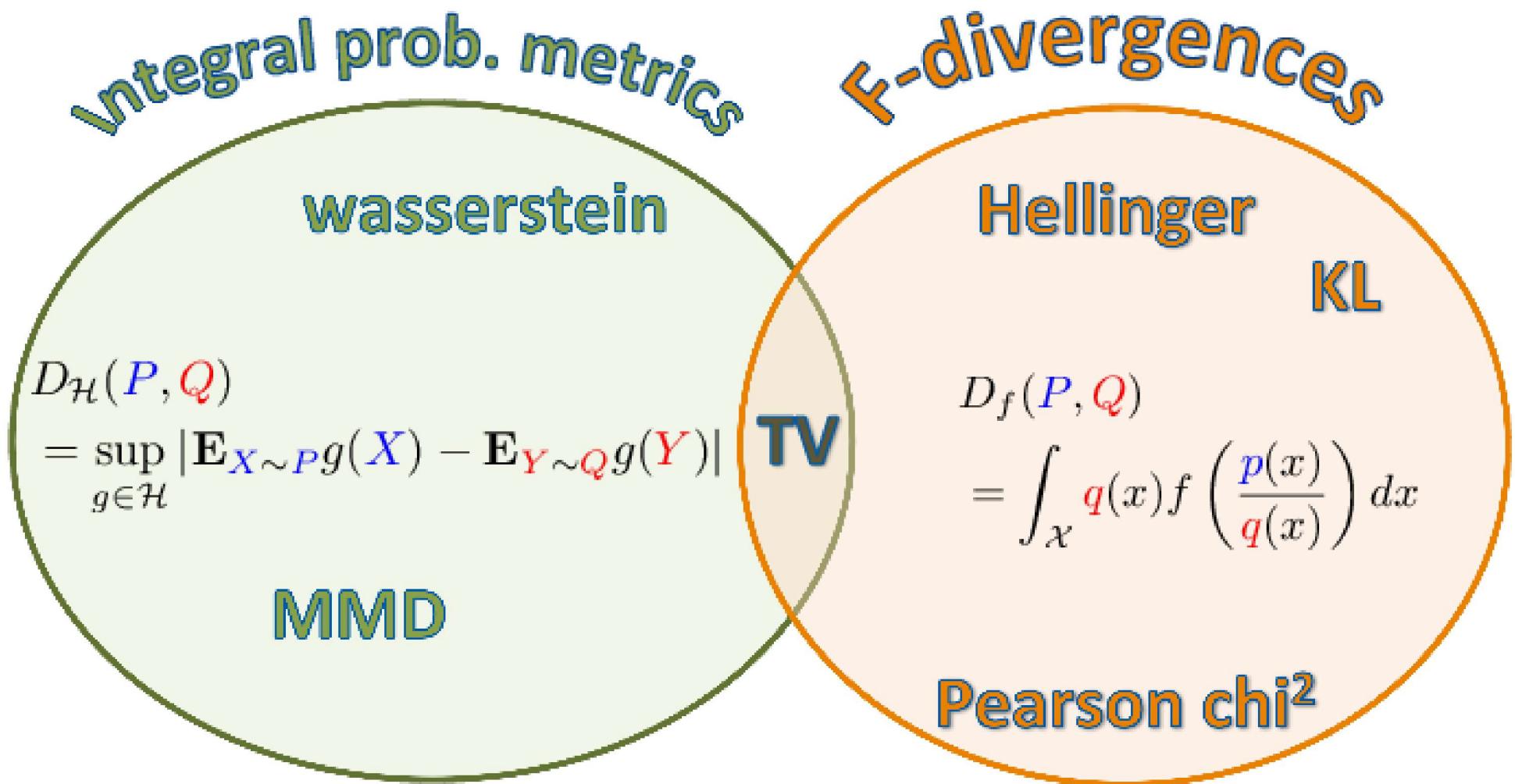
Divergences



Divergences



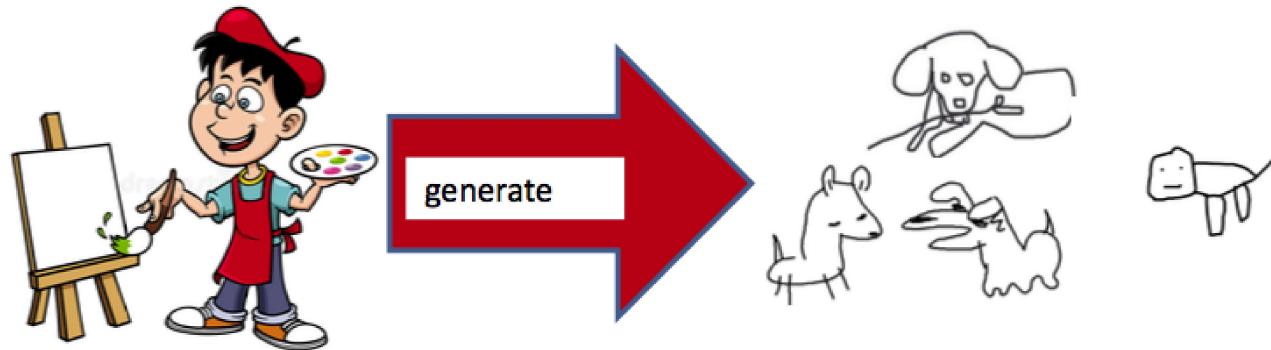
Divergences



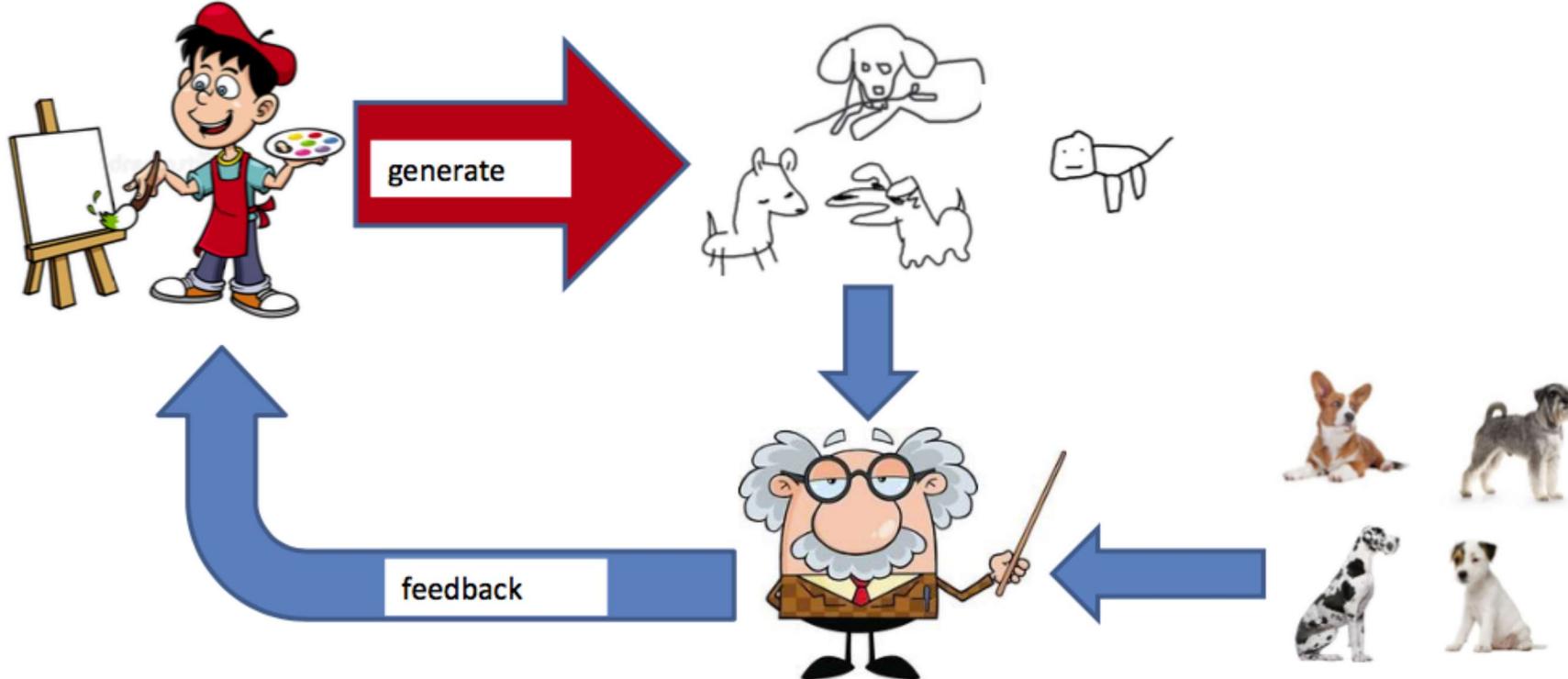
Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (2012)

Training Generative Adversarial Networks: Critics and Gradient Penalties

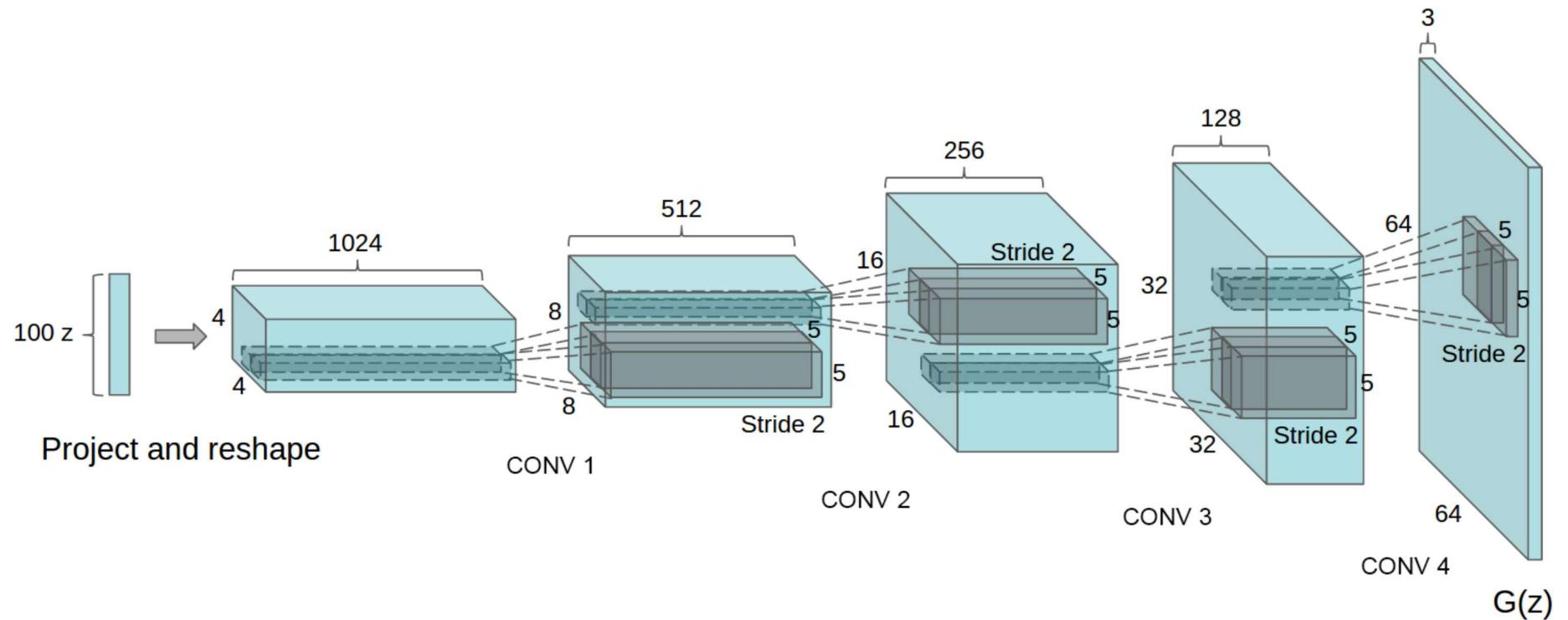
Visual notation: GAN setting



Visual notation: GAN setting



What I won't cover: the generator



Radford, Metz, Chintala, ICLR 2016

F-divergence as critic

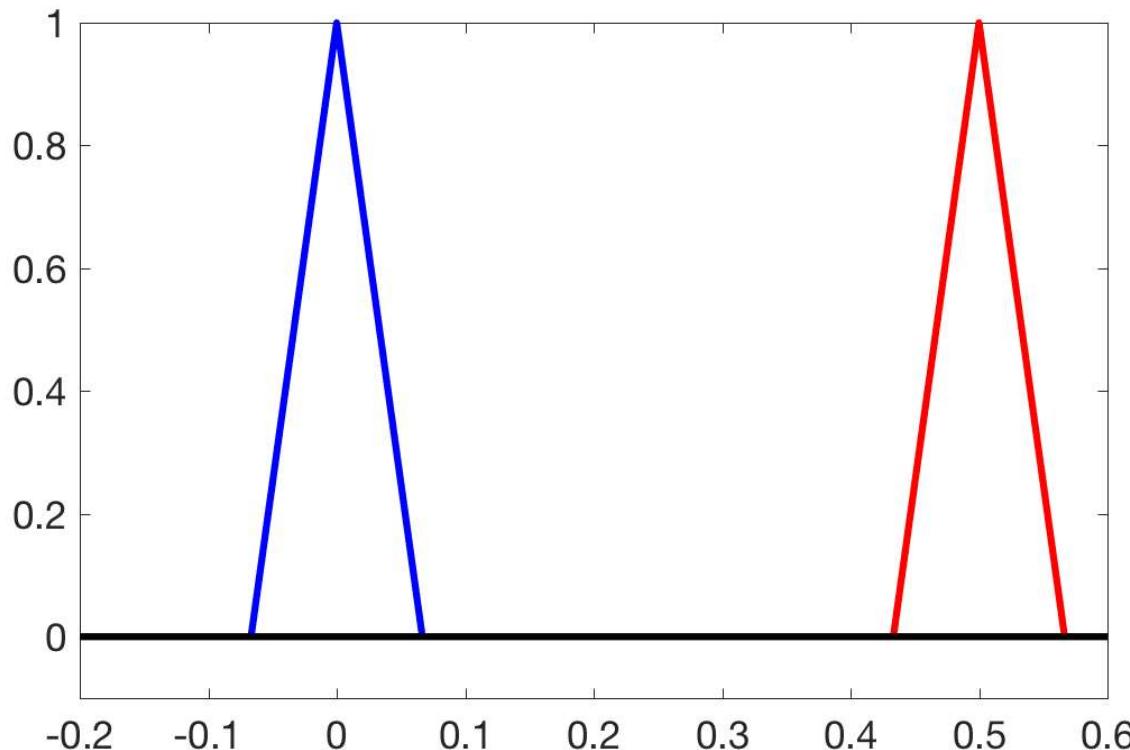


An **unhelpful critic?** Jensen-Shannon,

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

$$D_{JS}(P, Q) = \log 2$$



F-divergence as critic

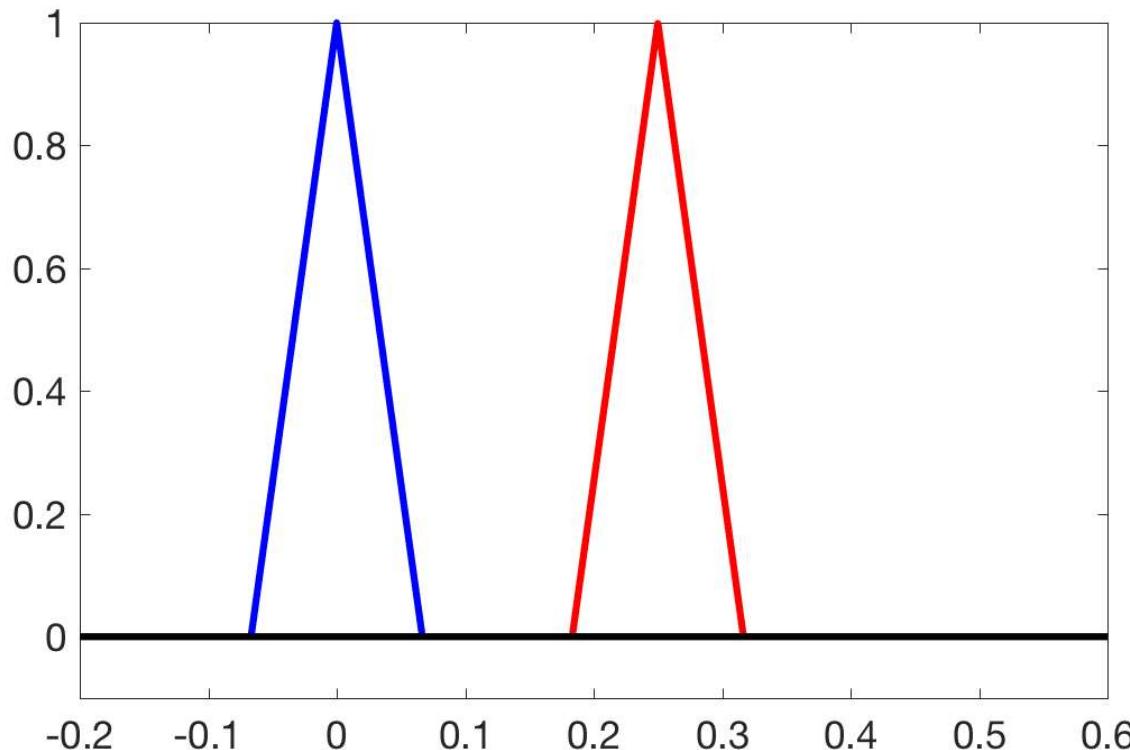


An **unhelpful critic?** Jensen-Shannon,

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

$$D_{JS}(P, Q) = \log 2$$



F-divergence as critic



An **unhelpful critic?** Jensen-Shannon,

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

What is done in practice?

F-divergence as critic



An **unhelpful critic?** Jensen-Shannon,

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

What is done in practice?

- Use a **variational approximation** to the critic, alternate generator and critic training (we will return to this!) Goodfellow et al. [NeurIPS 2014], Nowozin et al. [NeurIPS 2016]

F-divergence as critic



An **unhelpful critic?** Jensen-Shannon,

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

What is done in practice?

- Use a **variational approximation** to the critic, **alternate generator and critic training** (we will return to this!) Goodfellow et al. [NeurIPS 2014], Nowozin et al. [NeurIPS 2016]
- Add “**instance noise**” to the reference and generator observations Sonderby et al. [arXiv 2016], Arjovsky and Bottou [ICLR 2017]

F-divergence as critic



An **unhelpful critic?** Jensen-Shannon,

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

What is done in practice?

- Use a **variational approximation** to the critic, **alternate generator and critic training** (we will return to this!) Goodfellow et al. [NeurIPS 2014], Nowozin et al. [NeurIPS 2016]
- Add “**instance noise**” to the reference and generator observations Sonderby et al. [arXiv 2016], Arjovsky and Bottou [ICLR 2017]
 - ...or (approx. equivalently) a **data-dependent gradient penalty** for the variational critic (we will return to this!) Roth et al [NeurIPS 2017], Nagarajan and Kolter [NeurIPS 2017], Mescheder et al. [ICML 2018]

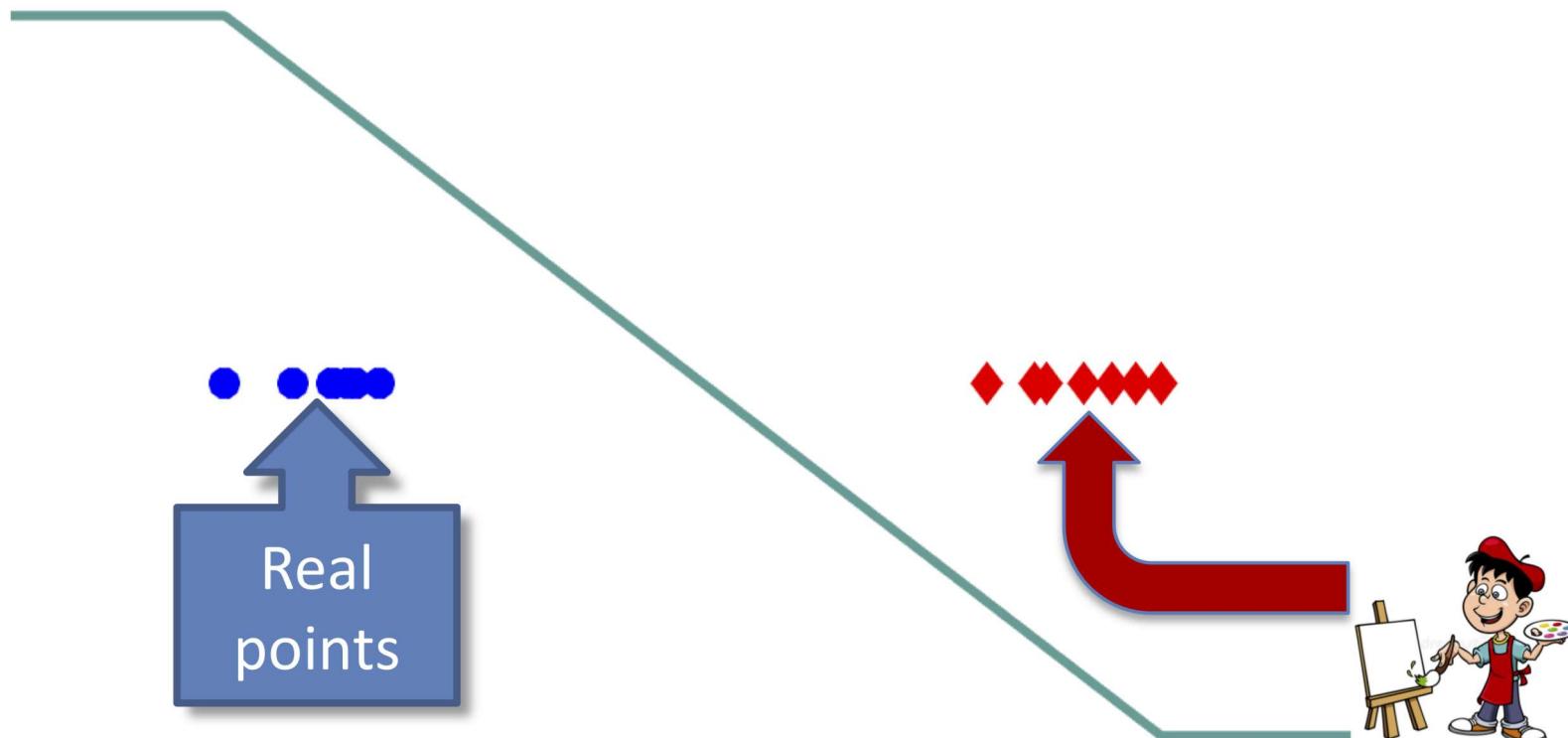
Wasserstein distance as critic



A helpful critic witness:

$$W_1(P, Q) = \sup_{\|f\|_L \leq 1} E_P f(X) - E_Q f(Y).$$
$$\|f\|_L := \sup_{x \neq y} |f(x) - f(y)| / \|x - y\|$$

$$W_1=0.88$$



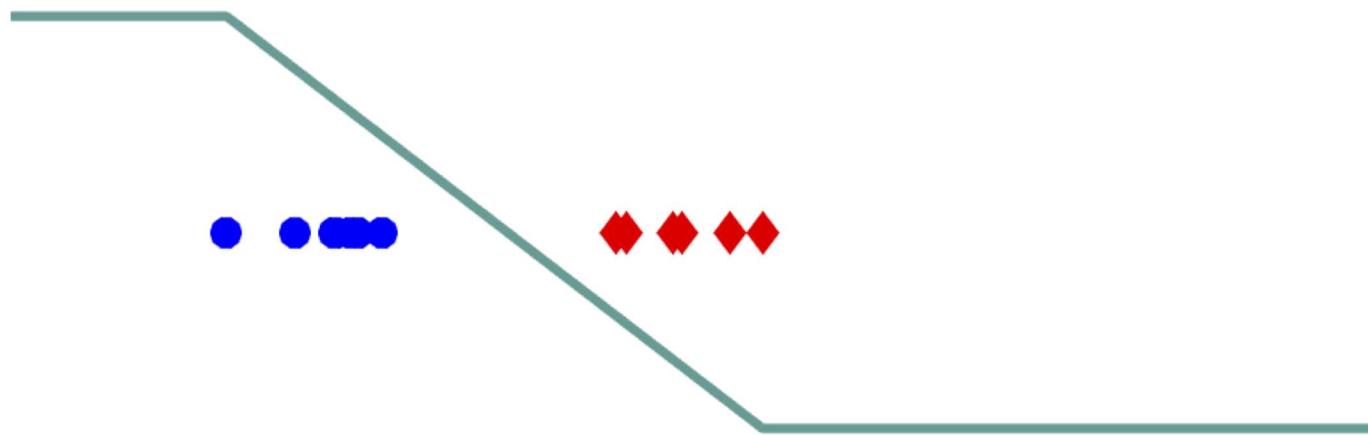
Wasserstein distance as critic



A helpful critic witness:

$$W_1(P, Q) = \sup_{\|f\|_L \leq 1} E_P f(X) - E_Q f(Y).$$
$$\|f\|_L := \sup_{x \neq y} |f(x) - f(y)| / \|x - y\|$$

$$W_1 = 0.65$$



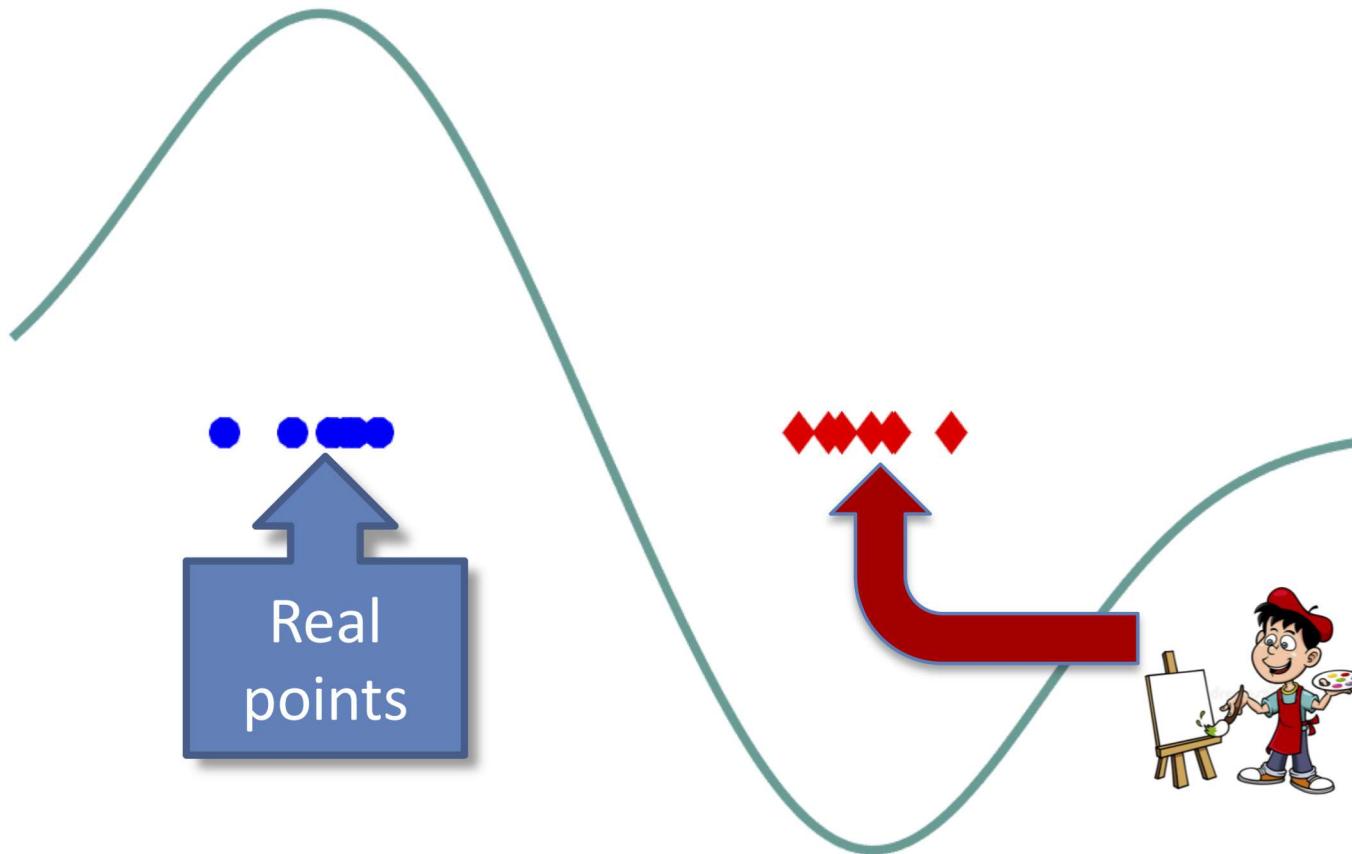
MMD as critic



A helpful critic witness:

$$MMD(P, Q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} E_P f(X) - E_Q f(Y).$$

MMD=1.8



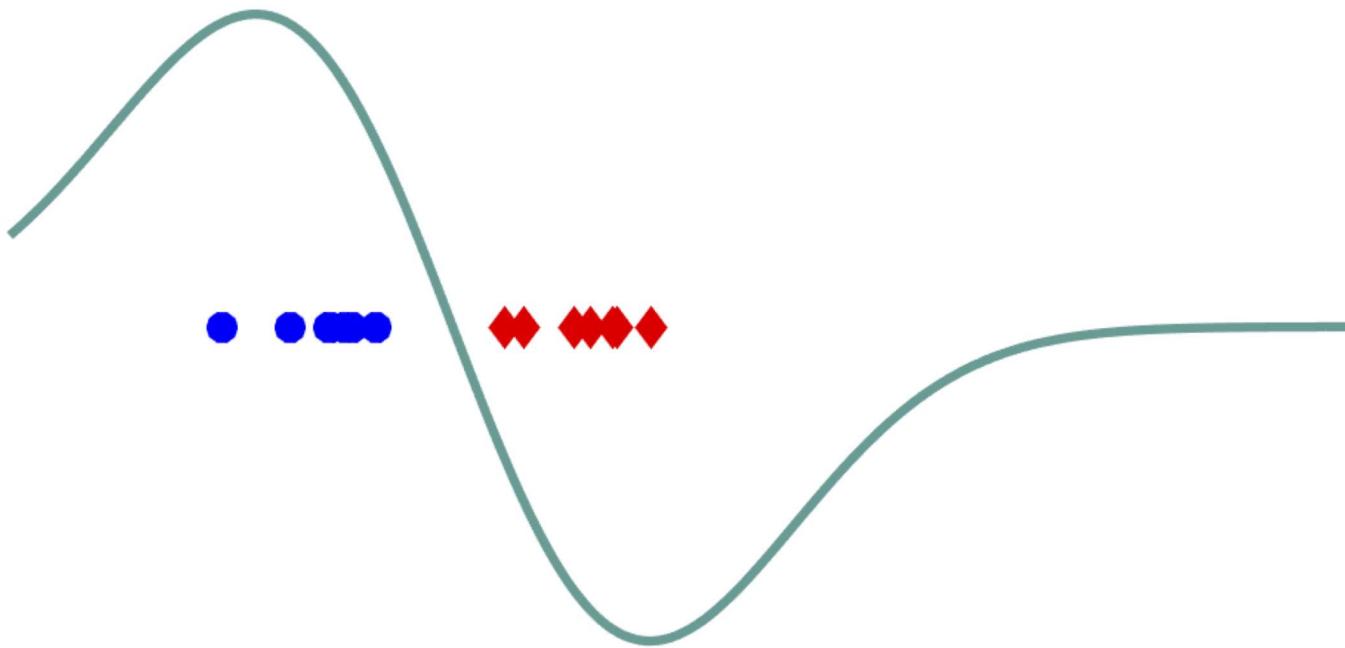
MMD as critic



A helpful critic witness:

$$MMD(P, Q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} E_P f(X) - E_Q f(Y)$$

MMD=1.1

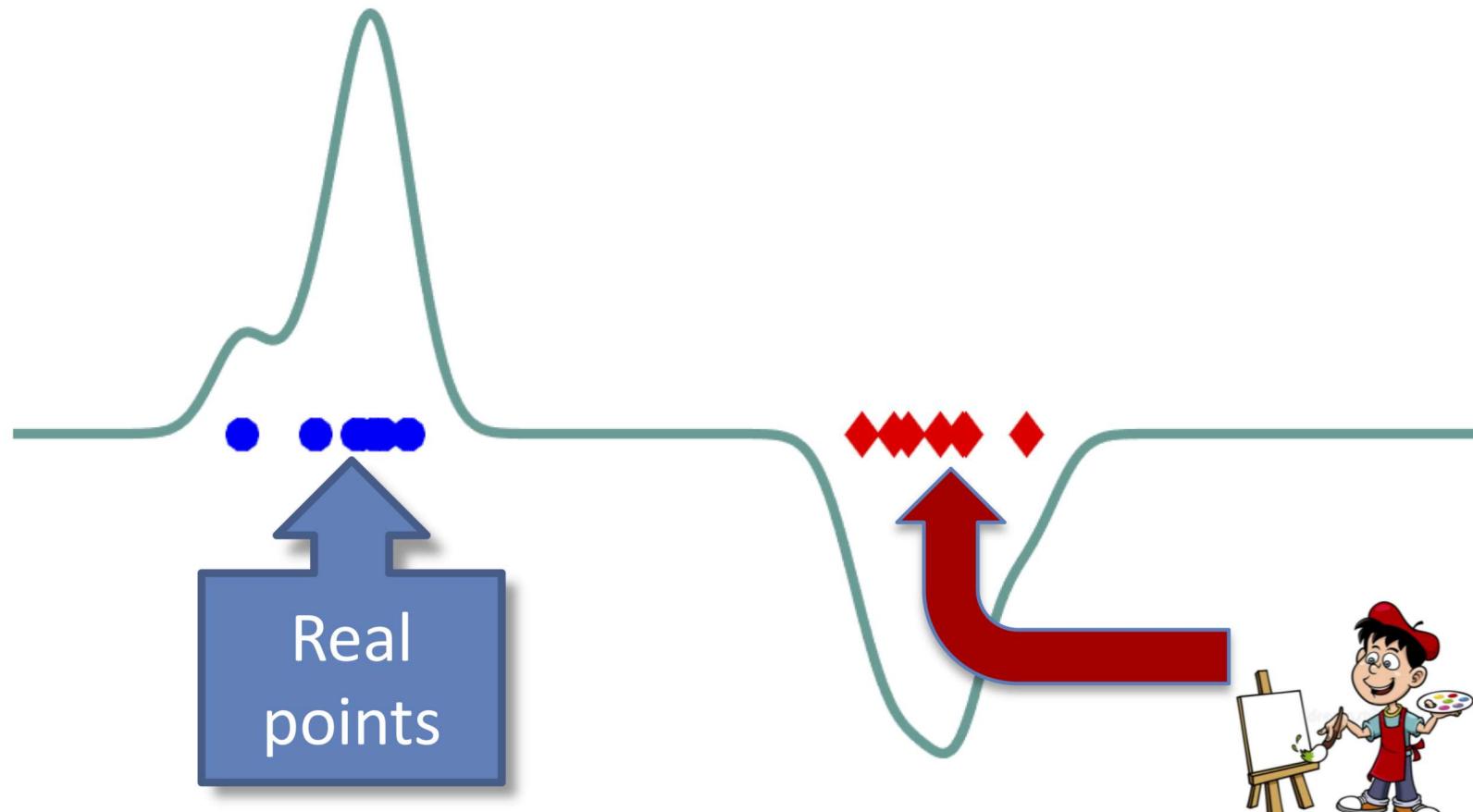


MMD as critic



An **unhelpful** critic witness:
 $MMD(P, Q)$ with a narrow kernel.

MMD=0.64

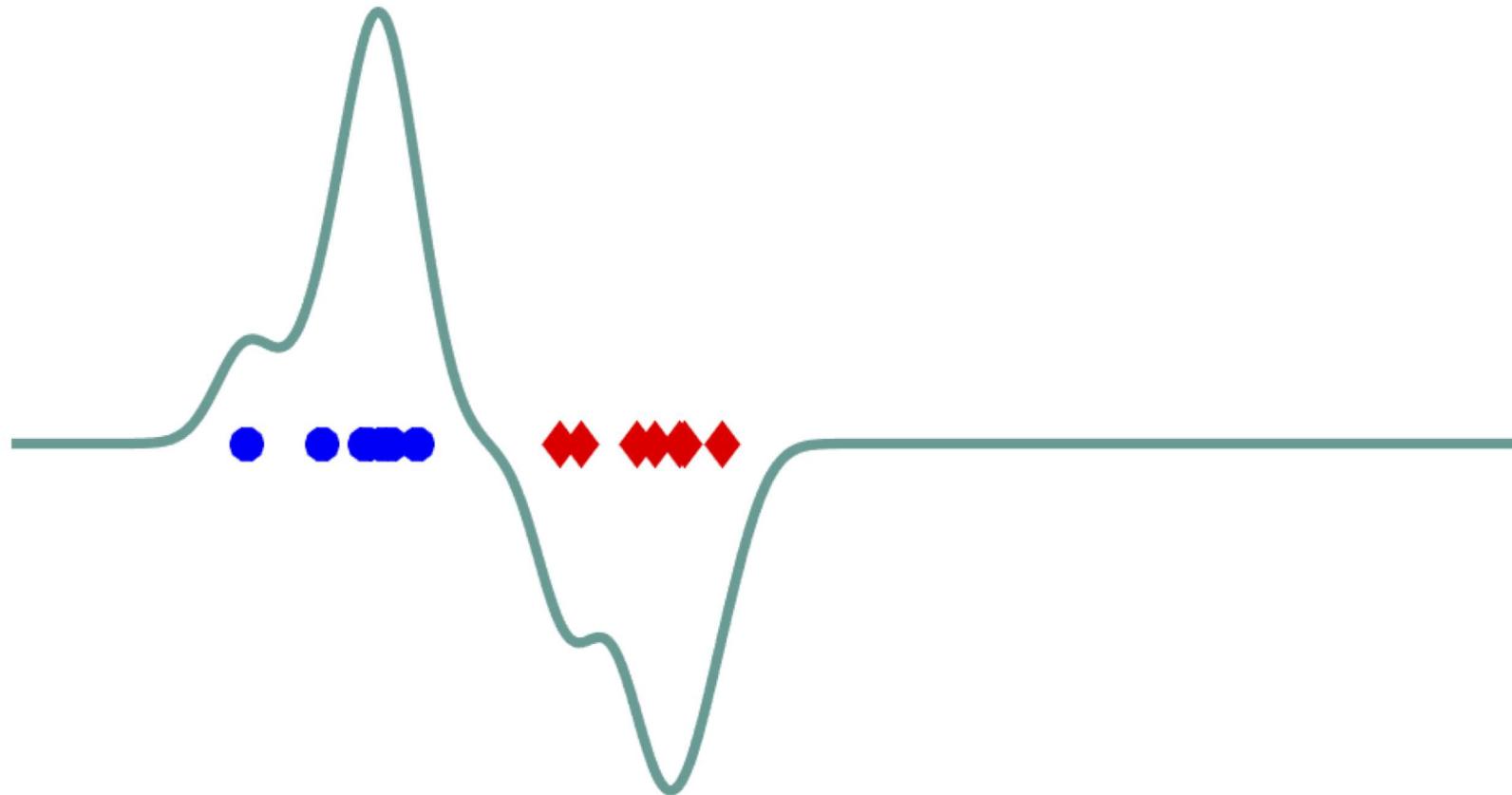


MMD as critic



An **unhelpful** critic witness:
 $MMD(P, Q)$ with a narrow kernel.

$MMD=0.64$



MMD for GAN critic

Can you use MMD as a critic to train GANs?

From ICML 2015:

Generative Moment Matching Networks

Yujia Li¹

Kevin Swersky¹

Richard Zemel^{1,2}

YUJIALI@CS.TORONTO.EDU

KSWERSKY@CS.TORONTO.EDU

ZEMEL@CS.TORONTO.EDU

¹Department of Computer Science, University of Toronto, Toronto, ON, CANADA

²Canadian Institute for Advanced Research, Toronto, ON, CANADA

From UAI 2015:

Training generative neural networks via Maximum Mean Discrepancy optimization

Gintare Karolina Dziugaite
University of Cambridge

Daniel M. Roy
University of Toronto

Zoubin Ghahramani
University of Cambridge

MMD for GAN critic

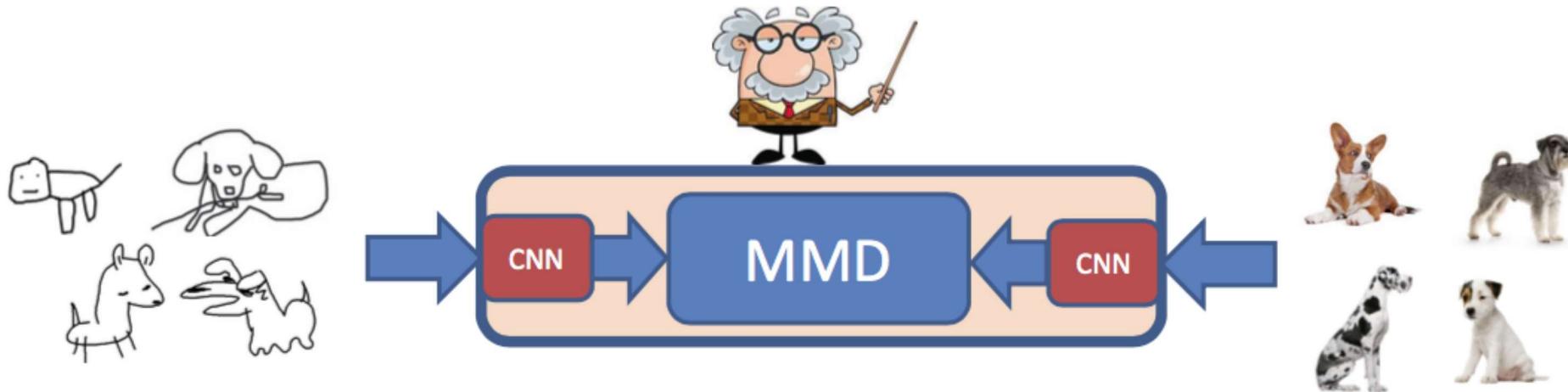
Can you use MMD as a critic to train GANs?



Need better image features.

CNN features for an MMD witness

- Add convolutional features!
- The **critic** (teacher) also needs to be trained.



$$\hat{\kappa}(x, y) = h_\psi^\top(x) h_\psi(y)$$

where $h_\psi(x)$ is a CNN map:

$$\hat{\kappa}(x, y) = k(h_\psi(x), h_\psi(y))$$

where $h_\psi(x)$ is a CNN map,

k is e.g. an exponentiated quadratic kernel

MMD Li et al., [NeurIPS 2017]

Cramer Bellemare et al. [2017]

Coulomb Unterthiner et al., [ICLR 2018]

Demystifying MMD GANs Binkowski,
Sutherland, Arbel, G., [ICLR 2018] 29/62

■ **Wasserstein GAN** Arjovsky

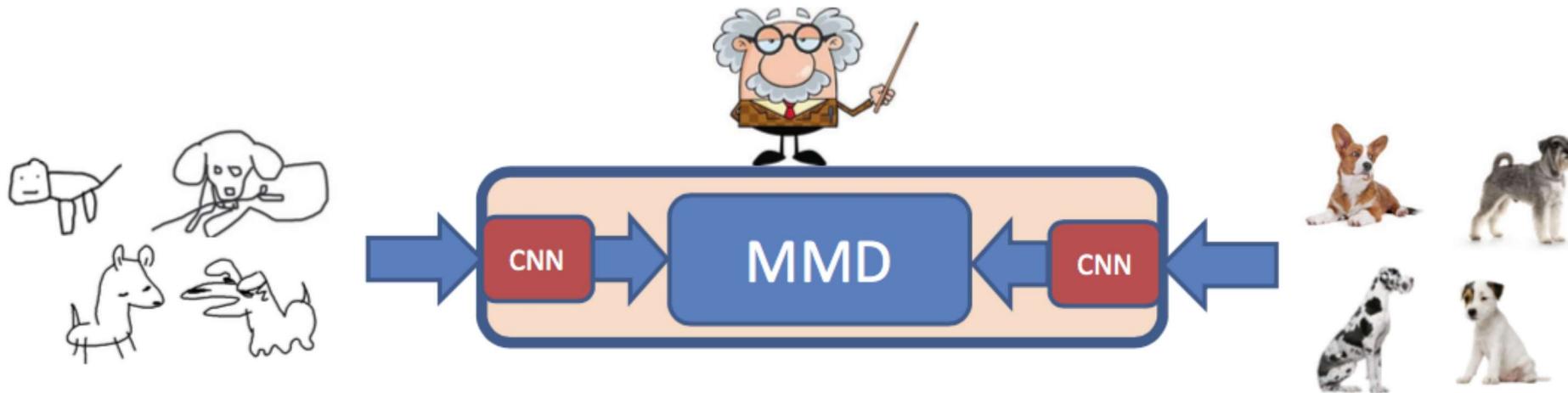
et al. [ICML 2017]

■ **WGANGP** Gulrajani et al.

[NeurIPS 2017]

CNN features for an MMD witness

- Add convolutional features!
- The **critic** (teacher) also needs to be trained.



$$\hat{\kappa}(x, y) = h_\psi^\top(x) h_\psi(y)$$

where $h_\psi(x)$ is a CNN map:

- **Wasserstein GAN** Arjovsky et al. [ICML 2017]
- **WGAN-GP** Gulrajani et al. [NeurIPS 2017]

$$\hat{\kappa}(x, y) = k(h_\psi(x), h_\psi(y))$$

where $h_\psi(x)$ is a CNN map,

k is e.g. an exponentiated quadratic

kernel

MMD Li et al., [NeurIPS 2017]

Cramer Bellemare et al. [2017]

Coulomb Unterthiner et al., [ICLR 2018]

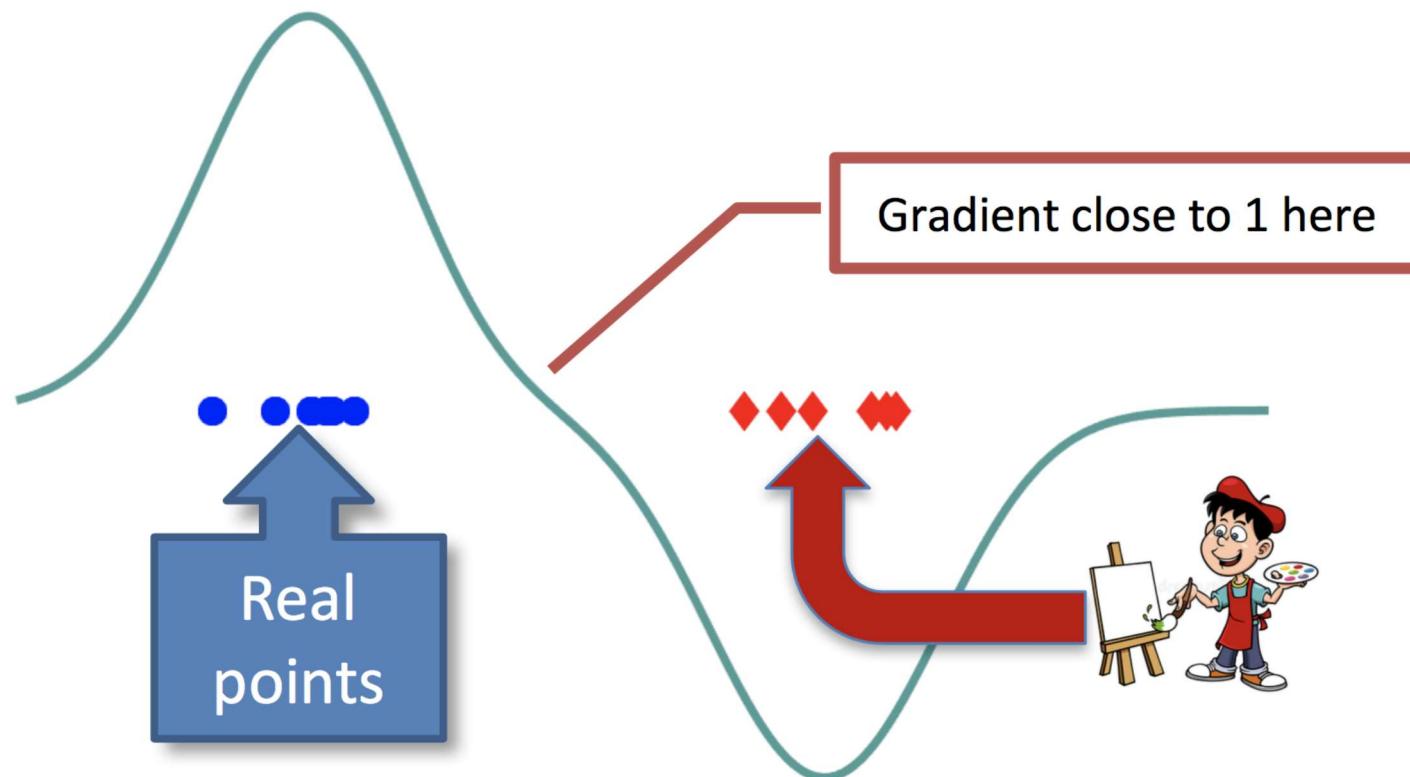
Demystifying MMD GANs Binkowski, Sutherland, Arbel, G., [ICLR 2018] 29/62

Gradient penalty: the optimisation viewpoint

WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]

WGAN-GP Gulrajani et al. [NeurIPS 2017]



WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]

WGAN-GP Gulrajani et al. [NeurIPS 2017]



- Given a generator G_θ with parameters θ to be trained.

Samples $Y \sim G_\theta(Z)$ where $Z \sim R$



- Given critic features h_ψ with parameters ψ to be trained. f_ψ a linear function, $\mathcal{R}(x, y) = h_\psi^\top(x)h_\psi(y)$.

WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]

WGAN-GP Gulrajani et al. [NeurIPS 2017]



- Given a generator G_θ with parameters θ to be trained.
Samples $\mathbf{Y} \sim G_\theta(\mathbf{Z})$ where $\mathbf{Z} \sim \mathcal{R}$



- Given critic features \mathbf{h}_ψ with parameters ψ to be trained. f_ψ a linear function, $\mathcal{K}(x, y) = \mathbf{h}_\psi^\top(\mathbf{x})\mathbf{h}_\psi(\mathbf{y})$.

WGAN-GP gradient penalty:

$$\max_{\psi} \mathbf{E}_{X \sim \mathcal{P}} f_\psi(\mathbf{X}) - \mathbf{E}_{Z \sim \mathcal{R}} f_\psi(G_\theta(\mathbf{Z})) + \lambda \mathbf{E}_{\tilde{\mathbf{X}}} \left(\left\| \nabla_{\tilde{\mathbf{X}}} f_\psi(\tilde{\mathbf{X}}) \right\| - 1 \right)^2$$

where

$$\tilde{\mathbf{X}} = \gamma \mathbf{x}_i + (1 - \gamma) G_\theta(\mathbf{z}_j)$$

$$\gamma \sim \mathcal{U}([0, 1]) \quad x_i \in \{\mathbf{x}_\ell\}_{\ell=1}^m \quad z_j \in \{\mathbf{z}_\ell\}_{\ell=1}^n$$

DiracGAN toy example

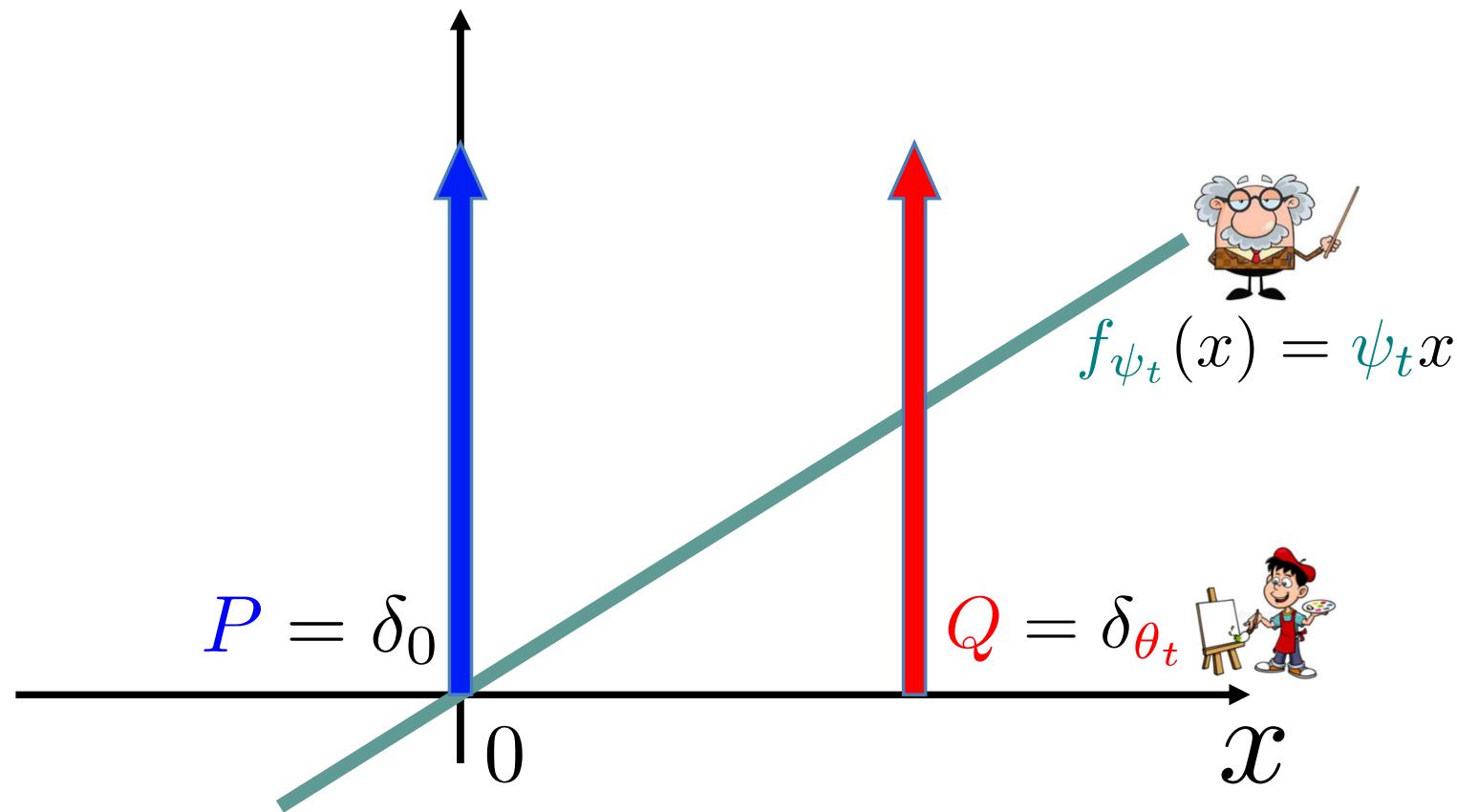
From ICML 2018:

Which Training Methods for GANs do actually Converge?

Lars Mescheder¹ Andreas Geiger^{1,2} Sebastian Nowozin³

Gives an **optimisation viewpoint** on gradient regularisation.

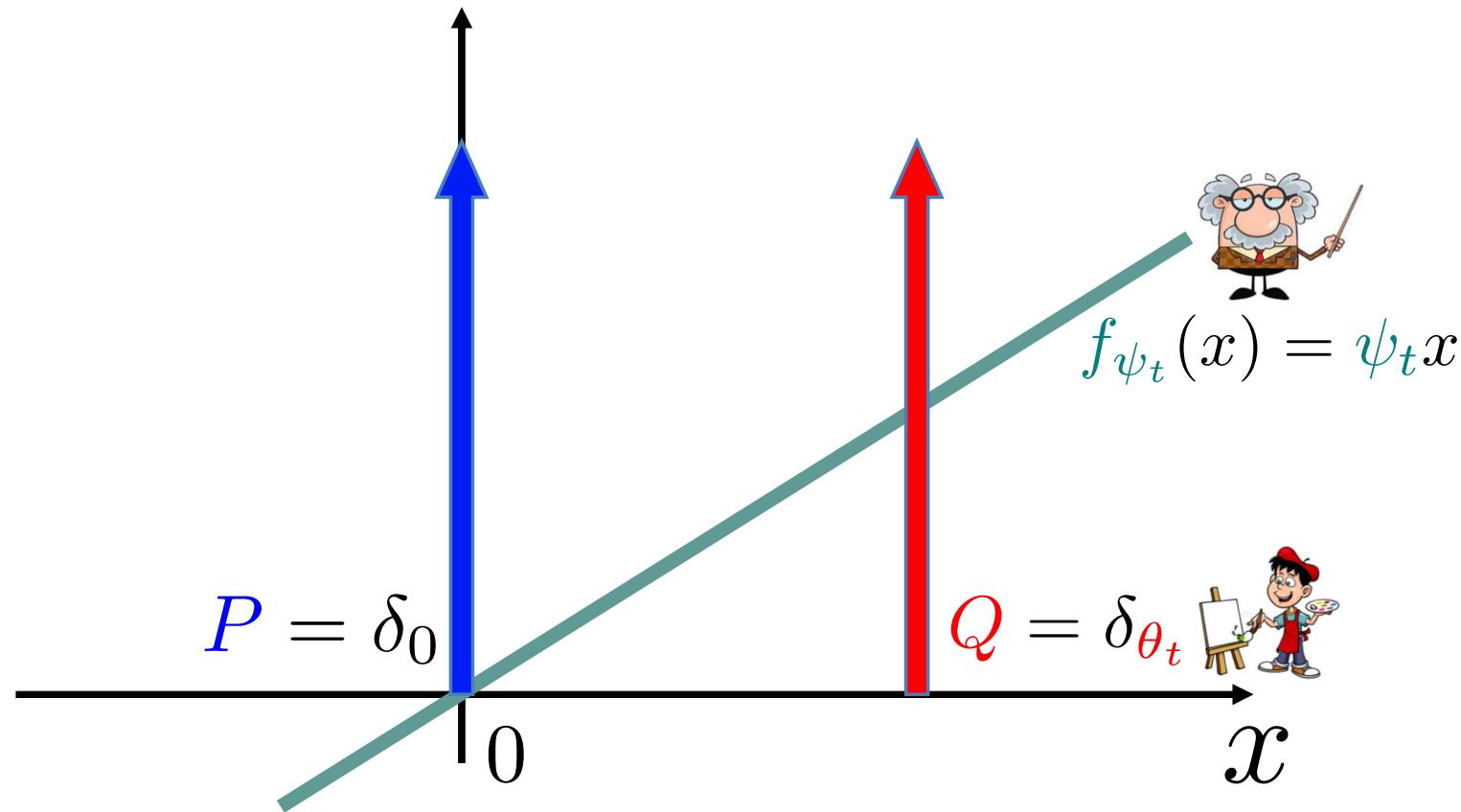
DiracGAN toy example



$$\begin{aligned} D(P, Q; \psi_t) &= \mathbf{E}_Q f_{\psi_t}(Y) - \mathbf{E}_P f_{\psi_t}(X) \\ &= \psi_t \theta_t \end{aligned}$$

DiracGAN toy example

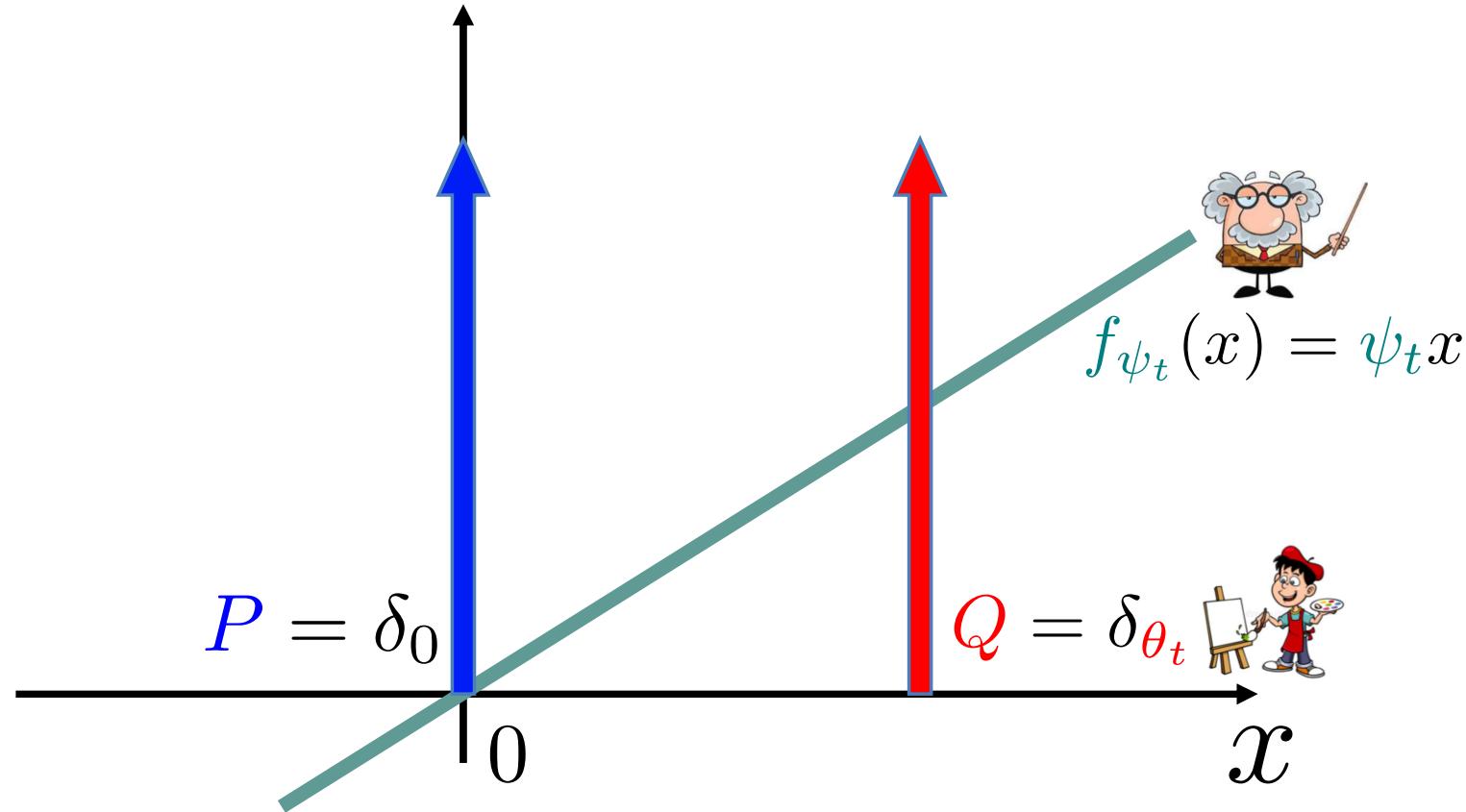
Gradient **descent** on **generator**:



$$\frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \frac{\partial}{\partial \theta} \psi_t \theta_t = \psi_t$$

DiracGAN toy example

Gradient **descent** on **generator**:



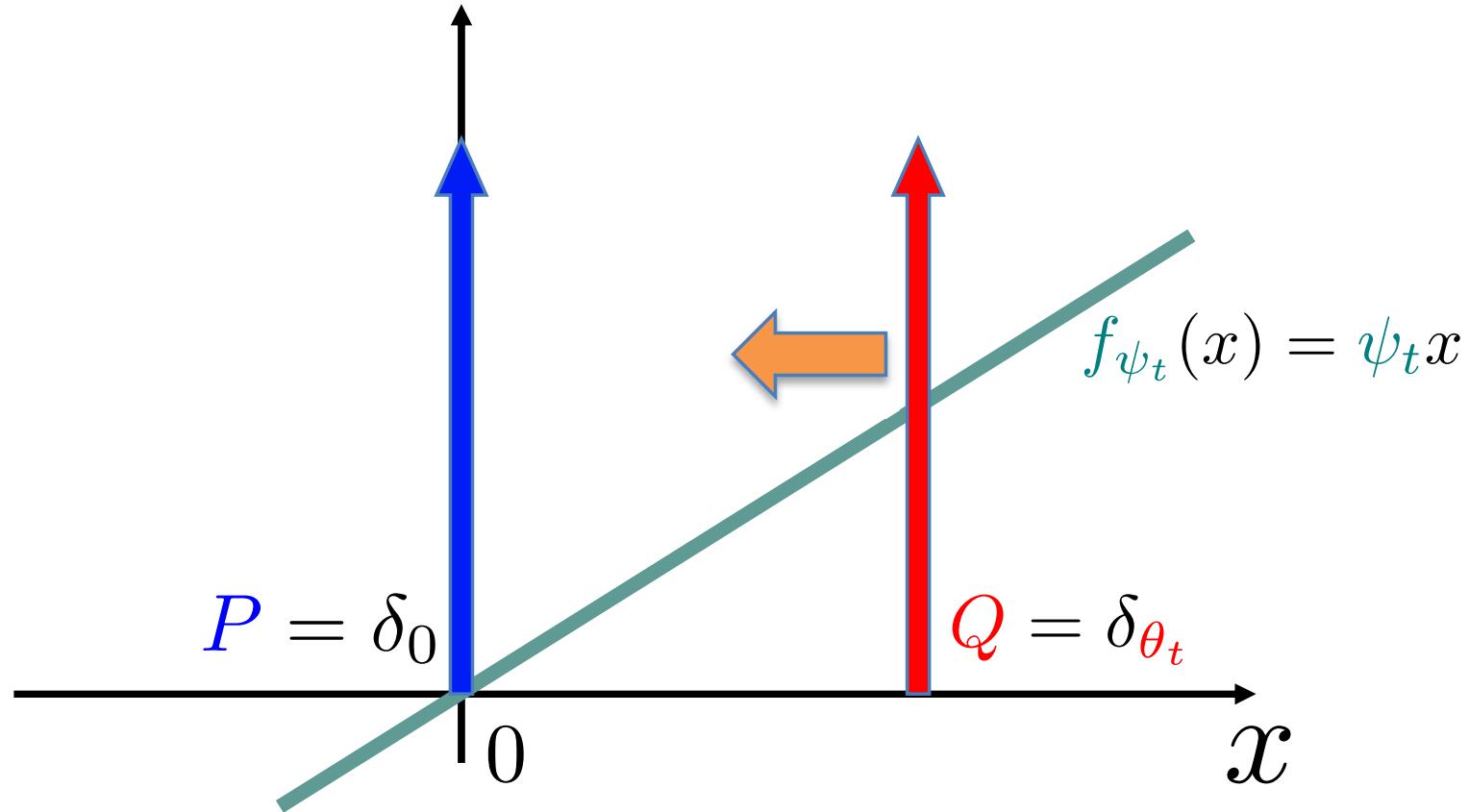
$$\frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \frac{\partial}{\partial \theta} \psi_t \theta_t = \psi_t$$

$$\theta_{t+1} = \theta_t - \gamma \frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \theta_t - \gamma \psi_t$$

for stepsize γ

DiracGAN toy example

Gradient **descent** on **generator**:

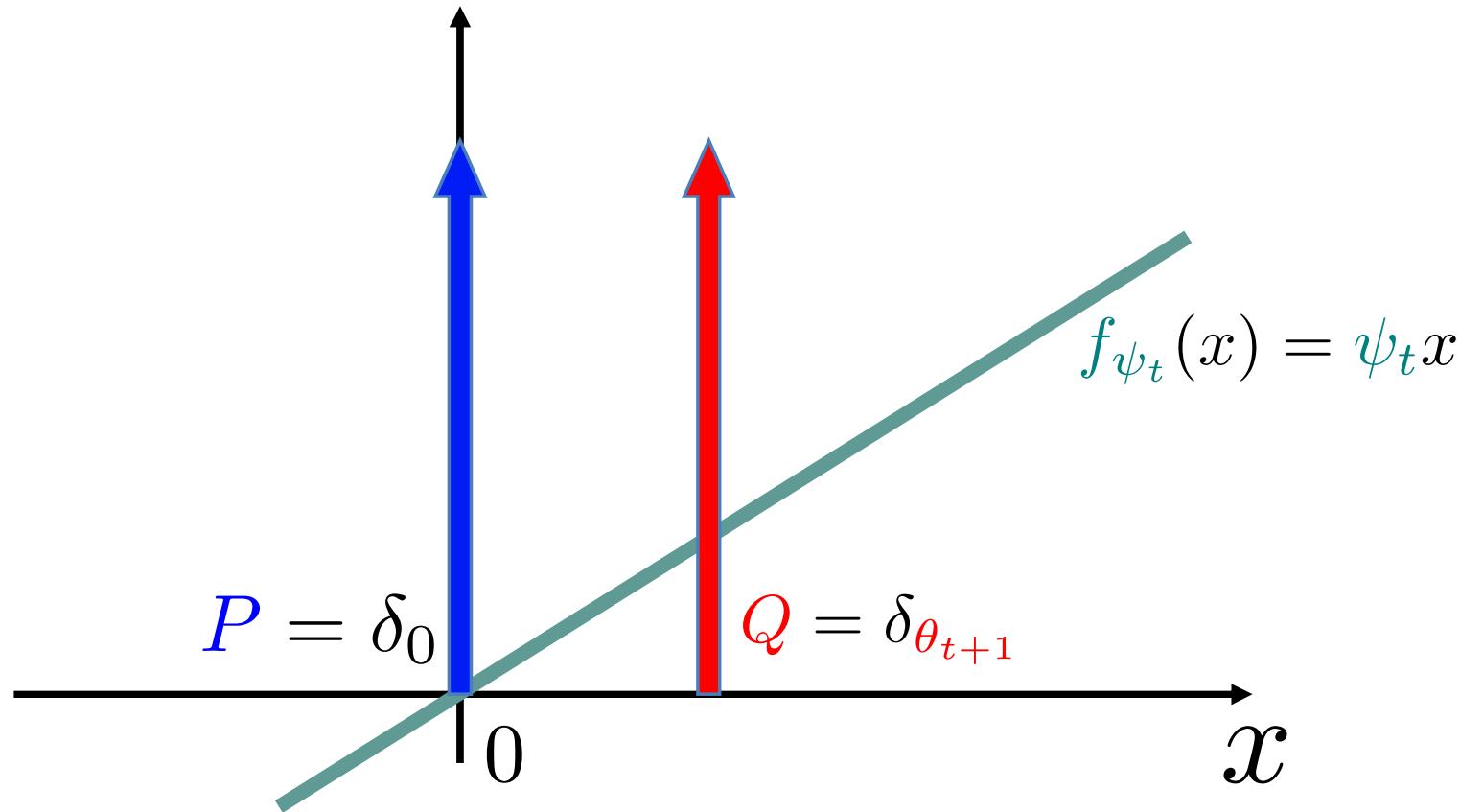


$$\frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \frac{\partial}{\partial \theta} \psi_t \theta_t = \psi_t$$

$$\theta_{t+1} = \theta_t - \gamma \frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \theta_t - \gamma \psi_t$$

DiracGAN toy example

Gradient **descent** on **generator**:

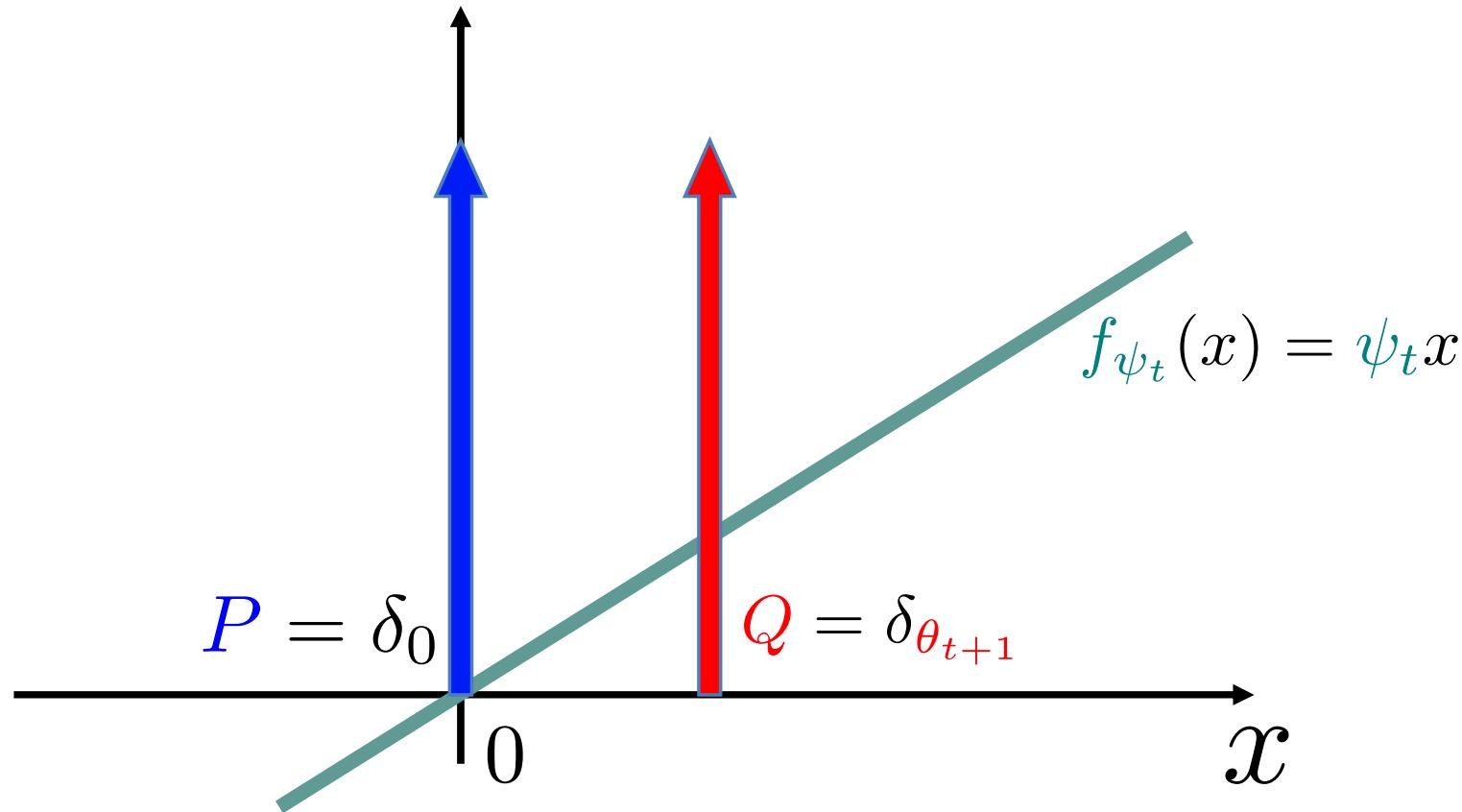


$$\frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \frac{\partial}{\partial \theta} \psi_t \theta_t = \psi_t$$

$$\theta_{t+1} = \theta_t - \gamma \frac{\partial}{\partial \theta} D(P, Q; \psi_t) = \theta_t - \gamma \psi_t$$

DiracGAN toy example

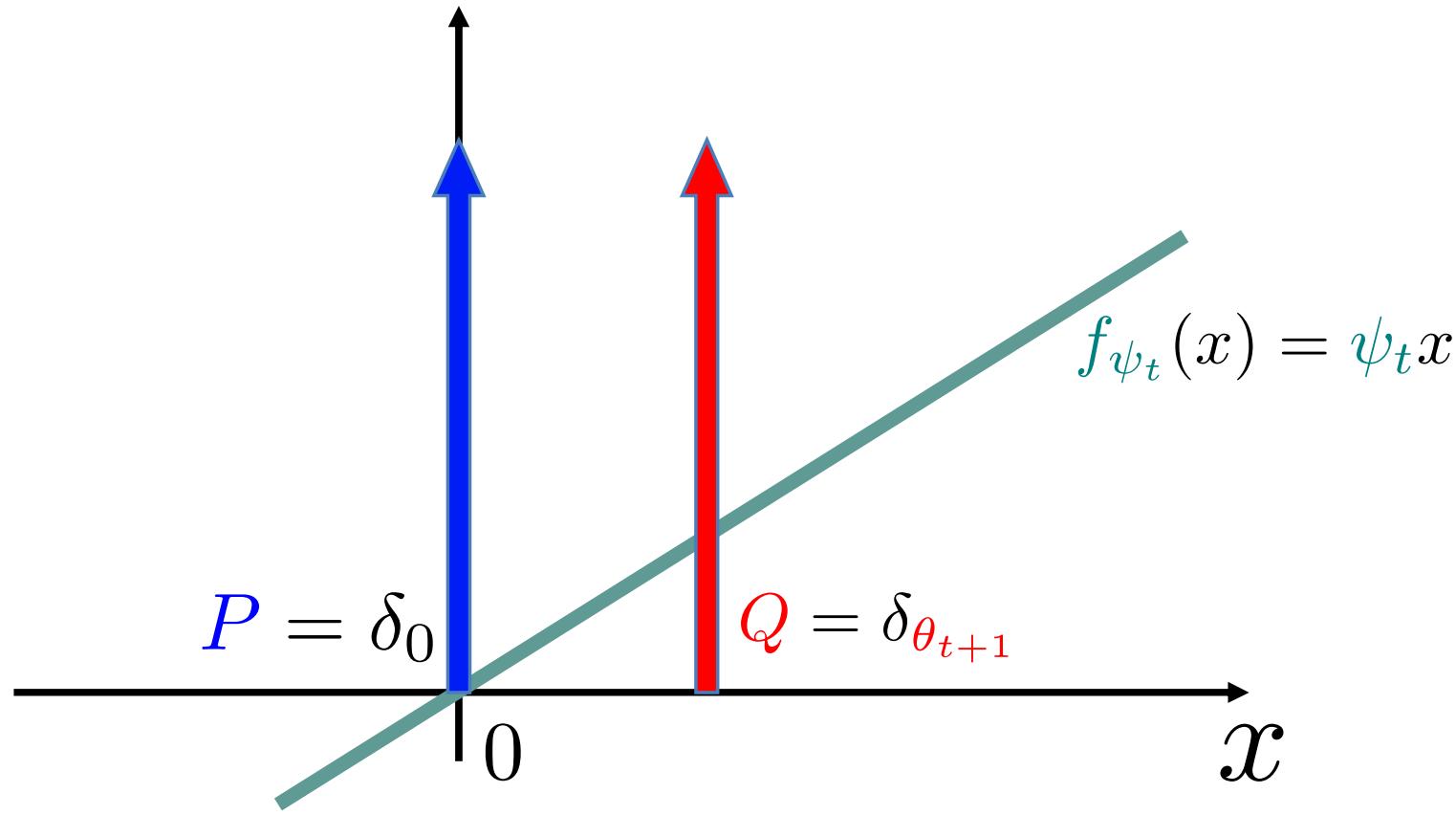
Gradient **ascent** on **critic**:



$$\frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \theta_{t+1}$$

DiracGAN toy example

Gradient **ascent** on **critic**:



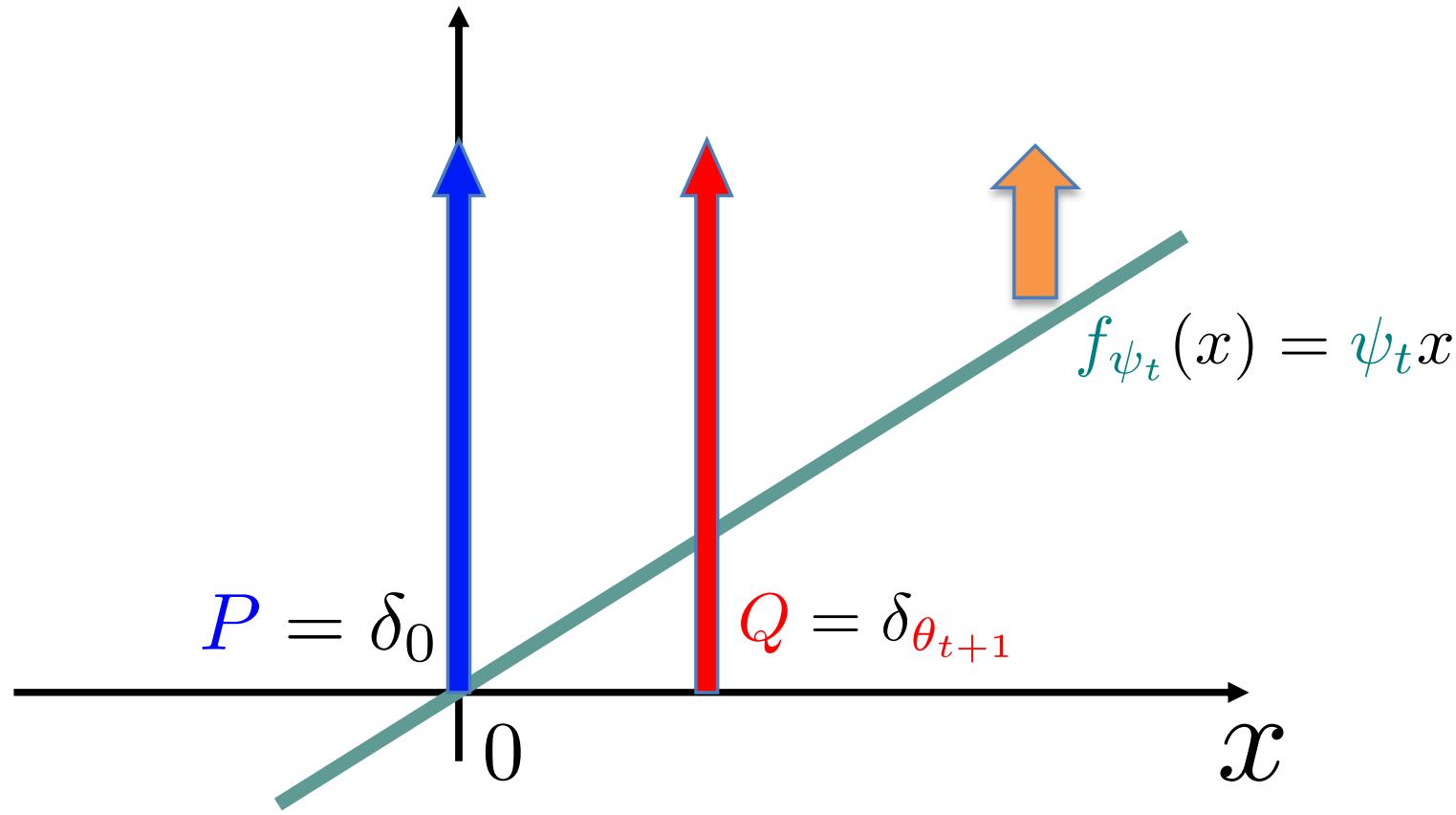
$$\frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \theta_{t+1}$$

$$\psi_{t+1} = \psi_t + \zeta \frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \psi_t + \zeta \theta_{t+1}$$

for stepsize ζ

DiracGAN toy example

Gradient **ascent** on **critic**:

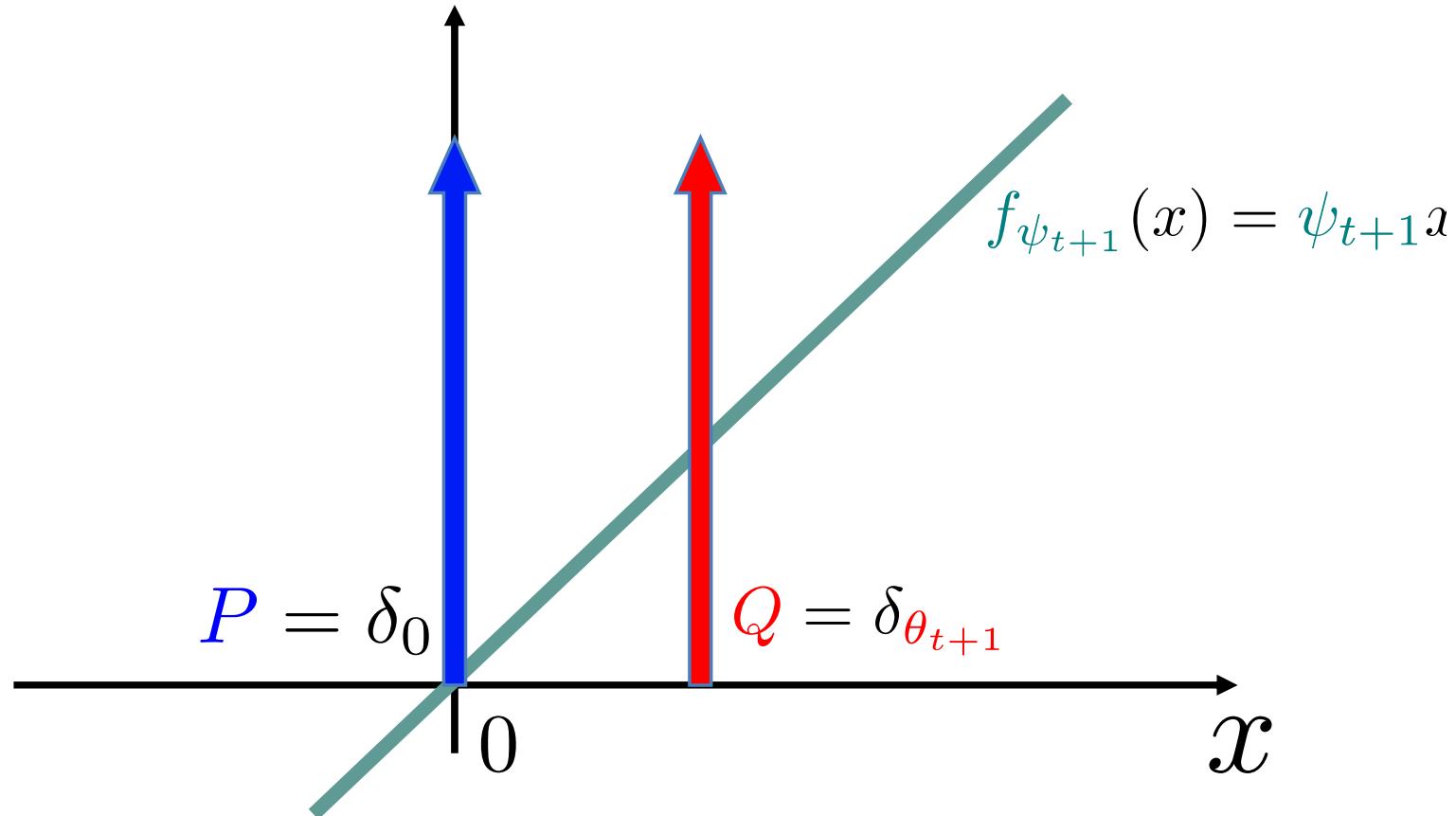


$$\frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \theta_{t+1}$$

$$\psi_{t+1} = \psi_t + \zeta \frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \psi_t + \zeta \theta_{t+1}$$

DiracGAN toy example

Gradient **ascent** on **critic**:



$$\frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \theta_{t+1}$$

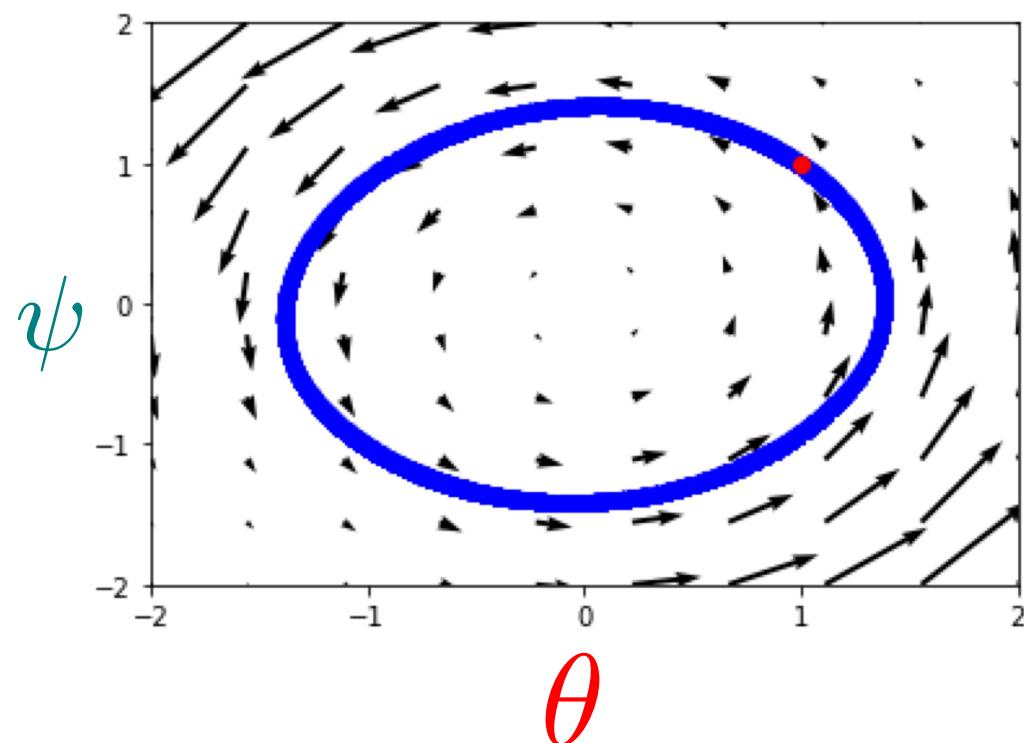
$$\psi_{t+1} = \psi_t + \zeta \frac{\partial}{\partial \psi} D(P, Q; \psi_t) = \psi_t + \zeta \theta_{t+1}$$

DiracGAN toy example

Idealised continuous system (infinitely small learning rate)

$$\begin{bmatrix} \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} -\nabla_{\psi} D(P, Q; \psi) \\ \nabla_{\theta} D(P, Q; \psi) \end{bmatrix}$$

Every integral curve $(\psi(t), \theta(t))$ of the gradient vector field satisfies $\psi^2(t) + \theta^2(t) = c$ for all $t \in [0, \infty)$.



WGAN toy example

WGAN-GP style gradient penalty may not converge near solution

Nagarajan and Kolter [NeurIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

Recall the WGAN-GP penalisation

$$\max_{\psi} \mathbf{E}_{X \sim P} f_{\psi}(X) - \mathbf{E}_{Z \sim R} f_{\psi}(G_{\theta}(Z)) + \lambda \mathbf{E}_{\tilde{X}} \left(\left\| \nabla_{\tilde{X}} f_{\psi}(\tilde{X}) \right\| - 1 \right)^2$$

WGAN toy example

WGAN-GP style gradient penalty may not converge near solution

Nagarajan and Kolter [NeurIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

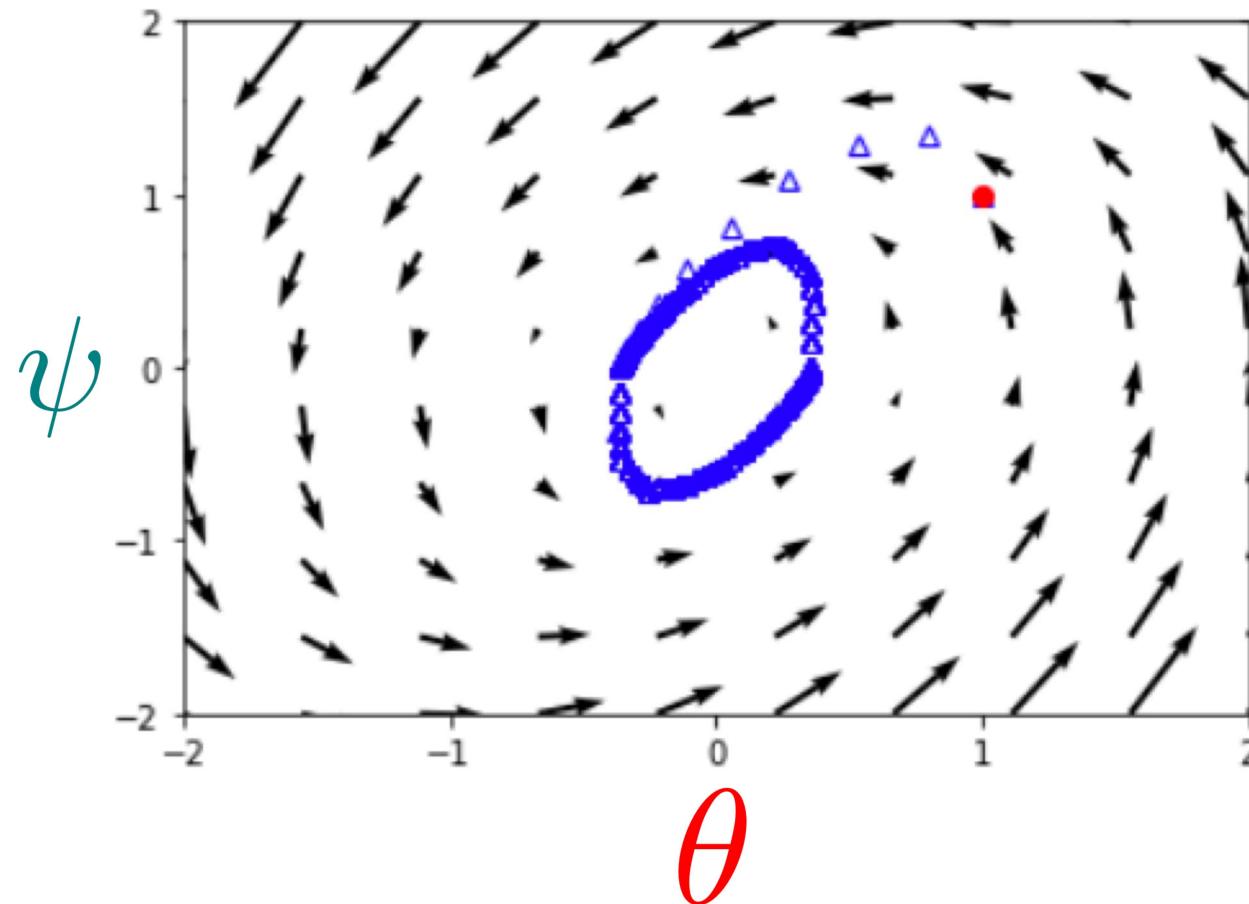


Figure from Mescheder et al. [ICML 2018]

WGAN toy example

WGAN-GP style gradient penalty may not converge near solution

Nagarajan and Kolter [NeurIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

A solution? Modified control of witness gradient

$$\max_{\psi} \mathbf{E}_{X \sim P} f_{\psi}(X) - \mathbf{E}_{Z \sim R} f_{\psi}(G_{\theta}(Z)) + \lambda \underbrace{\mathbf{E}_{\tilde{X}} \left\| \nabla_{\tilde{X}} f_{\psi}(\tilde{X}) \right\|^2}_{\text{new}}$$

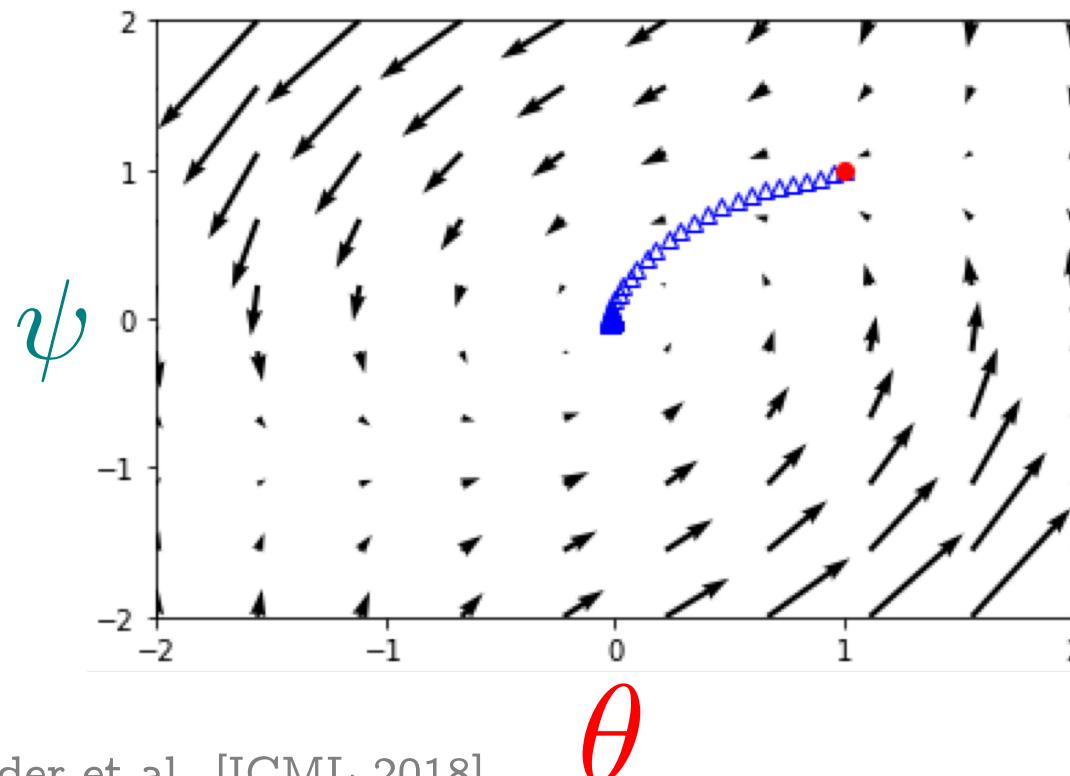
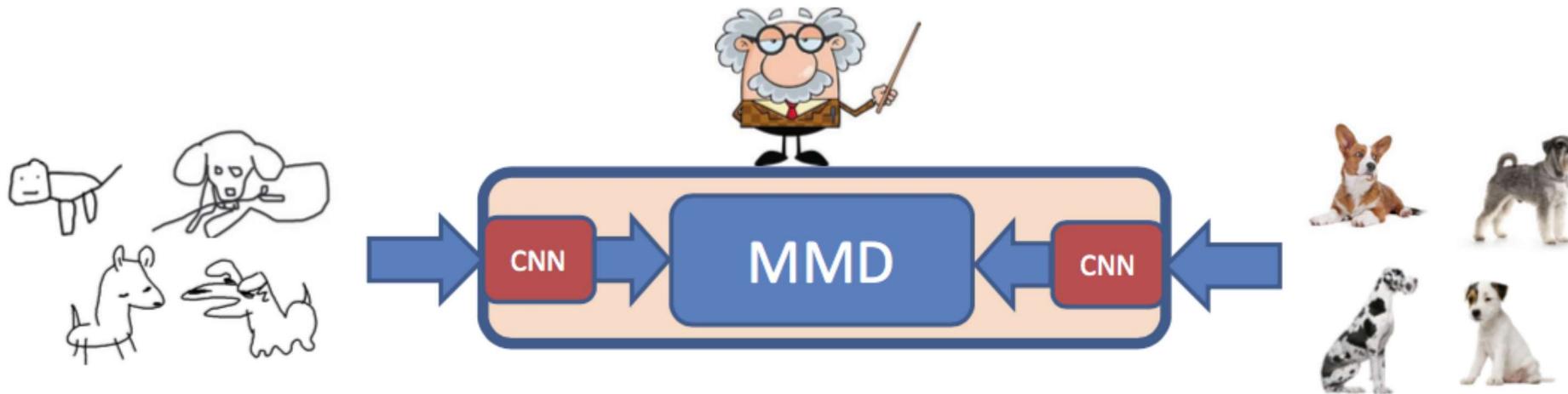


Figure from Mescheder et al. [ICML 2018]

Gradient penalty: the regularisation viewpoint

CNN features for an MMD witness

- Add convolutional features!
- The **critic** (teacher) also needs to be trained.



$$\mathcal{K}(x, y) = h_\psi^\top(x) h_\psi(y)$$

where $h_\psi(x)$ is a CNN map:

- **Wasserstein GAN** Arjovsky et al.
[ICML 2017]
- **WGANGP** Gulrajani et al.
[NeurIPS 2017]

$\mathcal{K}(x, y) = k(h_\psi(x), h_\psi(y))$
where $h_\psi(x)$ is a CNN map,

k is e.g. an exponentiated quadratic kernel

MMD Li et al., [NeurIPS 2017]

Cramer Bellemare et al. [2017]

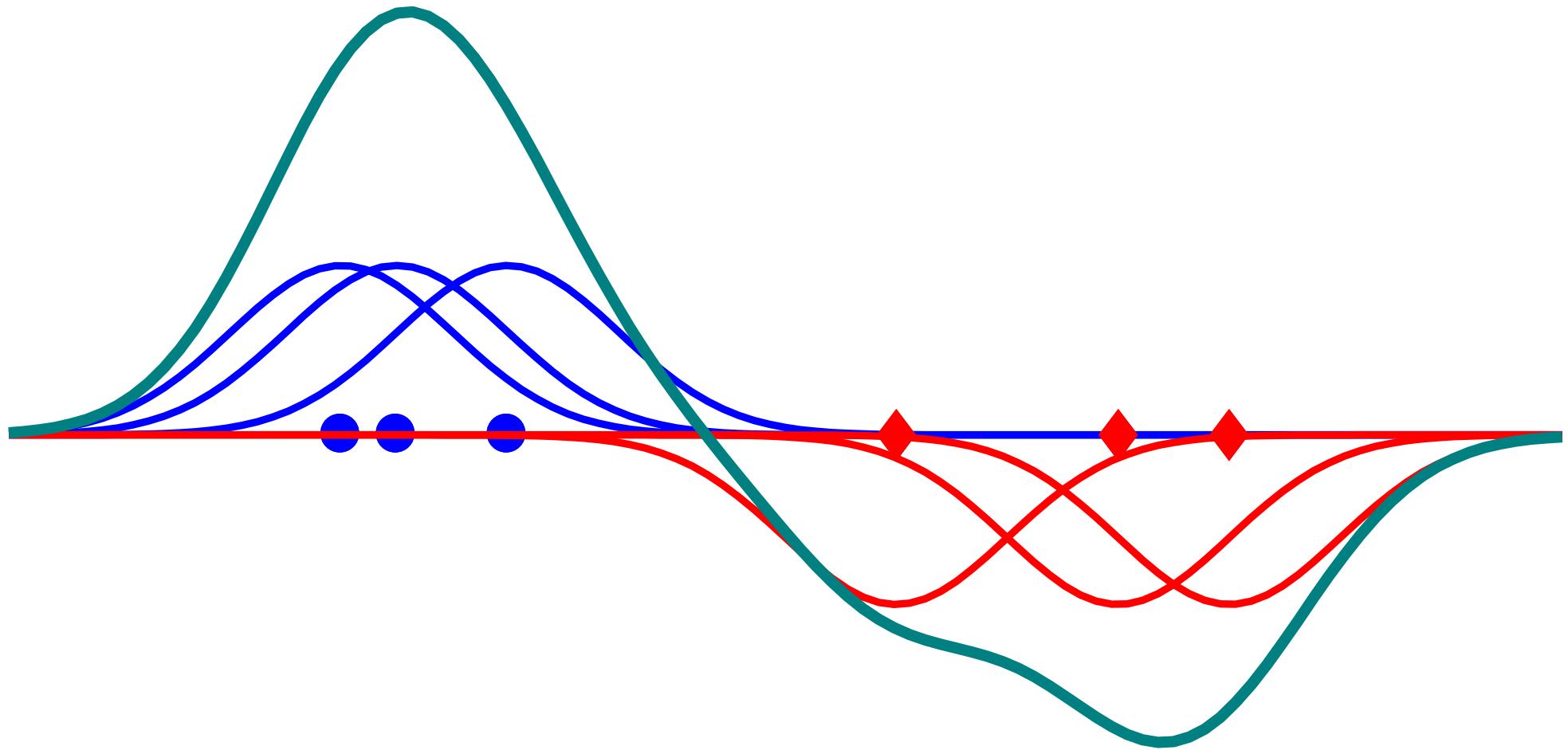
Coulomb Unterthiner et al., [ICLR 2018]

Demystifying MMD GANs Binkowski,
Sutherland, Arbel, G., [ICLR 2018]

Witness function, kernels on deep features

Reminder: witness function,

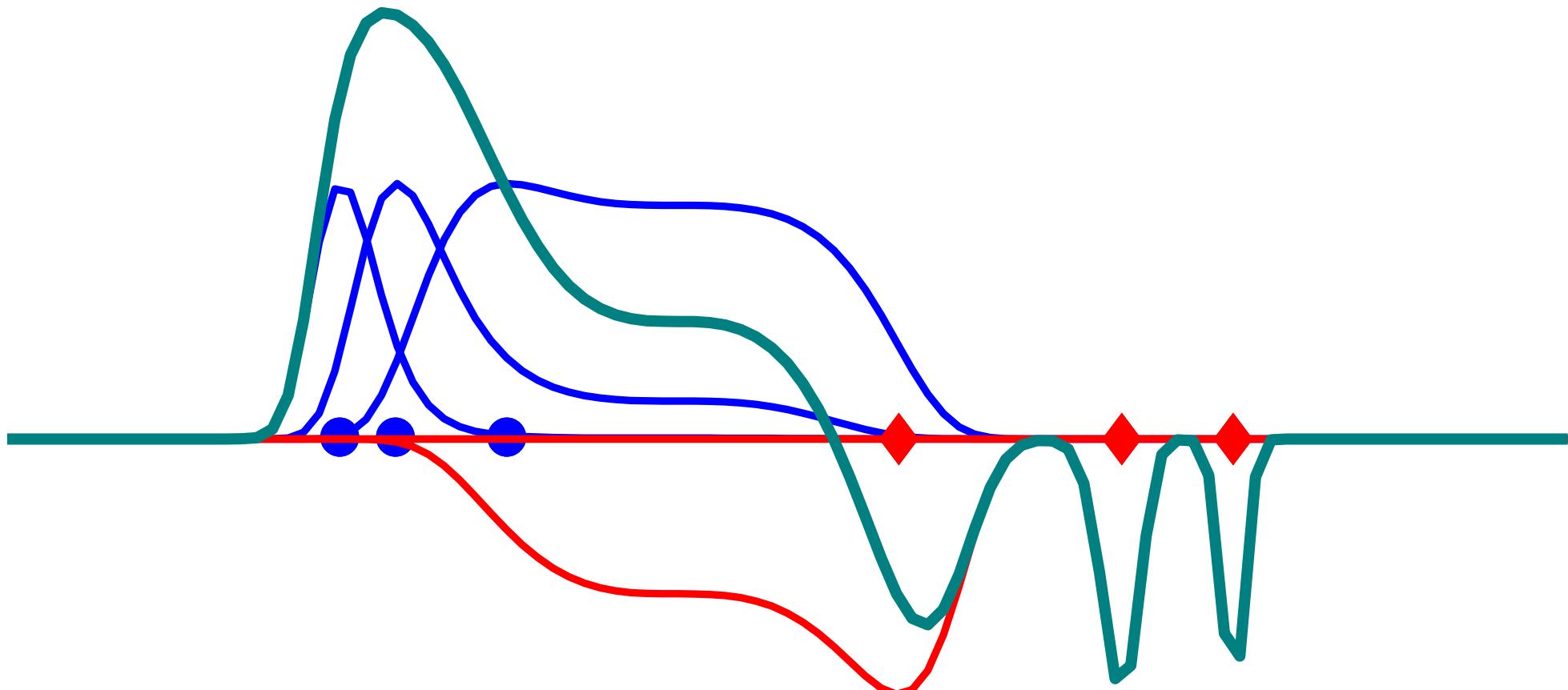
$k(x, y)$ is exponentiated quadratic



Witness function, kernels on deep features

Reminder: witness function,

$k(h_\psi(x), h_\psi(y))$ with nonlinear h_ψ and exp. quadratic k



Challenges for learned critic features

Learned critic features:

MMD with kernel $k(h_\psi(x), h_\psi(y))$ must give useful gradient to generator.

Challenges for learned critic features

Learned critic features:

MMD with kernel $k(h_\psi(x), h_\psi(y))$ must give useful gradient to generator.

Relation with test power?

If the MMD with kernel $k(h_\psi(x), h_\psi(y))$ gives a powerful test, will it be a good critic?

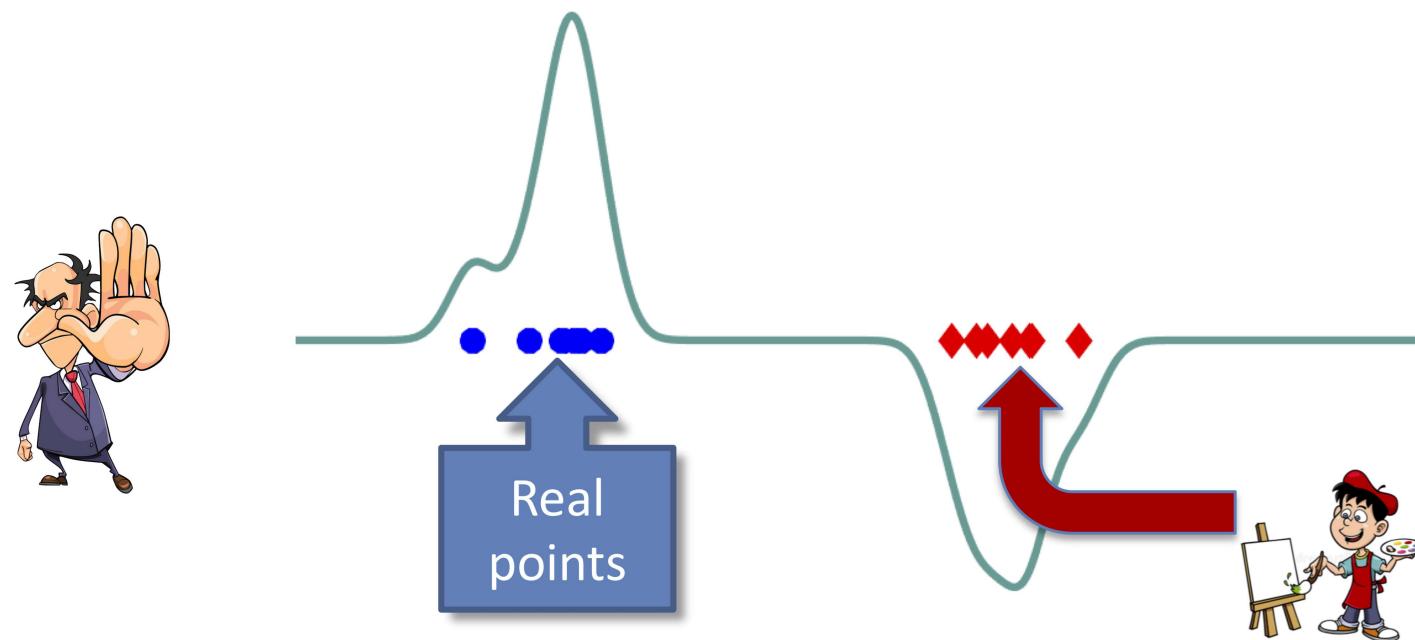
Challenges for learned critic features

Learned critic features:

MMD with kernel $k(h_\psi(x), h_\psi(y))$ must give useful gradient to generator.

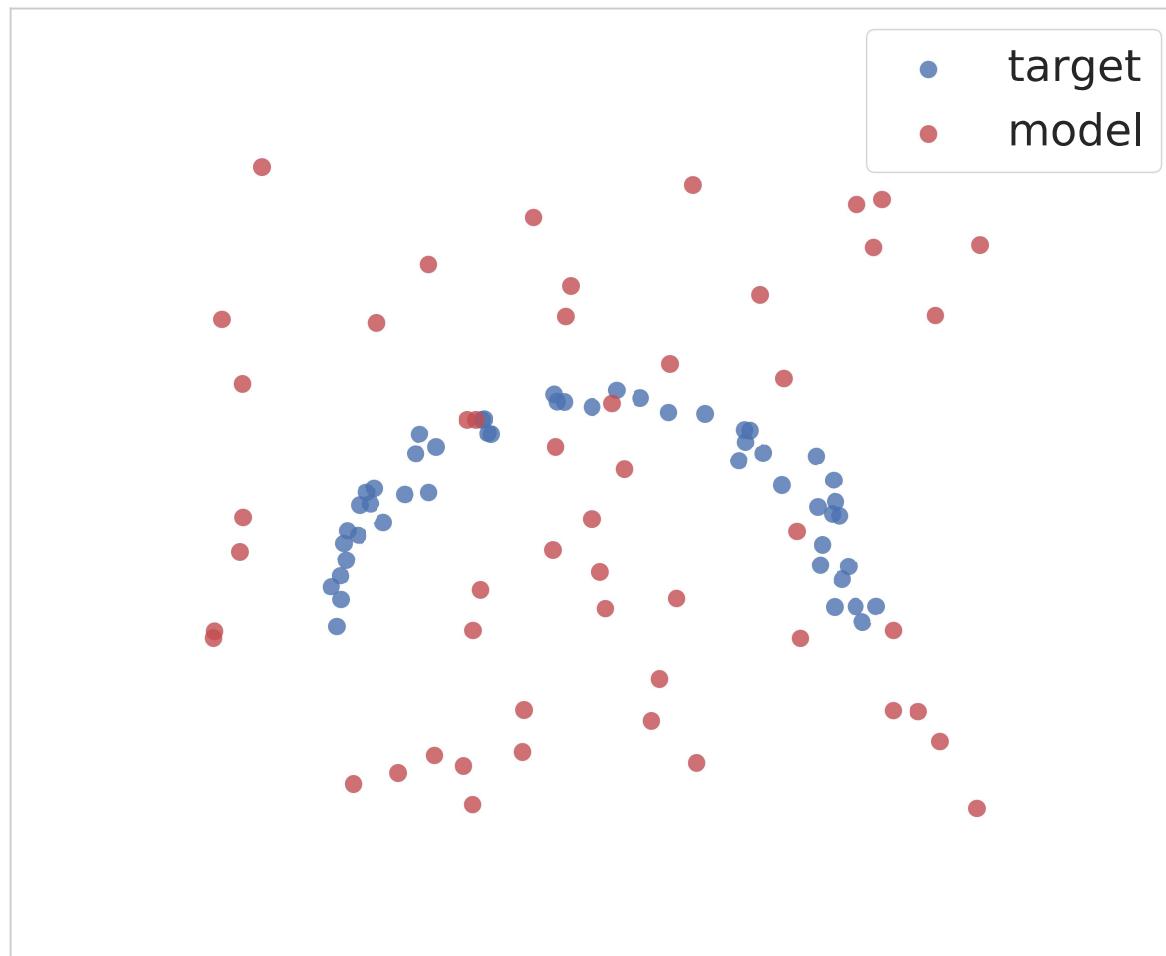
Relation with test power?

If the MMD with kernel $k(h_\psi(x), h_\psi(y))$ gives a powerful test, will it be a good critic?



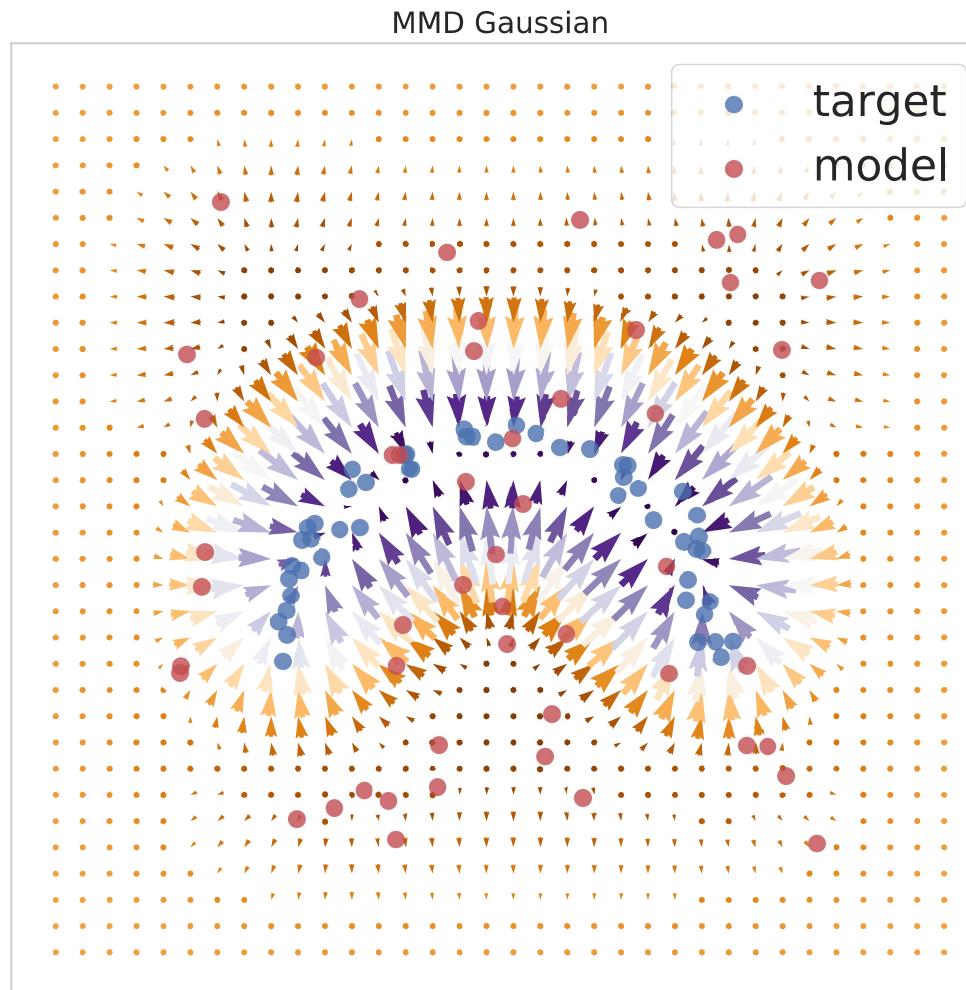
A simple 2-D example

Samples from target P and model Q



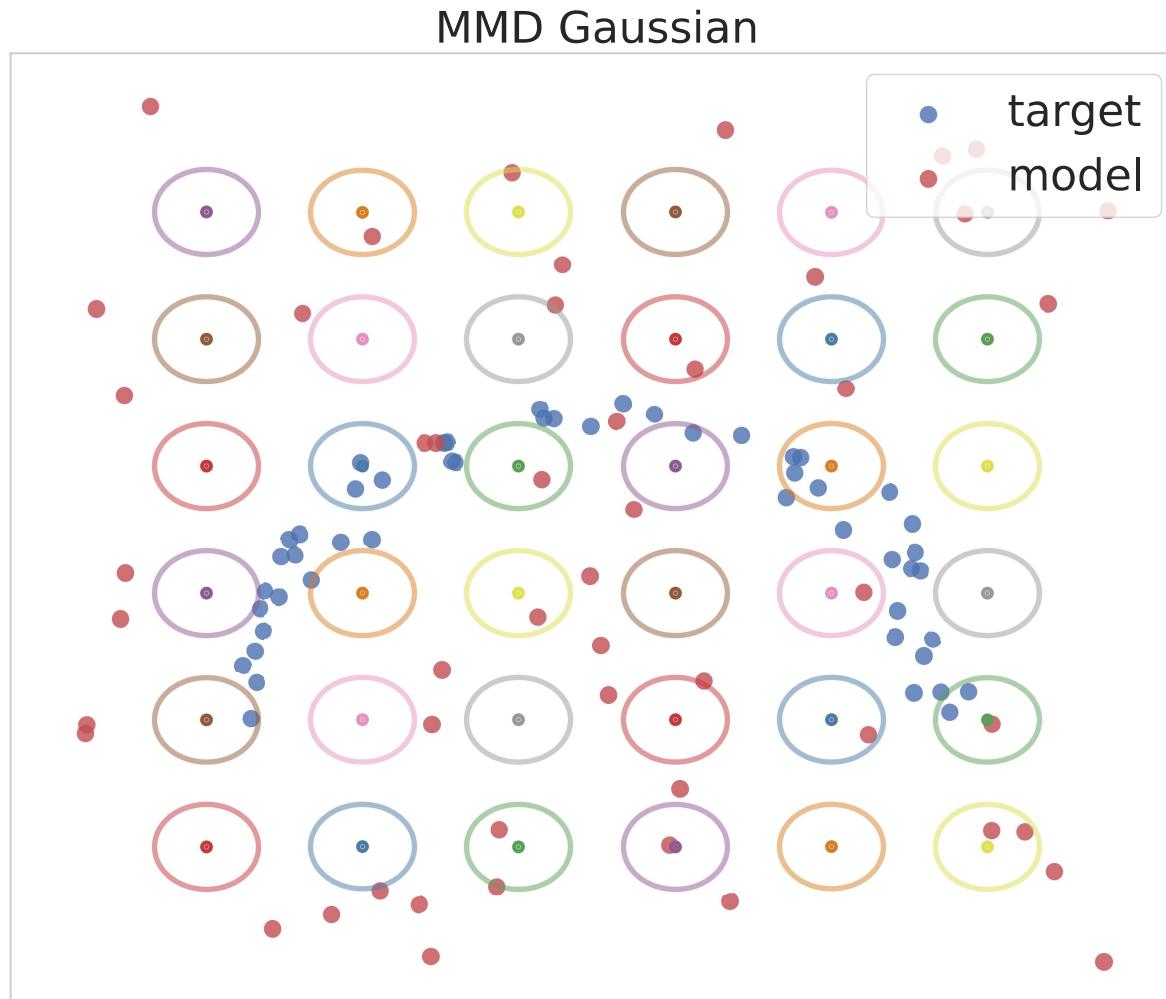
A simple 2-D example

Witness gradient, MMD with exp. quad. kernel $k(x, y)$



A simple 2-D example

What the kernels $k(x, y)$ look like



A data-adaptive gradient penalty: NeurIPS 2018

- New gradient regulariser Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
 - Also related to Sobolev GAN Mroueh et al. [ICLR 2018]
-

On gradient regularizers for MMD GANs

Michael Arbel

Gatsby Computational Neuroscience Unit
University College London
michael.n.arbel@gmail.com

Dougal J. Sutherland

Gatsby Computational Neuroscience Unit
University College London
dougal@gmail.com

Mikołaj Bińkowski

Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Arthur Gretton

Gatsby Computational Neuroscience Unit
University College London
arthur.gretton@gmail.com

A data-adaptive gradient penalty: NeurIPS 2018

- New gradient regulariser Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Also related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness constraint:

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

where

$$\|\mathbf{f}\|_S^2 = \|\mathbf{f}\|_{L_2(\mathbf{P})}^2 + \|\nabla \mathbf{f}\|_{L_2(\mathbf{P})}^2 + \lambda \|\mathbf{f}\|_k^2$$

The equation is annotated with three boxes and arrows: an orange box labeled "L₂ norm control" with an arrow pointing to the first term; a purple box labeled "Gradient control" with an arrow pointing to the second term; and an orange box labeled "RKHS smoothness" with an arrow pointing to the third term.

Maximise \widetilde{MMD} wrt critic features

A data-adaptive gradient penalty: NeurIPS 2018

- New gradient regulariser Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Also related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness constraint:

$$\widetilde{MMD} := \sup_{\|\mathbf{f}\|_S \leq 1} [\mathbb{E}_P f(\mathbf{X}) - \mathbb{E}_Q f(\mathbf{Y})]$$

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

A data-adaptive gradient penalty: NeurIPS 2018

- New gradient regulariser Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Also related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness constraint:

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

Maximise scaled MMD over critic features:

$$SMMD(\mathbf{P}, \lambda) = \sigma_{\mathbf{P}, \lambda} MMD$$

where

$$\sigma_{\mathbf{P}, \lambda}^2 = \lambda + \int k(h_\psi(\mathbf{x}), h_\psi(\mathbf{x})) d\mathbf{P}(\mathbf{x}) + \sum_{i=1}^d \int \partial_i \partial_{i+d} k(h_\psi(\mathbf{x}), h_\psi(\mathbf{x})) d\mathbf{P}(\mathbf{x})$$

Replace expensive constraint with cheap upper bound:

$$\|\mathbf{f}\|_S^2 \leq \sigma_{\mathbf{P}, \lambda}^{-1} \|\mathbf{f}\|_k^2$$

A data-adaptive gradient penalty: NeurIPS 2018

- New gradient regulariser Arbel, Sutherland, Binkowski, G. [NeurIPS 2018]
- Also related to Sobolev GAN Mroueh et al. [ICLR 2018]

Maximise scaled MMD over critic features:

$$SMMMD(\textcolor{blue}{P}, \lambda) = \sigma_{\textcolor{blue}{P}, \lambda} MMD$$

where

$$\sigma_{\textcolor{blue}{P}, \lambda}^2 = \lambda + \int \textcolor{brown}{k}(\textcolor{teal}{h}_\psi(\textcolor{blue}{x}), \textcolor{teal}{h}_\psi(\textcolor{blue}{x})) d\textcolor{blue}{P}(x) + \sum_{i=1}^d \int \partial_i \partial_{i+d} \textcolor{brown}{k}(\textcolor{teal}{h}_\psi(\textcolor{blue}{x}), \textcolor{teal}{h}_\psi(\textcolor{blue}{x})) d\textcolor{blue}{P}(x)$$

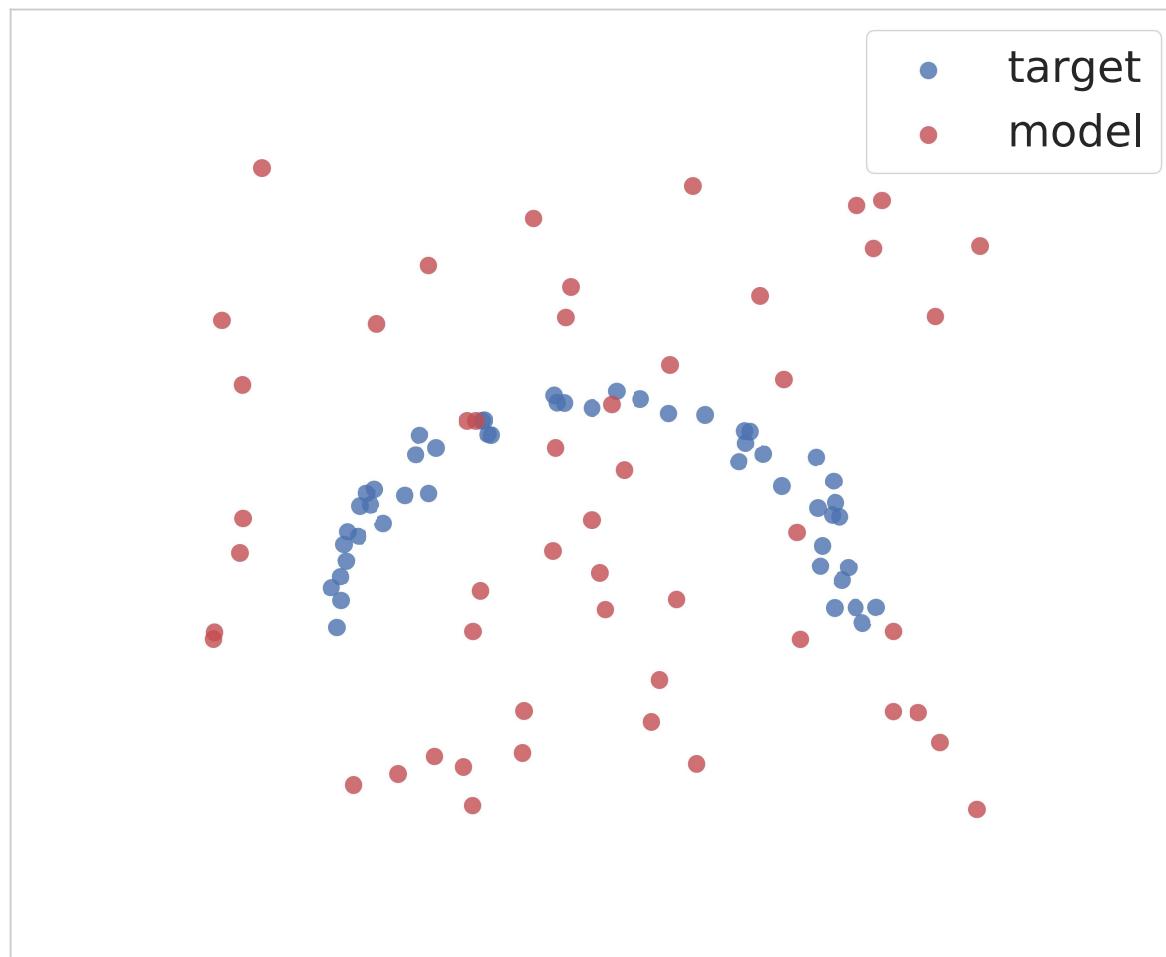
Replace expensive constraint with cheap upper bound:

$$\|\textcolor{teal}{f}\|_{\textcolor{brown}{S}}^2 \leq \sigma_{\textcolor{blue}{P}, \lambda}^{-1} \|\textcolor{teal}{f}\|_k^2$$

Idea: rather than regularise the critic or witness function, regularise features directly

Simple 2-D example revisited

Samples from target P and model Q



Simple 2-D example revisited

Use kernels $k(h_\psi(\mathbf{x}), h_\psi(\mathbf{y}))$ with features

$$h_\psi(\mathbf{x}) = L_3 \left(\begin{bmatrix} \mathbf{x} \\ L_2(L_1(\mathbf{x})) \end{bmatrix} \right)$$

where L_1, L_2, L_3 are fully connected with quadratic nonlinearity.

Simple 2-D example revisited

Witness gradient, **maximise** $SMMD(P, \lambda)$
to learn $h_\psi(x)$ for $k(h_\psi(x), h_\psi(y))$

Simple 2-D example revisited

What the kernels $k(h_\psi(x), h_\psi(y))$ look like

isolines movie, use Acrobat Reader to play

Our empirical observations

Data-adaptive critic loss:

- Witness function class for $SMMD(P, \lambda)$ depends on P .
 - Without data-dependent regularisation, maximising MMD over features h_ψ of kernel $k(h_\psi(x), h_\psi(y))$ can be **unhelpful**.
 - WGAN-GP is a pretty good data-dependent **regularisation strategy**
- Similar regularisation strategies apply to variational form in f-GANs

Roth et al [NeurIPS 2017, eq. 19 and 20]

Our empirical observations

Data-adaptive critic loss:

- Witness function class for $SMMD(P, \lambda)$ depends on P .
 - Without data-dependent regularisation, maximising MMD over features h_ψ of kernel $k(h_\psi(x), h_\psi(y))$ can be **unhelpful**.
 - WGAN-GP is a pretty good data-dependent **regularisation strategy**
- Similar regularisation strategies apply to variational form in f-GANs

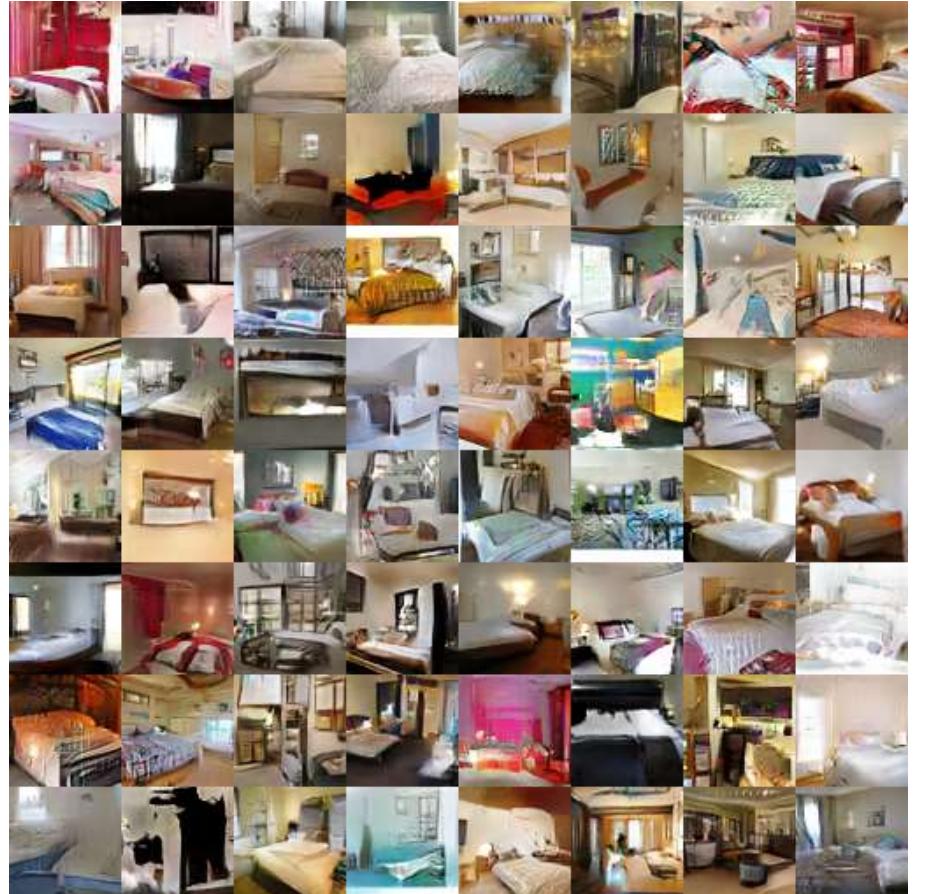
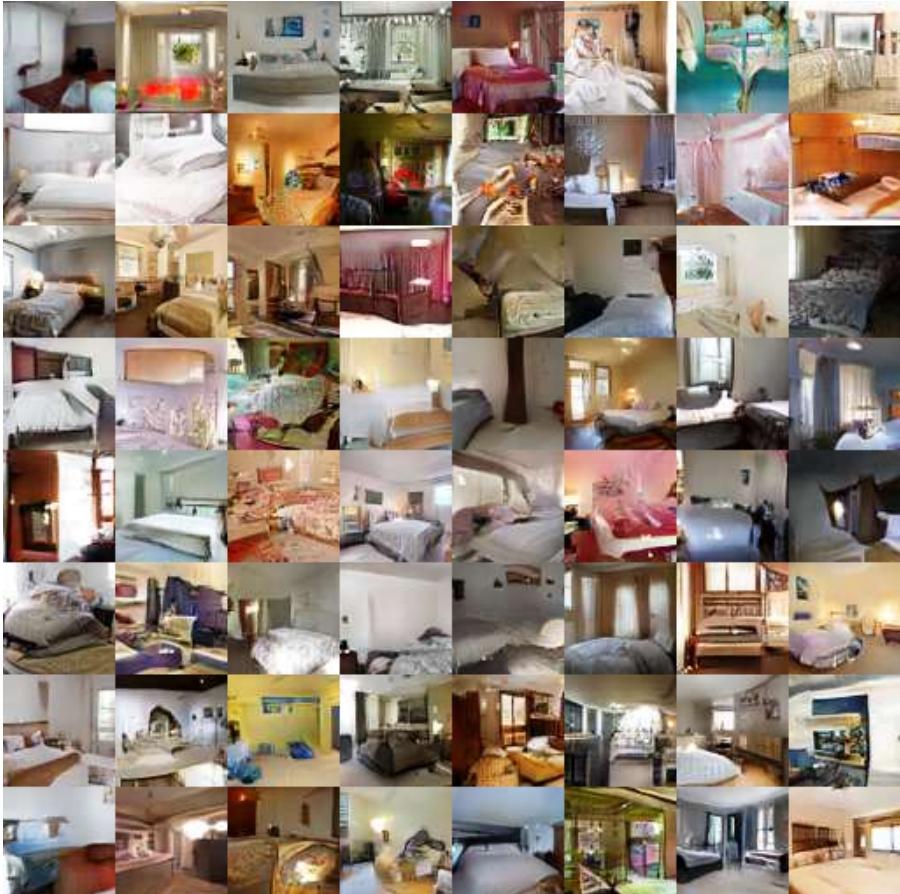
Roth et al [NeurIPS 2017, eq. 19 and 20]

Alternate critic and generator training:

- Weaker critics can give better signals to poor (early stage) generators.
- Incomplete training of the critic is also a **regularisation strategy**

Linear vs nonlinear kernels

- Critic features from DCGAN: an f -filter critic has f , $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64 .



$k(h_\psi(\mathbf{x}), h_\psi(\mathbf{y}))$, $f = 64$,
KID=3

$h_\psi^\top(\mathbf{x})h_\psi(\mathbf{y})$, $f = 64$, KID=4

Linear vs nonlinear kernels

- Critic features from DCGAN: an f -filter critic has f , $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64 .



$k(h_\psi(x), h_\psi(y))$, $f = 16$,
KID=9

$h_\psi^\top(x) h_\psi(y)$, $f = 16$, KID=37

The theory

Scaled MMD vs Wasserstein-1 (NeurIPS 18)

Let $k_\psi = \textcolor{brown}{k} \circ \textcolor{teal}{h}_\psi$.

Wasserstein-1 bounds SMMD,

$$SMMD(\textcolor{blue}{P}, \textcolor{red}{Q}) \leq \frac{\textcolor{brown}{Q}_k \kappa^L}{d_L \alpha^L} \mathcal{W}(\textcolor{blue}{P}, \textcolor{red}{Q})$$

■ Conditions on the neural network layers:

- $h_\psi : \mathcal{X} \rightarrow \mathbb{R}^s$ fully-connected L -layer network, Leaky-ReLU $_\alpha$ activations whose layers do not increase in width
- Width of ℓ th layer is d_ℓ .
- κ is the bound on condition number of the weight matrices W^ℓ

■ Conditions on the kernel and gradient regulariser:

- k satisfying mild smoothness conditions, summarised in $\textcolor{brown}{Q}_k < \infty$.
- μ is a probability measure with support over \mathcal{X} ,

$$\int k(x, x) d\mu(x) + \sum_{i=1}^d \int \partial_i \partial_{i+d} k(x, x) d\mu(x)$$

Scaled MMD vs Wasserstein-1 (NeurIPS 18)

Let $k_\psi = \textcolor{brown}{k} \circ \textcolor{teal}{h}_\psi$.

Wasserstein-1 bounds SMMD,

$$SMMD(\textcolor{blue}{P}, \textcolor{red}{Q}) \leq \frac{\textcolor{brown}{Q}_k \kappa^L}{d_L \alpha^L} \mathcal{W}(\textcolor{blue}{P}, \textcolor{red}{Q})$$

■ Conditions on the neural network layers:

- $h_\psi : \mathcal{X} \rightarrow \mathbb{R}^s$ fully-connected $\textcolor{teal}{L}$ -layer network, Leaky-ReLU $_\alpha$ activations whose layers do not increase in width
- Width of ℓ th layer is $\textcolor{teal}{d}_\ell$.
- κ is the bound on condition number of the weight matrices W^ℓ

■ Conditions on the kernel and gradient regulariser:

- k satisfying mild smoothness conditions, summarised in $\textcolor{brown}{Q}_k < \infty$.
- μ is a probability measure with support over \mathcal{X} ,

$$\int k(x, x) d\mu(x) + \sum_{i=1}^d \int \partial_i \partial_{i+d} k(x, x) d\mu(x)$$

Scaled MMD vs Wasserstein-1 (NeurIPS 18)

Let $k_\psi = \mathbf{k} \circ h_\psi$.

Wasserstein-1 bounds SMMD,

$$SMMD(\mathbf{P}, \mathbf{Q}) \leq \frac{Q_k \kappa^L}{d_L \alpha^L} \mathcal{W}(\mathbf{P}, \mathbf{Q})$$

■ Conditions on the neural network layers:

- $h_\psi : \mathcal{X} \rightarrow \mathbb{R}^s$ fully-connected L -layer network, Leaky-ReLU $_\alpha$ activations whose layers do not increase in width
- Width of ℓ th layer is d_ℓ .
- κ is the bound on condition number of the weight matrices W^ℓ

■ Conditions on the kernel and gradient regulariser:

- \mathbf{k} satisfying mild smoothness conditions, summarised in $Q_k < \infty$.
- μ is a probability measure with support over \mathcal{X} ,

$$\int k(x, x) d\mu(x) + \sum_{i=1}^d \int \partial_i \partial_{i+d} k(x, x) d\mu(x)$$

Unbiased gradients of MMD, WGAN-GP (ICLR 18)

Subject to mild conditions on

- Critic mappings h_ψ (conditions hold for almost all feedforward networks: convolutions, max pooling, ReLU,...)
- kernel k (a growth assumption)
- Target distribution P , generator network $Y \sim G_\theta(Z)$ (densities not needed, second moments must exist),

Then for μ -almost all ψ, θ where μ is Lebesgue,

$$\mathbb{E}_{\substack{X \sim P \\ Z \sim R}} [\partial_{\psi, \theta} k(h_\psi(X), h_\psi(G_\theta(Z)))] = \partial_{\psi, \theta} \mathbb{E}_{\substack{X \sim P \\ Z \sim R}} [k(h_\psi(X), h_\psi(G_\theta(Z)))] .$$

and thus MMD gradients unbiased.

Also true for WGAN-GP.

Bias of MMD GAN critic (ICLR 18)

Gradient bias when critic trained on a separate dataset?

Recall definition of MMD for P vs Q

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

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

Define f_{tr} as discriminator witness trained on $\{x_i^{\text{tr}}\}_{i=1}^m \stackrel{\text{i.i.d.}}{\sim} P$,
 $\{y_i^{\text{tr}}\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} Q$.

Then

$$[\mathbf{E}_P f_{tr}(X) - \mathbf{E}_Q f_{tr}(Y)] \leq MMD(P, Q; F)$$

Downwards bias. Unless bias is in f_{tr} constant, biased gradients too.

Same true for WGAN-GP.

Bias of MMD GAN critic (ICLR 18)

Gradient bias when critic trained on a separate dataset?

Recall definition of MMD for P vs Q

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

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

Define f_{tr} as discriminator witness trained on $\{x_i^{\text{tr}}\}_{i=1}^m \stackrel{\text{i.i.d.}}{\sim} P$,
 $\{y_i^{\text{tr}}\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} Q$.

Then

$$[\mathbf{E}_P f_{tr}(X) - \mathbf{E}_Q f_{tr}(Y)] \leq MMD(P, Q; F)$$

Downwards bias. Unless bias is in f_{tr} constant, biased gradients too.

Same true for WGAN-GP.

Bias of MMD GAN critic (ICLR 18)

Gradient bias when critic trained on a separate dataset?

Recall definition of MMD for P vs Q

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

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

Define f_{tr} as discriminator witness trained on $\{x_i^{\text{tr}}\}_{i=1}^m \stackrel{\text{i.i.d.}}{\sim} P$,
 $\{y_i^{\text{tr}}\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} Q$.

Then

$$[\mathbf{E}_P f_{tr}(X) - \mathbf{E}_Q f_{tr}(Y)] \leq MMD(P, Q; F)$$

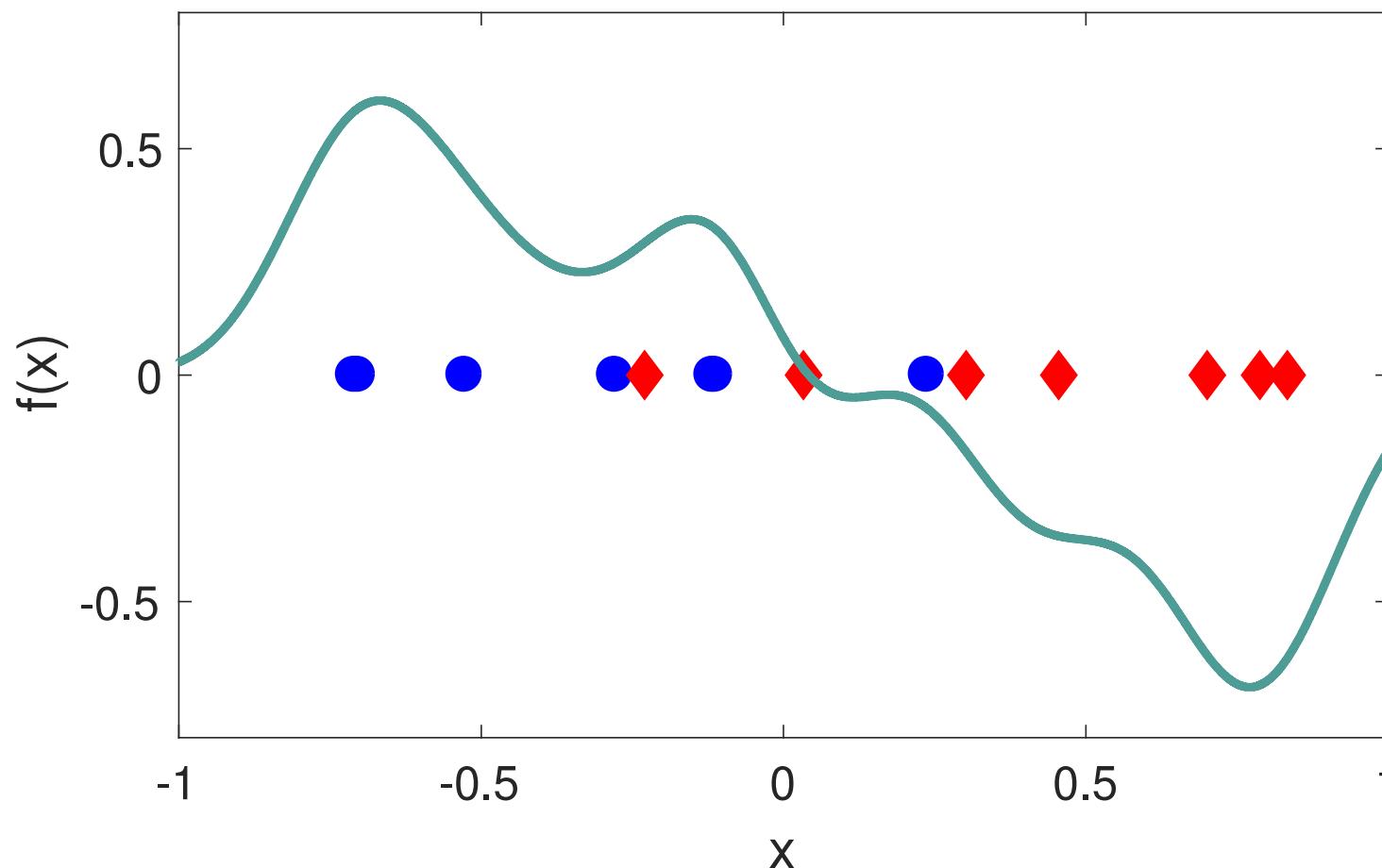
Downwards bias. Unless bias is in f_{tr} constant, biased gradients too.

Same true for WGAN-GP.

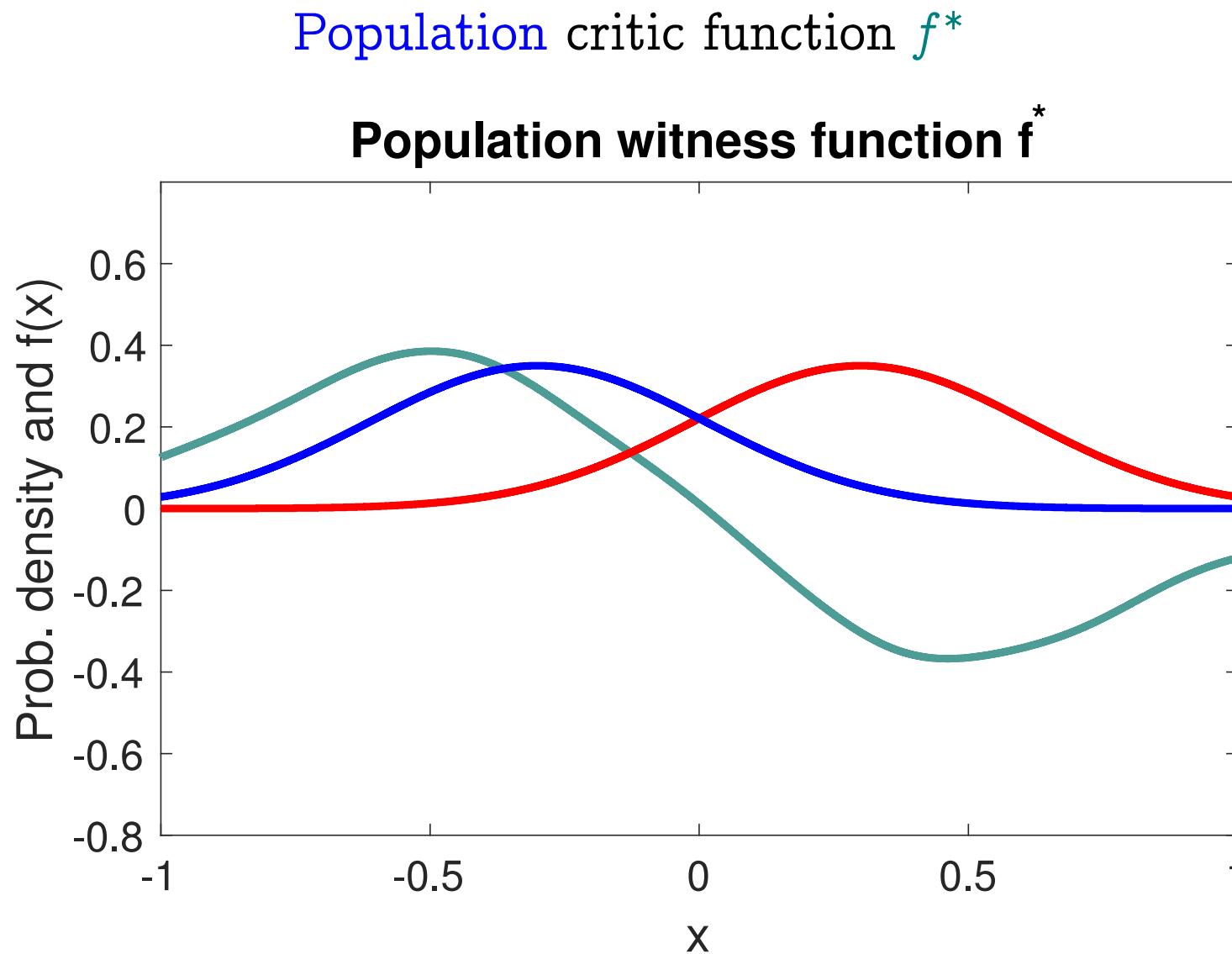
Bias of MMD GAN critic (ICLR 18)

Training minibatch critic function f_{tr}

Trained witness function f_{tr}

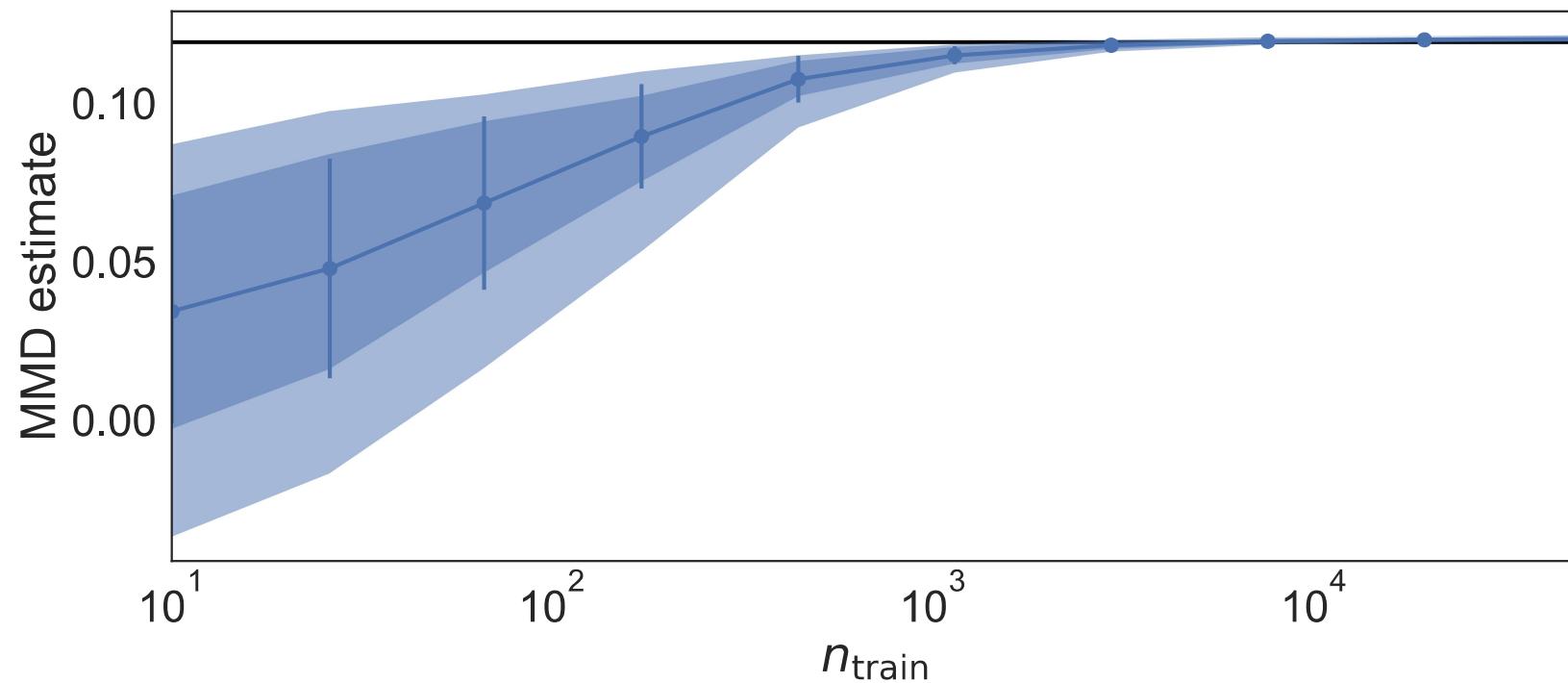


Bias of MMD GAN critic (ICLR 18)



Bias of MMD GAN critic (ICLR 18)

Bias in MMD vs training minibatch size:



Evaluation and experiments

Evaluation of GANs

The inception score? Salimans et al. [NeurIPS 2016]

Based on the classification output $p(y|x)$ of the inception model Szegedy et al. [ICLR 2014],

$$E_X \exp KL(P(y|X) || P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

Evaluation of GANs

The inception score? Salimans et al. [NeurIPS 2016]

Based on the classification output $p(y|x)$ of the inception model Szegedy et al. [ICLR 2014],

$$E_X \exp KL(P(y|X) || P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

Problem: relies on a trained classifier! Can't be used on new categories (celeb, bedroom...)

Evaluation of GANs

The Frechet inception distance? Heusel et al. [NeurIPS 2017]

Fits **Gaussians** to features in the inception architecture (pool3 layer):

$$FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}(\Sigma_P) + \text{tr}(\Sigma_Q) - 2\text{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)$$

where μ_P and Σ_P are the feature mean and covariance of P

Evaluation of GANs

The Frechet inception distance? Heusel et al. [NeurIPS 2017]

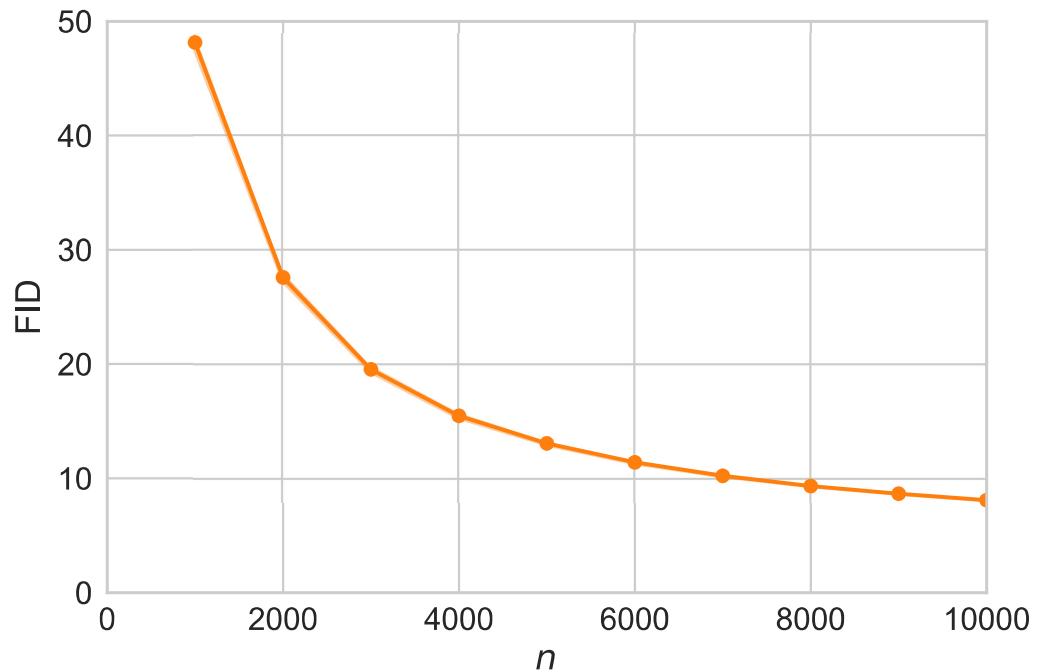
Fits Gaussians to features in the inception architecture (pool3 layer):

$$FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}(\Sigma_P) + \text{tr}(\Sigma_Q) - 2\text{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)$$

where μ_P and Σ_P are the feature mean and covariance of 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 theory.

Assume m samples from P and $n \rightarrow \infty$ samples from Q .

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2 ,

$$FID(\widehat{P}_1, Q) < FID(\widehat{P}_2, Q).$$

Evaluation of GANs

The FID can give the **wrong answer in theory**.

Assume m samples from P and $n \rightarrow \infty$ samples from Q .

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2 ,

$$FID(\widehat{P}_1, Q) < FID(\widehat{P}_2, Q).$$

Evaluation of GANs

The FID can give the **wrong answer in theory**.

Assume m samples from P and $n \rightarrow \infty$ samples from Q .

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2 ,

$$FID(\widehat{P}_1, Q) < FID(\widehat{P}_2, Q).$$

Evaluation of GANs

The FID can give the **wrong answer in theory**.

Assume m samples from P and $n \rightarrow \infty$ samples from Q .

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2 ,

$$FID(\widehat{P}_1, Q) < FID(\widehat{P}_2, Q).$$

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(0, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(1, .8\Sigma+.2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d))$$

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

For a random draw of C :

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

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

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C .

Evaluation of GANs

The FID can give the **wrong answer in practice**.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(0, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(1, .8\Sigma+.2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d))$$

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

For a random draw of C :

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

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

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C .

Evaluation of GANs

The FID can give the **wrong answer in practice**.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(0, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(1, .8\Sigma+.2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d))$$

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

For a random draw of C :

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

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

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C .

Evaluation of GANs

The FID can give the **wrong answer in practice**.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(0, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(1, .8\Sigma+.2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d))$$

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

For a random draw of C :

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

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

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C .

The kernel inception distance (KID)

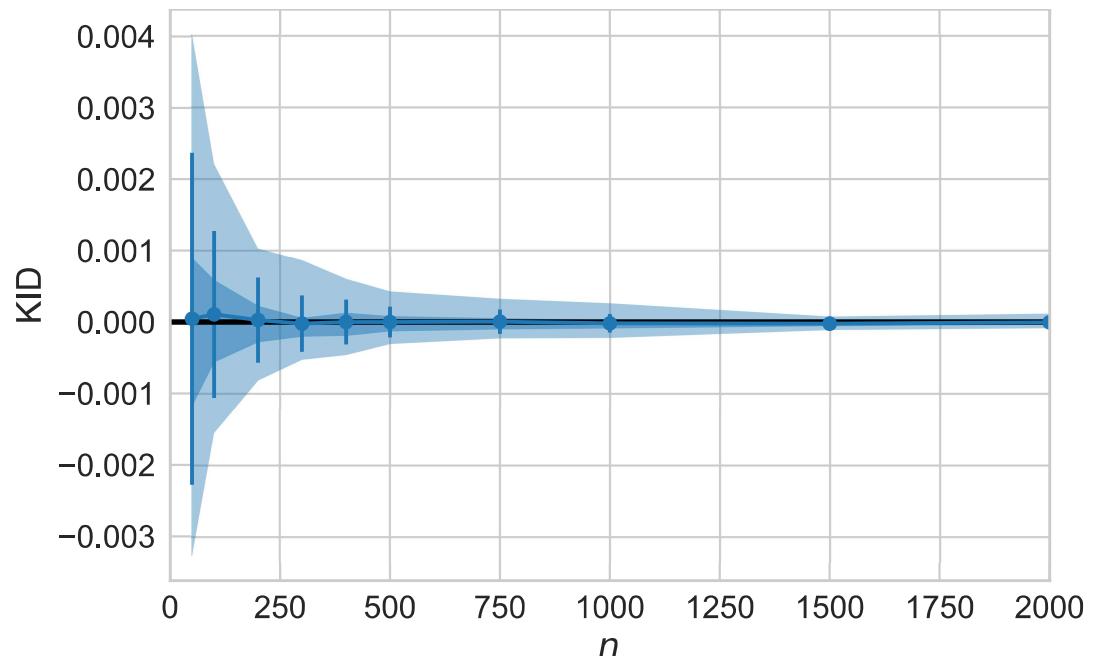
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1 \right)^3.$$

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



The kernel inception distance (KID)

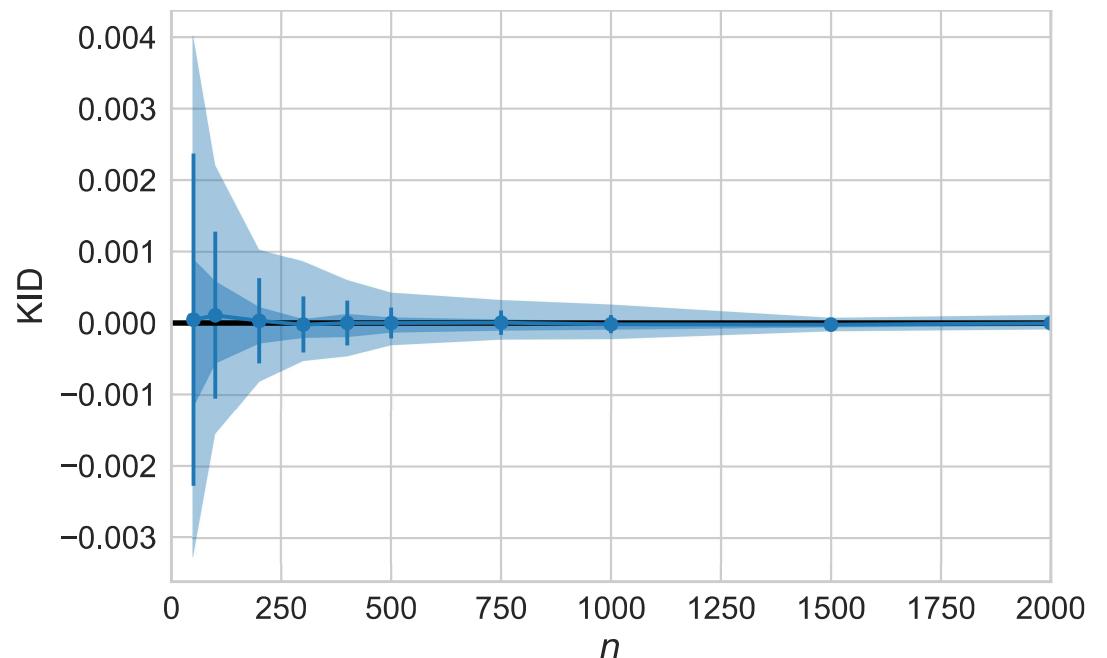
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1 \right)^3.$$

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



...“but isn't KID computationally costly?”

The kernel inception distance (KID)

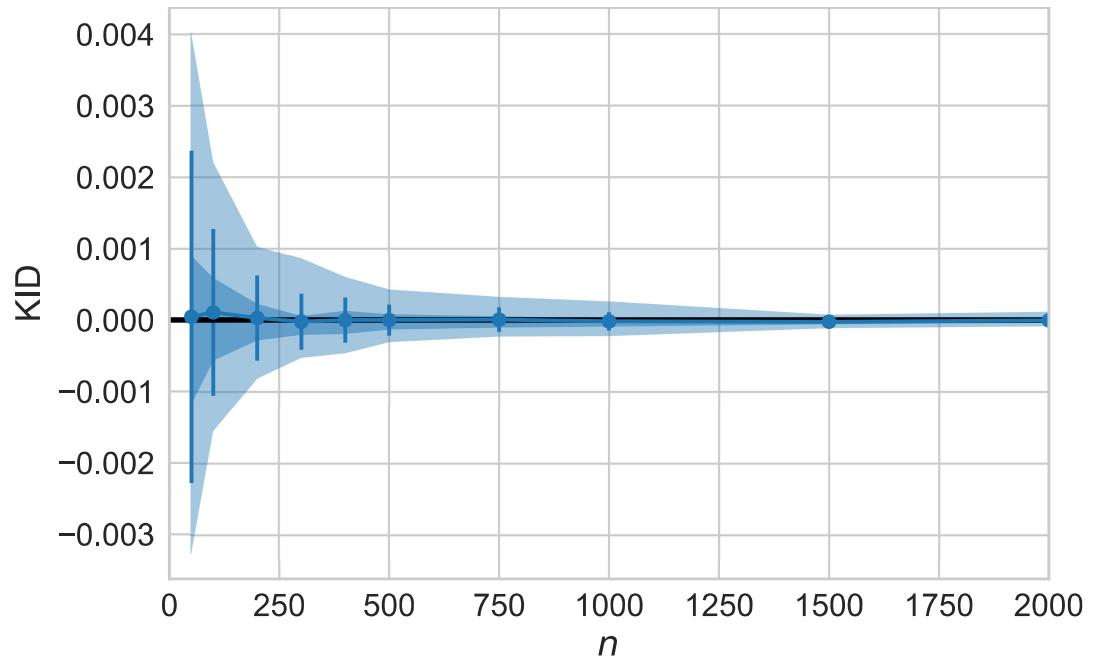
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1 \right)^3.$$

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



...“but isn't KID is computationally costly?”

“Block” KID implementation is cheaper than FID: see paper
(or use Tensorflow implementation)!

The kernel inception distance (KID)

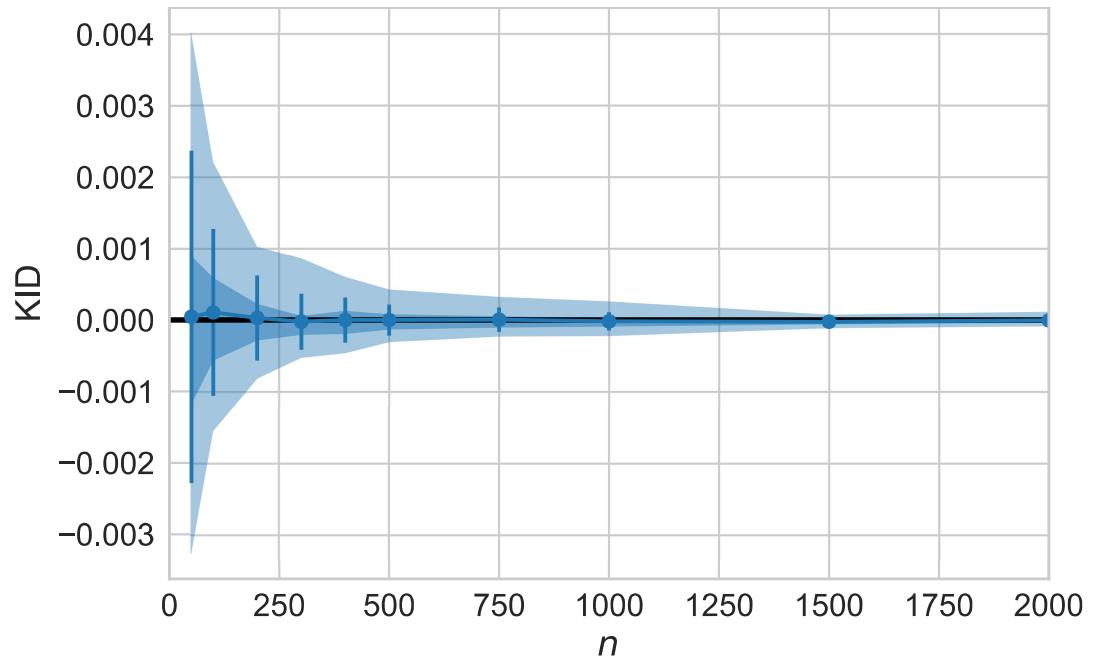
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1 \right)^3.$$

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



Also used for automatic learning rate adjustment: if $KID(\hat{P}_{t+1}, Q)$ not significantly better than $KID(\hat{P}_t, Q)$ then reduce learning rate.

[Bounliphone et al. ICLR 2016]

Benchmarks for comparison (all from ICLR 2018)

SPECTRAL NORMALIZATION FOR GENERATIVE ADVERSARIAL NETWORKS

Takeru Miyato¹, Toshiaki Kataoka¹, Masanori Koyama², Yuichi Yoshida³

{miyato, kataoka}@preferred.jp
koyama.masanori@gmail.com

yoshi@rii.ac.jp

Preferred Networks, Inc.¹Ritsumeikan University²National Institute of Informatics³

We
combine
with scaled
MMD

DEMYSTIFYING MMD GANS

Mikolaj Binkowski*

Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Our ICLR
2018
paper

Dougal J. Sutherland*, Michael Arbel & Arthur Gretton

Gatsby Computational Neuroscience Unit
University College London
dougal.sutherland,michael.n.arbel,arthur.gretton@gmail.com

SOBOLEV GAN

Youssef Mroueh[†], Chun-Liang Li^{0,*}, Tom Sercombe^{1,*}, Anant Raj^{0,*} & Yu Cheng¹

† IBM Research AI

◦ Carnegie Mellon University

◊ Max Planck Institute for Intelligent Systems

* denotes Equal Contribution

{mroueh, chengyu}@us.ibm.com, chunliang@cs.cmu.edu,
tom.sercombe@ibm.com, anant.raj@tuebingen.mpg.de

BOUNDARY-SEEKING GENERATIVE ADVERSARIAL NETWORKS

R Devon Hjelm*

MILA, University of Montréal, IVADO
erroneus@gmail.com

Athul Paul Jacob*

MILA, MSR, University of Waterloo
ap.jacob@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

Results: unconditional imagenet 64×64

KID scores:

- BGAN:
47
- SN-GAN:
44
- SMMD GAN:
35

ILSVRC2012 (ImageNet)
dataset, 1 281 167 images,
resized to 64×64 . 1000
classes.

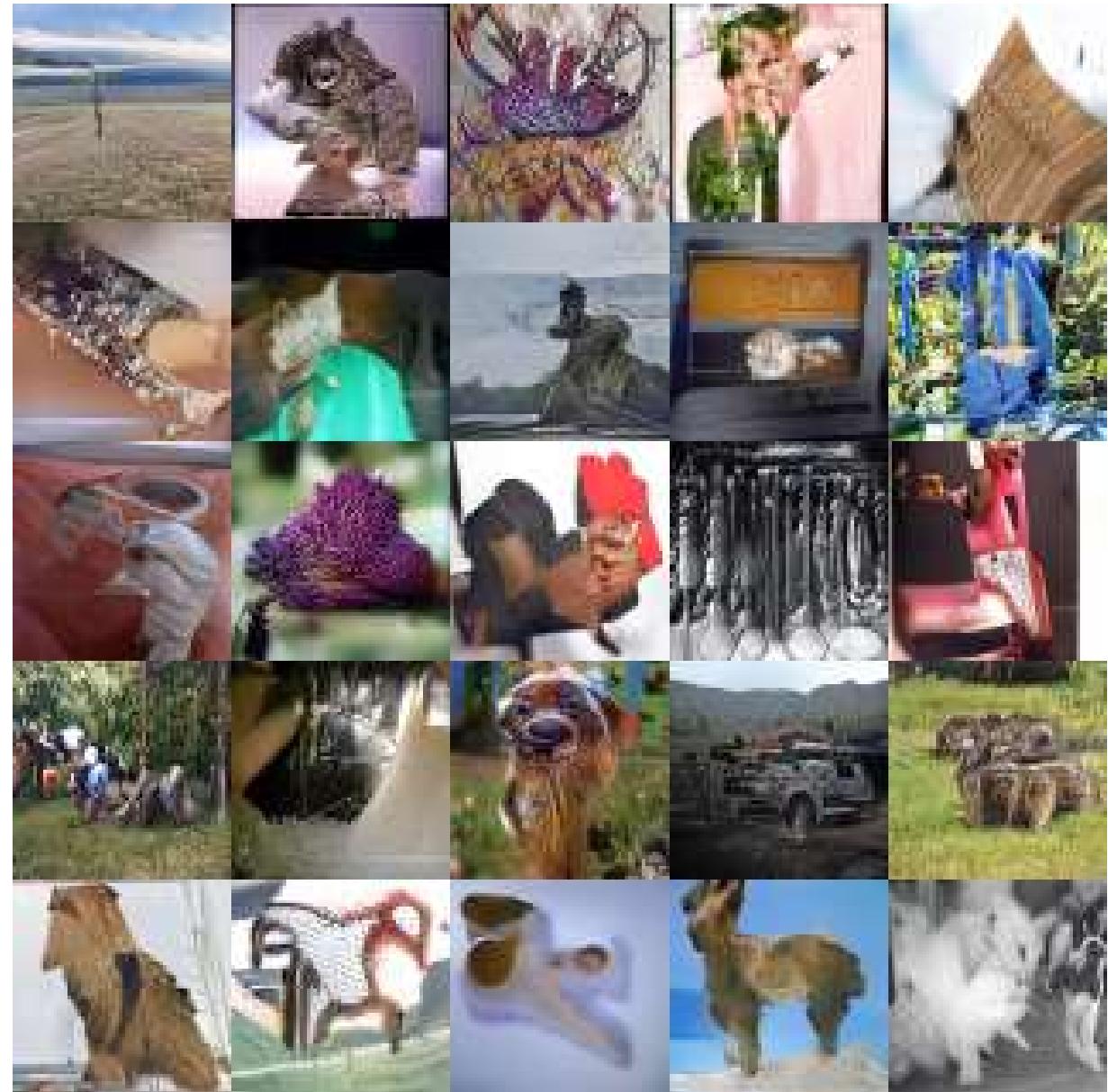


Results: unconditional imagenet 64×64

KID scores:

- BGAN:
47
- SN-GAN:
44
- SMMD GAN:
35

ILSVRC2012 (ImageNet)
dataset, 1 281 167 images,
resized to 64×64 . 1000
classes.



Results: unconditional imagenet 64×64

KID scores:

- BGAN:
47
- SN-GAN:
44
- SMMD GAN:
35

ILSVRC2012 (ImageNet)
dataset, 1 281 167 images,
resized to 64×64 . 1000
classes.



Summary

- GAN critics rely on two sources of regularisation
 - Regularisation by incomplete training
 - Data-dependent gradient regulariser
- Some advantages of hybrid kernel/neural features:
 - MMD loss still a valid critic when features not optimal (unlike WGAN-GP)
 - Kernel features do some of the “work”, so simpler h_ψ features possible.

“Demystifying MMD GANs,” including KID score, ICLR 2018:

<https://github.com/mbinkowski/MMD-GAN>

Gradient regularised MMD, NeurIPS 2018:

<https://github.com/MichaelArbel/Scaled-MMD-GAN>

Post-credit scene: MMD flow

From NeurIPS 2019:

Maximum Mean Discrepancy Gradient Flow

Michael Arbel

Gatsby Computational Neuroscience Unit
University College London
michael.n.arbel@gmail.com

Anna Korba

Gatsby Computational Neuroscience Unit
University College London
a.korba@ucl.ac.uk

Adil Salim

Visual Computing Center
KAUST
adil.salim@kaust.edu.sa

Arthur Gretton

Gatsby Computational Neuroscience Unit
University College London
arthur.gretton@gmail.com

Questions?

