Better gradient regularisation for MMD GANs

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Theoretical Foundations and Applications of Deep Generative Models, ICML 2018

Comparing two samples

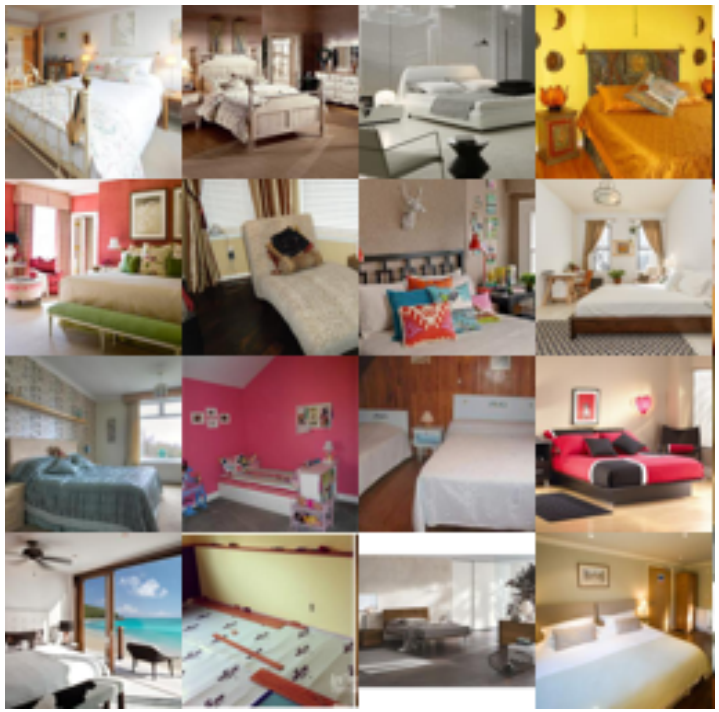
- **Given:** Samples from unknown distributions P and Q.
- **Goal:** do P and Q differ?

$\sim P$

$\sim Q$

Task: training 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 MMD to train a GAN

Binkowski, Sutherland, Arbel, G., [ICLR 2018];
Arbel, Sutherland, Binkowski, G., [arXiv May 2018]

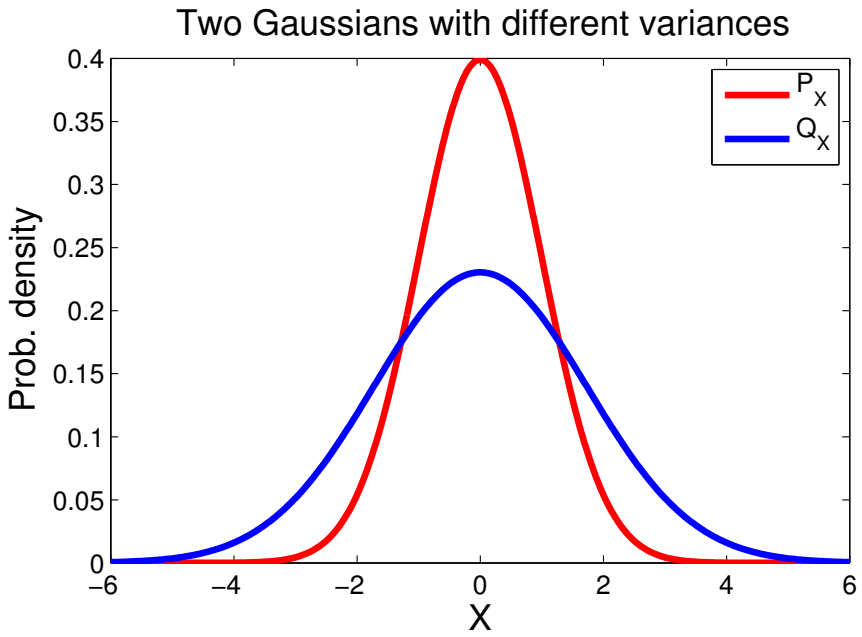
Outline

- Maximum Mean Discrepancy (MMD)...
  - ...as a difference in feature means
  - ...as an integral probability metric (not just a technicality!)
- Training GANs using MMD
  - Gradient penalty for training critic features
  - Comparison with WGAN-GP penalty
  - Evaluating GAN performance. Problems with Inception and FID.
  - Gradient bias for MMD GAN? For WGAN-GP? (not covered)

Maximum Mean Discrepancy

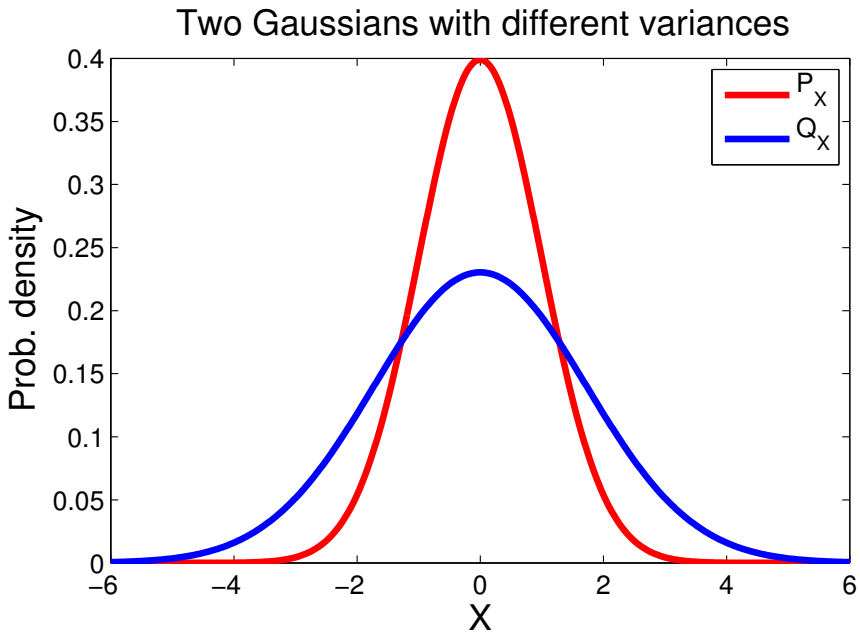
Feature mean difference

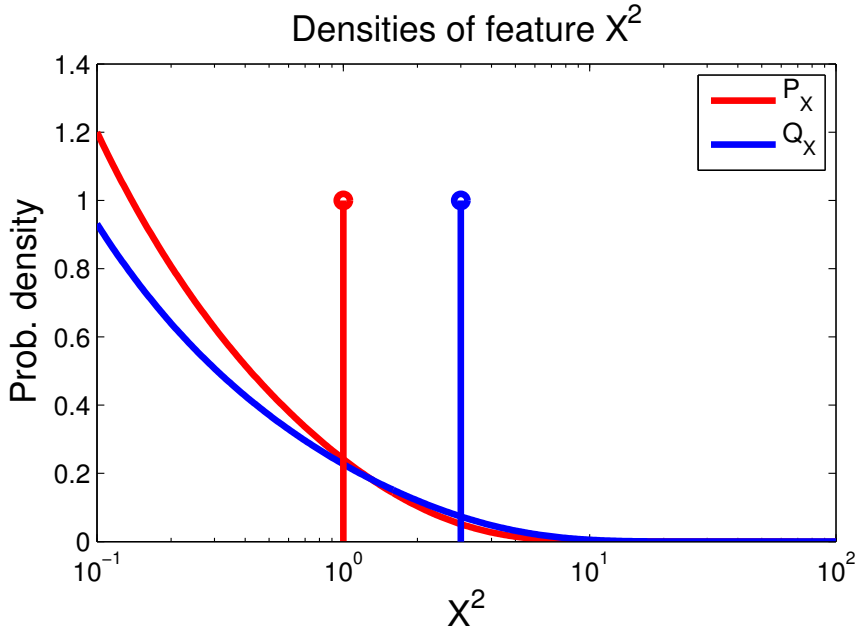
- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

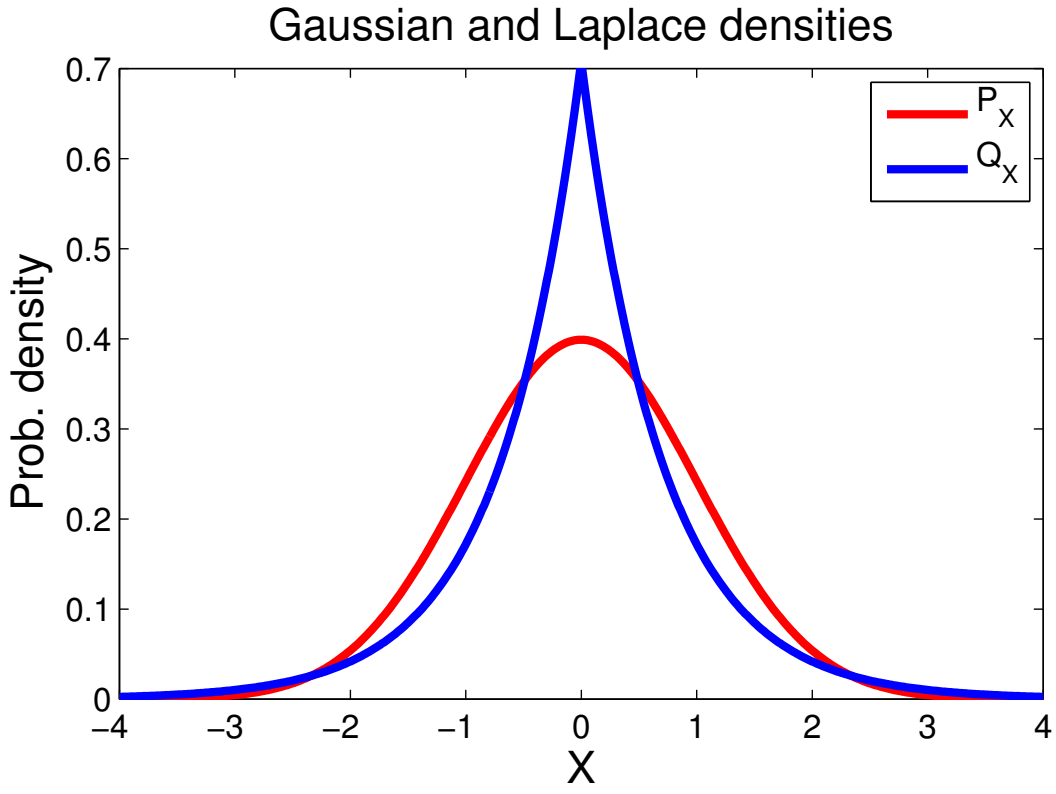
- Two Gaussians with same means, different variance
- Idea: look at difference in **means of features** of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$





Feature mean difference

- Gaussian and Laplace distributions
- Same mean and same variance
- Difference in means using **higher order features**...RKHS



Infinitely many features using kernels

Kernels: dot products of features

Feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in \ell_2$$

For positive definite k,

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

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

Infinitely many features using kernels

Kernels: dot products of features

Feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in \ell_2$$

For positive definite k,

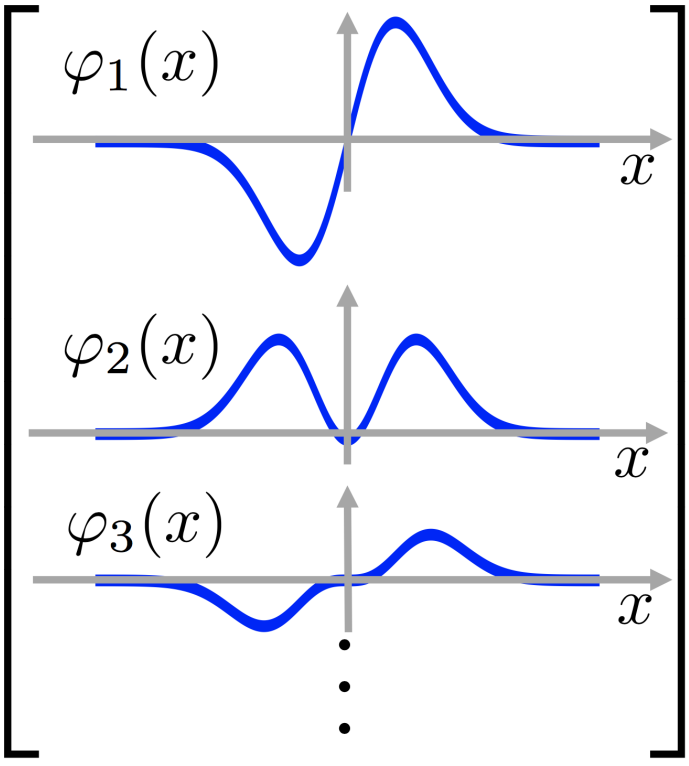
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

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

Exponentiated quadratic kernel

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

$$\varphi(x) = \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \\ \vdots \end{bmatrix}$$



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

Infinitely many features of distributions

Given P a Borel **probability measure** on $\mathcal{X}$, define feature map of probability P,

$$\mu_P = [\ldots \mathbf{E}_P [\varphi_i(X)] \ldots]$$

For positive definite $k(x, x')$,

$$\langle \mu_P, \mu_Q \rangle_{\mathcal{F}} = \mathbf{E}_{P,Q} k(x, y)$$

for $x \sim P$ and $y \sim Q$.

Fine print: feature map $\varphi(x)$ must be Bochner integrable for all probability measures considered. Always true if kernel bounded.

Infinitely many features of distributions

Given P a Borel **probability measure** on $\mathcal{X}$, define feature map of probability P,

$$\mu_P = [\ldots \mathbf{E}_P [\varphi_i(X)] \ldots]$$

For positive definite $k(x, x')$,

$$\langle \mu_P, \mu_Q \rangle_{\mathcal{F}} = \mathbf{E}_{P,Q} k(x, y)$$

for $x \sim P$ and $y \sim Q$.

Fine print: feature map $\varphi(x)$ must be Bochner integrable for all probability measures considered. Always true if kernel bounded.

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|_{\mathcal{F}}^2 \\
&= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} \\
&= \underbrace{\mathbf{E}_P k(X, X')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(Y, Y')}_{\text{(a)}} - 2 \underbrace{\mathbf{E}_{P,Q} k(X, Y)}_{\text{(b)}}
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|_{\mathcal{F}}^2 \\
&= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} \\
&= \underbrace{\mathbf{E}_P k(X, X')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(Y, Y')}_{\text{(a)}} - 2 \underbrace{\mathbf{E}_{P,Q} k(X, Y)}_{\text{(b)}}
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between **feature means**:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|_{\mathcal{F}}^2 \\
&= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} \\
&= \underbrace{\mathbf{E}_P k(X, X')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(Y, Y')}_{\text{(a)}} - 2\underbrace{\mathbf{E}_{P,Q} k(X, Y)}_{\text{(b)}}
\end{aligned}
$$

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

Illustration of MMD

- Dogs ($= P$) and fish ($= Q$) example revisited
- Each entry is one of $k(\text{dog}_i, \text{dog}_j)$, $k(\text{dog}_i, \text{fish}_j)$, or $k(\text{fish}_i, \text{fish}_j)$

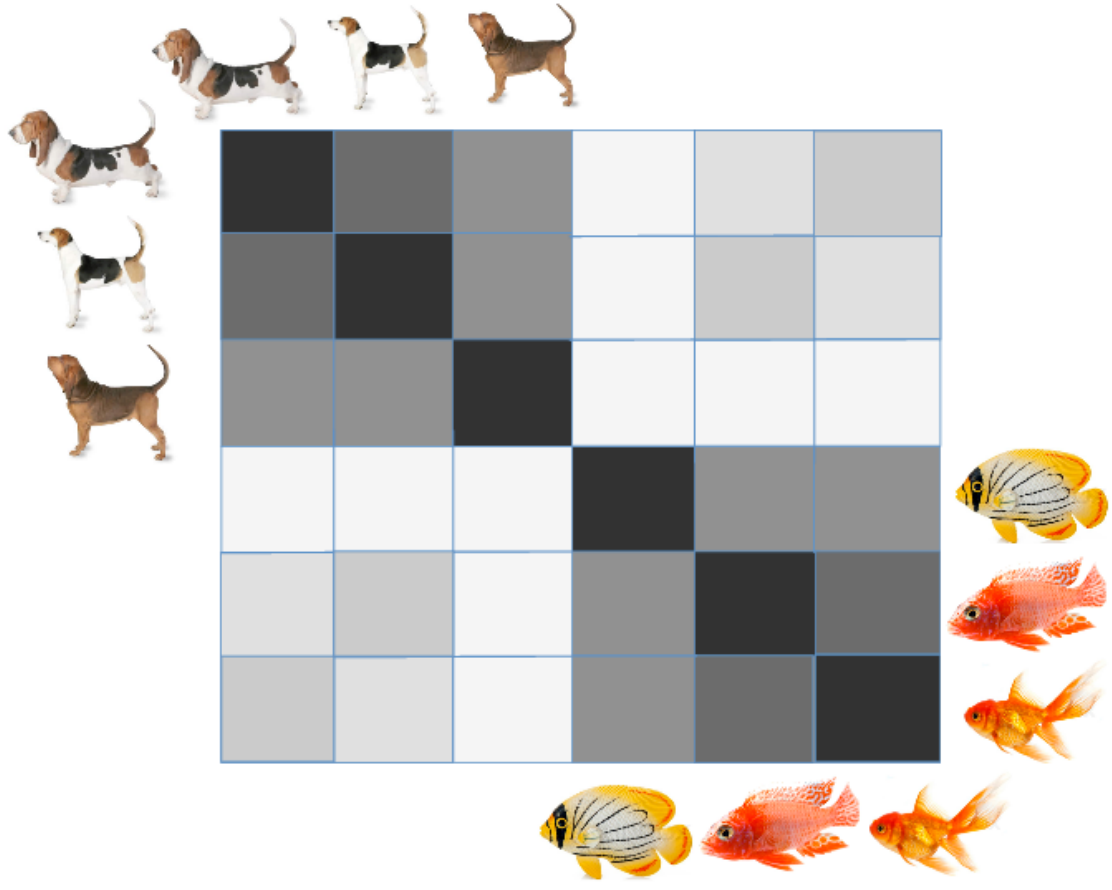
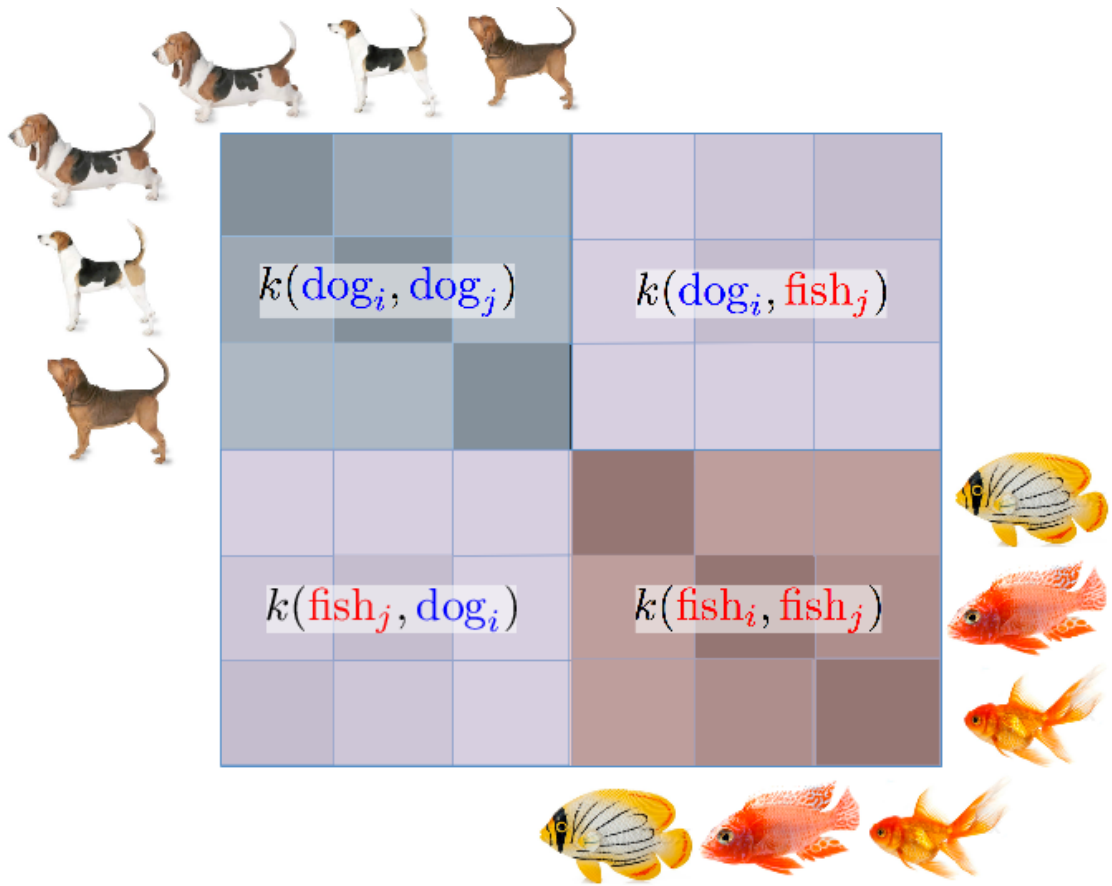


Illustration of MMD

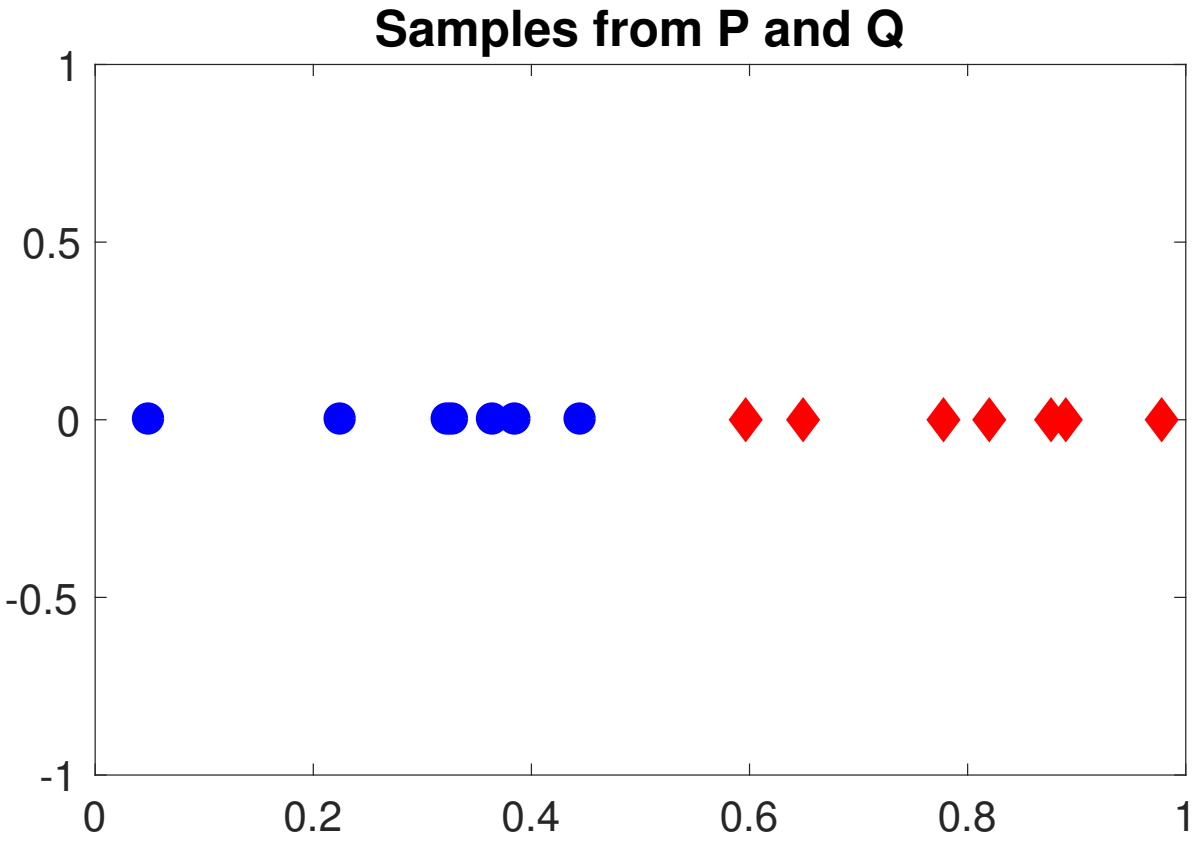
The maximum mean discrepancy:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\text{dog}_i, \text{dog}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\text{fish}_i, \text{fish}_j) - \frac{2}{n^2} \sum_{i,j} k(\text{dog}_i, \text{fish}_j)$$



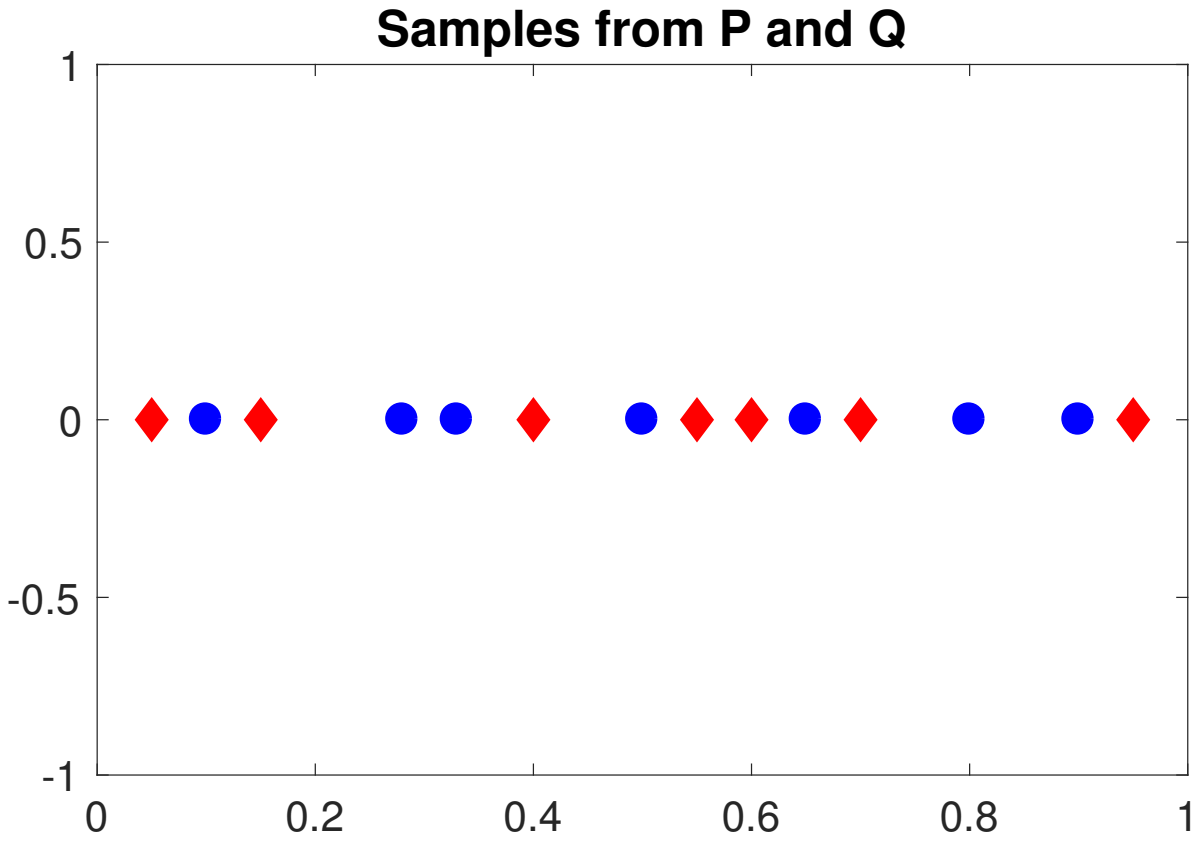
MMD as an integral probability metric

Are P and Q different?



MMD as an integral probability metric

Are P and Q different?

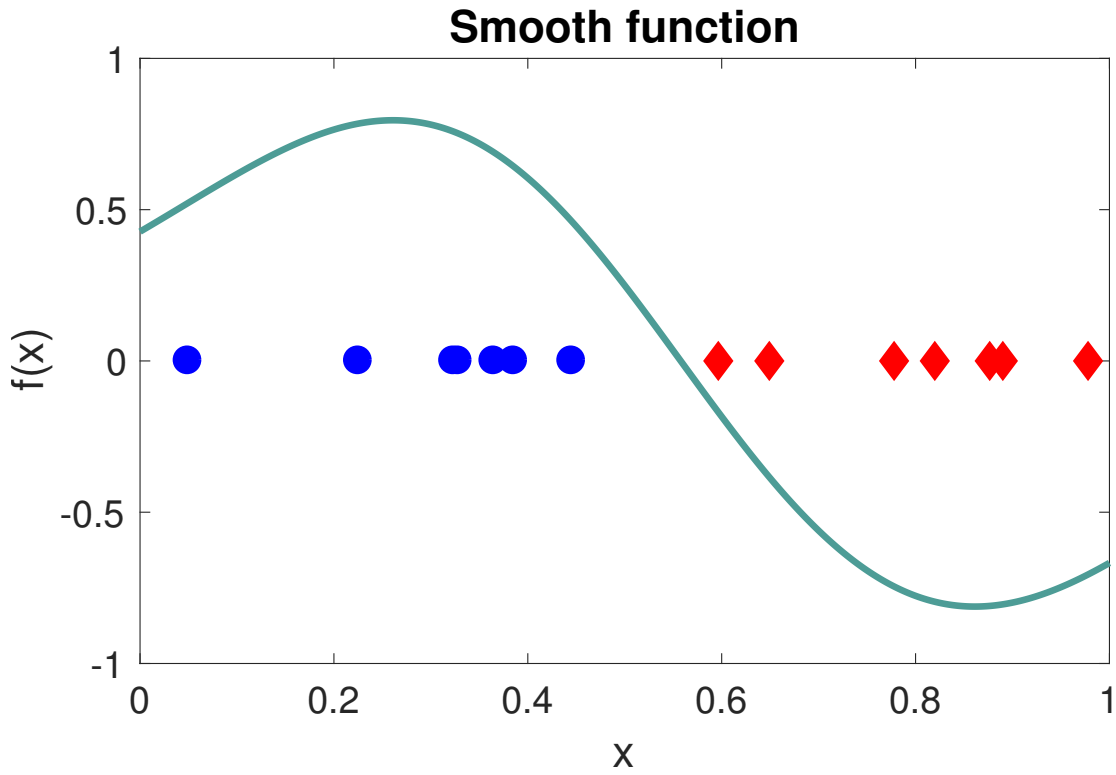


MMD as an integral probability metric

Integral probability metric:

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

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

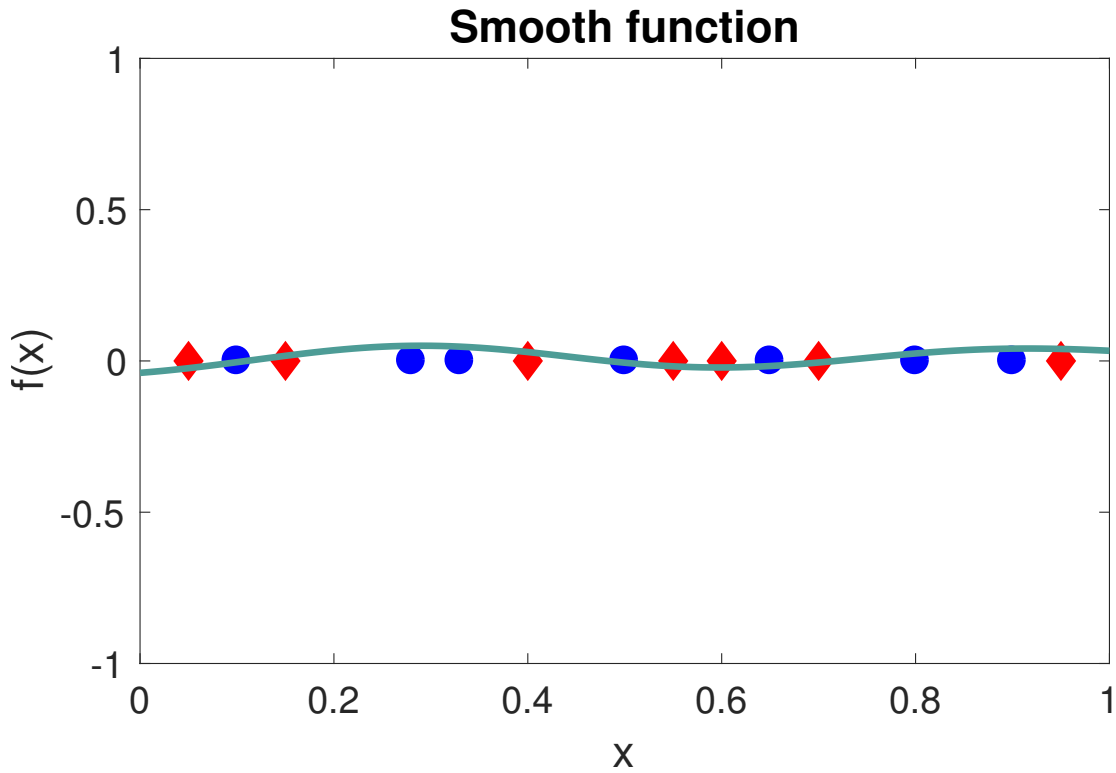


MMD as an integral probability metric

Integral probability metric:

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

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

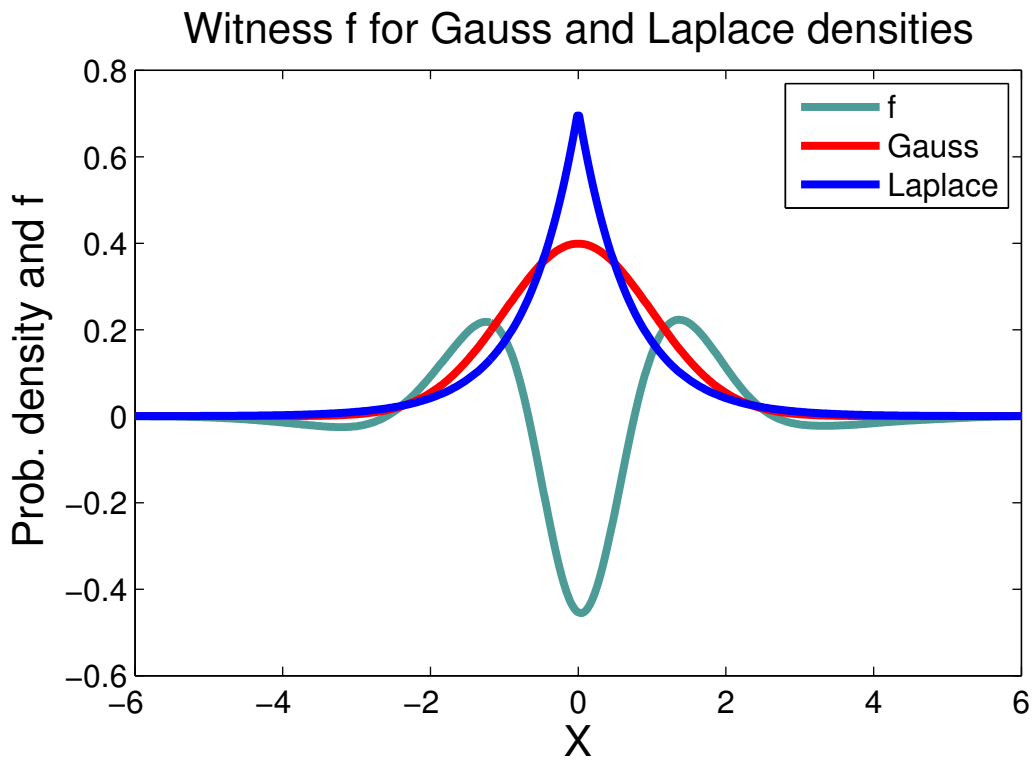
(F = unit ball in RKHS $\mathcal{F}$)

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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



MMD as an integral probability metric

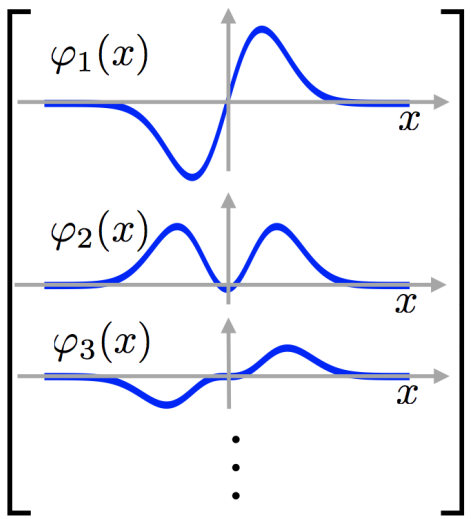
Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

(F = 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)

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

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

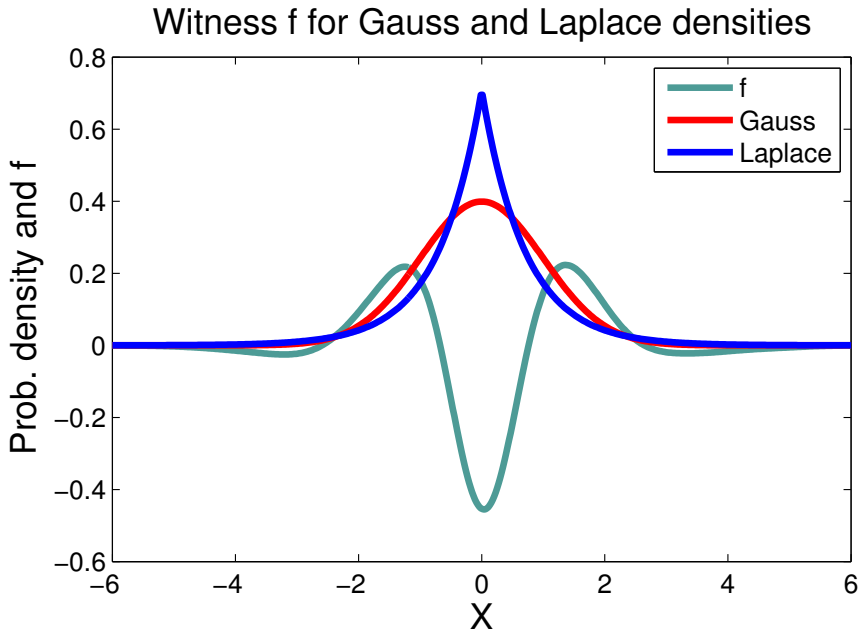
Other choices for witness function class:

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

Integral prob. metric vs feature difference

The MMD:

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



Integral prob. metric vs feature difference

The MMD:

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

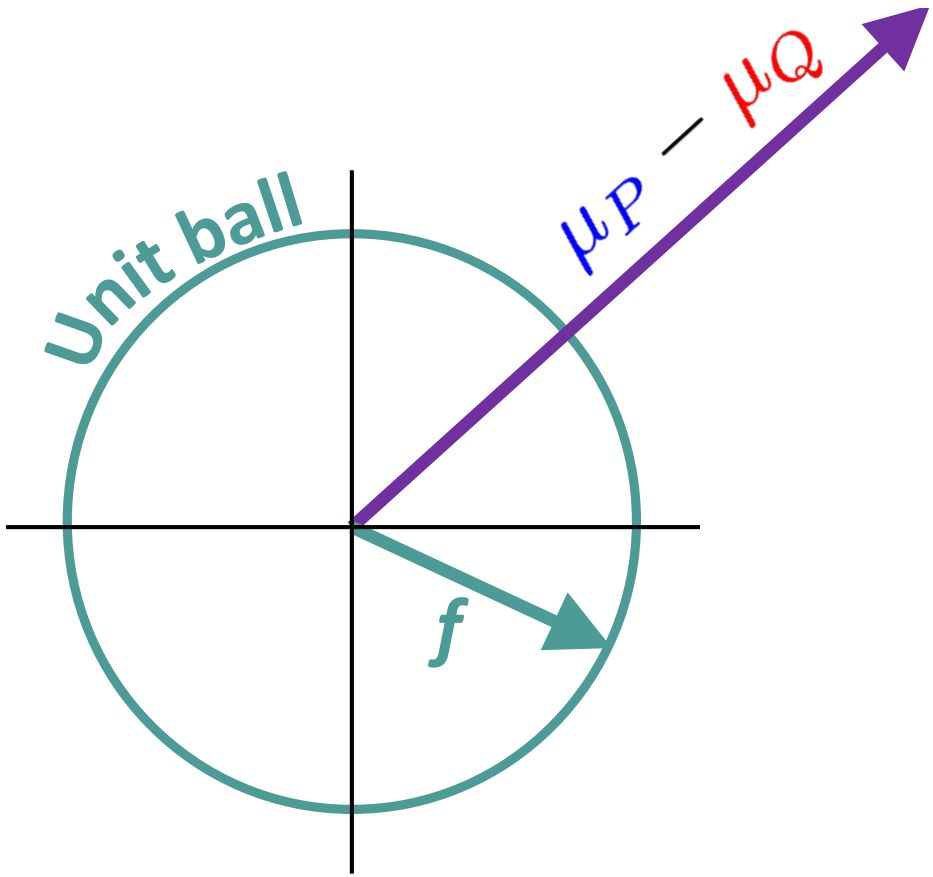
use

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

Integral prob. metric vs feature difference

The MMD:

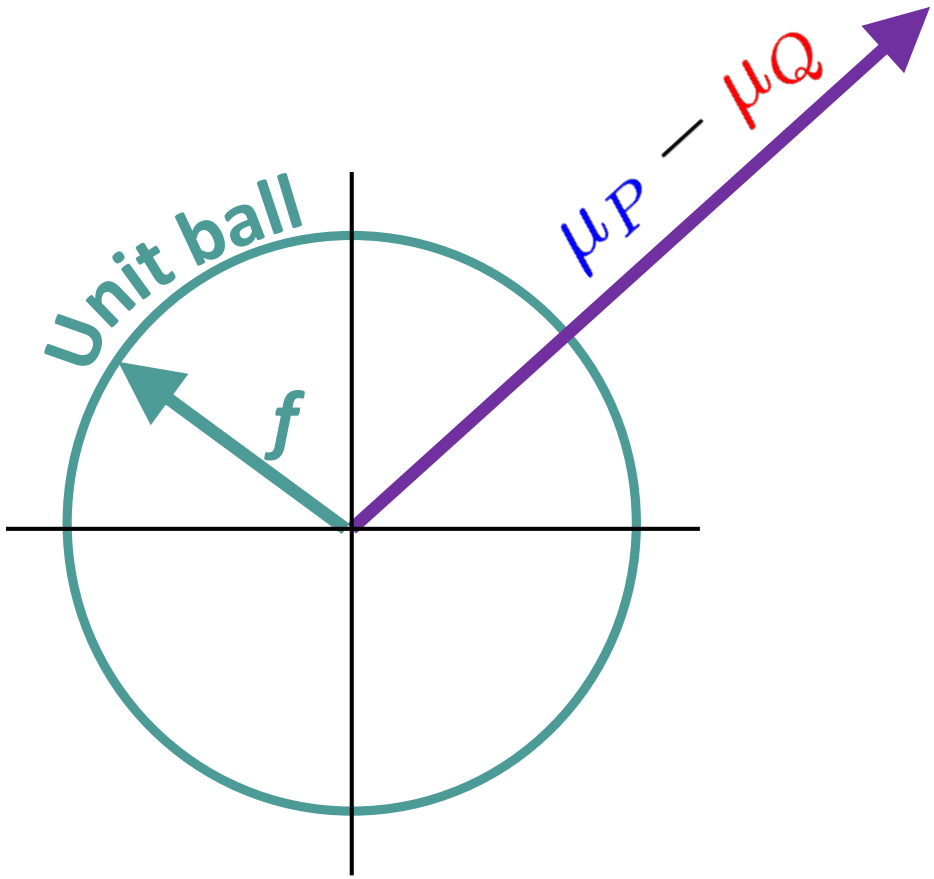
$$MMD(P, Q; F)$$
$$= \sup_{f \in F} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$
$$= \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature difference

The MMD:

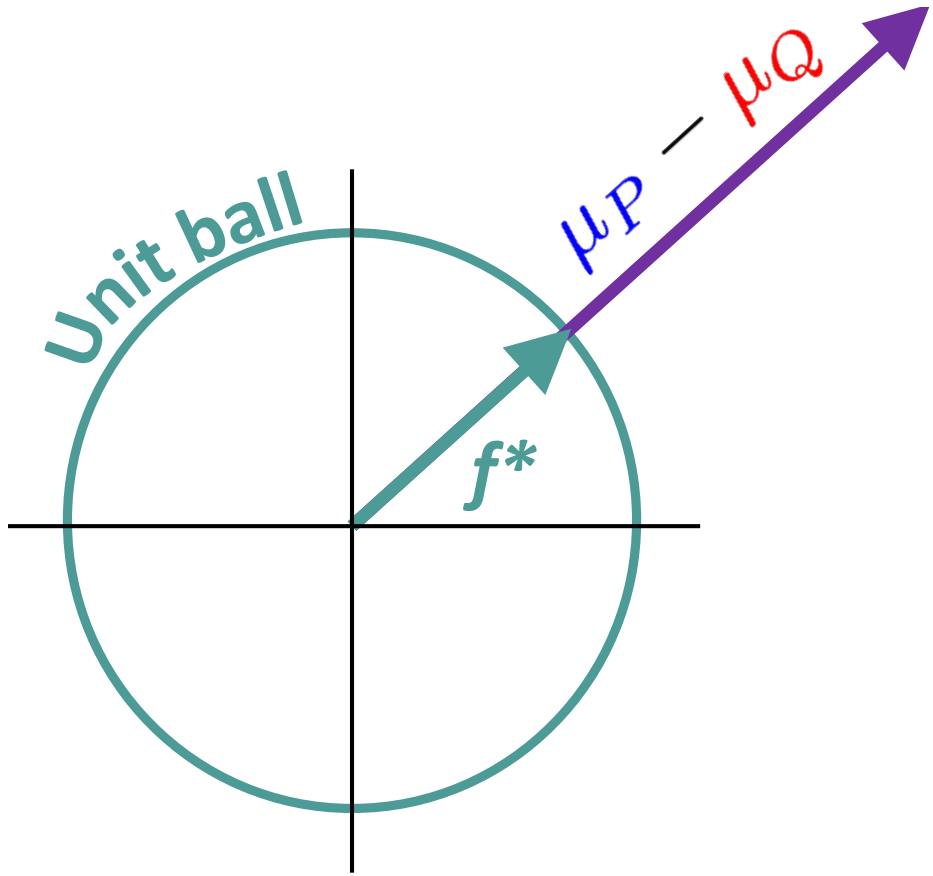
$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{f \in F} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
&= \sup_{f \in F} \left\langle f, \mu_P - \mu_Q \right\rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature difference

The MMD:

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



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

Integral prob. metric vs feature difference

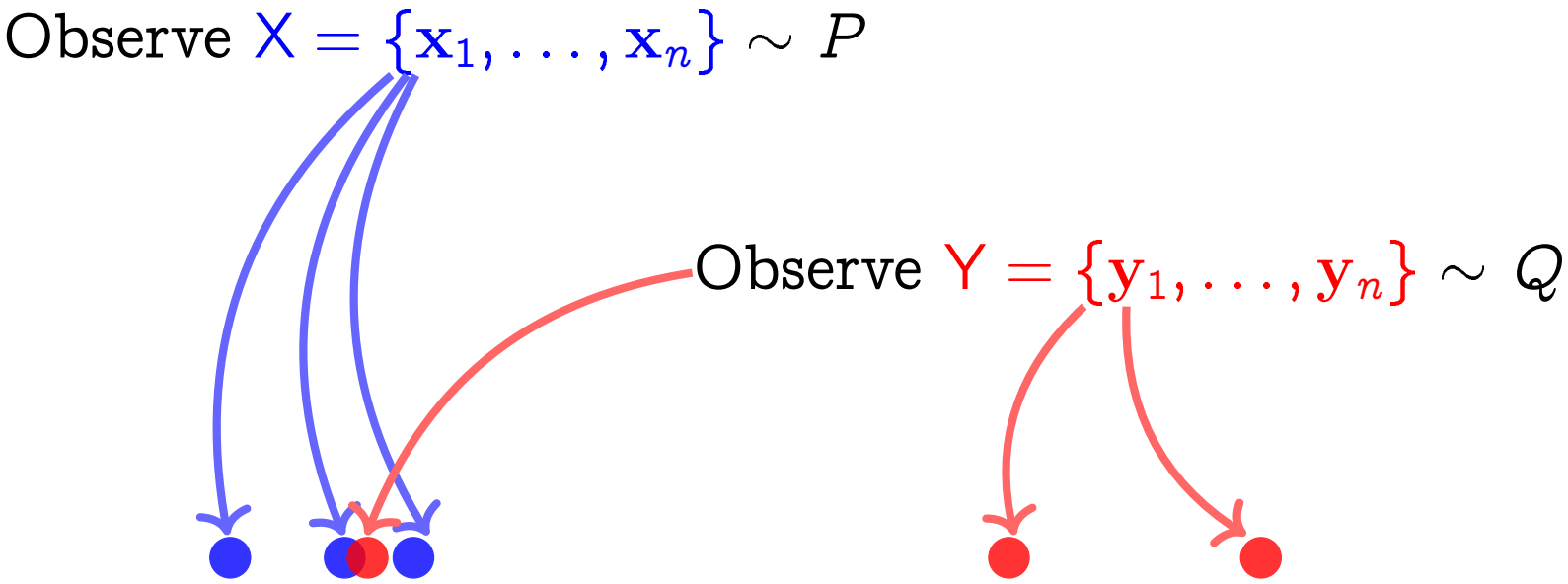
The MMD:

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

Function view and feature view equivalent

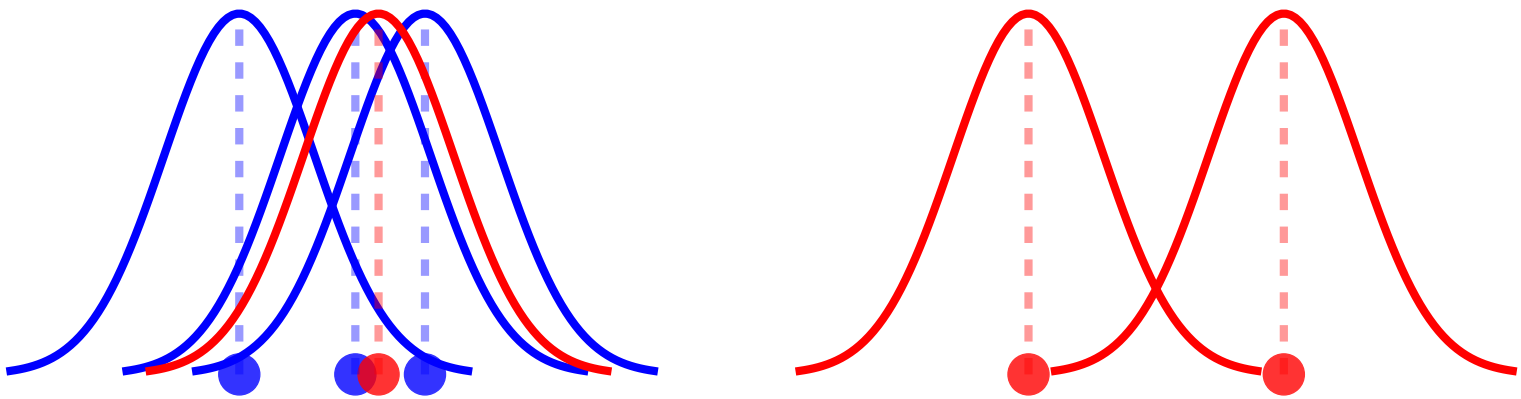
Construction of MMD witness

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



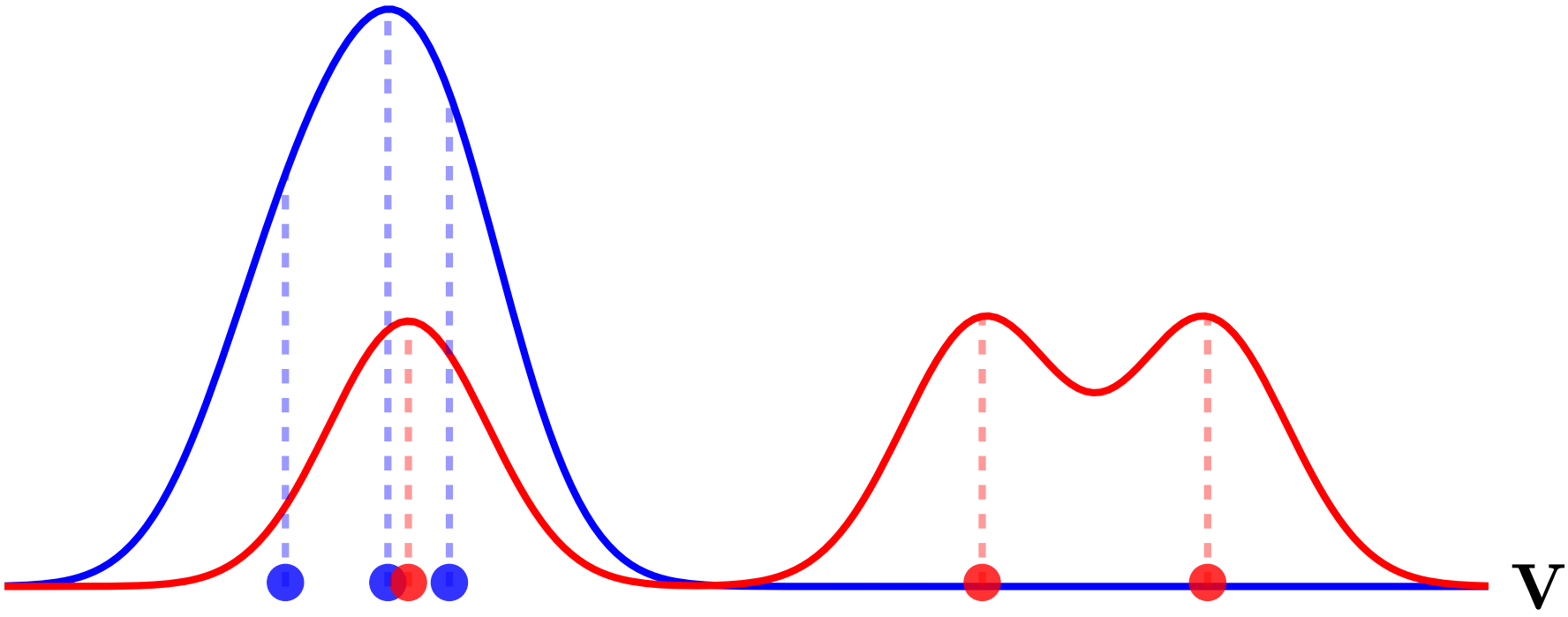
Construction of MMD witness

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



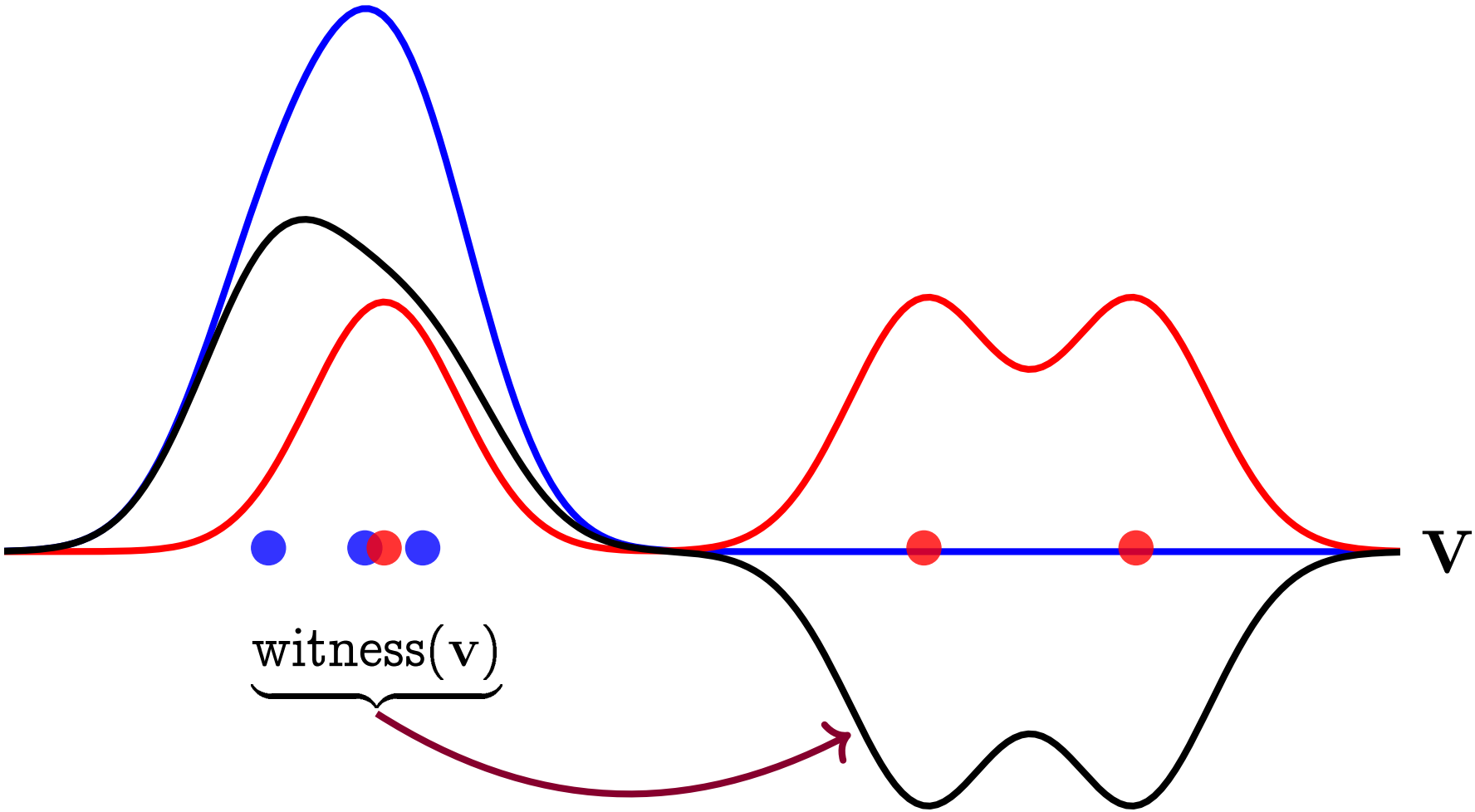
Construction of MMD witness

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



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

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

Derivation of empirical witness function

Recall the witness function expression

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

The empirical feature mean for P

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

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

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

The empirical witness function at v

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

Derivation of empirical witness function

Recall the witness function expression

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

The empirical feature mean for P

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

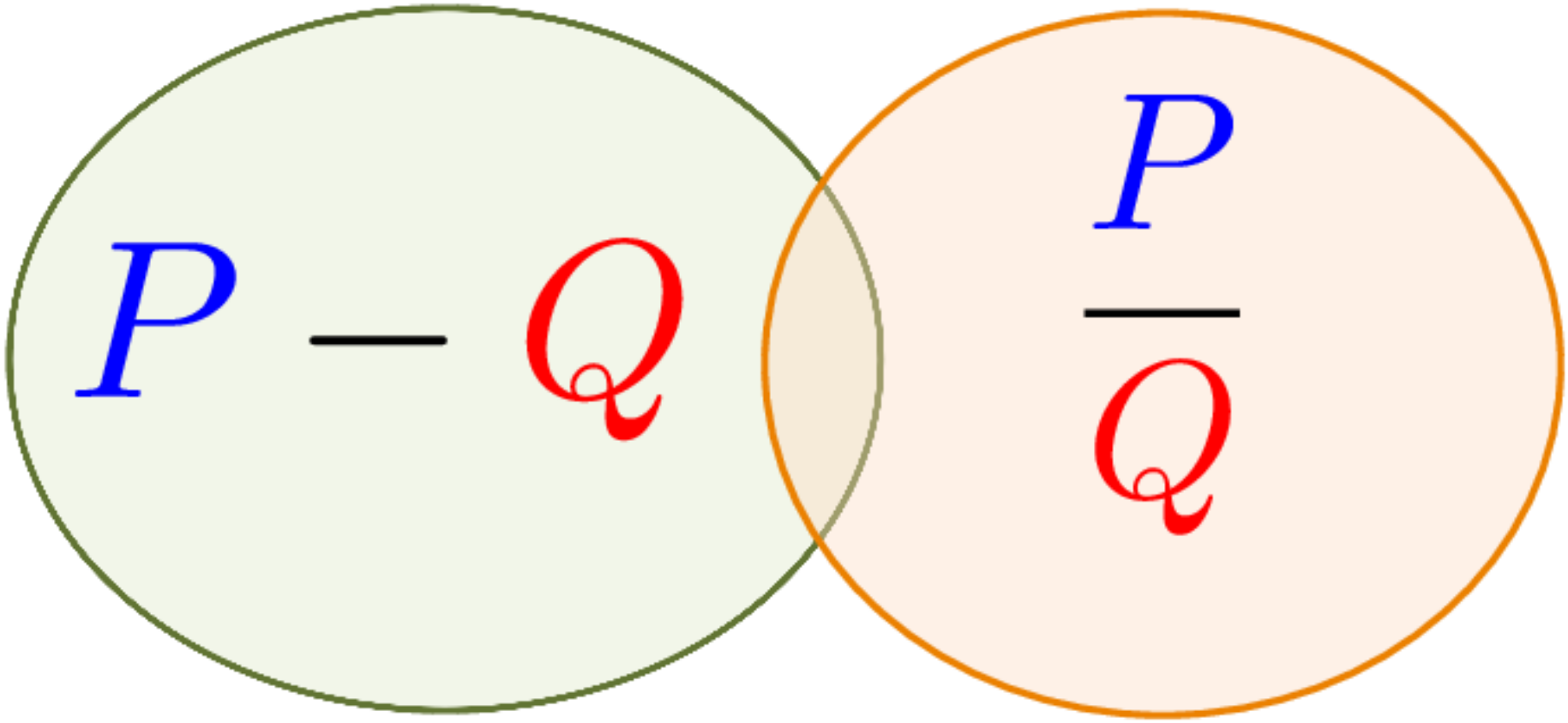
The empirical witness function at v

$$\begin{aligned} f^*(v) &= \langle f^*, \varphi(v) \rangle_{\mathcal{F}} \\ &\propto \langle \widehat{\mu}_P - \widehat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \\ &= \frac{1}{n} \sum_{i=1}^{n} k(x_i, v) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, v) \end{aligned}$$

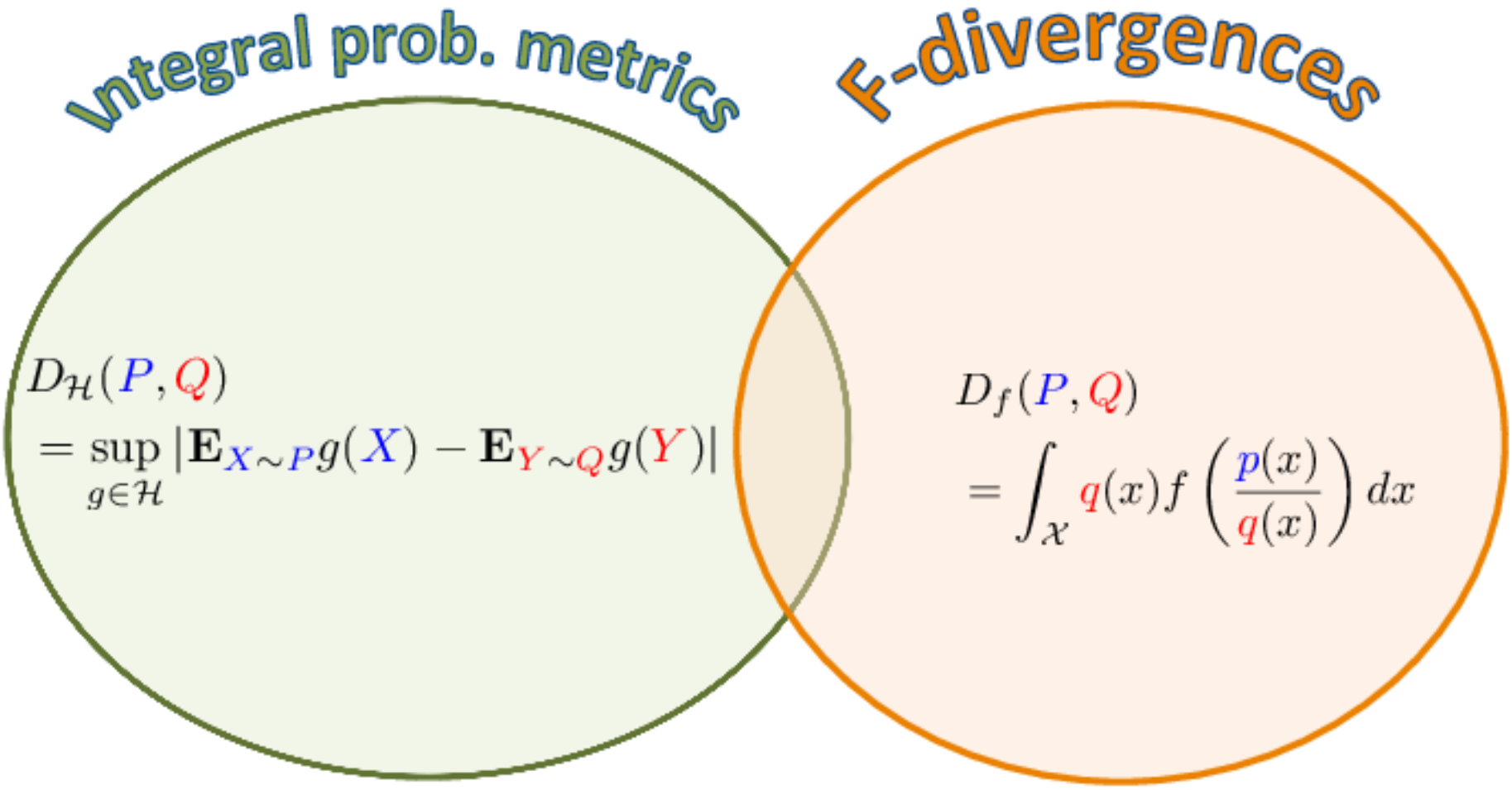
Don't need explicit feature coefficients $f^* := \begin{bmatrix} f_1^* & f_2^* & \dots \end{bmatrix}$

Interlude: divergence measures

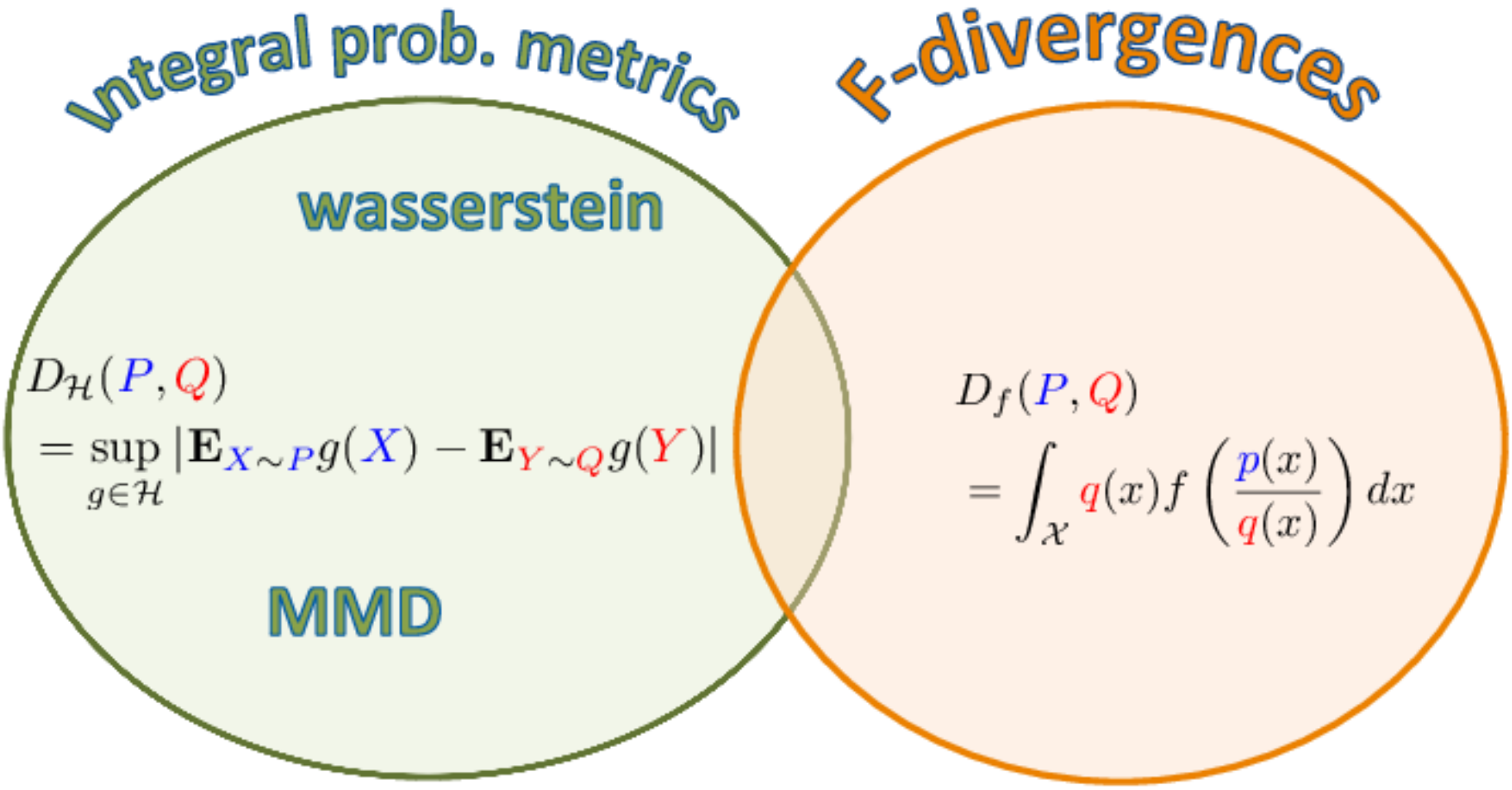
Divergences



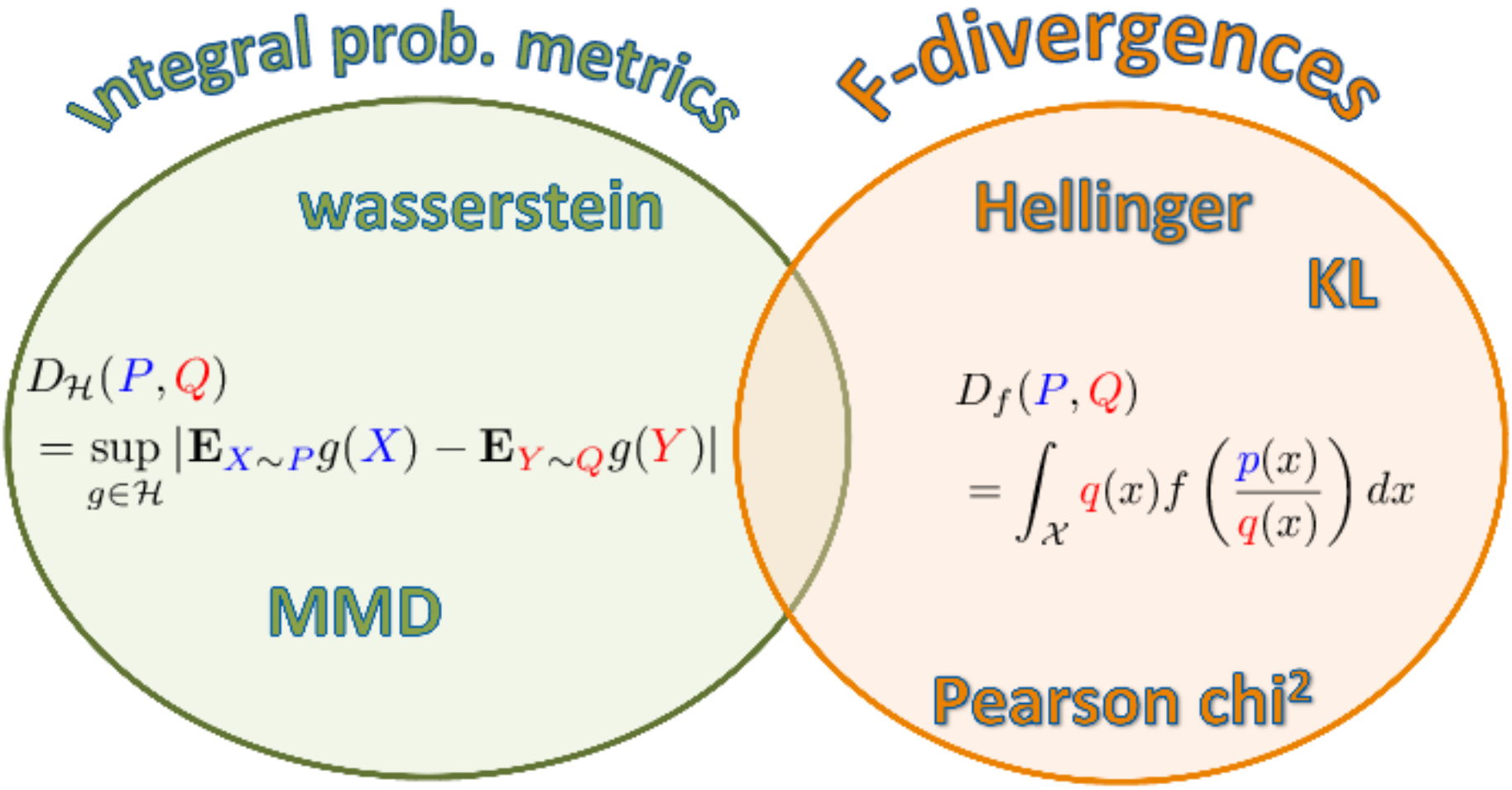
Divergences



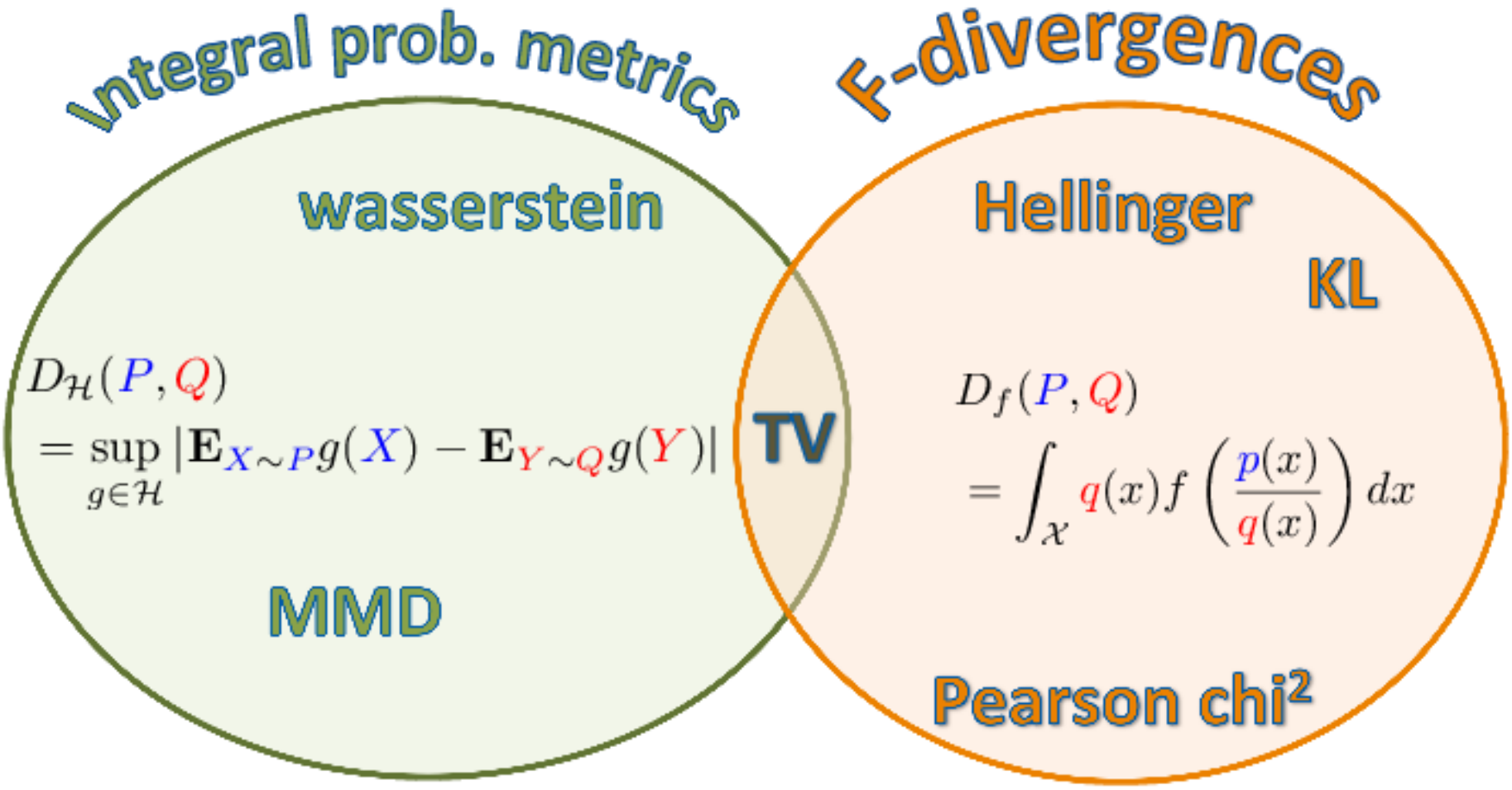
Divergences



Divergences



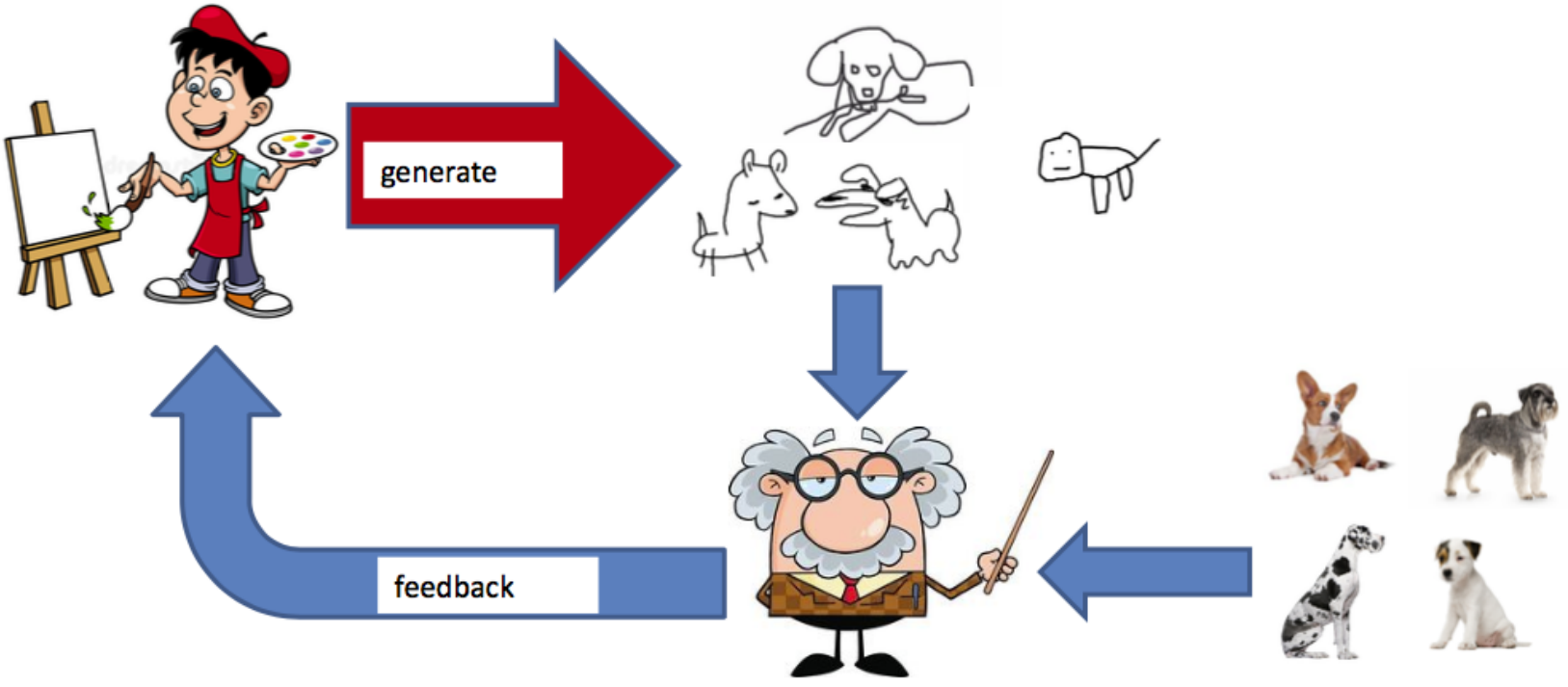
Divergences



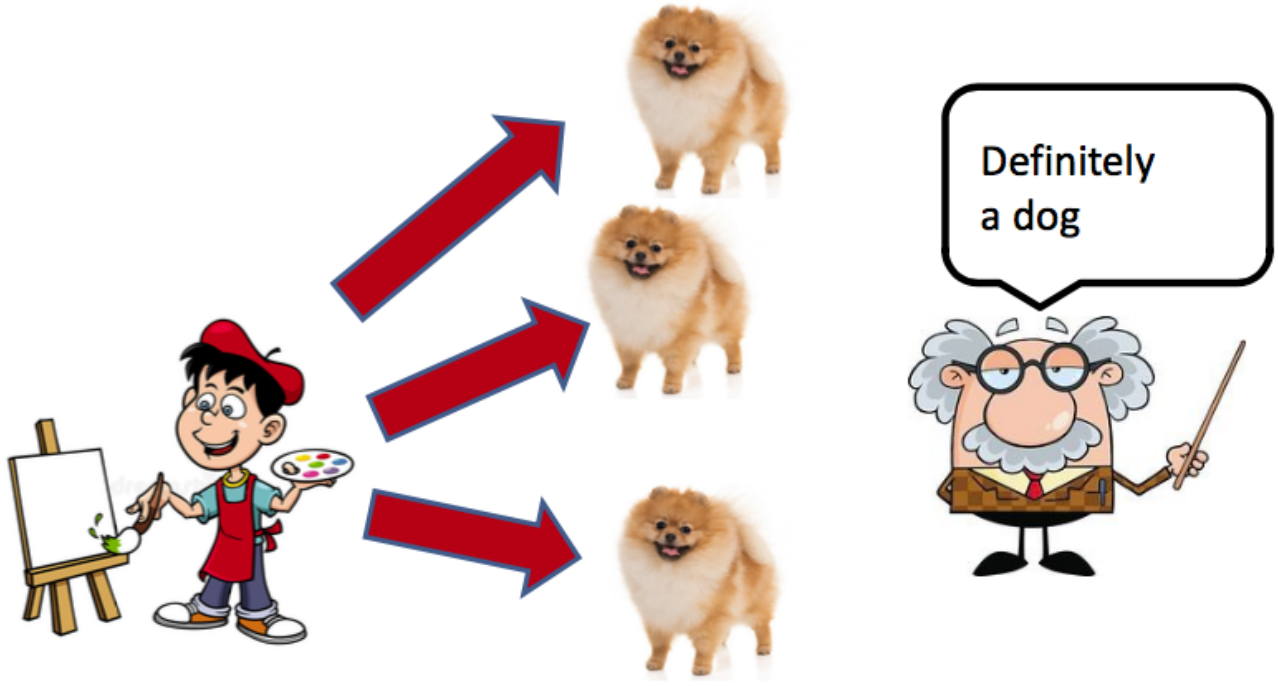
Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet [EJS 2012]

Training GANs with MMD

Notation: Generative Adversarial Network (GAN)?



Why is classification not enough?



Classification **not** enough!

Need to compare **sets**

(otherwise student can just produce the same dog over and over)

MMD for GAN critic

Can you use MMD as a critic to train GANs?

From ICML 2015:

Generative Moment Matching Networks

Yujia Li[1] YUJIALI@CS.TORONTO.EDU
Kevin Swersky[1] KSWERSKY@CS.TORONTO.EDU
Richard Zemel[1,2] ZEMEL@CS.TORONTO.EDU
[1]Department of Computer Science, University of Toronto, Toronto, ON, CANADA
[2]Canadian 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

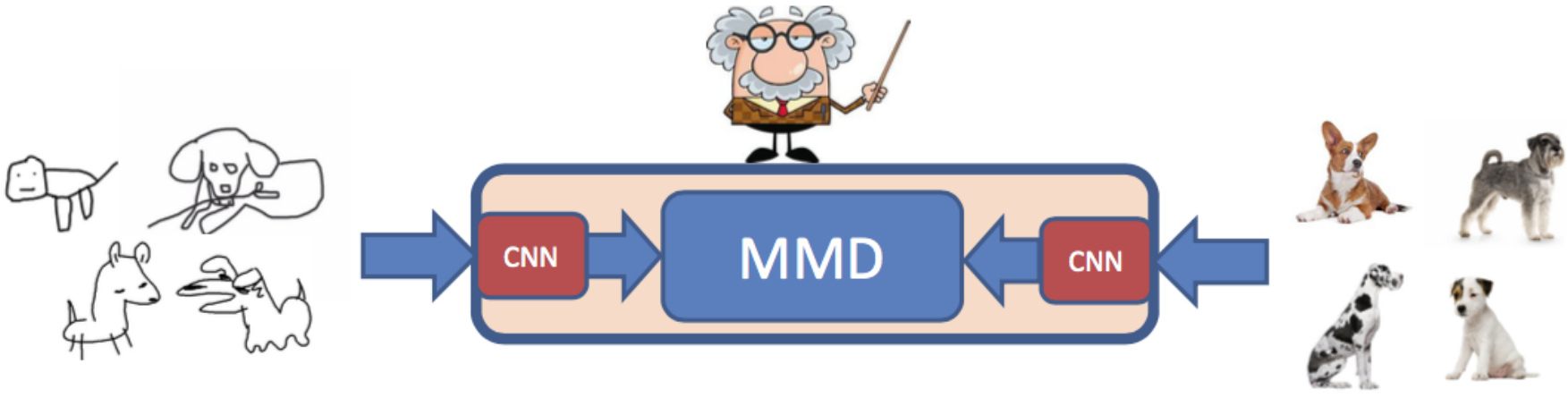
MMD for GAN critic

Can you use MMD as a critic to train GANs?

Need better image features.

How to improve the critic witness

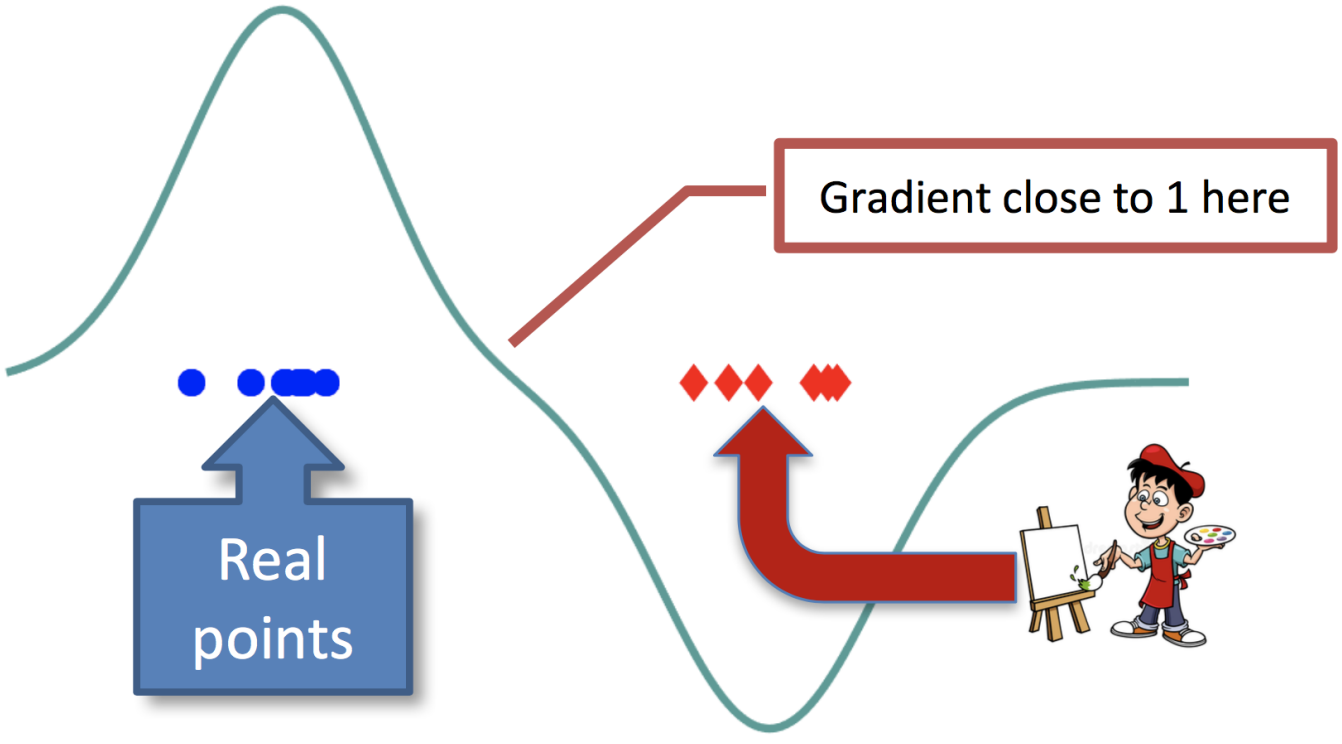
- Add convolutional features!
- The critic (teacher) also needs to be trained.
- How to regularise?



MMD GAN Li et al., [NIPS 2017]
Coulomb GAN Unterthiner et al., [ICLR 2018]

WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017].
WGAN-GP Gukrajani et al. [NIPS 2017]



WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017].
WGAN-GP Gukrajani et al. [NIPS 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 of h_ψ.

WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017].
WGAN-GP Gukrajani et al. [NIPS 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 of h_ψ.

WGAN-GP gradient penalty:

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

where

$$\widetilde{X} = \gamma x_i + (1-\gamma) G_\psi(z_j)$$

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

The (W)MMD

Train MMD critic features with the witness function gradient penalty

Binkowski, Sutherland, Arbel, G. [ICLR 2018], Bellemare et al. [2017] for energy distance:

$$\max_{\psi} MMD^2(h_\psi(X), h_\psi(G_\theta(Z))) + \lambda \mathbf{E}_{\widetilde{X}} \left(\left\| \nabla_{\widetilde{X}} f_\psi(\widetilde{X}) \right\| - 1 \right)^2$$

where

$$f_\psi(\cdot) = \frac{1}{m} \sum_{i=1}^{m} k(h_\psi(x_i), \cdot) - \frac{1}{n} \sum_{j=1}^{n} k(h_\psi(G_\theta(z_j)), \cdot)$$

New

$$\widetilde{X} = \gamma x_i + (1-\gamma) G_\psi(z_j)$$

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

Remark by Bottou et al. (2017): this modifies the function class. So critic is not an MMD in RKHS $\mathcal{F}$.

MMD for GAN critic: revisited

From ICLR 2018:

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
University College London
{dougal,michael.n.arbel,arthur.gretton}@gmail.com

MMD for GAN critic: revisited

Samples are better!

MMD for GAN critic: revisited

Samples are better!

Can we do better still?

Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

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

The Dirac-GAN

$$P = \delta_0 \qquad Q = \delta_\theta \qquad f_\psi(x) = \psi \cdot x$$

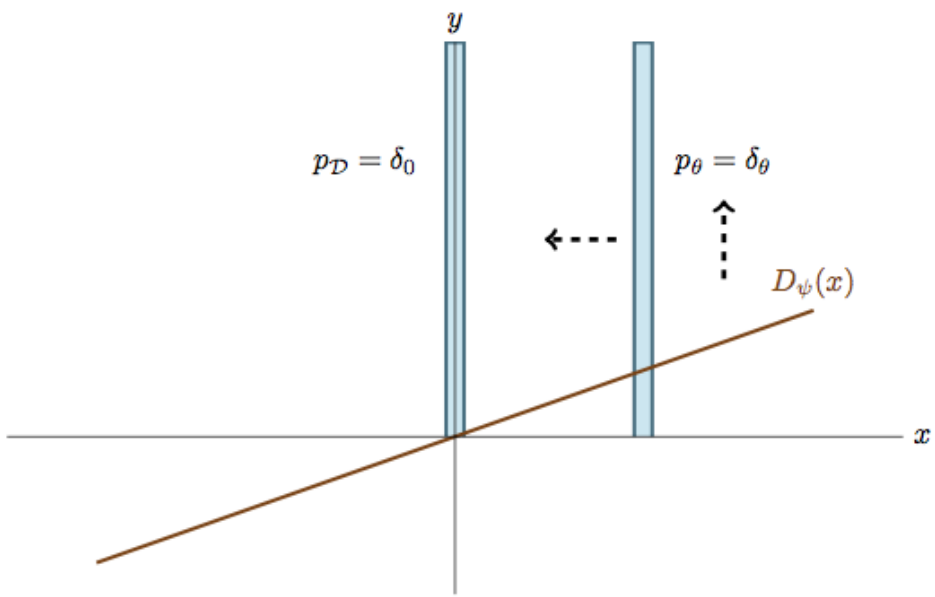


Figure from Mescheder et al. [ICML 2018]

Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

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

The Dirac-GAN

$$P = \delta_0 \qquad Q = \delta_\theta \qquad f_\psi(x) = \psi \cdot x$$

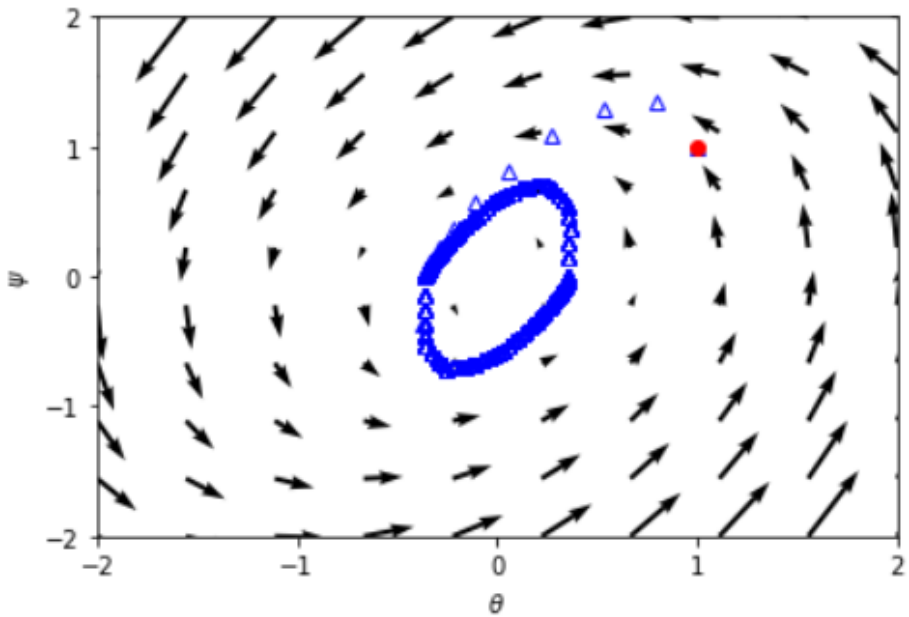
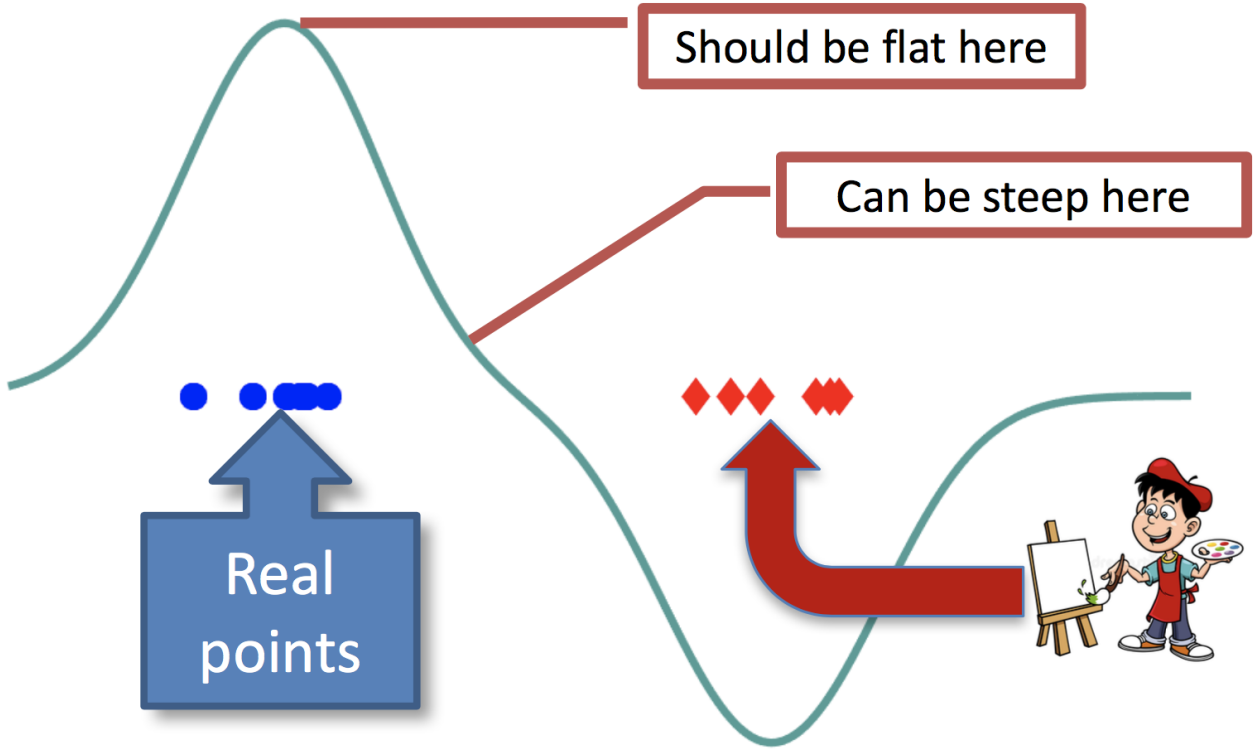


Figure from Mescheder et al. [ICML 2018]

A better gradient penalty

- New MMD GAN witness regulariser

  Arbel, Sutherland, Binkowski, G. [arxiv, May 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]



A better gradient penalty

- New MMD GAN witness regulariser

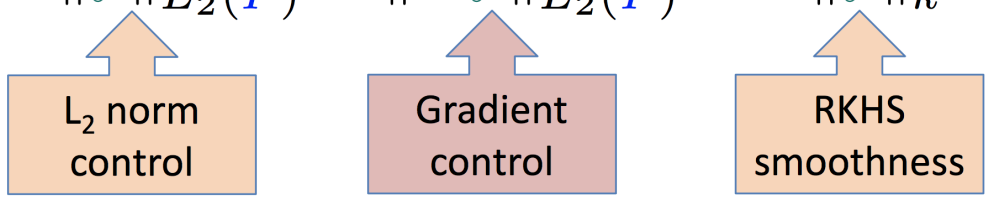
  Arbel, Sutherland, Binkowski, G. [arxiv, May 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

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

where

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



A better gradient penalty

- New MMD GAN witness regulariser

  Arbel, Sutherland, Binkowski, G. [arxiv, May 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

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

where

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

L_2 norm control

Gradient control

RKHS smoothness

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

A better gradient penalty

- New MMD GAN witness regulariser

  Arbel, Sutherland, Binkowski, G. [arxiv, May 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

$$SMMD = \sigma_{k,P,\lambda}\ MMD$$

where

$$\sigma_{k,P,\lambda} = \left(\lambda + \int k(x,x)\,dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x,y)\ dP(x) \right)^{-1/2}$$

Replace expensive constraint with cheap upper bound:

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

A better gradient penalty

- New MMD GAN witness regulariser

  Arbel, Sutherland, Binkowski, G. [arxiv, May 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

$$SMMD = \sigma_{k,P,\lambda}\ MMD$$

where

$$\sigma_{k,P,\lambda} = \left(\lambda + \int k(x,x)\,dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x,y)\ dP(x) \right)^{-1/2}$$

Replace expensive constraint with cheap upper bound:

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

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

Evaluation and experiments

Evaluation of GANs

The inception score? Salimans et al. [NIPS 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. [NIPS 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. [NIPS 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. [NIPS 2017]

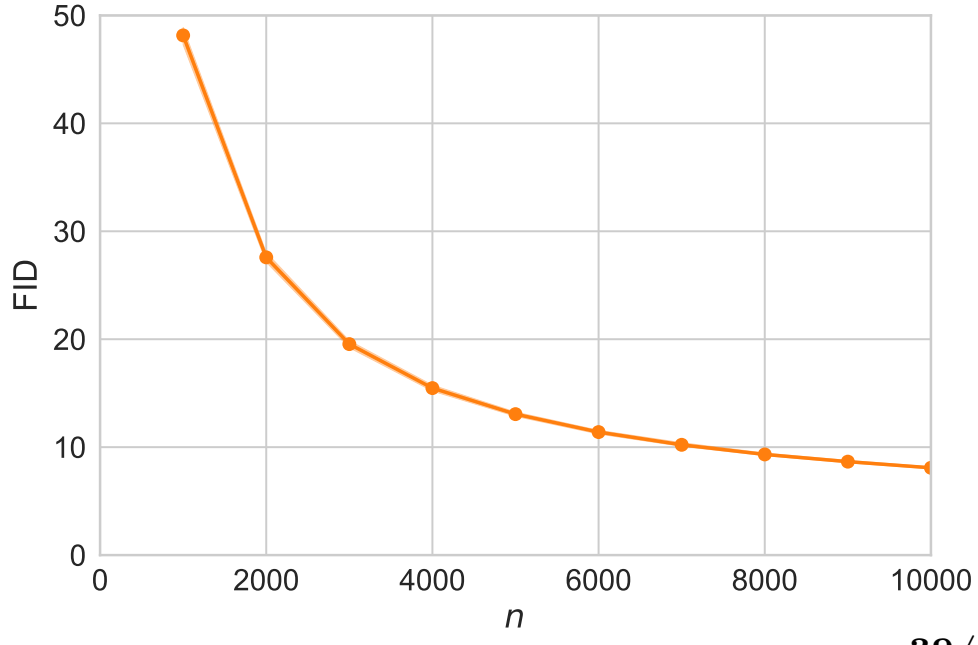
Fits Gaussians to features in the inception architecture (pool3 layer):

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

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 $n \to \infty$ samples from Q.

Given two alternatives:

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

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 $n \to \infty$ samples from Q.

Given two alternatives:

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

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 $n \to \infty$ samples from Q.

Given two alternatives:

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

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 $n \to \infty$ samples from Q.

Given two alternatives:

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

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

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

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

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

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

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

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

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

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)

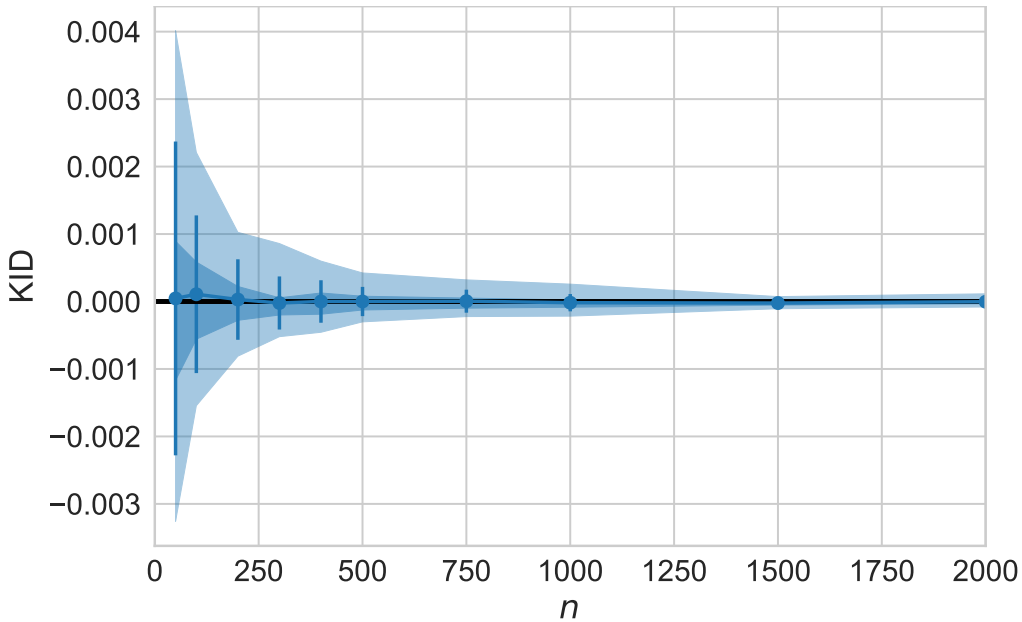
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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)

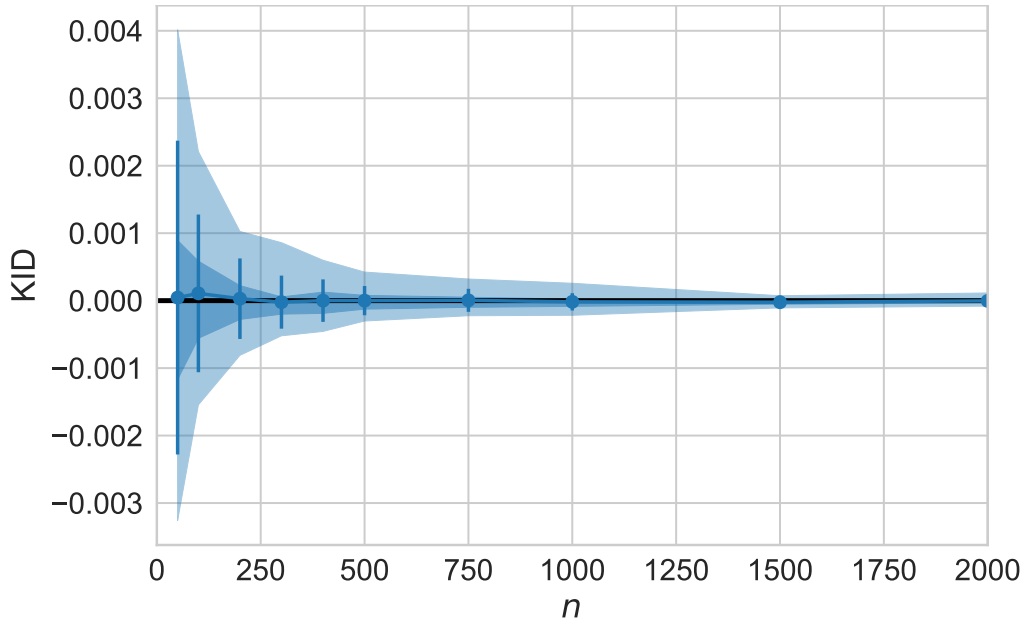
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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 is computationally costly?”

The kernel inception distance (KID)

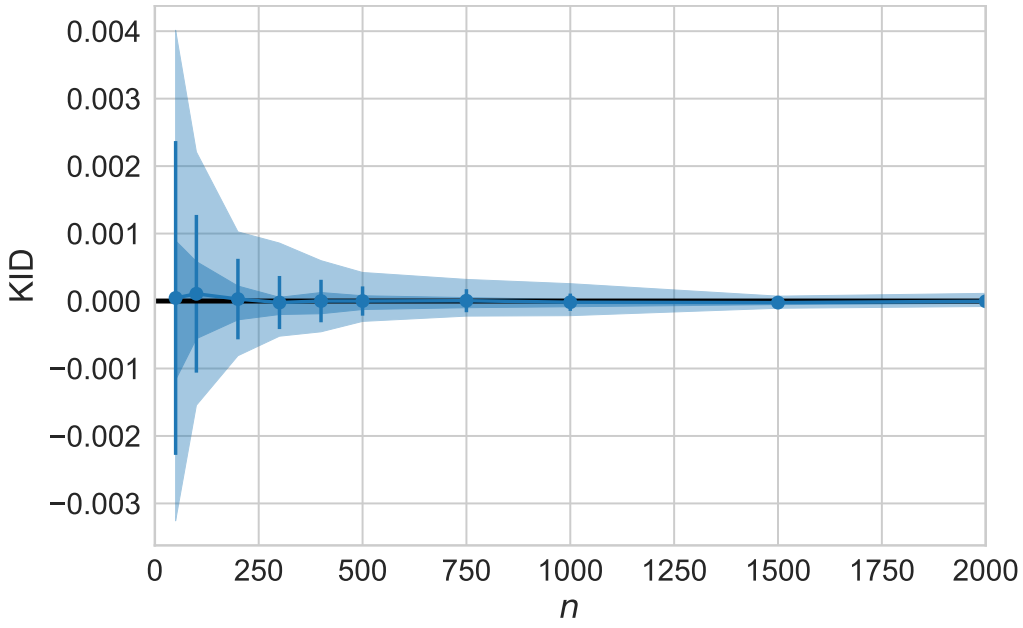
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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 our code)!

The kernel inception distance (KID)

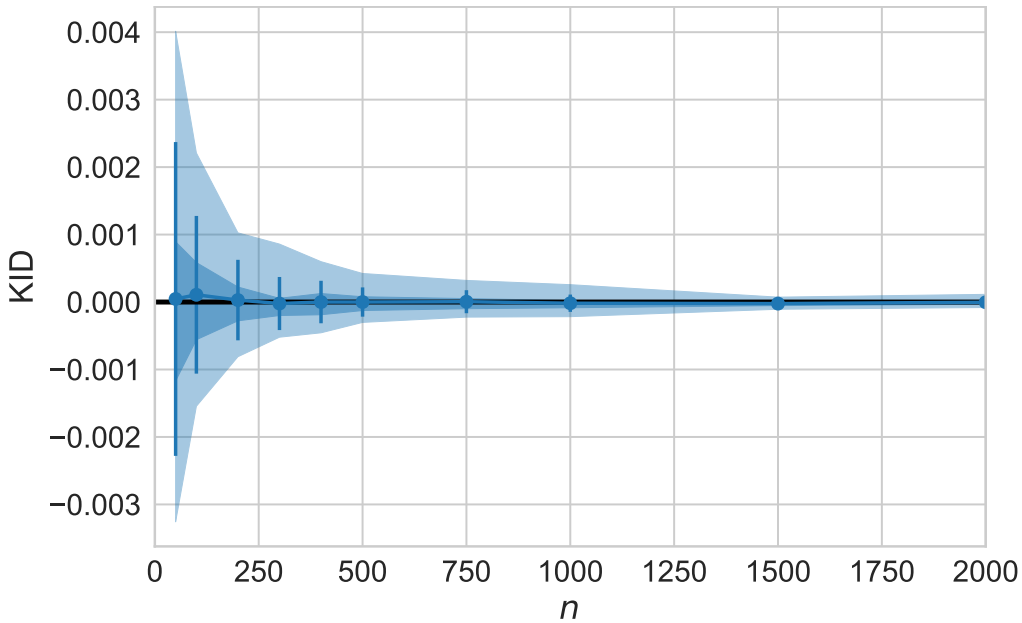
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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 $KID(\widehat{P}_{t+1}, Q)$ not significantly better than $KID(\widehat{P}_t, Q)$ then reduce learning rate.

[Bounliphone et al. ICLR 2016]

Related: "An empirical study on evaluation metrics of generative adversarial networks", Xu et al. [arxiv, June 2018]

Benchmarks for comparison (all from 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
yosh 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

Results: what does MMD buy you?

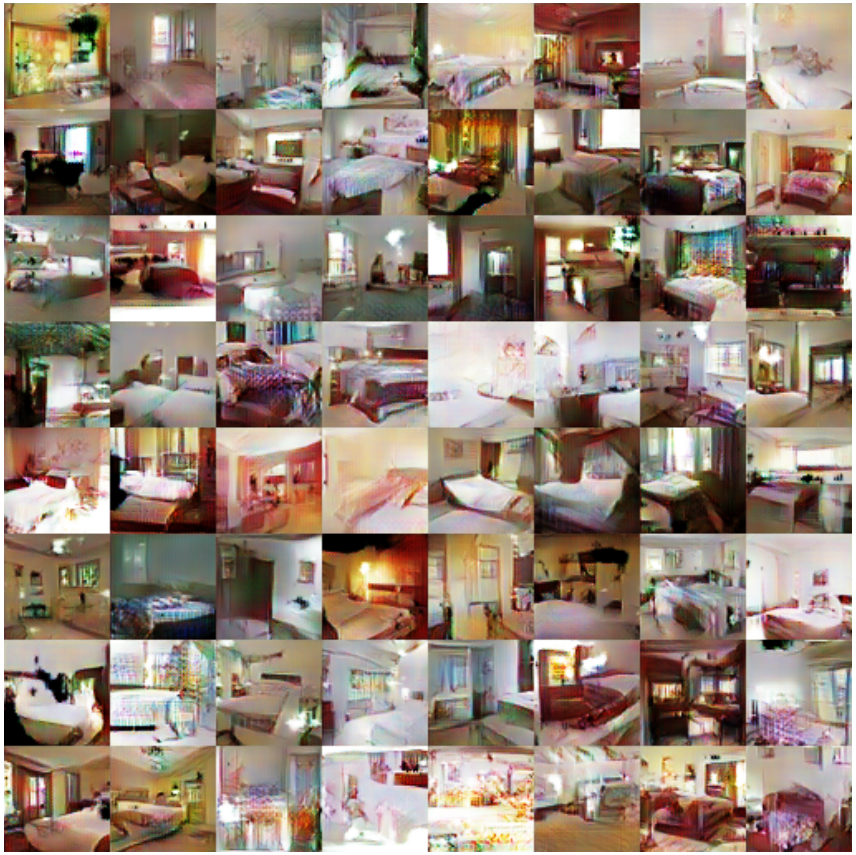
- Critic features from DCGAN: an f-filter critic has f, $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64.

MMD GAN samples, $f = 64$,
FID=32, KID=3

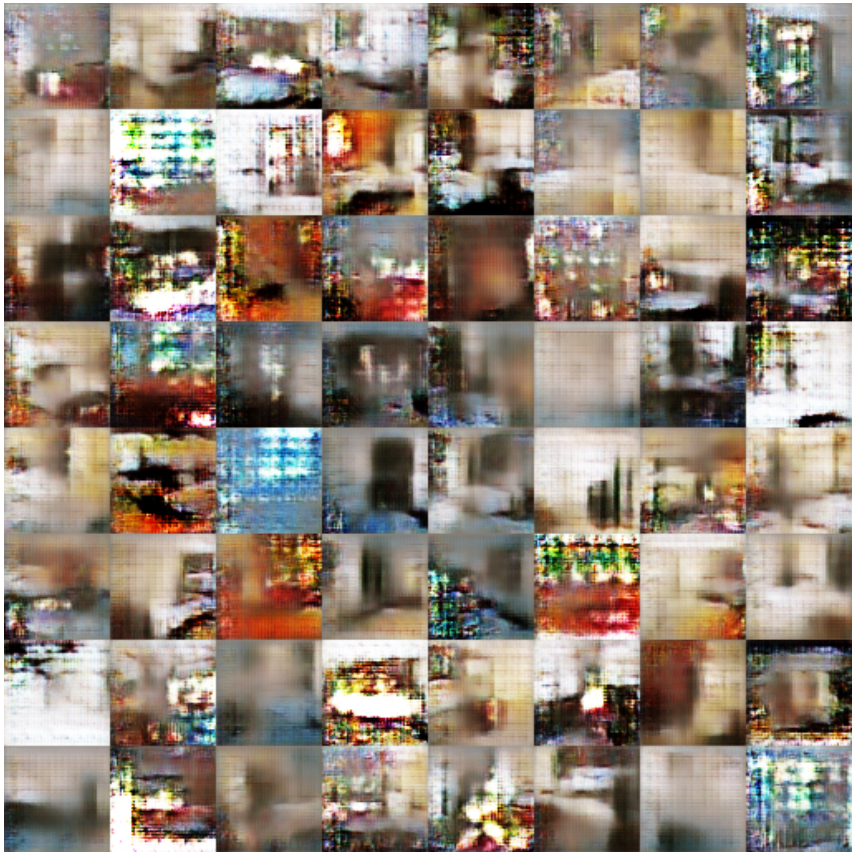
WGAN samples, $f = 64$,
FID=41, KID=4

Results: what does MMD buy you?

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



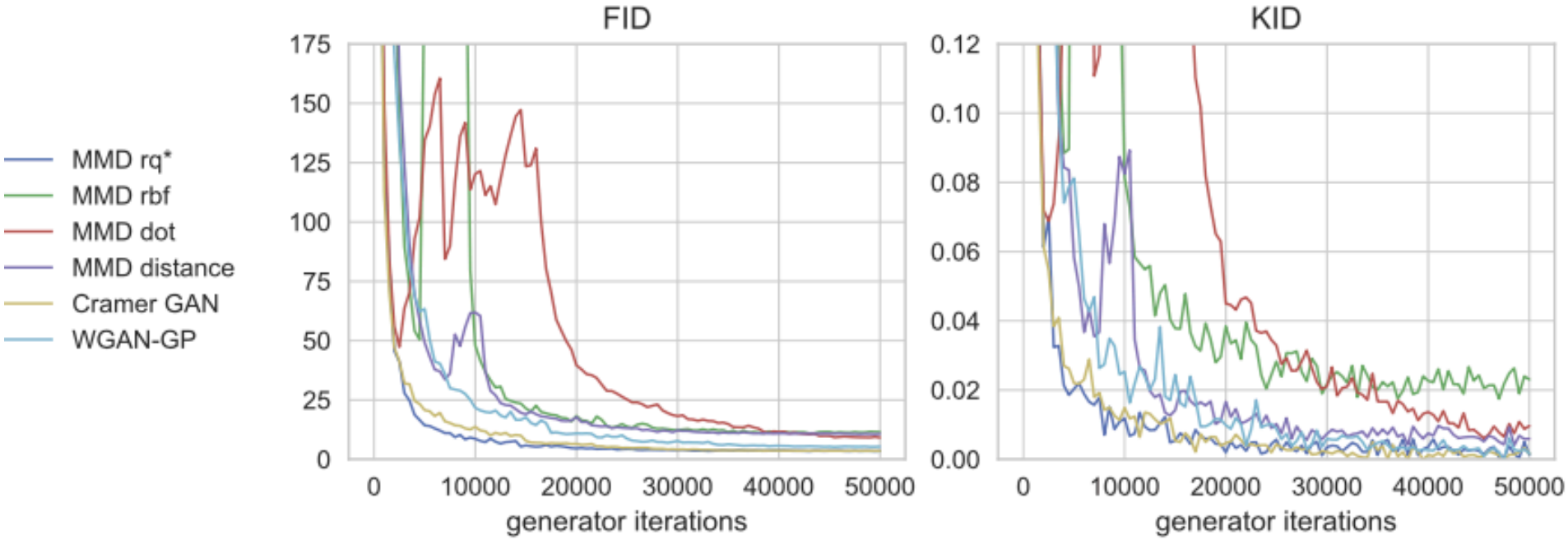
MMD GAN samples, $f = 16$,
FID=86, KID=9



WGAN samples, $f = 16$,
$f = 64$, FID=293, KID=37

The kernel inception distance (KID)

Faster training: performance scores vs generator iterations on MNIST



Results: celebrity faces 160×160

KID (FID) scores:

- Sobolev GAN: 14 (20)
- SN-GAN: 18 (28)
- Old MMD GAN: 13 (21)
- SMMD GAN: 6 (12)

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

Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
  47 (44)
- SN-GAN:
  44 (48)
- SMMD GAN:
  35 (37)

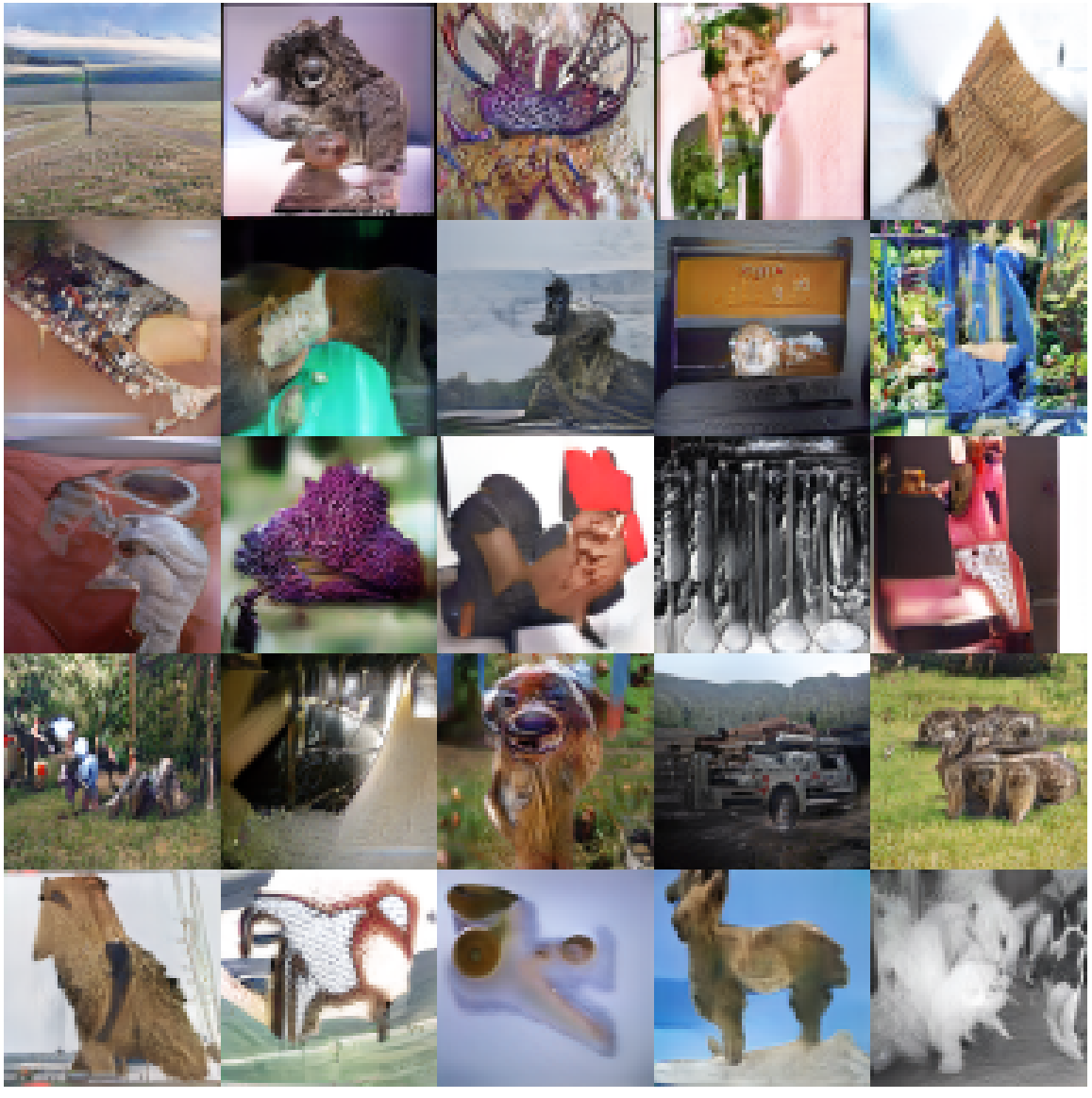
ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. Around 20 000 classes.

Results: imagenet 64×64

KID (FID) scores:

- BGAN: 47 (44)
- SN-GAN: 44 (48)
- SMMD GAN: 35 (37)

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



Results: imagenet 64×64

KID (FID) scores:

- BGAN: 47 (44)
- SN-GAN: 44 (48)
- SMMD GAN: 35 (37)

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

Summary

- MMD critic gives state-of-the-art performance for GAN training (FID and KID)
  - use convolutional input features
  - train with gradient regulariser
- Faster training, simpler critic network
- Reasons for good performance:
  - Unlike WGAN-GP, MMD loss still a valid critic when features not optimal
  - Kernel features do some of the "work", so simpler h_ψ features possible.
  - Better gradient/feature regulariser gives better critic

Code for "Demystifying MMD GANs," ICLR 2018, including KID score: https://github.com/mbinkowski/MMD-GAN

Code for new SMMD:
https://github.com/MichaelArbel/Scaled-MMD-GAN

Questions?

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
University College London
{dougal,michael.n.arbel,arthur.gretton}@gmail.com

arXiv.org > stat > arXiv:1805.11565

Statistics > Machine Learning

On gradient regularizers for MMD GANs

Michael Arbel, Dougal J. Sutherland, Mikołaj Bińkowski, Arthur Gretton

(Submitted on 29 May 2018)

Gradient bias for MMD GANs

MMD and WGAN-GP have unbiased gradients

Recall definitions:

- Generator G_ψ with parameters ψ. Samples $Y \sim G_\psi(Z)$ where $Z \sim R$
- Critic features h_θ with parameters θ.

Subject to mild conditions on the mappings h_θ and G_ψ, the kernel k, and the distributions P and R, for μ-almost all ψ, θ where μ is Lebesgue,

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

and thus MMD gradients unbiased.

Same true for WGAN-GP.

MMD and WGAN-GP have unbiased gradients

Recall definitions:

- Generator G_ψ with parameters ψ. Samples $Y \sim G_\psi(Z)$ where $Z \sim R$
- Critic features h_θ with parameters θ.

Subject to mild conditions on the mappings h_θ and G_ψ, the kernel k, and the distributions P and R, for μ-almost all ψ, θ where μ is Lebesgue,

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

and thus MMD gradients unbiased.

Same true for WGAN-GP.

Bias of MMD GAN critic

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 $\{x_i^{\text{tr}}\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} P$, $\{y_i^{\text{tr}}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} Q$.

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

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 $\{x_i^{\text{tr}}\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} P$, $\{y_i^{\text{tr}}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} Q$.

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

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 $\{x_i^{\text{tr}}\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} P$, $\{y_i^{\text{tr}}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} Q$.

Then

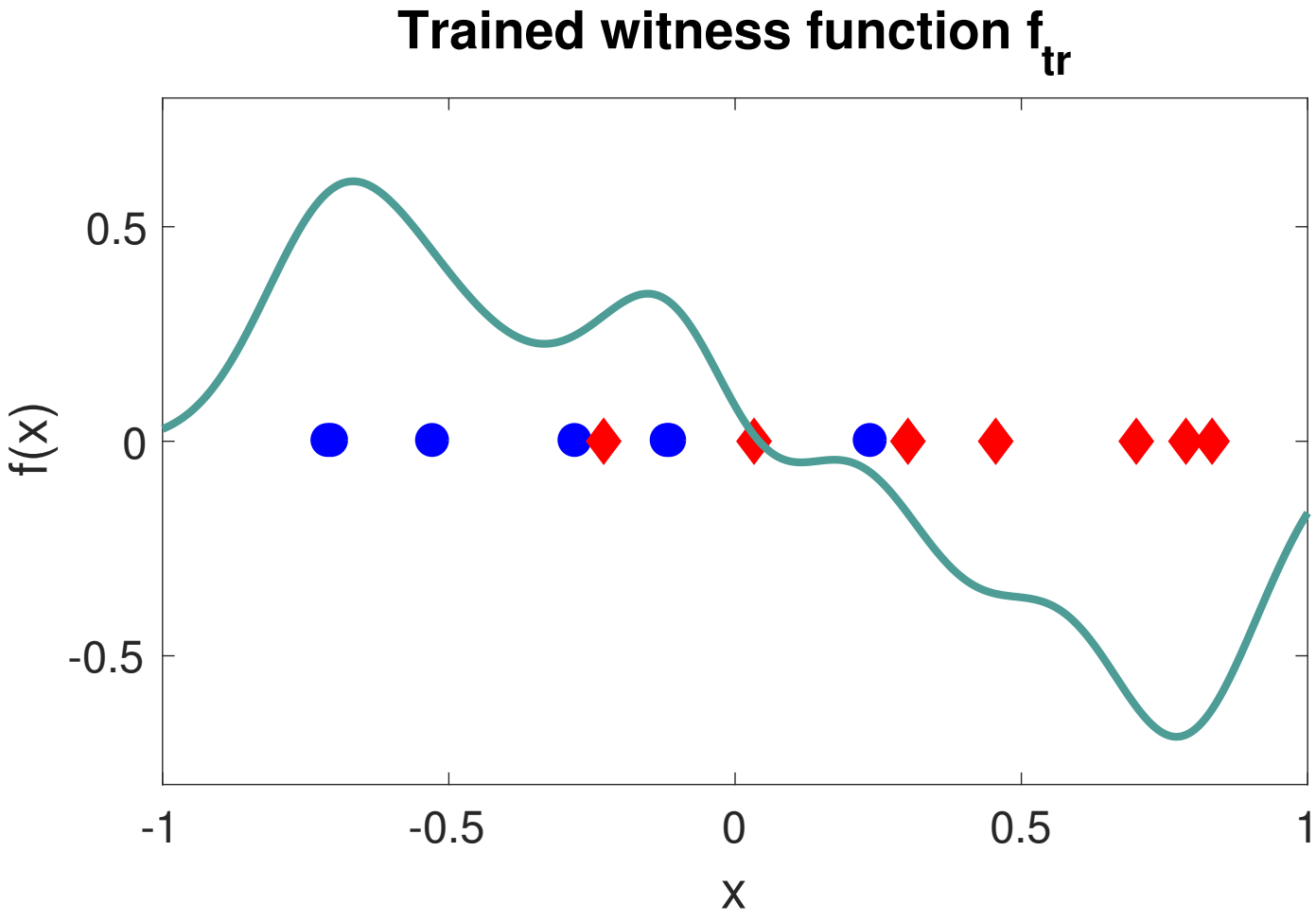
$$[\mathbf{E}_P f_{tr}(X) - \mathbf{E}_Q f_{tr}(Y)] \leq MMD(P, Q; F)$$

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

Same true for WGAN-GP.

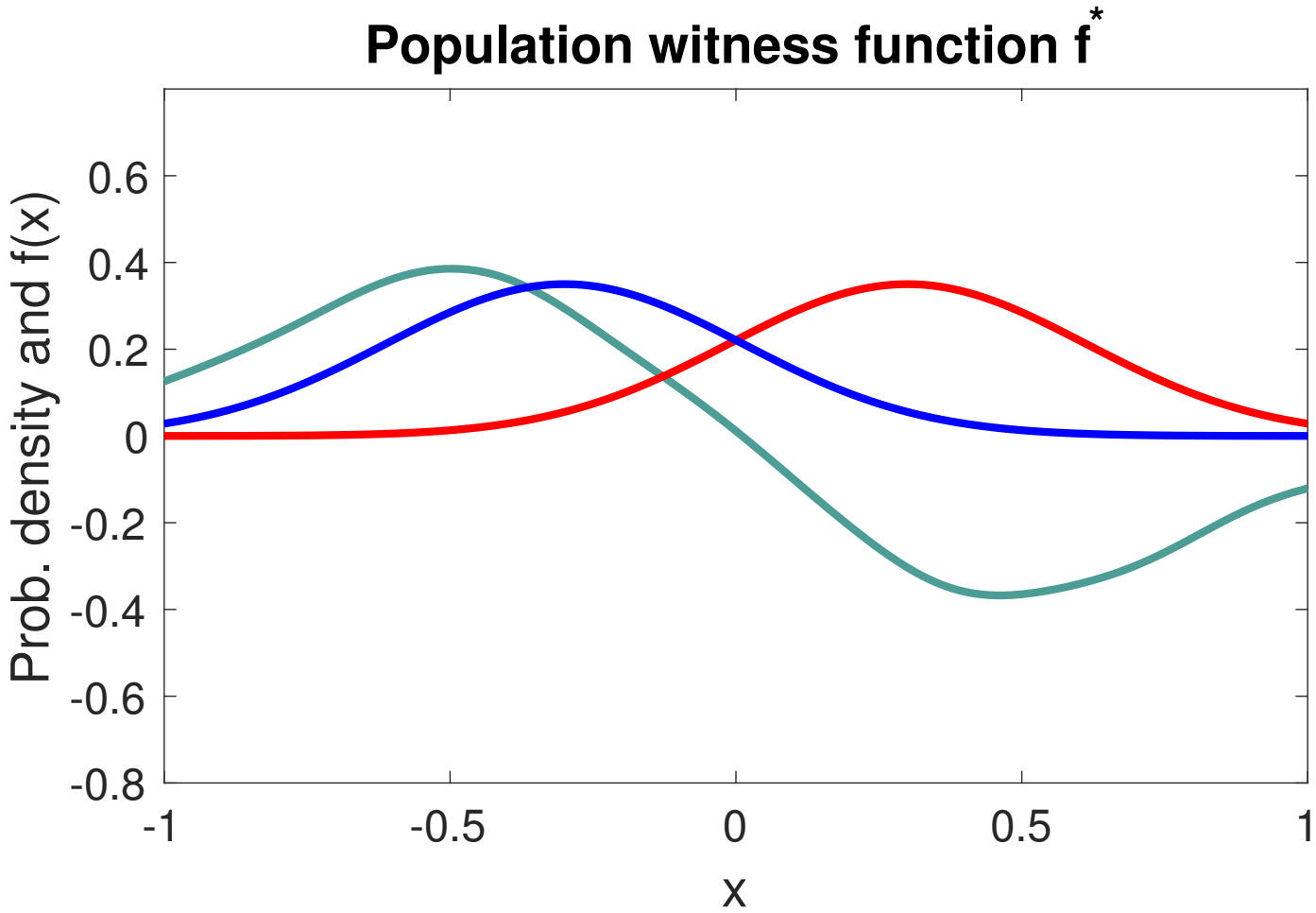
Bias of MMD GAN critic

Training minibatch critic function f_{tr}



Bias of MMD GAN critic

Population critic function f^*



Bias of MMD GAN critic

Bias in MMD vs training minibatch size: