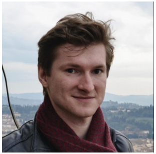
GANs with integral probability metrics: some results and conjectures

Arthur Gretton


Gatsby Computational Neuroscience Unit,
University College London

University of Oxford, 2020

A motivation: comparing two samples

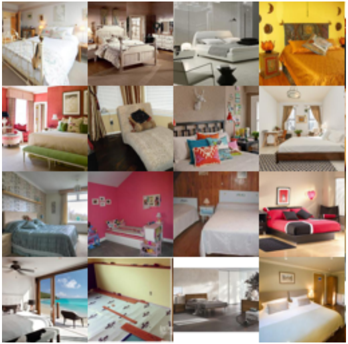
- **Given:** Samples from unknown distributions P and Q.
- **Goal:** do P and Q differ?

$$\sim P$$

$$\sim Q$$

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

(Binkowski, Sutherland, Arbel, G., ICLR 2018),
(Arbel, Sutherland, Binkowski, G., NeurIPS 2018)

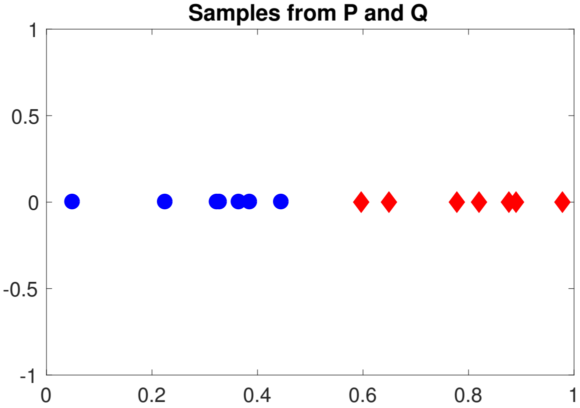
Outline

- 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

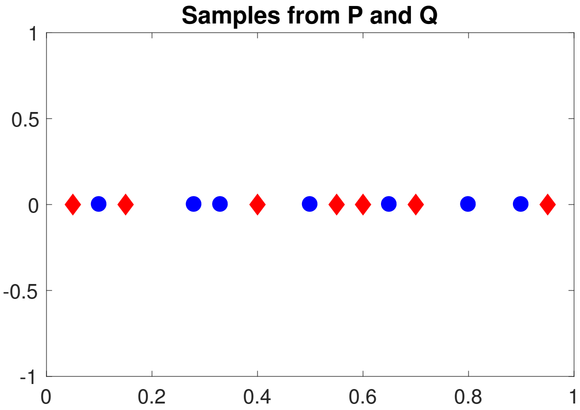
Integral probability metrics

Are P and Q different?



Integral probability metrics

Are P and Q different?

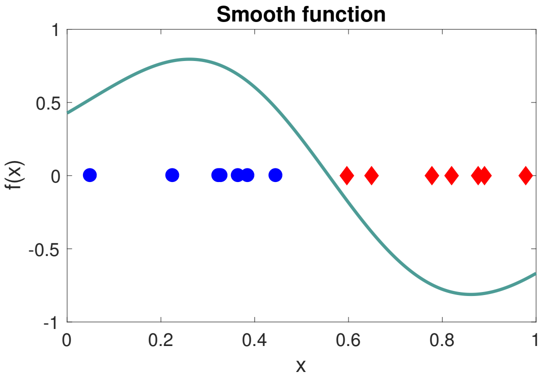


Integral probability metrics

Integral probability metric:

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

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

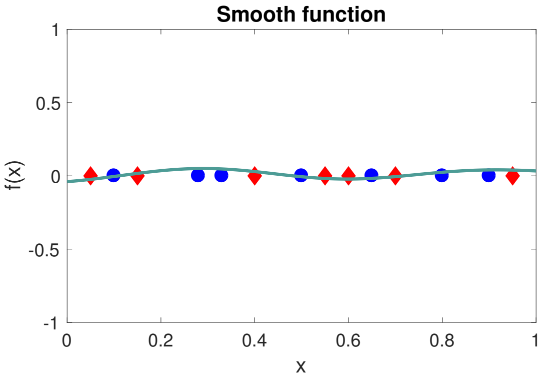


Integral probability metrics

Integral probability metric:

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

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



The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

The MMD: an integral probability metric

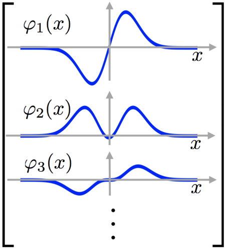
Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

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



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

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,

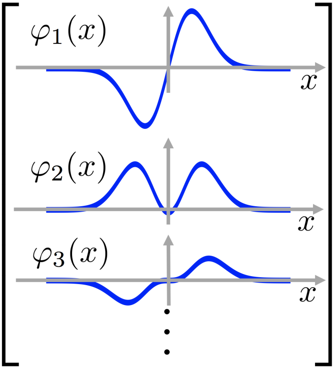
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

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

Exponentiated quadratic kernel

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

$$\varphi(x) =$$



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

The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

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]
- Lipschitz (Wasserstein distances) [Dudley, 2002]

- Energy distance is a special case [Sejdinovic, Sriperumbudur, G. Fukumizu, 2013]

The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

Expectations of functions are linear combinations of expected features

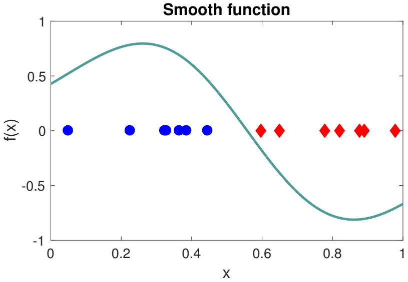
$$\mathbf{E}_P(f(X)) = \langle f, \mathbf{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

Integral prob. metric vs feature mean difference

The MMD:

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



Integral prob. metric vs feature mean difference

The MMD:

use

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

$$= \sup_{\|f\|\leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right]$$

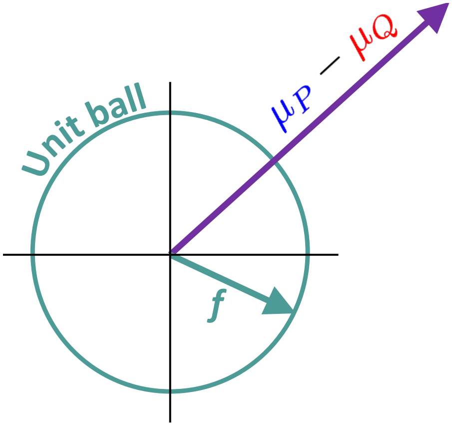
$$= \sup_{\|f\|\leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$

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

Integral prob. metric vs feature mean difference

The MMD:

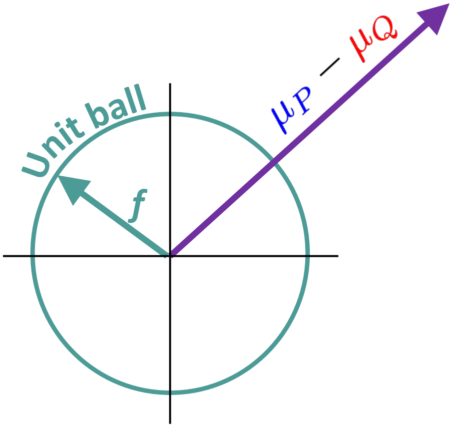
$$MMD(P, Q; F) = \sup_{\|f\| \leq 1} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)] = \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature mean difference

The MMD:

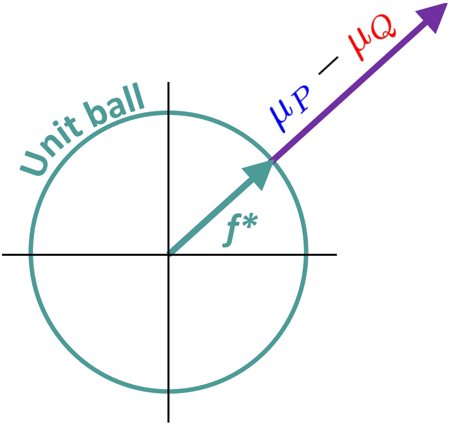
$$
\begin{aligned}
& MMD(P, Q; F) \\
&= \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
&= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature mean difference

The MMD:

$$
\begin{aligned}
& MMD(P, Q; F) \\
&= \sup_{\|f\| \leq 1} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)] \\
&= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



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

Integral prob. metric vs feature mean difference

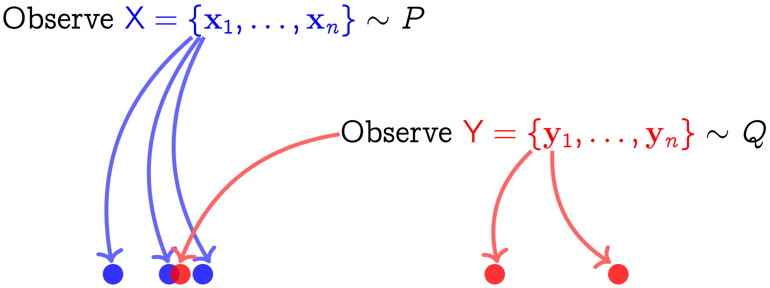
The MMD:

$$
\begin{aligned}
& MMD(P, Q; F) \\
&= \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
&= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \| \mu_P - \mu_Q \|
\end{aligned}
$$

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

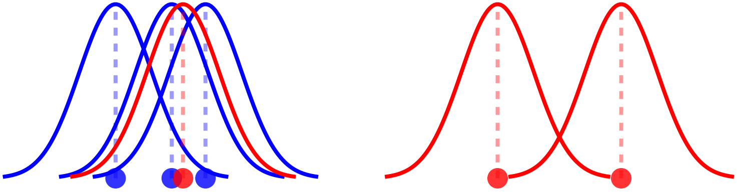
Construction of MMD witness

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



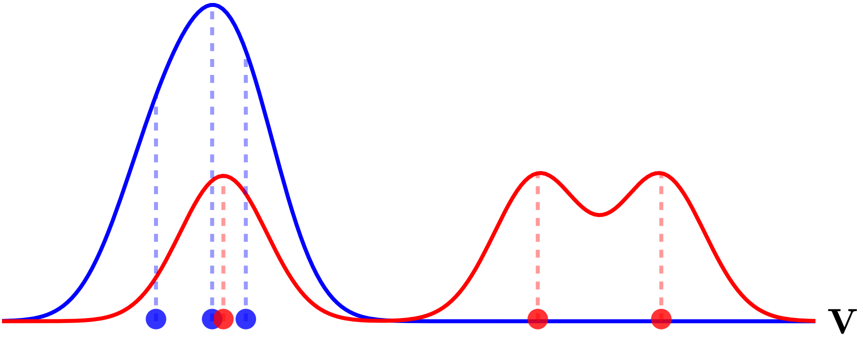
Construction of MMD witness

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



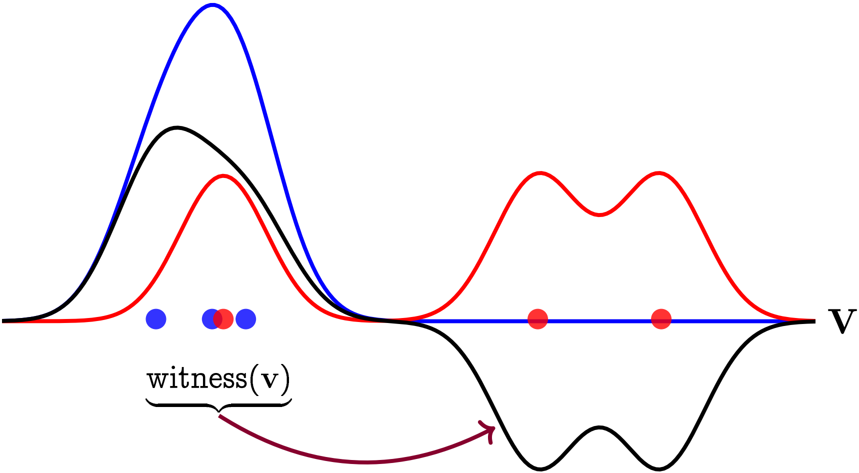
Construction of MMD witness

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

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

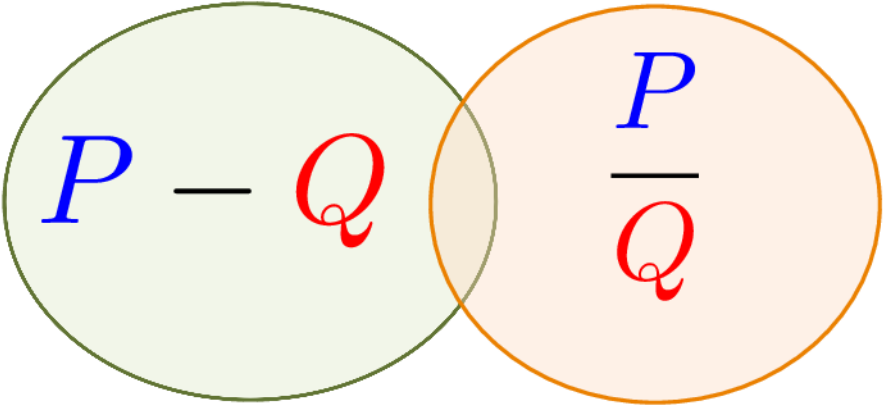
The empirical witness function at v

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

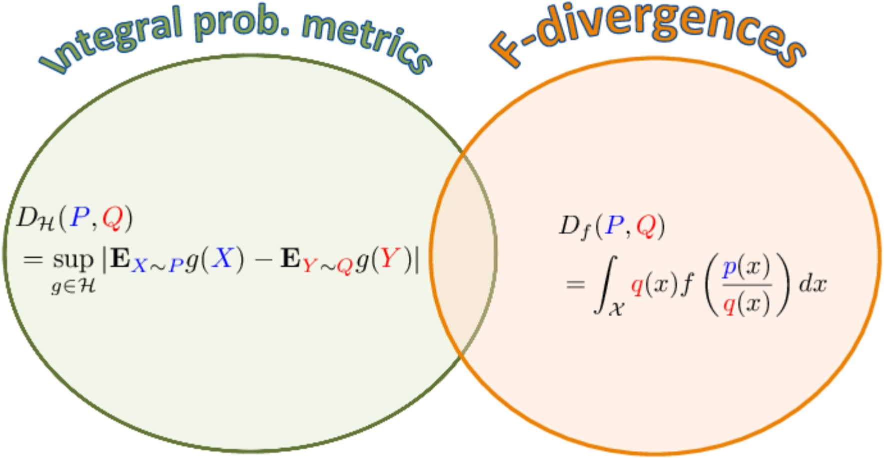
Don't need explicit feature coefficients $f^* := \begin{bmatrix} f_1^* & f_2^* & \dots \end{bmatrix}$

Interlude: divergence measures

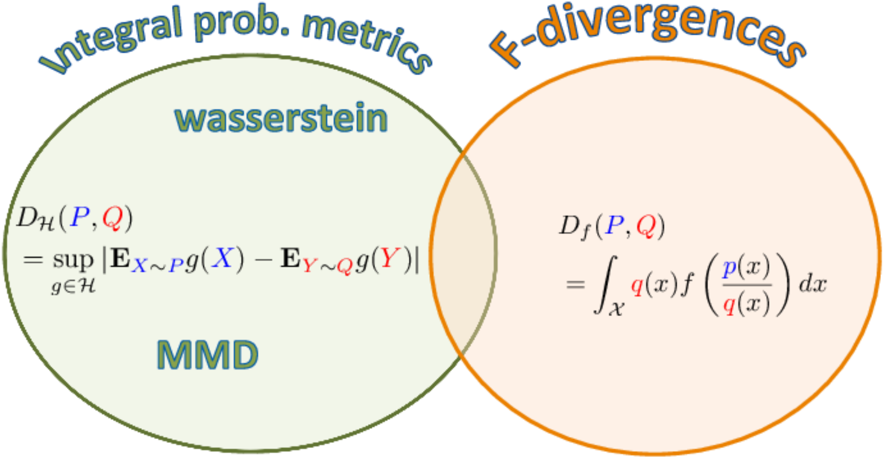
Divergences



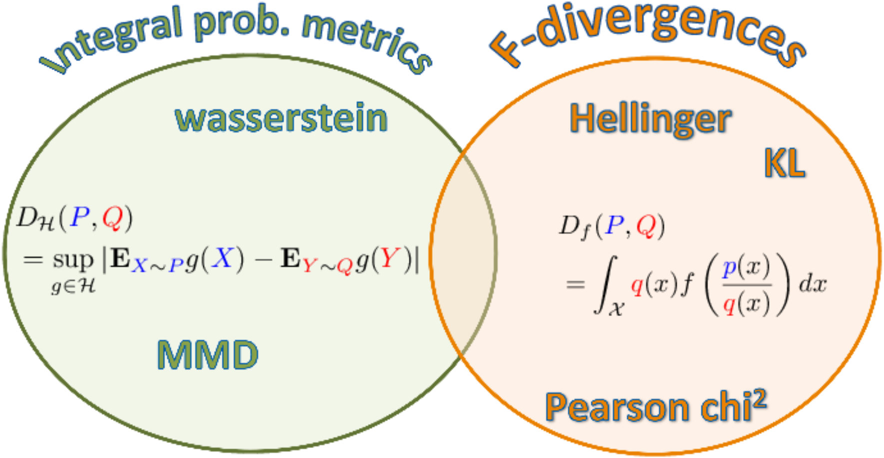
Divergences



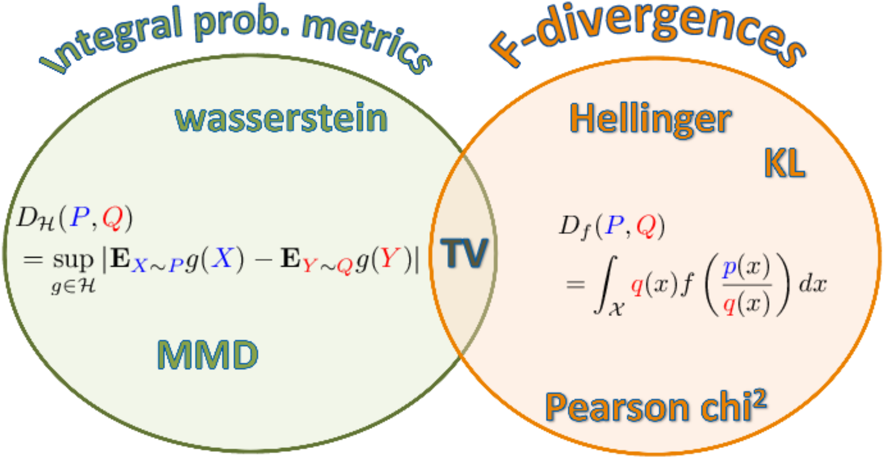
Divergences



Divergences



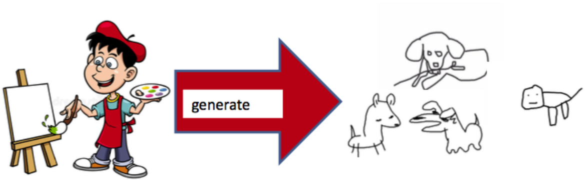
Divergences



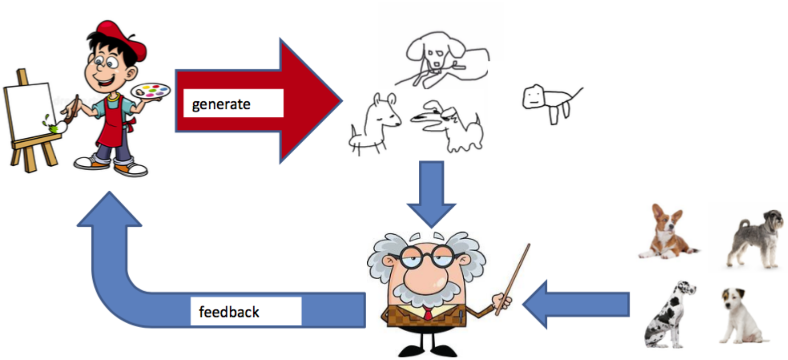
Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (2012)

Training Generative Adversarial Networks: Critics and Gradient Penalties

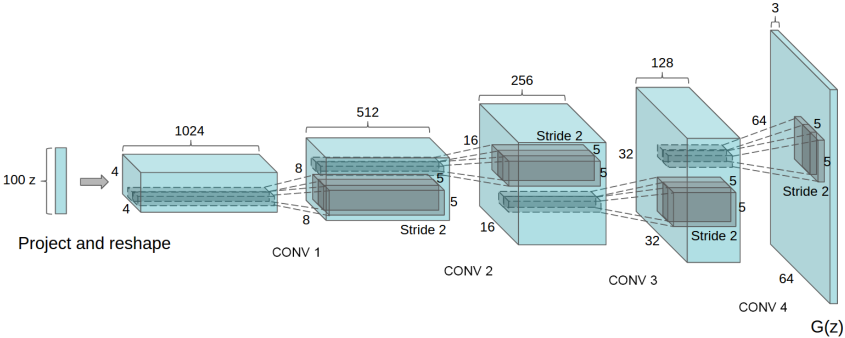
Visual notation: GAN setting



Visual notation: GAN setting



What I won't cover: the generator



Radford, Metz, Chintala, ICLR 2016

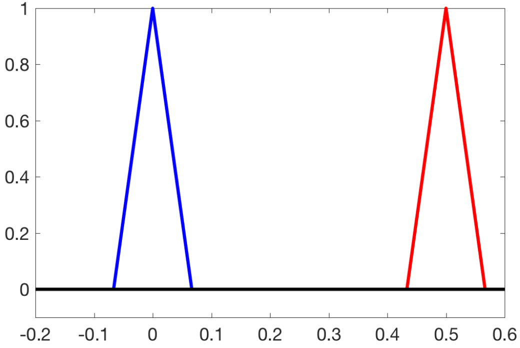
F-divergence as critic

An unhelpful critic? Jensen-Shannon,

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

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

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



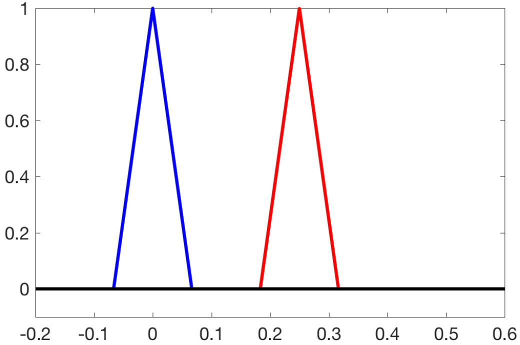
F-divergence as critic

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]

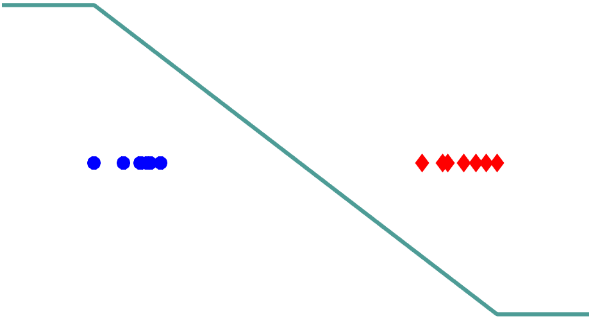
Wasserstein distance as critic

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



Santambrogio, Optimal Transport for Applied Mathematicians (2015, Section 5.4)
G Peyré, M Cuturi, Computational Optimal Transport (2019)
M. Cuturi, J. Solomon, NeurIPS tutorial (2017)

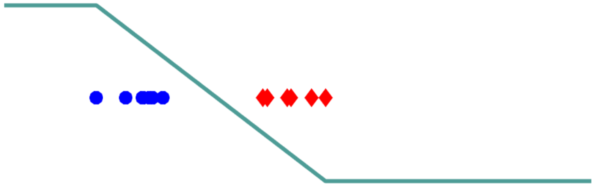
Wasserstein distance as critic

A helpful critic witness:

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

$\|f\|_L := \sup_{x \neq y} |f(x) - f(y)| / \|x - y\|$

$W_1 = 0.65$



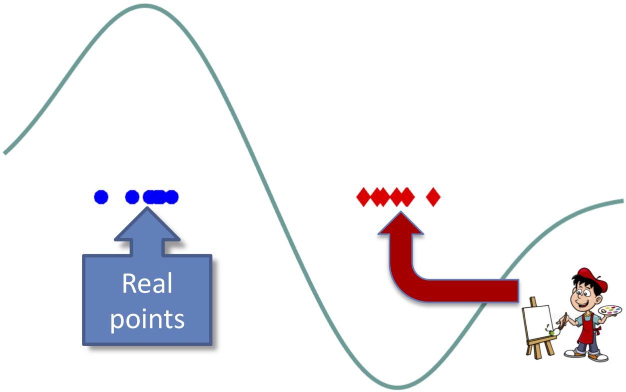
Santambrogio, Optimal Transport for Applied Mathematicians (2015, Section 5.4)
G Peyré, M Cuturi, Computational Optimal Transport (2019)
M. Cuturi, J. Solomon, NeurIPS tutorial (2017)

MMD as critic

A helpful critic witness:

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

MMD=1.8

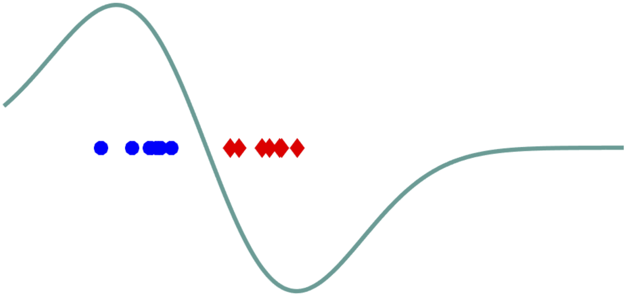


MMD as critic

A helpful critic witness:

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

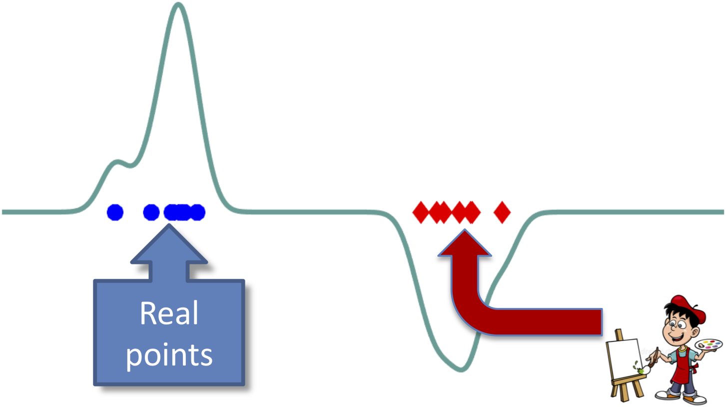
MMD=1.1



MMD as critic

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

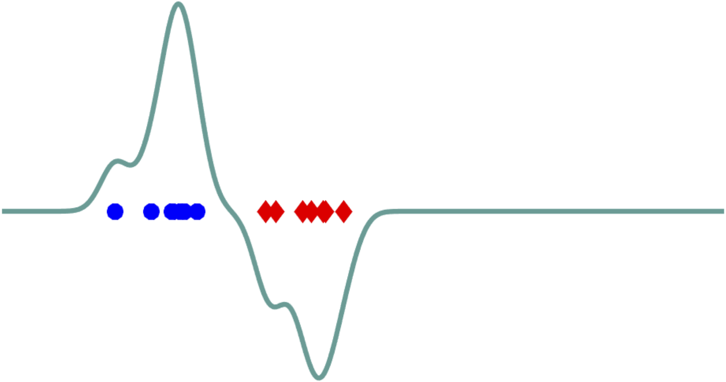
MMD=0.64



MMD as critic

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

MMD=0.64



f-divergences $(\phi - divergences)$

The ϕ-divergences

Define the ϕ-divergence(f-divergence):

$$D_\phi(P, Q) = \int \phi\left(\frac{dP}{dQ}\right) dQ = \int \phi\left(\frac{p(x)}{q(x)}\right) q(x)dx$$

where ϕ is convex, lower-semicontinuous, $\phi(1) = 0$.

- Example: $\phi(x) = -\log(x)$ gives reverse KL divergence,

$$D_{KL}(Q, P) = \int \log\left(\frac{q(x)}{p(x)}\right) q(x)dx$$

The ϕ-divergences

Define the ϕ-divergence(f-divergence):

$$D_\phi(P, Q) = \int \phi\left(\frac{dP}{dQ}\right) dQ = \int \phi\left(\frac{p(x)}{q(x)}\right) q(x) dx$$

where ϕ is convex, lower-semicontinuous, $\phi(1) = 0$.

- Example: $\phi(x) = -\log(x)$ gives reverse KL divergence,

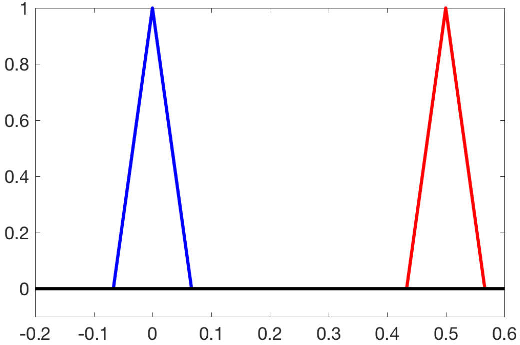
$$D_{KL}(Q, P) = \int \log\left(\frac{q(x)}{p(x)}\right) q(x) dx$$

How do ϕ-divergences behave?

Simple example: disjoint support, revisited.

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

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

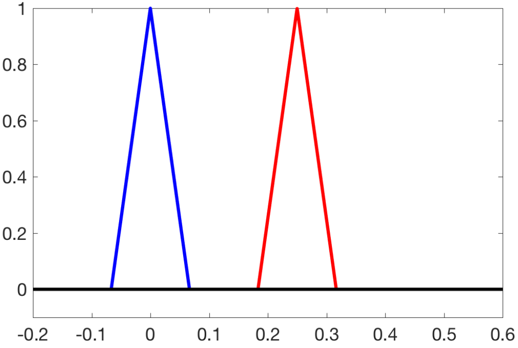


How do ϕ-divergences behave?

Simple example: disjoint support, revisited.

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

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



ϕ-divergences in practice

Background: the Fenchel dual

- Conjugate (fenchel) dual:

$$\phi^*(v) = \sup_{u \in \Re} \{uv - \phi(u)\}.$$

  - v is slope of ϕ
  - u is the argument of ϕ where it has slope v.

$$\partial \phi^*(v) = u$$

  - $\phi^*(v)$ is the negative of the intercept of the line with slope v, tangent to $\phi(u)$ at u.

ϕ-divergences in practice

Background: the Fenchel dual

- Conjugate (fenchel) dual:

$$\phi^*(v) = \sup_{u \in \Re} \{uv - \phi(u)\}.$$

  - v is slope of ϕ
  - u is the argument of ϕ where it has slope v.

$$\partial\phi^*(v) = u$$

  - $\phi^*(v)$ is the negative of the intercept of the line with slope v, tangent to $\phi(u)$ at u.

- For a convex l.s.c. ϕ we have

$$\phi^{**}(v) = \phi(v) = \sup_{u \in \Re} \{uv - \phi^*(u)\}$$

ϕ-divergences in practice

Background: the Fenchel dual

- Conjugate (fenchel) dual:

$$\phi^*(v) = \sup_{u \in \Re} \{uv - \phi(u)\}.$$

  - v is slope of ϕ
  - u is the argument of ϕ where it has slope v.

$$\partial \phi^*(v) = u$$

  - $\phi^*(v)$ is the negative of the intercept of the line with slope v, tangent to $\phi(u)$ at u.

- Reverse KL:

$$\phi(u) = -\log(u) \qquad \phi^*(v) = \begin{cases} -1 - \log v & v < 0 \\ \infty & v \geq 0 \end{cases}$$

A variational lower bound

How to compute ϕ-divergences in practice:

$$D_\phi(P, Q) = \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

A variational lower bound

How to compute ϕ-divergences in practice:

$$D_\phi(P, Q) = \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz$$
$$= \int q(z) \underbrace{\sup_{f_z}\left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right)}_{\phi\left(\frac{p(z)}{q(z)}\right)}$$

$\phi^*(u)$ is dual of $\phi(u)$.

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

A variational lower bound

How to compute ϕ-divergences in practice:

$$D_\phi(P, Q) = \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz$$
$$= \int q(z) \sup_{f_z} \left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right)$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

A variational lower bound

How to compute ϕ-divergences in practice:

$$\begin{aligned} D_\phi(P, Q) &= \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz \\ &= \int q(z) \sup_{f_z} \left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right) \\ &\geq \sup_{f \in \mathcal{H}} \mathbf{E}_P f(X) - \mathbf{E}_Q \phi^*\left(f(Y)\right) \end{aligned}$$

(restrict the function class)

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

A variational lower bound

How to compute ϕ-divergences in practice:

$$D_\phi(P, Q) = \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz$$
$$= \int q(z) \sup_{f_z}\left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right)$$
$$\geq \sup_{f \in \mathcal{H}} \mathbf{E}_P f(X) - \mathbf{E}_Q \phi^*\left(f(Y)\right)$$

(restrict the function class)

Optimum $f_z^\diamond$ has property

$$\frac{p(z)}{q(z)} = \partial\phi^*(f_z^\diamond) \iff f_z^\diamond = \partial\phi\left(\frac{p(z)}{q(z)}\right).$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

ϕ-divergences in practice

Case of the reverse KL

$$D_{KL}(Q, P) = \int q(z) \log\left(\frac{q(z)}{p(z)}\right) dz$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

ϕ-divergences in practice

Case of the reverse KL

$$D_{KL}(Q, P) = \int q(z) \log\left(\frac{q(z)}{p(z)}\right) dz$$

$$\geq \sup_{f<0, f\in\mathcal{H}} \mathbf{E}_P f(X) + \mathbf{E}_Q \underbrace{\log(-f(Y)) + 1}_{-\phi^*(f(Y))}$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

ϕ-divergences in practice

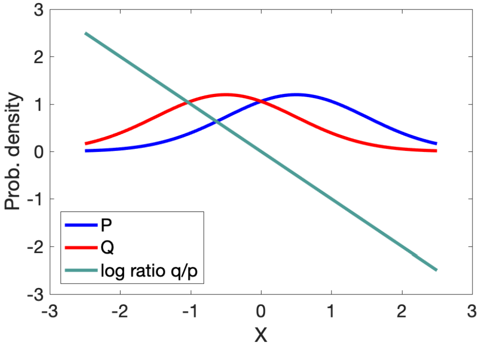
Case of the reverse KL

$$D_{KL}(Q, P) = \int q(z) \log\left(\frac{q(z)}{p(z)}\right) dz$$

$$\geq \sup_{f<0, f\in\mathcal{H}} \mathbf{E}_P f(X) + \mathbf{E}_Q \log(-f(Y)) + 1$$

Bound tight when:

$$f^\diamond(z) = -\frac{q(z)}{p(z)}$$



Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

ϕ-divergences in practice

Case of the reverse KL

$$D_{KL}(Q, P) = \int q(z) \log\left(\frac{q(z)}{p(z)}\right) dz$$

$$\geq \sup_{f<0, f\in\mathcal{H}} \mathbf{E}_P f(X) + \mathbf{E}_Q \log(-f(Y)) + 1$$

$$\approx \sup_{f<0, f\in\mathcal{H}} \left[\frac{1}{n}\sum_{j=1}^{n} f(x_i) + \frac{1}{n}\sum_{i=1}^{n} \log(-f(y_i))\right] + 1$$

$x_i \overset{\text{i.i.d.}}{\sim} P$

$y_i \overset{\text{i.i.d.}}{\sim} Q$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

ϕ-divergences in practice

Case of the reverse KL

$$D_{KL}(Q, P) = \int q(z) \log \left(\frac{q(z)}{p(z)} \right) dz$$

$$\geq \sup_{f<0, f\in\mathcal{H}} \mathbf{E}_P f(X) + \mathbf{E}_Q \log\left(-f(Y)\right) + 1$$

$$\approx \sup_{f<0, f\in\mathcal{H}} \left[\frac{1}{n} \sum_{j=1}^{n} f(x_i) + \frac{1}{n} \sum_{i=1}^{n} \log(-f(y_i)) \right] + 1$$

This is a

KL

Approximate

Lower-bound

Estimator.

ϕ-divergences in practice

Case of the reverse KL

$$D_{KL}(Q, P) = \int q(z) \log\left(\frac{q(z)}{p(z)}\right) dz$$

$$\geq \sup_{f<0, f\in\mathcal{H}} \mathbf{E}_P f(X) + \mathbf{E}_Q \log\left(-f(Y)\right) + 1$$

$$\approx \sup_{f<0, f\in\mathcal{H}} \left[\frac{1}{n}\sum_{j=1}^{n} f(x_i) + \frac{1}{n}\sum_{i=1}^{n} \log(-f(y_i))\right] + 1$$

This is a

K

A

L

E

ϕ-divergences in practice

Case of the reverse KL

$$D_{KL}(Q, P) = \int q(z) \log\left(\frac{q(z)}{p(z)}\right) dz$$

$$\geq \sup_{f<0, f\in\mathcal{H}} \mathbf{E}_P f(X) + \mathbf{E}_Q \log\left(-f(Y)\right) + 1$$

$$\approx \sup_{f<0, f\in\mathcal{H}} \left[\frac{1}{n}\sum_{j=1}^{n} f(x_i) + \frac{1}{n}\sum_{i=1}^{n} \log(-f(y_i))\right] + 1$$

The KALE divergence

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

How does the KALE divergence behave?

$$KALE(Q,P) = \sup_{f<0, f\in\mathcal{H}} E_P f(X) + E_Q \log(-f(Y)) + 1$$

$$f = -\exp\langle w, \phi(x)\rangle_{\mathcal{F}}$$

$$\|w\|^2_{\mathcal{F}} \quad \text{penalized :}$$

How does the KALE divergence behave?

$$KALE(Q, P) = \sup_{f<0, f \in \mathcal{H}} E_P f(X) + E_Q \log(-f(Y)) + 1$$

$$f = -\exp \langle w, \phi(x) \rangle_{\mathcal{F}}$$

$$\|w\|_{\mathcal{F}}^2 \quad \text{penalized : KALE smoothie}$$

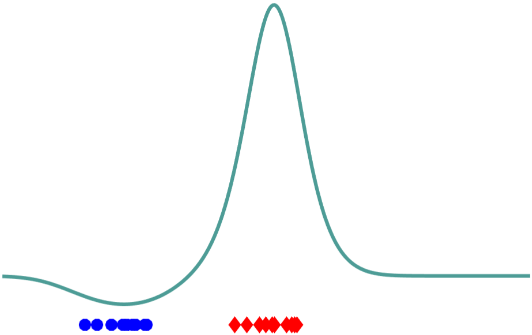
How does the KALE divergence behave?

$$KALE(Q,P) = \sup_{f<0, f\in\mathcal{H}} E_P f(X) + E_Q \log(-f(Y)) + 1$$

$$f = -\exp \langle w, \phi(x) \rangle_{\mathcal{F}}$$

$$\|w\|^2_{\mathcal{F}} \quad \text{penalized : KALE smoothie}$$

$$KALE(Q,P) = 0.18$$



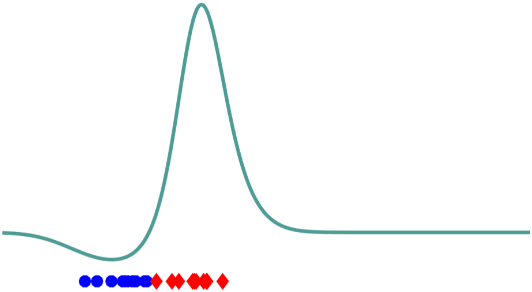
How does the KALE divergence behave?

$$KALE(Q,P) = \sup_{f<0, f\in\mathcal{H}} E_P f(X) + E_Q \log(-f(Y)) + 1$$

$$f = -\exp\langle w, \phi(x)\rangle_{\mathcal{F}}$$

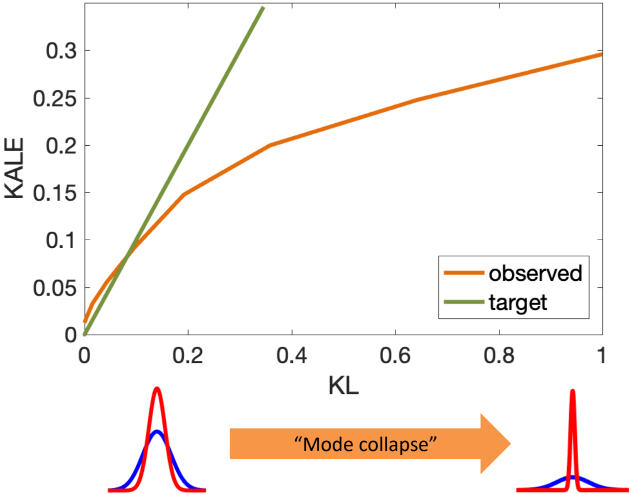
$$\|w\|_{\mathcal{F}}^2 \quad \text{penalized : KALE smoothie}$$

$$KALE(Q,P){=}0.12$$



The KALE smoothie and "mode collapse"

- Two Gaussians with same means, different variance



Example thanks to M. Arbel and M. Rosca

Gradient penalty: the regularisation viewpoint

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

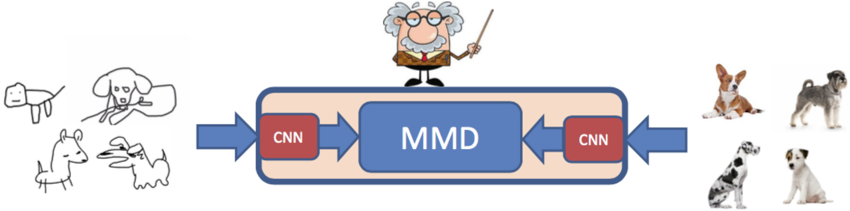
MMD for GAN critic

Can you use MMD as a critic to train GANs?

Need better image features.

CNN features for IPM witness functions

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



$\mathfrak{K}(x, y) = h_\psi{}^\top(x) h_\psi(y)$
where $h_\psi(x)$ is a CNN map:

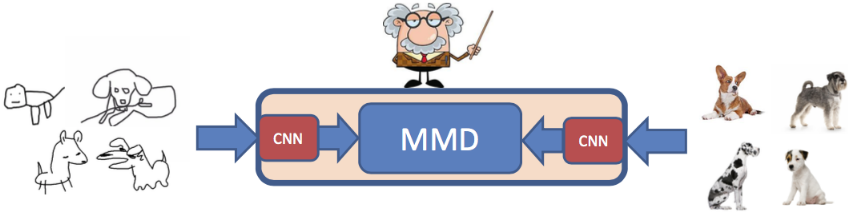
- Wasserstein GAN Arjovsky et al. [ICML 2017]
- WGAN-GP Gulrajani et al. [NeurIPS 2017]

$\mathfrak{K}(x, y) = k(h_\psi(x), h_\psi(y))$
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]

CNN features for IPM witness functions

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



$\mathfrak{K}(x, y) = h_\psi{}^\top(x) h_\psi(y)$
where $h_\psi(x)$ is a CNN map:

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

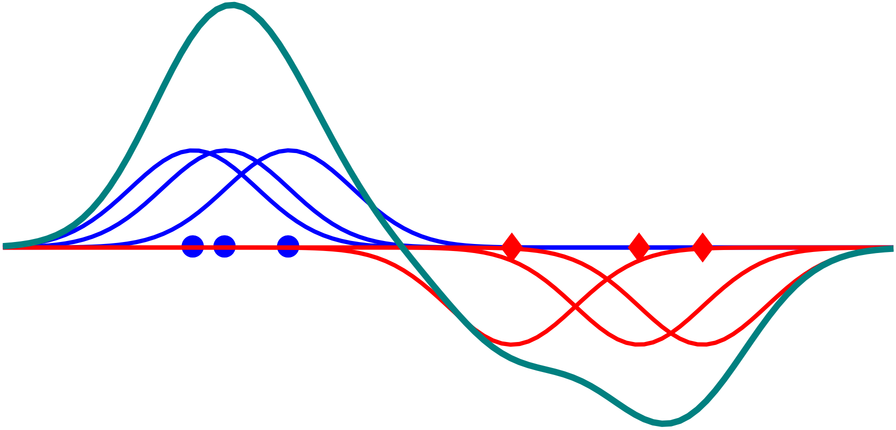
$\mathfrak{K}(x, y) = k(h_\psi(x), h_\psi(y))$
where $h_\psi(x)$ is a CNN map,
k is e.g. an exponentiated quadratic kernel

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,

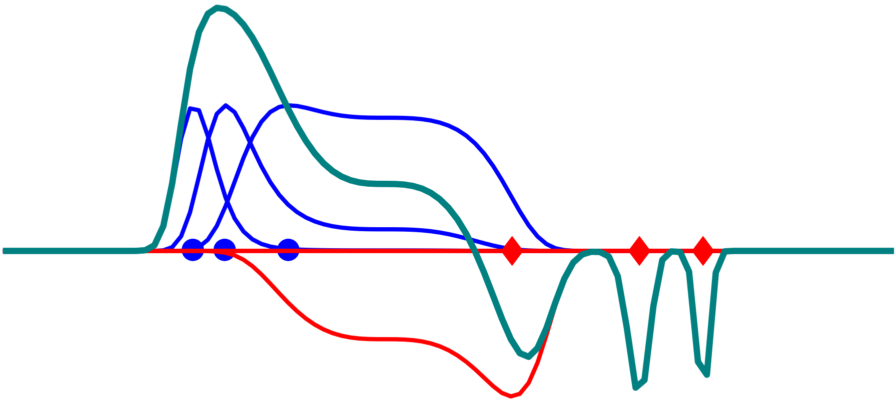
$k(x, y)$ is exponentiated quadratic



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?

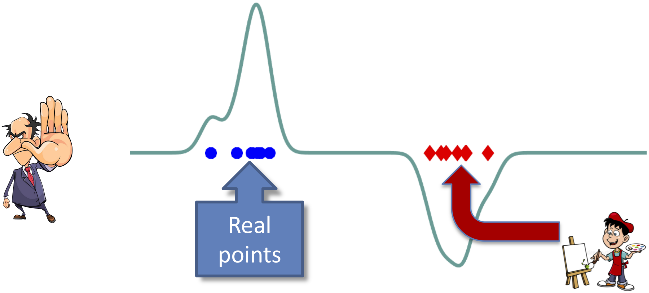
Challenges for learned critic features

Learned critic features:

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

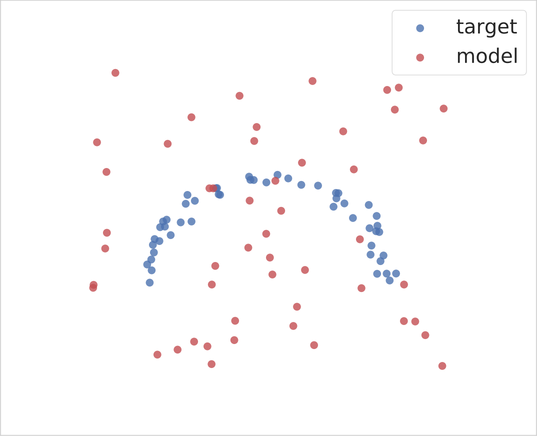
Relation with test power?

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



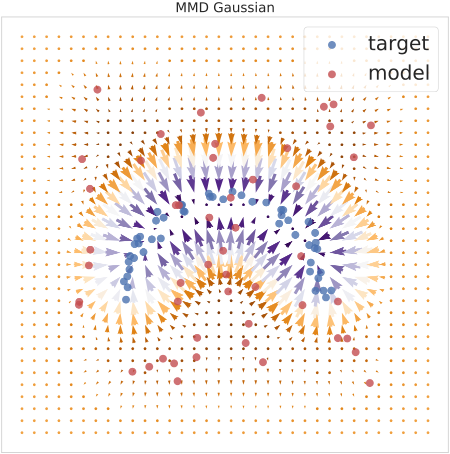
A simple 2-D example

Samples from target P and model Q



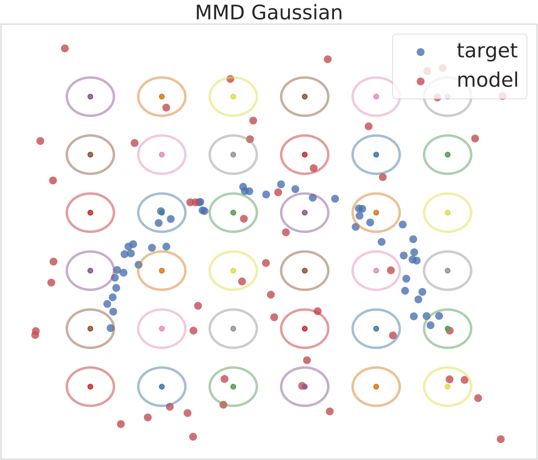
A simple 2-D example

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:

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

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

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:

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

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:

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

Maximise scaled MMD over critic features:

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

where

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

Replace expensive constraint with cheap upper bound:

$$\|f\|_S^2 \leq \sigma^{-1}_{P,\lambda} \; \|f\|_k^2$$

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:

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

where

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

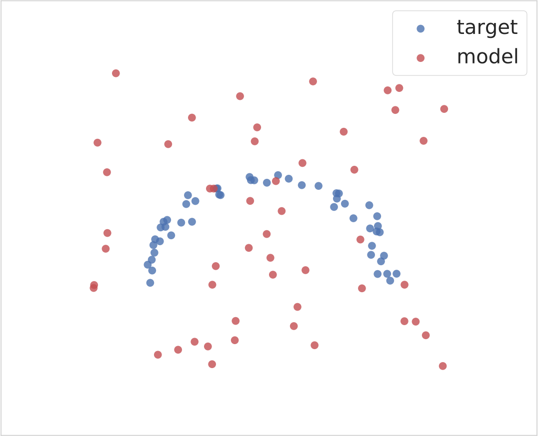
Replace expensive constraint with cheap upper bound:

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

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 $k(h_\psi(x), h_\psi(y))$ with features

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

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 $k(h_\psi(x), h_\psi(y))$

vector field movie, use Acrobat Reader to play

Simple 2-D example revisited

What the kenels $k(h_\psi(x), h_\psi(y))$ 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 kenels

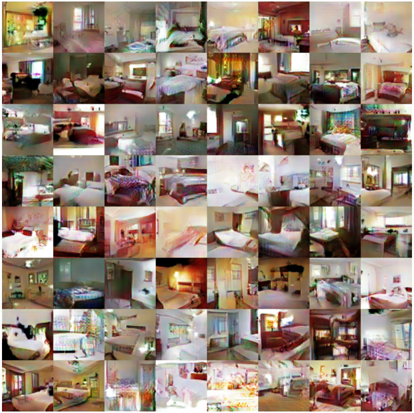
- Critic features from DCGAN: an f-filter critic has f, $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64.

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

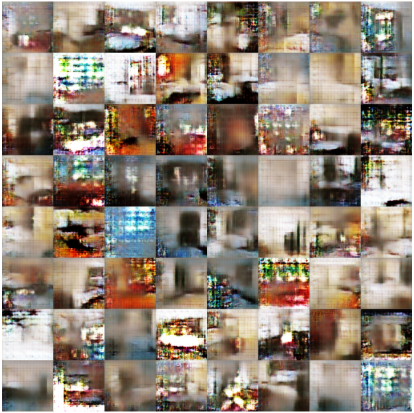
$h_\psi^\top(x) h_\psi(y)$, $f = 64$, KID=4

Linear vs nonlinear kenels

- 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 for MMD GANs

Scaled MMD vs Wasserstein-1 (NeurIPS 18)

Let $k_\psi = k \circ h_\psi$.

Wasserstein-1 bounds SMMD,

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

- Conditions on the neural network layers:
  - $h_\psi : \mathcal{X} \to \Re^s$ fully-connected L-layer network, Leaky-ReLU$_\alpha$ activations whose layers do not increase in width
  - Width of ℓth 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 $Q_k < \infty$.
  - μ is a probabilty 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 $k_\psi = k \circ h_\psi$.

Wasserstein-1 bounds SMMD,

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

- Conditions on the neural network layers:
  - $h_\psi : \mathcal{X} \rightarrow \Re^s$ fully-connected L-layer network, Leaky-ReLU$_\alpha$ activations whose layers do not increase in width
  - Width of ℓth 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 $Q_k < \infty$.
  - μ is a probabilty 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 $k_\psi = k \circ h_\psi$.

Wasserstein-1 bounds SMMD,

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

- Conditions on the neural network layers:
  - $h_\psi : \mathcal{X} \to \Re^s$ fully-connected L-layer network, Leaky-ReLU$_\alpha$ activations whose layers do not increase in width
  - Width of ℓth 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 $Q_k < \infty$.
  - μ is a probabilty 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,

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

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

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

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

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

Then

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

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

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

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

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

Then

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

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

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

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

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

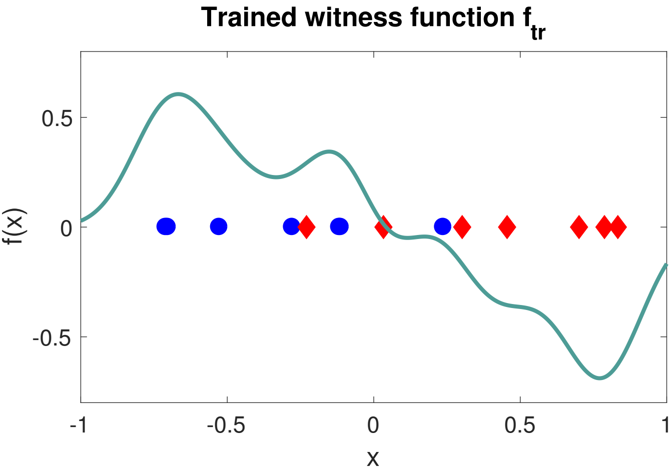
$$\left[\mathbf{E}_P f_{tr}(X) - \mathbf{E}_Q f_{tr}(Y) \right] \leq MMD(P, Q; F)$$

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

Same true for WGAN-GP.

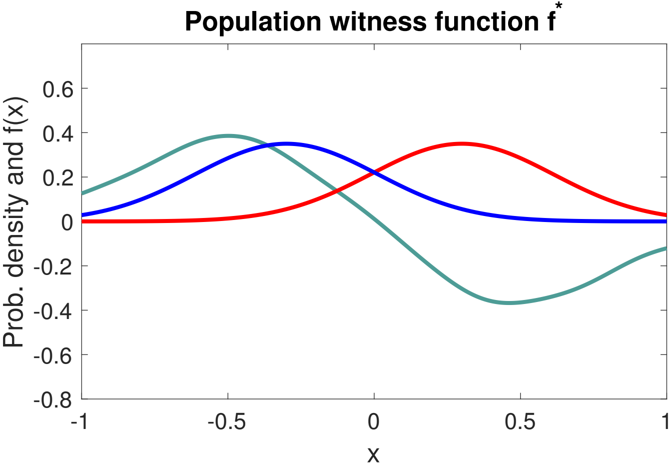
Bias of MMD GAN critic (ICLR 18)

Training minibatch critic function f_{tr}



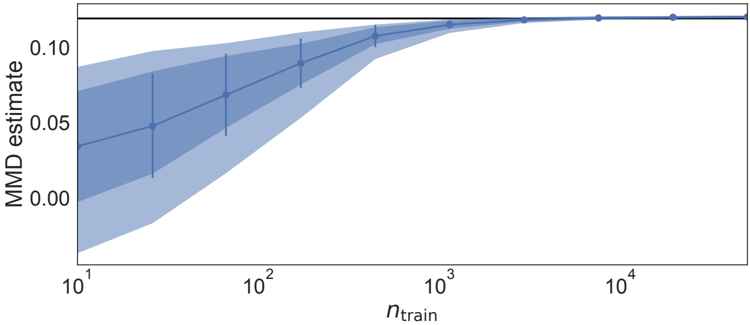
Bias of MMD GAN critic (ICLR 18)

Population critic function f^*



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]

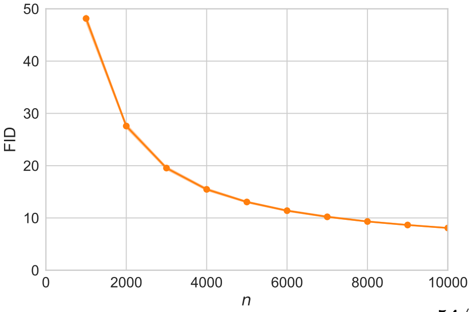
Fits Gaussians to features in the inception architecture (pool3 layer):

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

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

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

where $\Sigma = \frac{4}{d} C C^T$, 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)

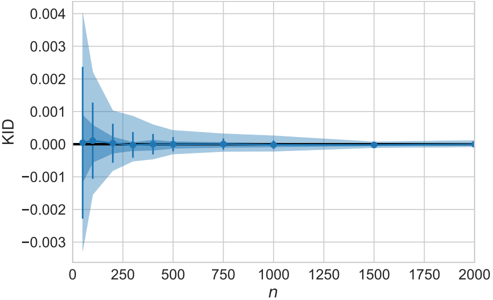
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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)

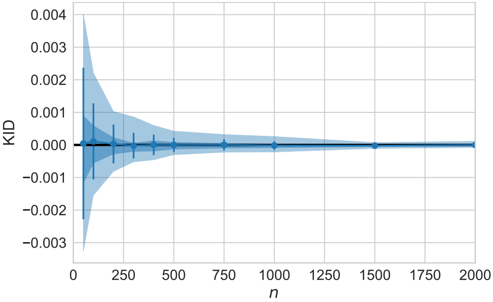
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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)

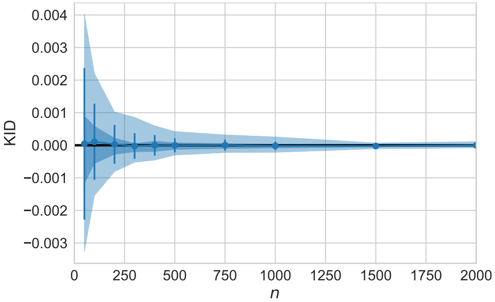
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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 Tensorflow implementation)!

The kernel inception distance (KID)

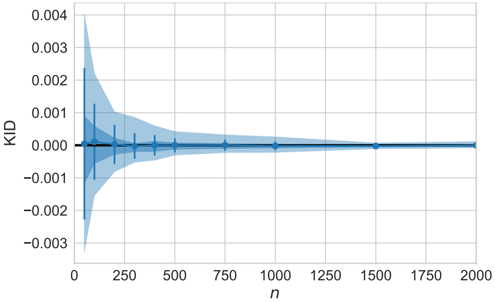
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

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 $KID(\widehat{P}_{t+1}, Q)$ 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
oyama masanori@gmail.com
yos 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.sercu1@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: unconditional imagenet 64×64

KID scores:

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

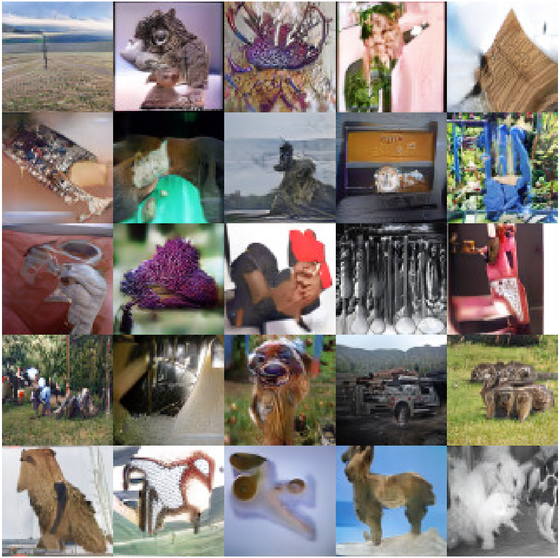
ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. 1000 classes.

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?


THE KERNEL
BREWERY LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT