Probability Divergences and Generative Models

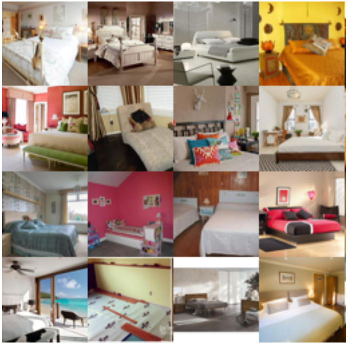
Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

MLSS Taipei, 2021

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

Role of divergence $D(P, Q)$?

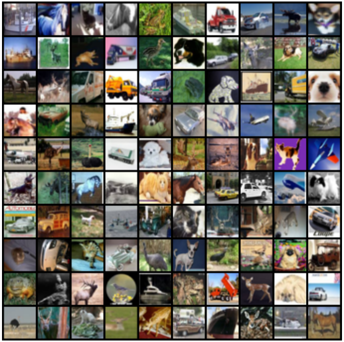
Testing for differences in samples

Given samples $X \sim P$ and $Y \sim Q$, are P and Q distinguishable (via $D(P, Q)$)?

- Application: detecting domain shift (did I train for the right task?)

CIFAR-10 test set (Krizhevsky 2009)

$X \sim P$



CIFAR-10.1 (Recht+ ICML 2019)

$Y \sim Q$

Outline

- Integral probability metrics (MMD, Wasserstein)

- ϕ-divergences (f-divergences) and a variational lower bound (KL)

- Generalized energy-based models
  - "Like a GAN" but incorporate critic into sample generation
  - Perform better than using generator alone

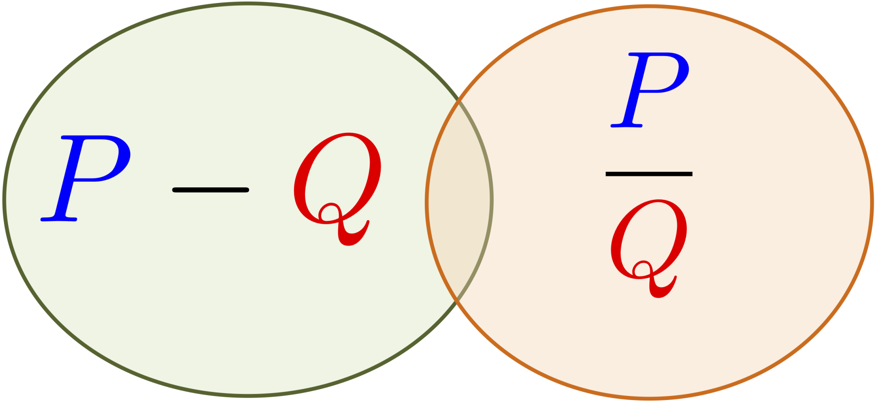
  Arbel, Zhou, G., Generalized Energy Based Models (ICLR 2021)

- Comparing samples with MMD

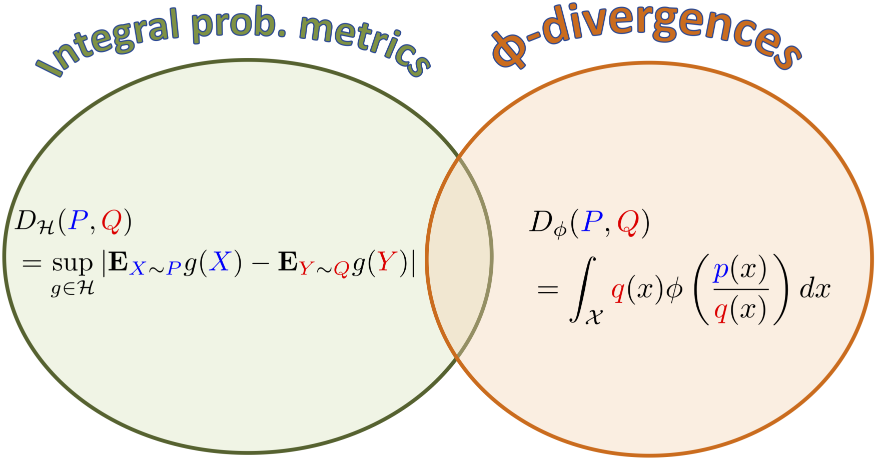
  Liu, Xu, Lu, Zhang, G. Sutherland, Learning Deep Kernels for Non-Parametric Two-Sample Tests (ICML 2020)

Divergence measures (critics)

Divergences



Divergences



$$D_{\mathcal{H}}(P, Q) = \sup_{g \in \mathcal{H}} \left| \mathbf{E}_{X \sim P} g(X) - \mathbf{E}_{Y \sim Q} g(Y) \right|$$

$$D_{\phi}(P, Q) = \int_{\mathcal{X}} q(x) \phi \left(\frac{p(x)}{q(x)} \right) dx$$

The Integral Probability Metrics

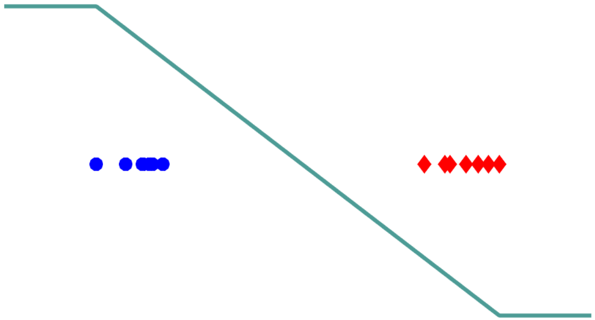
Wasserstein distance

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)

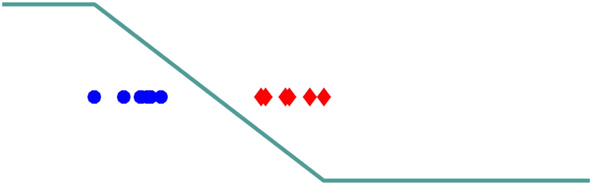
Wasserstein distance

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)

The Maximum Mean Discrepancy

Maximum mean discrepancy: smooth function for P vs Q

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

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

The Maximum Mean Discrepancy

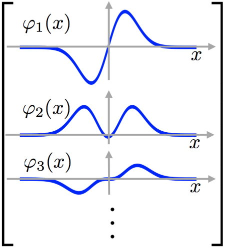
Maximum mean discrepancy: smooth function for P vs Q

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

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

Functions are linear combinations of features:

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



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

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,

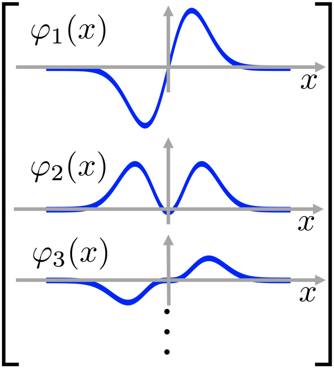
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

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

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

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

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

- 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

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

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

Expectations of functions are linear combinations of expected features

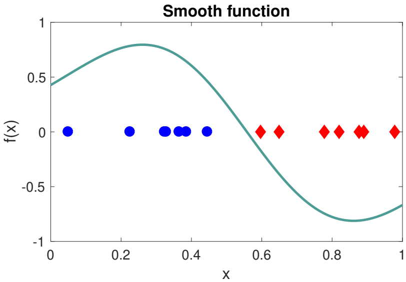
$$\mathrm{E}_P(f(X)) = \langle f, \mathrm{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

Integral prob. metric vs feature mean difference

The MMD:

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



Integral prob. metric vs feature mean difference

The MMD:

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

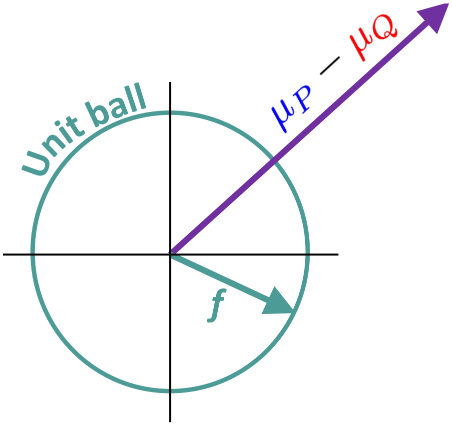
use

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

Integral prob. metric vs feature mean difference

The MMD:

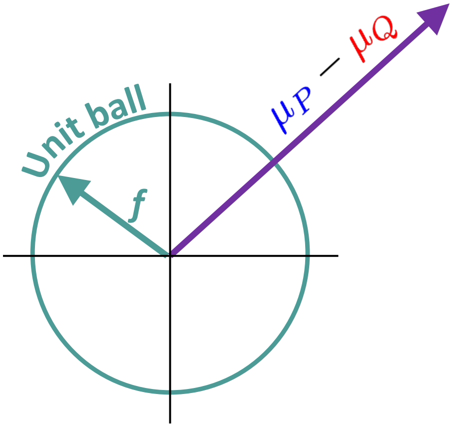
$$MMD(P, Q; F)$$
$$= \sup_{\|f\| \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right]$$
$$= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature mean difference

The MMD:

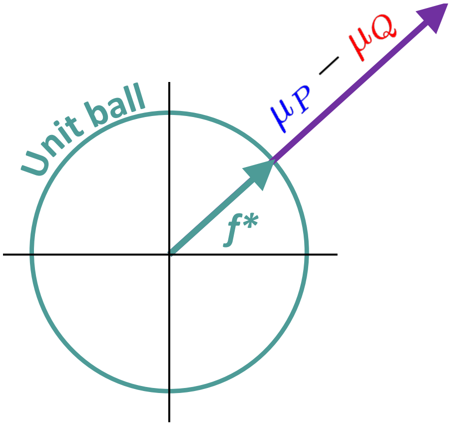
$$MMD(P, Q; F)$$
$$= \sup_{\|f\| \leq 1} [\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)]$$
$$= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature mean difference

The MMD:

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



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

Integral prob. metric vs feature mean difference

The MMD:

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

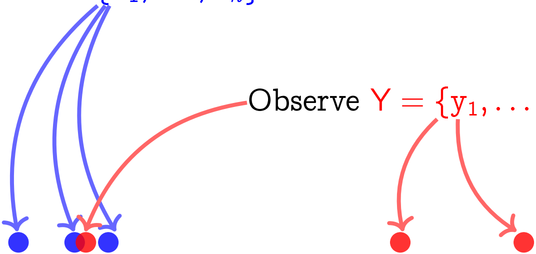
IPM view equivalent to feature mean difference (kernel case only)

Construction of MMD witness

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

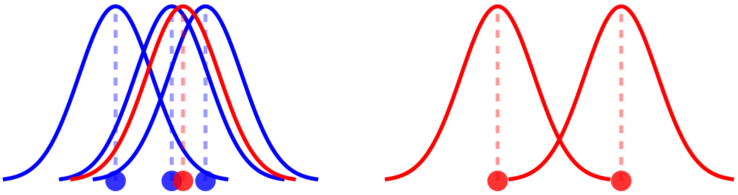
Observe $\mathsf{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\} \sim P$

Observe $\mathsf{Y} = \{\mathrm{y}_1, \ldots, \mathrm{y}_n\} \sim Q$



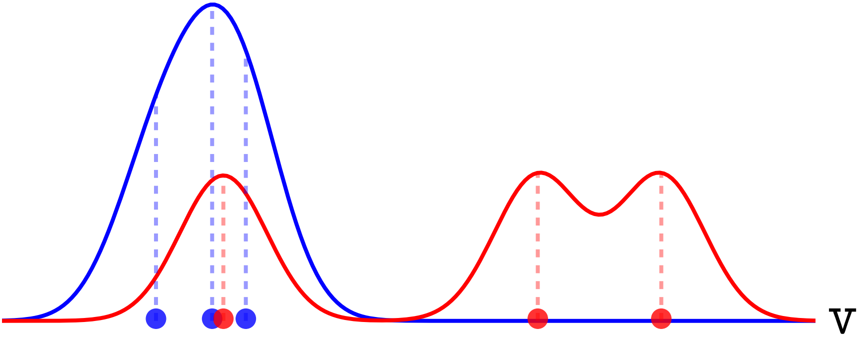
Construction of MMD witness

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



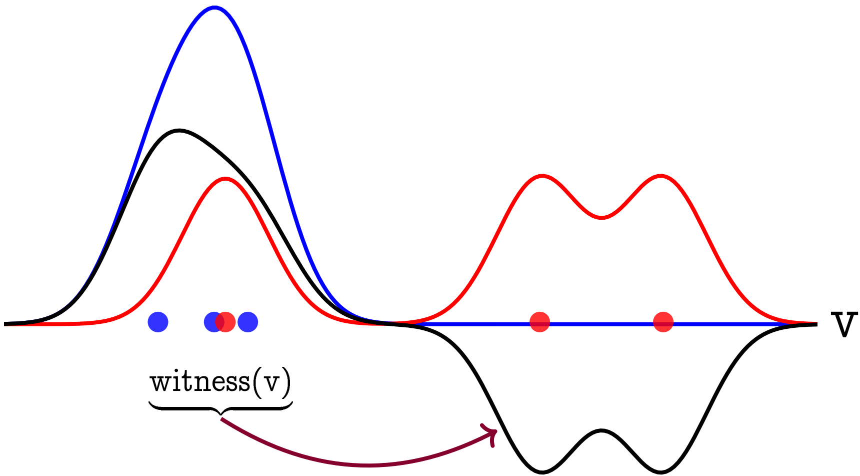
Construction of MMD witness

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



Construction of MMD witness

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



Derivation of empirical witness function

Recall the witness function expression

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

Derivation of empirical witness function

Recall the witness function expression

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

The empirical feature mean for P

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

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

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

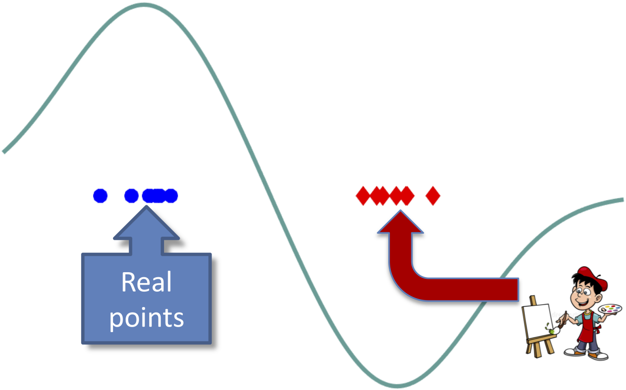
Don't need explicit feature coefficients $f^* := \begin{bmatrix} f_1^* & f_2^* & \dots \end{bmatrix}$

Maximum mean discrepancy

A helpful critic:

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

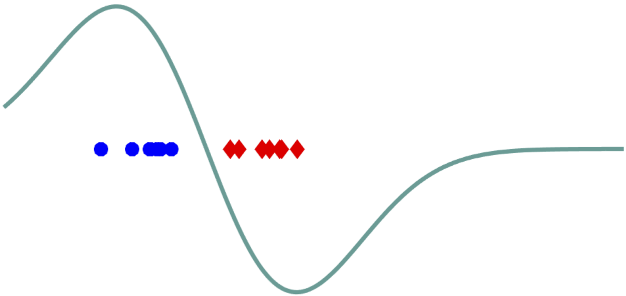
MMD=1.8



Maximum mean discrepancy

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

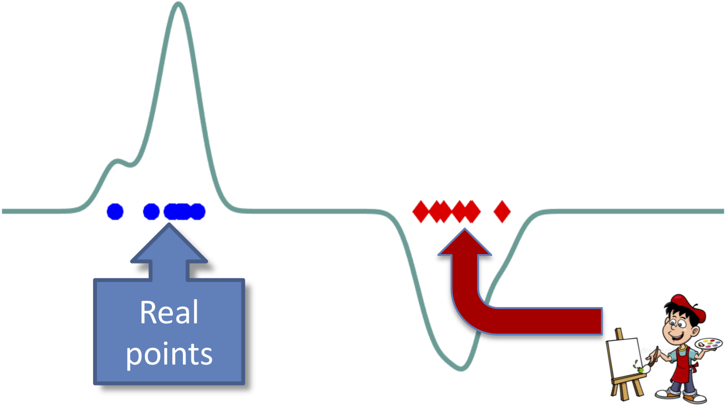
MMD=1.1



Maximum mean discrepancy

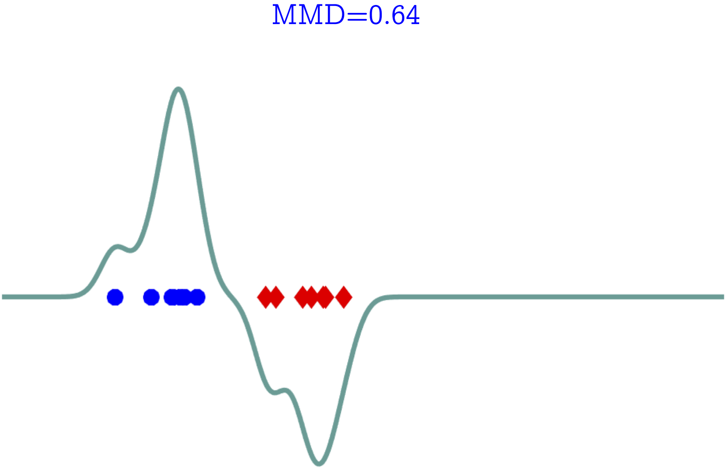
An unhelpful critic:
$MMD(P, Q)$ with a narrow kernel.

MMD=0.64

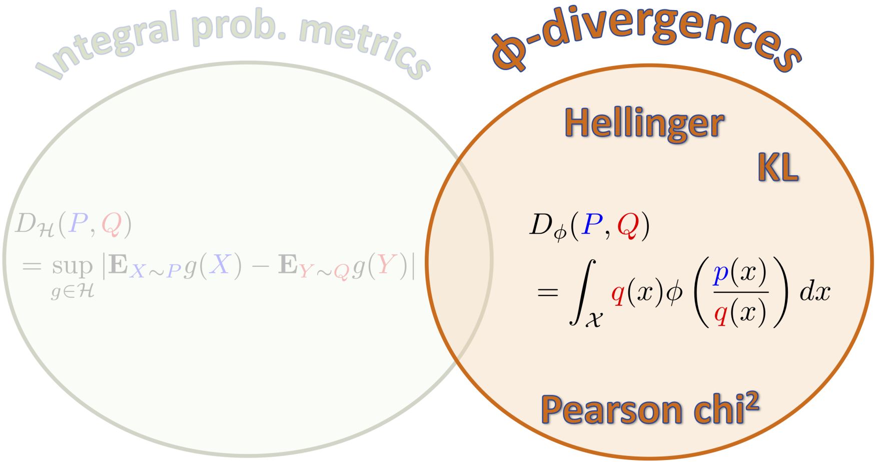


Maximum mean discrepancy

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



The ϕ-divergences



The ϕ-divergences

Define the ϕ-divergence(f-divergence):

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

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

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

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

The ϕ-divergences

Define the ϕ-divergence(f-divergence):

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

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

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

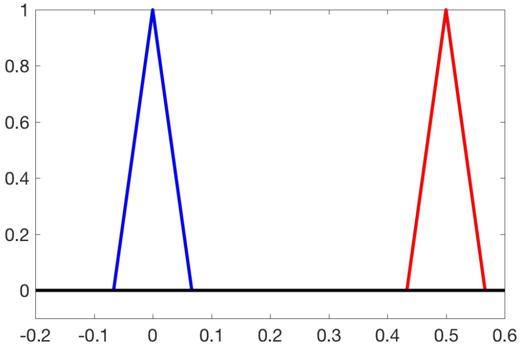
$$\begin{aligned} D_{KL}(P, Q) &= \int \log\left(\frac{p(z)}{q(z)}\right) p(z)dz \\ &= \int \left(\frac{p(z)}{q(z)}\right) \log\left(\frac{p(z)}{q(z)}\right) q(z)dz \end{aligned}$$

Are ϕ-divergences good critics?

Simple example: disjoint support.

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

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

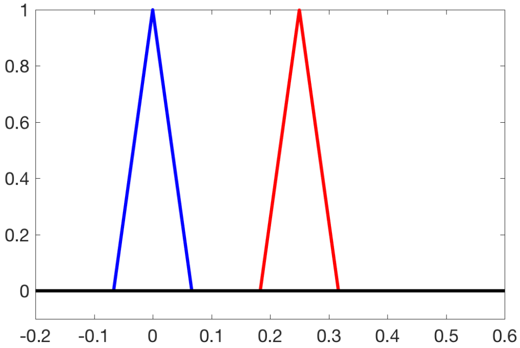


Are ϕ-divergences good critics?

Simple example: disjoint support.

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

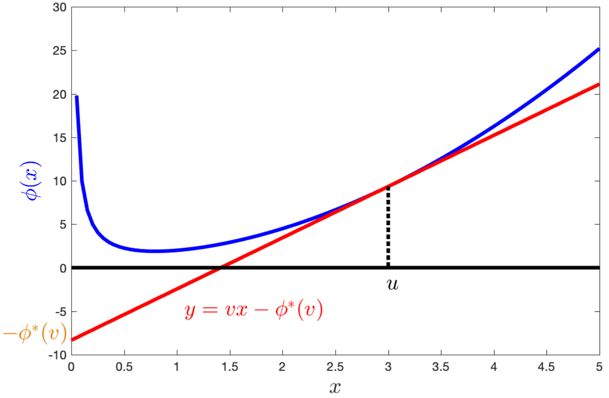
$$D_{KL}(P, Q) = \infty \qquad D_{JS}(P, Q) = \log 2$$



ϕ-divergences in practice

Background: the conjugate (Fenchel) dual

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



- $\phi^*(v)$ is negative intercept of tangent to ϕ with slope v

ϕ-divergences in practice

Background: the conjugate (Fenchel) dual

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

- For a convex l.s.c. ϕ we have

$$\phi^{**}(x) = \phi(x) = \sup_{v \in \mathbb{R}} \{xv - \phi^*(v)\}$$

ϕ-divergences in practice

Background: the conjugate (Fenchel) dual

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

- For a convex l.s.c. ϕ we have

$$\phi^{**}(x) = \phi(x) = \sup_{v \in \mathbb{R}} \{xv - \phi^*(v)\}$$

- KL divergence:

$$\phi(x) = x \log(x) \qquad \phi^*(v) = \exp(v - 1)$$

A variational lower bound

A lower-bound ϕ-divergence approximation:

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

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

A variational lower bound

A lower-bound ϕ-divergence approximation:

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

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

$\phi^*(v)$ is dual of $\phi(x)$.

A variational lower bound

A lower-bound ϕ-divergence approximation:

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

(restrict the function class)

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

A variational lower bound

A lower-bound ϕ-divergence approximation:

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

(restrict the function class)

Bound tight when:

$$f^\diamond(z) = \partial\phi\left(\frac{p(z)}{q(z)}\right)$$

if ratio defined.

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

Case of the KL

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

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

Case of the KL

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

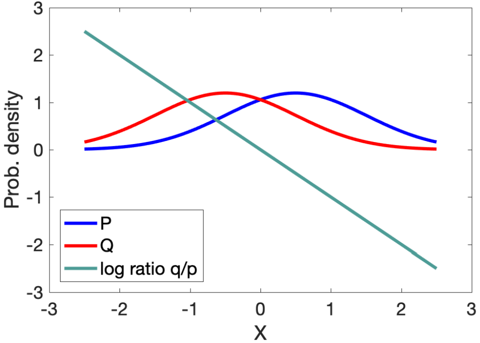
$$\geq \sup_{f \in \mathcal{H}} -\mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \underbrace{\exp\left(-f(Y)\right)}_{\phi^*(-f(Y)+1)}$$

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

Case of the KL

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

$$\geq \sup_{f \in \mathcal{H}} -\mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp(-f(Y))$$



Bound tight when:

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

if ratio defined.

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

Case of the KL

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

$$\geq \sup_{f \in \mathcal{H}} -\mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(-f(Y)\right)$$

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

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

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

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

Case of the KL

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

$$\geq \sup_{f \in \mathcal{H}} -\mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(-f(Y)\right)$$

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

This is a

KL

Approximate

Lower-bound

Estimator.

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

Case of the KL

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

$$\geq \sup_{f \in \mathcal{H}} -\mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(-f(Y)\right)$$

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

This is a

K

A

L

E

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

Case of the KL

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

$$\geq \sup_{f \in \mathcal{H}} -\mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(-f(Y)\right)$$

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

The KALE divergence

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

Empirical properties of KALE

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

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

$\|w\|^2_{\mathcal{H}}$ penalized :

Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (arXiv, 2021, Section 2)

Empirical properties of KALE

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

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

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

Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (arXiv, 2021, Section 2)

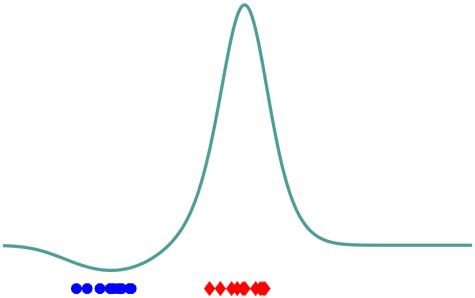
Empirical properties of KALE

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

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

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

$$KALE(Q, P; \mathcal{H}) = 0.18$$



Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (arXiv, 2021, Section 2)

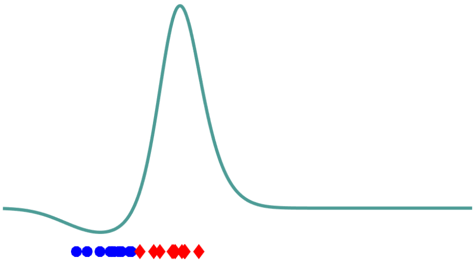
Empirical properties of KALE

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

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

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

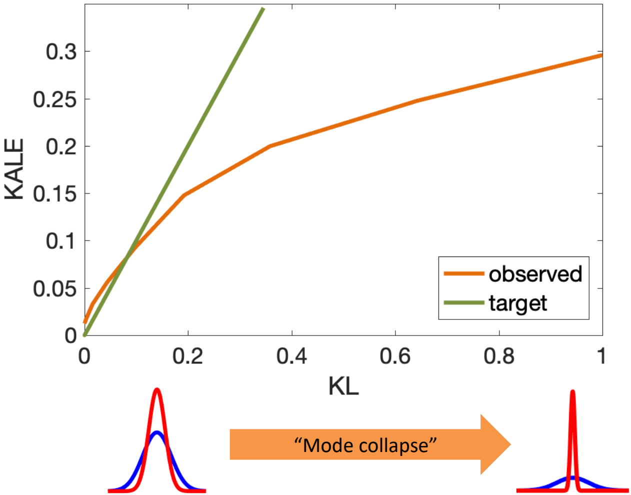
$$KALE(Q, P; \mathcal{H}) = 0.12$$



Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (arXiv, 2021, Section 2)

The KALE smoothie and “mode collapse”

- Two Gaussians with same means, different variance



Example thanks to M. Arbel and M. Rosca

Topological properties of KALE (1)

Key requirements on $\mathcal{H}$ and $\mathcal{X}$:

- Compact domain $\mathcal{X}$,
- $\mathcal{H}$ dense in the space $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$ wrt $\|\cdot\|_\infty$.
- If $f \in \mathcal{H}$ then $-f \in \mathcal{H}$ and $cf \in \mathcal{H}$ for $0 \leq c \leq C_{\max}$.

Theorem: $KALE(P, Q; \mathcal{H}) \geq 0$ and $KALE(P, Q; \mathcal{H}) = 0$ iff $P = Q$.

Zhang, Liu, Zhou, Xu, and He. "On the Discrimination-Generalization Tradeoff in GANs" (ICLR 2018, Corollary 2.4; Theorem B.1)
Arbel, Liang, G. (ICLR 2021, Proposition 1)

Topological properties of KALE (1)

Key requirements on $\mathcal{H}$ and $\mathcal{X}$:

- Compact domain $\mathcal{X}$,
- $\mathcal{H}$ dense in the space $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$ wrt $\|\cdot\|_\infty$.
- If $f \in \mathcal{H}$ then $-f \in \mathcal{H}$ and $cf \in \mathcal{H}$ for $0 \leq c \leq C_{\mathrm{max}}$.

Theorem: $KALE(P, Q; \mathcal{H}) \geq 0$ and $KALE(P, Q; \mathcal{H}) = 0$ iff $P = Q$.

$\mathcal{H}$ dense in $C(\mathcal{X})$ for $\mathcal{X} \subset \mathbb{R}^d$ when:

$$\mathcal{H} = \mathrm{span}\{\sigma(w\top x + b) : [w, b] \in \Theta\}$$

$\sigma(u) = \max\{u, 0\}^\alpha$, $\alpha \in \mathbb{N}$, and $\{\lambda\theta : \lambda \geq 0, \theta \in \Theta\} = \mathbb{R}^{d+1}$.

Zhang, Liu, Zhou, Xu, and He. "On the Discrimination-Generalization Tradeoff in GANs" (ICLR 2018, Corollary 2.4; Theorem B.1)
Arbel, Liang, G. (ICLR 2021, Proposition 1)

Topological properties of KALE (2)

Additional requirement: all functions in $\mathcal{H}$ Lipschitz in their inputs with constant L

Theorem: $KALE(P, Q^n; \mathcal{H}) \to 0$ iff $Q^n \to P$ under the weak topology.

Liu, Bousquet, Chaudhuri. "Approximation and Convergence Properties of Generative Adversarial Learning" (NeurIPS 2017); Arbel, Liang, G. (ICLR 2021, Proposition 1)

Topological properties of KALE (2)

Additional requirement: all functions in $\mathcal{H}$ Lipschitz in their inputs with constant L

Theorem: $KALE(P, Q^n; \mathcal{H}) \to 0$ iff $Q^n \to P$ under the weak topology.

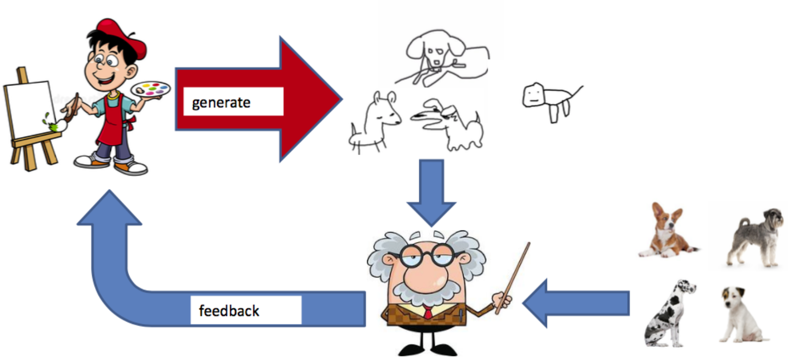
Partial proof idea:

$$\begin{aligned}
KALE(P, Q; \mathcal{H}) &= -\int f \, dP - \int \exp(-f) \, dQ + 1 \\
&= \int f(x) \, dQ(x) - f(x') \, dP(x') \\
&\quad - \int \underbrace{(\exp(-f) + f - 1)}_{\geq 0} \, dQ \\
&\leq \int f(x) \, dQ(x) - f(x') \, dP(x') \leq LW_1(P, Q)
\end{aligned}$$

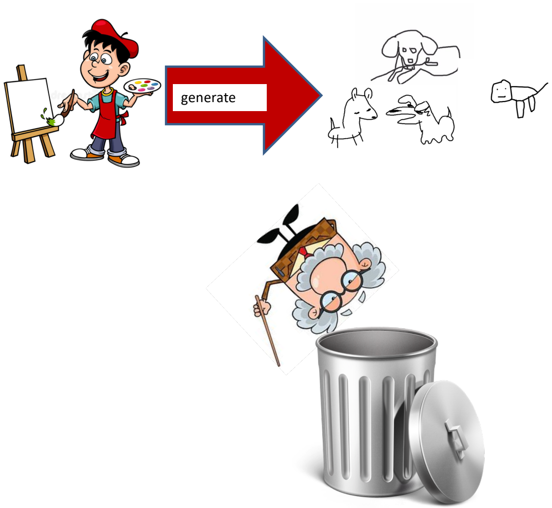
Liu, Bousquet, Chaudhuri. “Approximation and Convergence Properties of Generative Adversarial Learning” (NeurIPS 2017); Arbel, Liang, G. (ICLR 2021, Proposition 1)

How to train your ~~GAN~~
Generalized Energy-Based Model

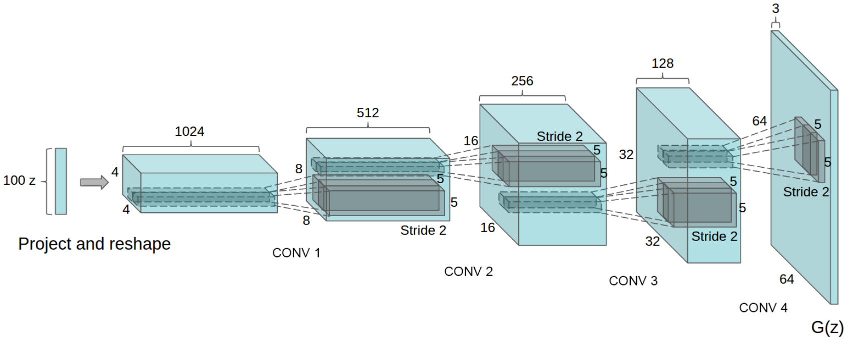
Visual notation: GAN setting



Visual notation: GAN setting



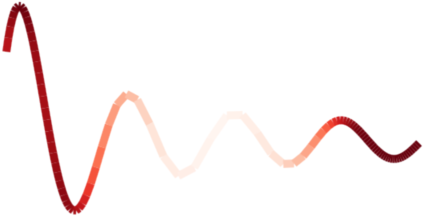
Reminder: the generator



Radford, Metz, Chintala, ICLR 2016

Generalized energy-based models: illustration

Target distribution P

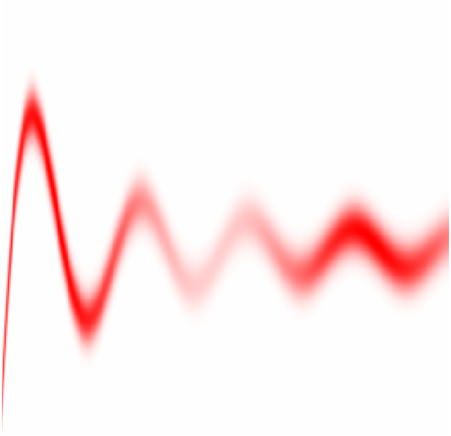


$$z \sim Unif[0,1]$$
$$\tilde{z} = \tau(z)$$
$$X = G_{\theta^\star}(\tilde{z}), \quad X_1 = \tilde{z}$$

Example thanks to M. Arbel

Generalized energy-based models: illustration

EBM approximation to target:



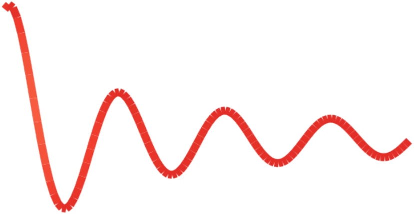
$$p(X) \propto \exp(-E(X))$$

$$E(X) = \frac{1}{2\sigma^2} \| G_\theta(X_1) - X \|^2 + A_\theta(X_1)$$

Example thanks to M. Arbel

Generalized energy-based models: illustration

GAN (generator) distribution Q_θ

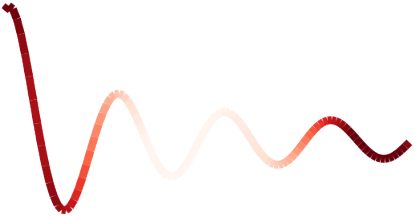


Generator	Critic
$z \sim unif[0,1]$	$MLP(X)$
$X = B_\theta(z)$	

Example thanks to M. Arbel

Generalized energy-based models: illustration

Mass of GEBM corrected by critic



Generator

$z \sim unif[0, 1]$

$X = B_\theta(z)$

Re-weight using importance weights defined by energy:

$$w(x) \propto \exp(-E(x))$$

Example thanks to M. Arbel

Generalized energy-based models

Define a model $Q_{B_\theta, E}$ as follows:

- Sample from generator with parameters θ

$$X \sim Q_\theta \quad \Longleftrightarrow \quad X = B_\theta(Z), \quad Z \sim \eta$$

- Reweight the samples according to importance weights:

$$f_{Q,E}(x) = \frac{\exp(-E(x))}{Z_{Q_\theta,E}}, \qquad Z_{Q,E} = \int \exp(-E(x))\, d\,Q_\theta(x),$$

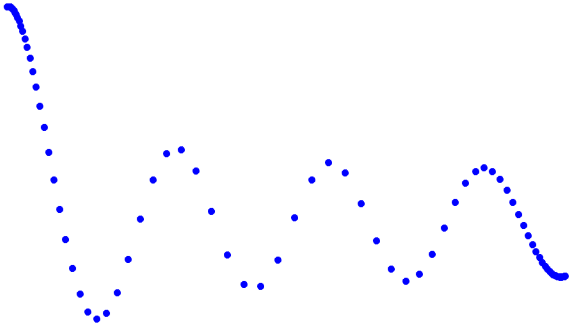
  where $E \in \mathcal{E}$, the energy function class.

  $f_{Q,E}(x)$ is Radon-Nikodym derivative of $Q_{B_\theta, E}$ wrt Q_θ.

- When Q_θ has density wrt Lebesgue on $\mathcal{X}$, this is a standard energy-based model.

The energy function, on our example

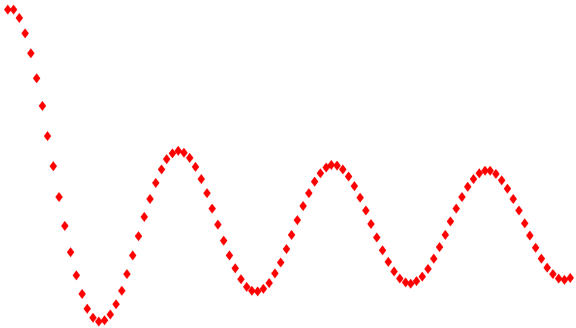
Target distribution P



Example thanks to M. Arbel

The energy function, on our example

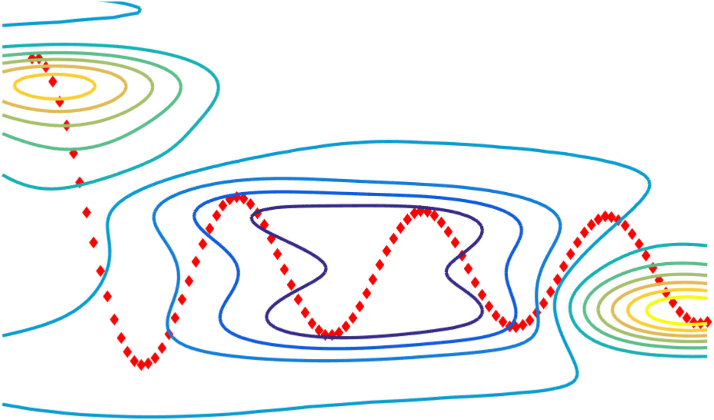
GAN (generator) Q_θ, correct support but wrong mass



Example thanks to M. Arbel

The energy function, on our example

Log energy function and Q_θ



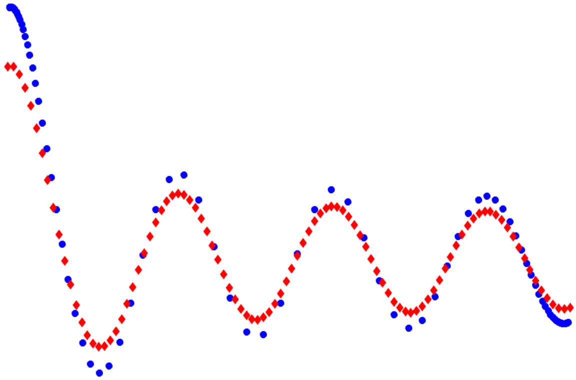
Key:

- Orange: increase mass
- Blue: reduce mass

Example thanks to M. Arbel

The energy function, on our example

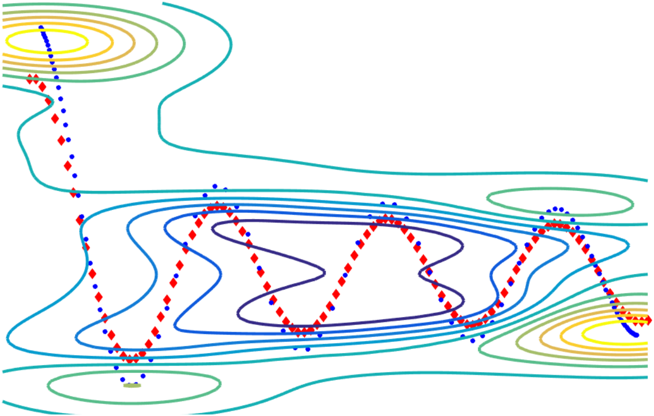
Target distribution P and GAN (generator) Q_θ, wrong support and wrong mass



Example thanks to M. Arbel

The energy function, on our example

Log energy function, P, and Q_θ



Key:

- Orange: increase weight
- Blue: reduce weight

Example thanks to M. Arbel

How do we learn the energy E?

How do we learn the energy E?

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log Z_{Q,E}$$

- When $KL(P, Q_\theta)$ well defined, above is Donsker-Varadhan lower bound on KL
  - tight when $E(z) = -\log(p(z)/q(z))$.
- However, Generalized Log-Likelihood still defined when P and Q_θ mutually singular (as long as E smooth)!

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log \int \exp(-E)\, dQ_{\theta}$$

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log \int \exp(-E)\, dQ_\theta$$

One last trick...(convexity of exponential)

$$-\log \int \exp(-E)\, dQ_\theta \geq -c - e^{-c} \int \exp(-E)\, dQ_\theta + 1$$

tight whenever $c = \log \int \exp(-E)\, dQ_\theta$.

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log \int \exp(-E)\, dQ_\theta$$

One last trick...(convexity of exponential)

$$-\log \int \exp(-E)\, dQ_\theta \geq -c - e^{-c} \int \exp(-E)\, dQ_\theta + 1$$

tight whenever $c = \log \int \exp(-E)\, dQ_\theta$.

Generalized Log-Likelihood has the lower bound:

$$\mathcal{L}_{P,Q}(E) \geq -\int (E + c)\, dP - \int \exp(-E - c)\, dQ_\theta + 1$$

$$:= \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$$

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log \int \exp(-E)\, dQ_\theta$$

One last trick...(convexity of exponential)

$$-\log \int \exp(-E)\, dQ_\theta \geq -c - e^{-c} \int \exp(-E)\, dQ_\theta + 1$$

tight whenever $c = \log \int \exp(-E)\, dQ_\theta$.

Generalized Log-Likelihood has the lower bound:

$$\mathcal{L}_{P,Q}(E) \geq -\int (E + c)\, dP - \int \exp(-E - c)\, dQ_\theta + 1$$

$$:= \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$$

This is the KALE! with function class $\mathcal{E} + \mathbb{R}$.

KALE and the energy function

Fit the model using Generalized Log-Likelihood:

$$\mathcal{L}_{P,Q}(E) := \int \log(f_{Q,E})\, dP = -\int E\, dP - \log \int \exp(-E)\, dQ_\theta$$

One last trick...(convexity of exponential)

$$-\log \int \exp(-E)\, dQ_\theta \geq -c - e^{-c} \int \exp(-E)\, dQ_\theta + 1$$

tight whenever $c = \log \int \exp(-E)\, dQ_\theta$.

Generalized Log-Likelihood has the lower bound:

$$\begin{aligned}\mathcal{L}_{P,Q}(E) &\geq -\int (E + c)\, dP - \int \exp(-E - c)\, dQ_\theta + 1 \\ &:= \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})\end{aligned}$$

Jointly maximizing yields the maximum likelihood energy E^* and corresponding $c^* = \log \int \exp(-E)\, dQ_\theta$.

Training the base measure (generator)

Recall the generator:

$$X = B_\theta(Z), \quad Z \sim \eta$$

Define: $\mathcal{K}(\theta) := \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$

Training the base measure (generator)

Recall the generator:

$$X = B_\theta(Z), \quad Z \sim \eta$$

Define: $\mathcal{K}(\theta) := \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$

Theorem: $\mathcal{K}$ is lipschitz and differentiable for almost all $\theta \in \Theta$ with:

$$\nabla \mathcal{K}(\theta) = Z_{Q,E^*}^{-1} \int \nabla_x E^*(B_\theta(z)) \nabla_\theta B_\theta(z) \exp(-E^*(B_\theta(z)))\eta(z)dz.$$

where E^* achieves supremum in $\mathcal{F}(P, Q; \mathcal{E} + \mathbb{R})$.

Training the base measure (generator)

Recall the generator:

$$X = B_\theta(Z), \quad Z \sim \eta$$

Define: $\mathcal{K}(\theta) := \mathcal{F}(P, Q_\theta; \mathcal{E} + \mathbb{R})$

Theorem: $\mathcal{K}$ is lipschitz and differentiable for almost all $\theta \in \Theta$ with:

$$\nabla \mathcal{K}(\theta) = Z_{Q,E^*}^{-1} \int \nabla_x E^*(B_\theta(z)) \nabla_\theta B_\theta(z) \exp(-E^*(B_\theta(z)))\eta(z) dz.$$

where E^* achieves supremum in $\mathcal{F}(P, Q; \mathcal{E} + \mathbb{R})$.

Assumptions:

- Functions in $\mathcal{E}$ parametrized by $\psi \in \Psi$, where Ψ compact,
  - jointly continous w.r.t. (ψ, x), L-lipschitz and L-smooth w.r.t. x.
- $(\theta, z) \mapsto B_\theta(z)$ jointly continuous wrt (θ, z), $z \mapsto B_\theta(z)$ uniformly Lipschitz w.r.t. z, lipschitz and smooth wrt θ (see paper: constants depend on z)

Sampling from the model

Consider end-to-end model $Q_{B_\theta,E}$, where recall that
$X = B_\theta(Z), \quad Z \sim \eta,$

$$f_{B,E}(x) := \frac{\exp(-E(x))}{Z_{Q,E}}$$

Sampling from the model

Consider end-to-end model $Q_{B_\theta,E}$, where recall that
$X = B_\theta(Z), \quad Z \sim \eta,$

$$f_{B,E}(x) := \frac{\exp(-E(x))}{Z_{Q,E}}$$

For a test function g,

$$\int g(x)dQ_{B,E}(x) = \int g(B(z))f_{B,E}(B(z))\eta(z)dz$$

Posterior latent distribution therefore

$$\nu_{B,E}(z) = \eta(z)f_{B,E}(B(z))$$

Sampling from the model

Consider end-to-end model $Q_{B_\theta,E}$, where recall that
$X = B_\theta(Z), \quad Z \sim \eta,$

$$f_{B,E}(x) := \frac{\exp(-E(x))}{Z_{Q,E}}$$

For a test function g,

$$\int g(x)\, dQ_{B,E}(x) = \int g(B(z)) f_{B,E}(B(z)) \eta(z)\, dz$$

Posterior latent distribution therefore

$$\nu_{B,E}(z) = \eta(z) f_{B,E}(B(z))$$

Sample $z \sim \nu_{B,E}$ via Langevin diffusion-derived algorithms (MALA, ULA, HMC,...) to exploit gradient information.

Generate new samples in $\mathcal{X}$ via

$$X \sim Q_{B,E} \quad \Longleftrightarrow \quad Z \sim \nu_{B,E}, \quad X = B_\theta(Z).$$

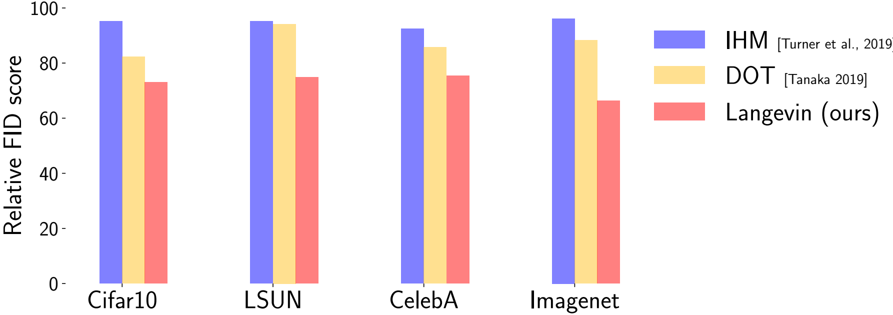
Experiments

Examples: sampling at modes

Tempered GEBM Cifar10 samples at different stages of sampling using a Kinetic Langevin Algorithm (KLA). Early samples $\rightarrow$ late samples. Model run at *low temperature* ($\beta = 100$) for better quality samples.

Sampling at modes: results

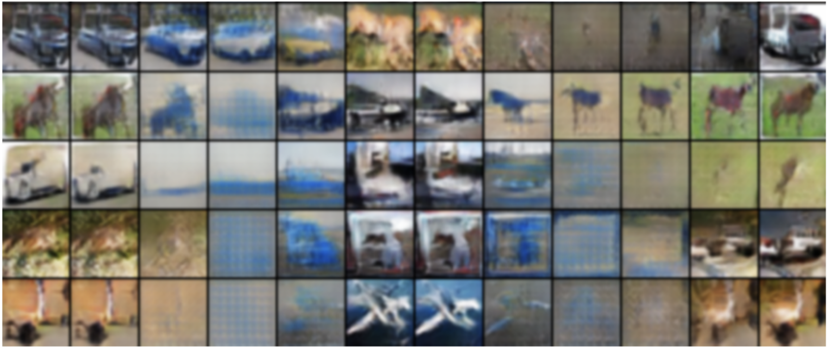
The relative FID score: $\frac{\mathrm{FID}(Q_{B_\theta,E})}{\mathrm{FID}(B_\theta)}$



For a given generator B_θ and energy E, samples always better (FID score) than generator alone.

Examples: moving between modes

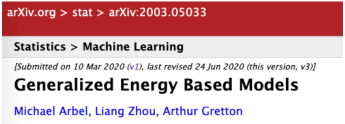
Tempered GEBM Cifar10 samples at different stages of sampling using KLA. Early samples $\rightarrow$ late samples.
Model run at *lower friction* (but still low temperature, $\beta = 100$) for mode exploration.



Summary

- Generalized energy based model:
  - End-to-end model incorporating generator and critic
  - Always better samples than generator alone.
- ICLR 2021

https://github.com/MichaelArbel/GeneralizedEBM



Summary

- Generalized energy based model:
  - End-to-end model incorporating generator and critic
  - Always better samples than generator alone.
- ICLR 2021

https://github.com/MichaelArbel/GeneralizedEBM

arXiv.org > stat > arXiv:2003.05033

Statistics > Machine Learning

[Submitted on 10 Mar 2020 (v1), last revised 24 Jun 2020 (this version, v3)]

Generalized Energy Based Models

Michael Arbel, Liang Zhou, Arthur Gretton

NeurIPS 2020:

arXiv.org > cs > arXiv:2003.06060

Computer Science > Machine Learning

[Submitted on 12 Mar 2020 (v1), last revised 24 Mar 2020 (this version, v2)]

Your GAN is Secretly an Energy-based Model and You Should use Discriminator Driven Latent Sampling

Tong Che, Ruixiang Zhang, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, Yoshua Bengio

ICLR 2021:

arXiv.org > cs > arXiv:2012.00780

Computer Science > Machine Learning

[Submitted on 1 Dec 2020 (v1), last revised 5 Jun 2021 (this version, v4)]

Refining Deep Generative Models via Discriminator Gradient Flow

Abdul Fatir Ansari, Ming Liang Ang, Harold Soh

ICLR 2021:

arXiv.org > cs > arXiv:2010.00654

Computer Science > Machine Learning

[Submitted on 1 Oct 2020 (v1), last revised 9 Feb 2021 (this version, v2)]

VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models

Zhisheng Xiao, Karsten Kreis, Jan Kautz, Arash Vahdat

How to find the best kernel for MMD

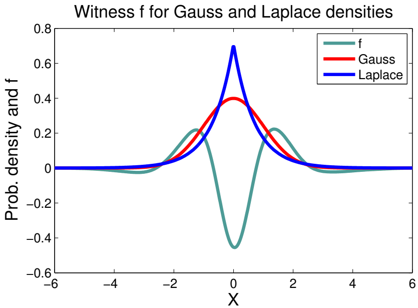
Integral prob. metric vs feature difference

The MMD:

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

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

$$= \|\mu_P - \mu_Q\|_{\mathcal{F}}$$



The maximum mean discrepancy

The maximum mean discrepancy in terms of kernel means:

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

The maximum mean discrepancy

The maximum mean discrepancy in terms of kernel means:

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

The maximum mean discrepancy

The maximum mean discrepancy in terms of kernel means:

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

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

The maximum mean discrepancy

The maximum mean discrepancy in terms of kernel means:

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

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 $k(\text{fish}_i, \text{fish}_j)$

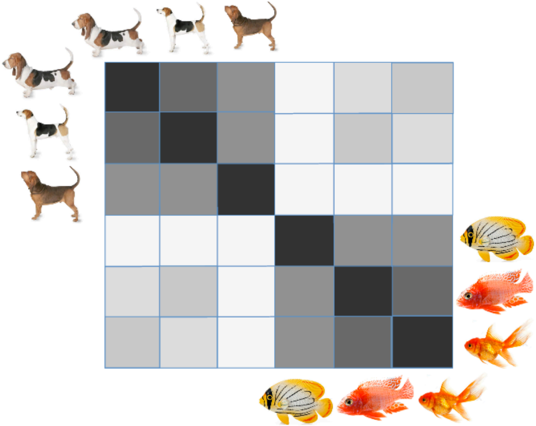
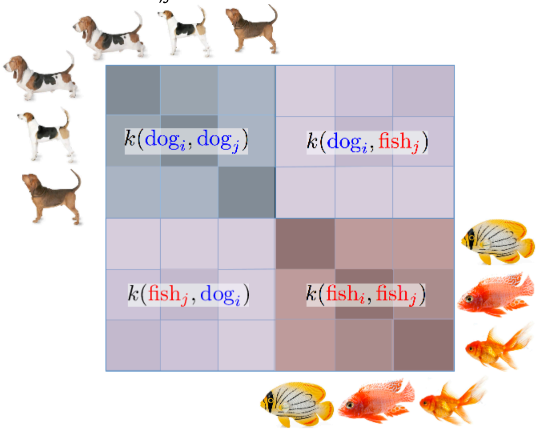


Illustration of MMD

The maximum mean discrepancy:

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



A statistical test using MMD

The empirical MMD:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{y}_i, \mathrm{y}_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, \mathrm{y}_j)$$

How does this help decide whether $P = Q$?

A statistical test using MMD

The empirical MMD:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{y}_i, \mathrm{y}_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, \mathrm{y}_j)$$

Perspective from statistical hypothesis testing:

- Null hypothesis $\mathcal{H}_0$ when $P = Q$
  - should see $\widehat{MMD}^2$ "close to zero".
- Alternative hypothesis $\mathcal{H}_1$ when $P \neq Q$
  - should see $\widehat{MMD}^2$ "far from zero"

A statistical test using MMD

The empirical MMD:

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)$$

Perspective from statistical hypothesis testing:

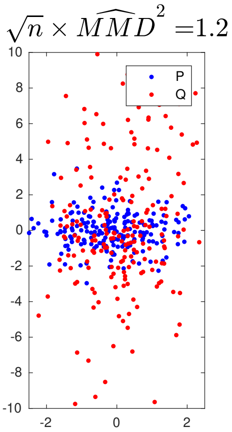
- Null hypothesis $\mathcal{H}_0$ when $P = Q$
  - should see $\widehat{MMD}^2$ "close to zero".
- Alternative hypothesis $\mathcal{H}_1$ when $P \neq Q$
  - should see $\widehat{MMD}^2$ "far from zero"

Want Threshold c_α for $\widehat{MMD}^2$ to get false positive rate α

Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Draw $n = 200$ i.i.d samples from P and Q

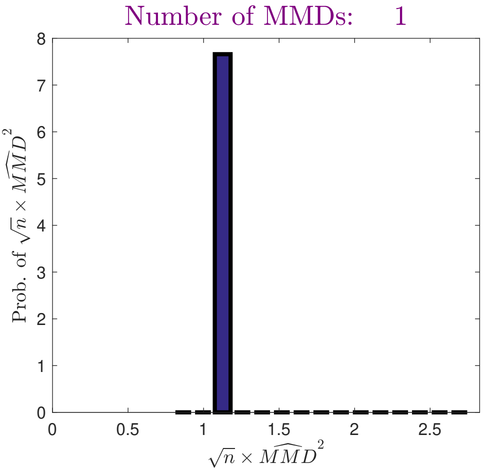
- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.2$

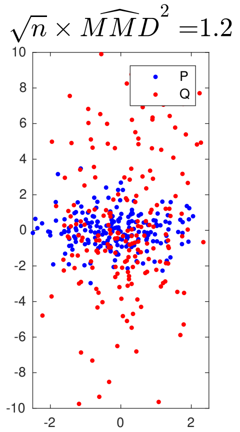


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Draw $n = 200$ i.i.d samples from P and Q

- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.2$

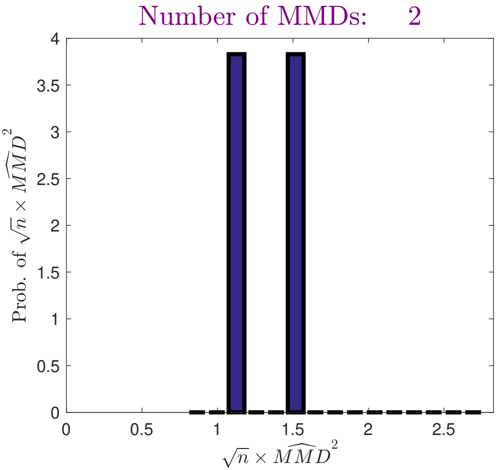


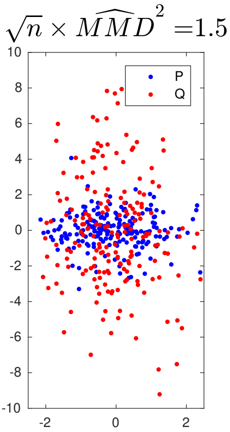


Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Draw $n = 200$ new samples from P and Q

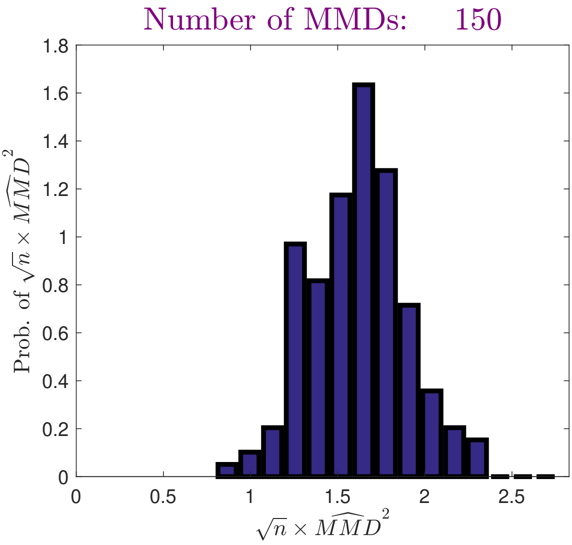
- Laplace with different y-variance.
- $\sqrt{n} \times \widehat{MMD}^2 = 1.5$





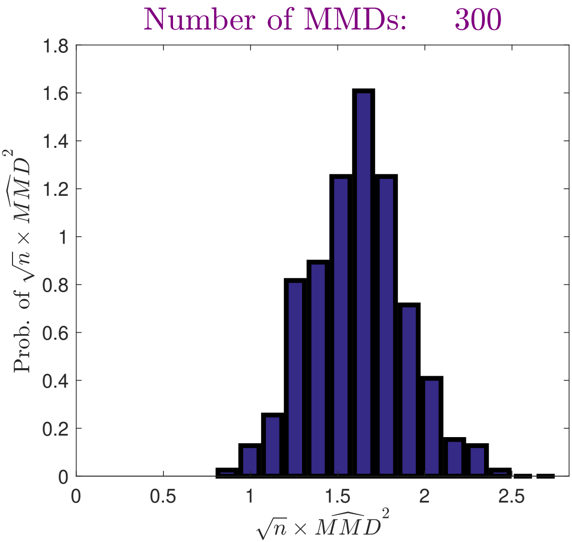
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 150 times . . .



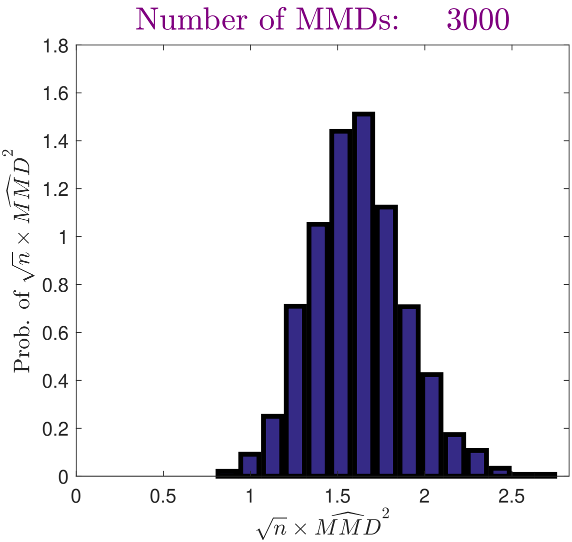
Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 300 times . . .



Behaviour of $\widehat{MMD}^2$ when $P \neq Q$

Repeat this 3000 times . . .

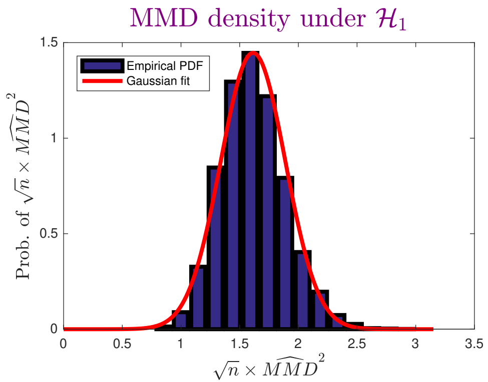


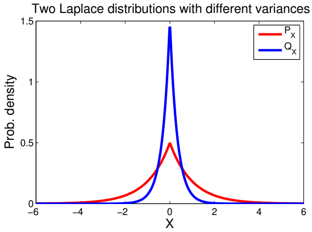
Asymptotics of $\widehat{MMD}^2$ when $P \neq Q$

When $P \neq Q$, statistic is asymptotically normal,

$$\frac{\widehat{\text{MMD}}^2 - \text{MMD}^2(P, Q)}{\sqrt{V_n(P, Q)}} \xrightarrow{D} \mathcal{N}(0, 1),$$

where variance $V_n(P, Q) = O\left(n^{-1}\right)$.



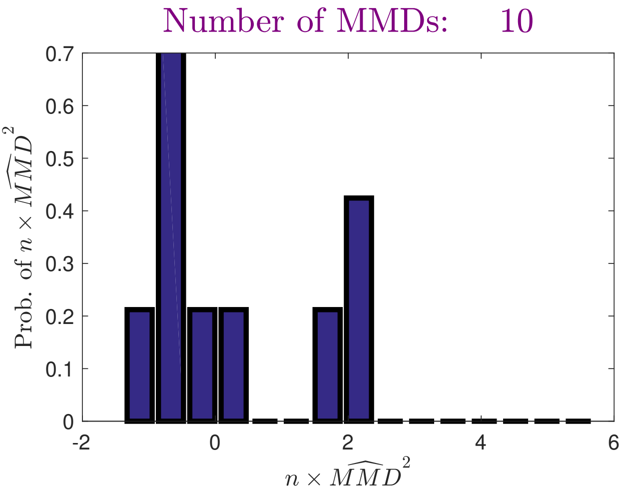


Behaviour of $\widehat{MMD}^2$ when $P = Q$

What happens when P and Q are the same?

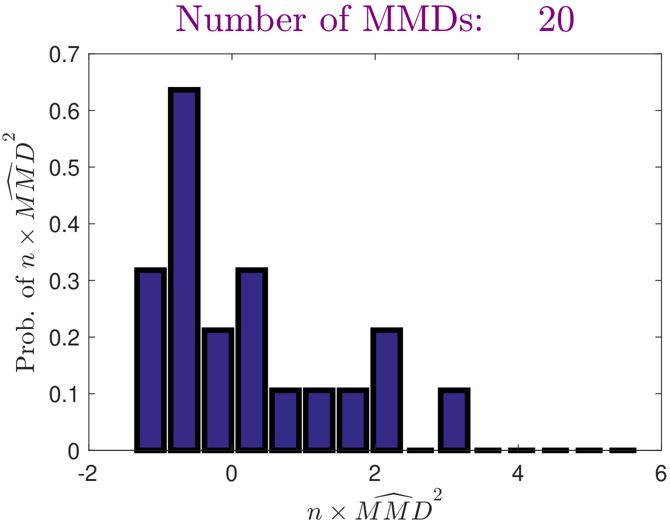
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0, 1)$



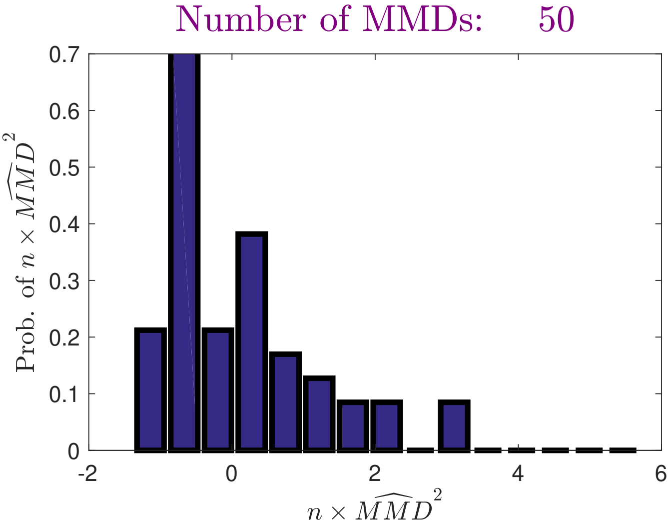
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0,1)$



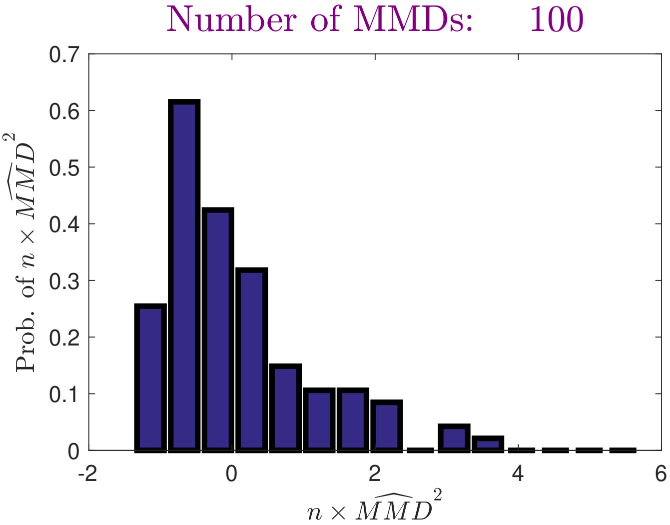
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0, 1)$



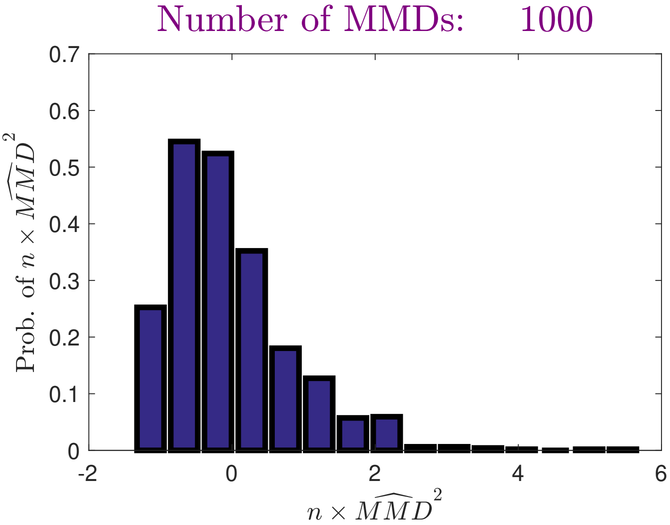
Behaviour of $\widehat{MMD}^2$ when $P = Q$

- Case of $P = Q = \mathcal{N}(0, 1)$



Behaviour of $\widehat{MMD}^2$ when $P = Q$

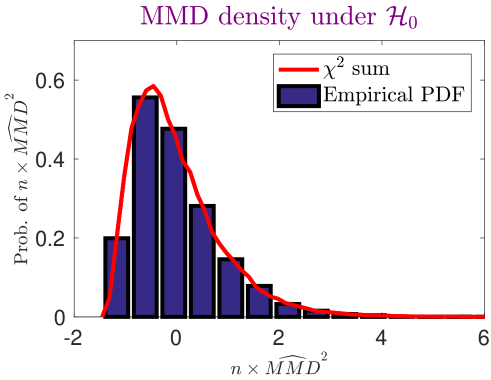
- Case of $P = Q = \mathcal{N}(0, 1)$



Asymptotics of $\widehat{MMD}^2$ when $P = Q$

Where $P = Q$, statistic has asymptotic distribution

$$n\widehat{\text{MMD}}^2 \sim \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2 \right]$$



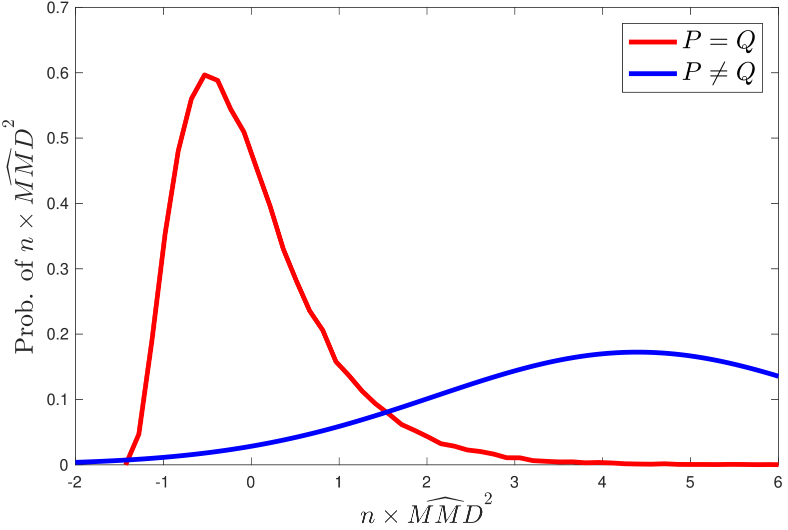
where

$$\lambda_i \psi_i(x') = \int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) dP(x)$$

$$z_l \sim \mathcal{N}(0, 2) \quad \text{i.i.d.}$$

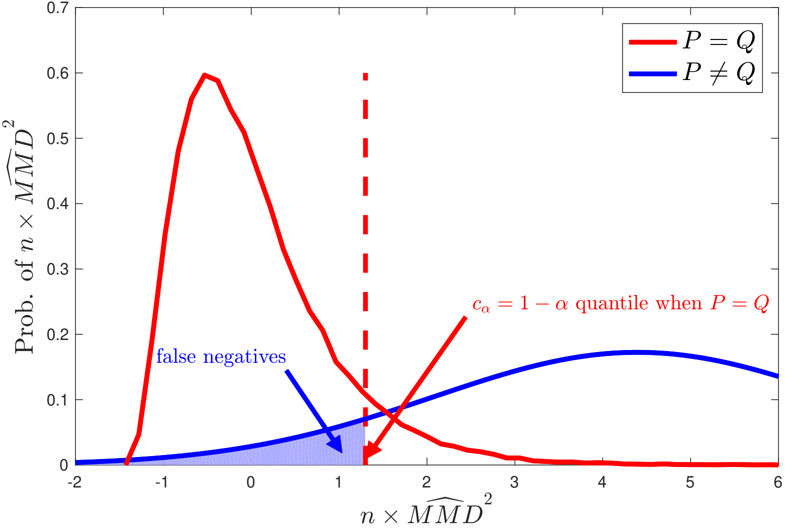
A statistical test

A summary of the asymptotics:



A statistical test

Test construction: (G., Borgwardt, Rasch, Schoelkopf, and Smola, JMLR 2012)



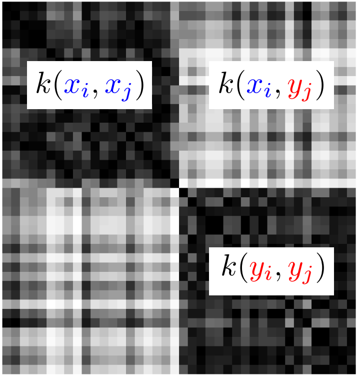
How do we get test threshold c_α?

Original empirical MMD for dogs and fish:

$X = [$  $\ldots\ ]$

$Y = [$  $\ldots\ ]$

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)$$

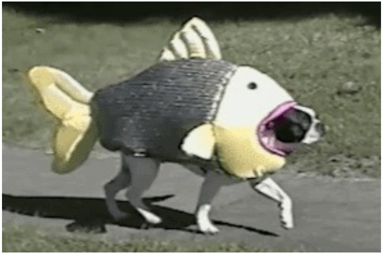


How do we get test threshold c_α?

Permuted dog and fish samples (merdogs):

$\widetilde{X} = [$  $\ldots]$

$\widetilde{Y} = [$  $\ldots]$



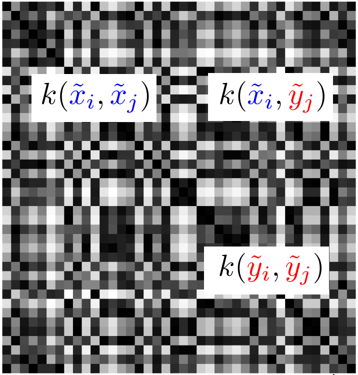
How do we get test threshold c_α?

Permuted dog and fish samples (merdogs):

$\widetilde{X} = [\ \ldots\ ]$

$\widetilde{Y} = [\ \ldots\ ]$

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



Permutation simulates

$P = Q$

The best test for the job

- A test's power depends on $k(x, x')$, P, and Q (and n)
- With characteristic kernel, MMD test has power $\to 1$ as $n \to \infty$ for any (fixed) problem
  - But, for many P and Q, will have terrible power with reasonable n!

The best test for the job

- A test's power depends on $k(x, x')$, P, and Q (and n)
- With characteristic kernel, MMD test has power $\to 1$ as $n \to \infty$ for any (fixed) problem
  - But, for many P and Q, will have terrible power with reasonable n!
- You *can* choose a good kernel for a given problem
- You *can't* get one kernel that has good finite-sample power for all problems
  - No one test can have all that power

Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$

Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

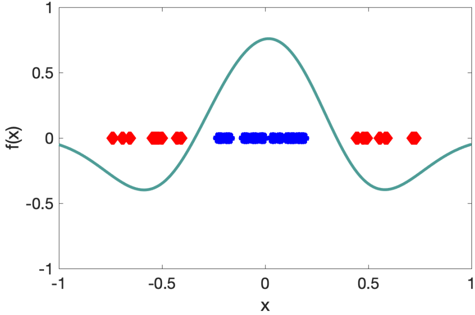
- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...

Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...

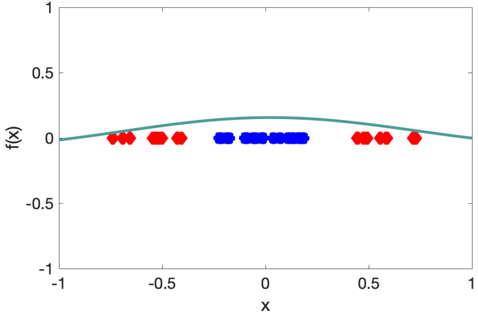


Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...

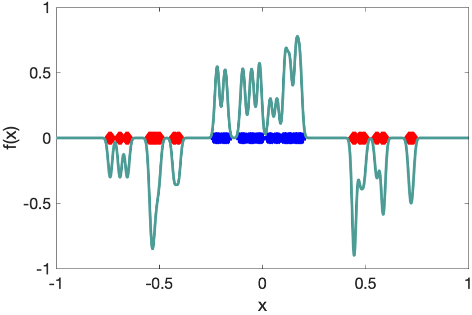


Choosing a kernel for the test

- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...



Choosing a kernel for the test

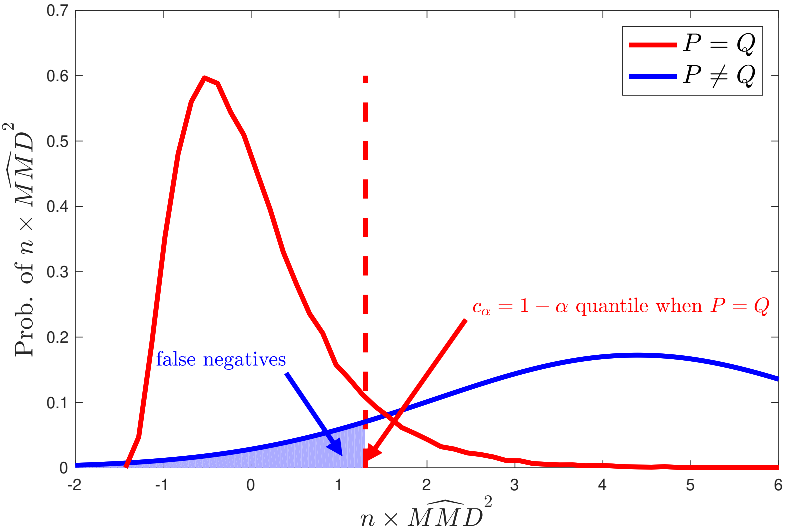
- Simple choice: exponentiated quadratic

$$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$$

- *Characteristic:* for any σ: for any P and Q, power $\to 1$ as $n \to \infty$
- But choice of σ is very important for finite n...
- ...and some problems (e.g. images) might have no good choice for σ

Graphical illustration

- Maximising test power same as minimizing false negatives



Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\text{MMD}}^2 > \hat{c}_\alpha\right)$$
$$\to \Phi\left(\frac{\text{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P,Q)}}\right)$$

where

- Φ is the CDF of the standard normal distribution.
- $\hat{c}_\alpha$ is an estimate of c_α test threshold.

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

$$\to \Phi\left(\underbrace{\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}}_{O(n^{1/2})} - \underbrace{\frac{c_\alpha}{n\sqrt{V_n(P,Q)}}}_{O(n^{-1/2})}\right)$$

For large n, second term negligible!

Optimizing kernel for test power

The power of our test ($\Pr_1$ denotes probability under $P \neq Q$):

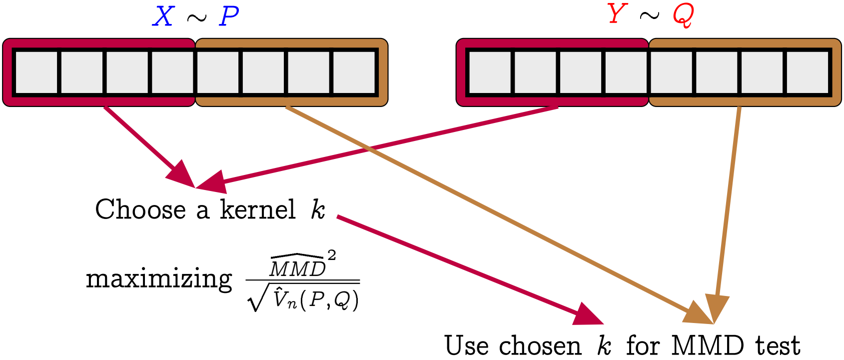
$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

$$\to \Phi\left(\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} - \frac{c_\alpha}{n\sqrt{V_n(P,Q)}}\right)$$

To maximize test power, maximize

$$\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}$$

Data splitting

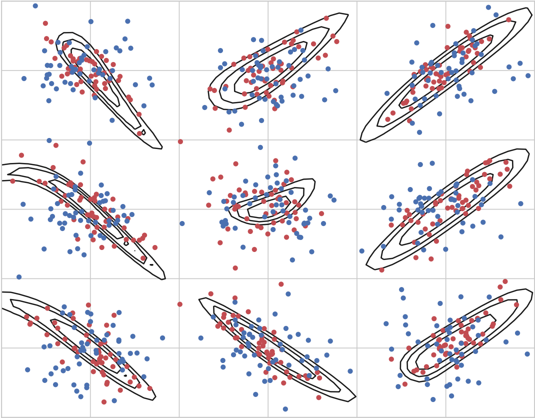


Learning a kernel helps a lot

Kernel with deep learned features:

$$k_\theta(x, y) = [(1-\epsilon)\kappa(\Phi_\theta(x), \Phi_\theta(y)) + \epsilon]\, q(x, y)$$

κ and q are Gaussian kernels



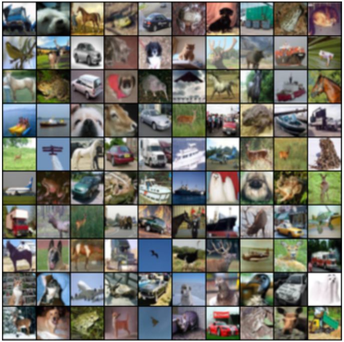
Learning a kernel helps a lot

Kernel with deep learned features:

$$k_\theta(x, y) = [(1-\epsilon)\kappa(\Phi_\theta(x), \Phi_\theta(y)) + \epsilon]\, q(x, y)$$

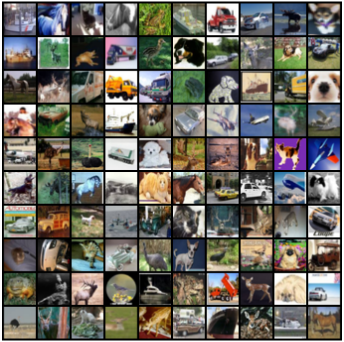
κ and q are Gaussian kernels

- CIFAR-10 vs CIFAR-10.1, null rejected 75% of time



CIFAR-10 test set (Krizhevsky 2009)

$X \sim P$



CIFAR-10.1 (Recht+ ICML 2019)

$Y \sim Q$

Learning a kernel helps a lot

Kernel with deep learned features:

$$k_\theta(x, y) = [(1 - \epsilon)\kappa(\Phi_\theta(x), \Phi_\theta(y)) + \epsilon]\ q(x, y)$$

κ and q are Gaussian kernels

- CIFAR-10 vs CIFAR-10.1, null rejected 75% of time

arXiv.org > stat > arXiv:2002.09116

Statistics > Machine Learning

[Submitted on 21 Feb 2020]

Learning Deep Kernels for Non-Parametric Two-Sample Tests

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, D. J. Sutherland

ICML 2020

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT

Post-credit scene: MMD flow

From NeurIPS 2019:

Maximum Mean Discrepancy Gradient Flow

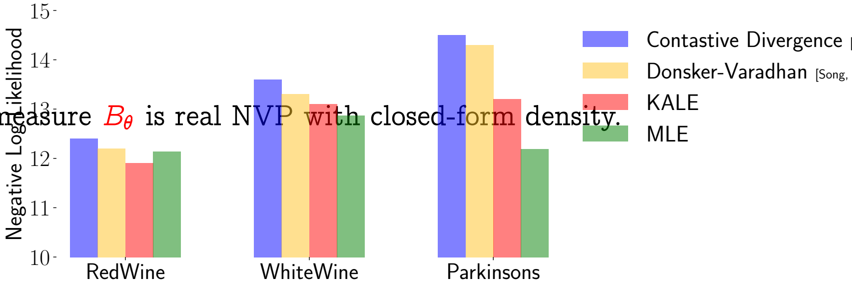
Michael Arbel
Gatsby Computational Neuroscience Unit
University College London
michael.n.arbel@gmail.com

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

Sanity check: reduction to EBM case



Base measure B_θ is real NVP with closed-form density.