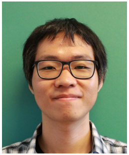
Interpretable comparison of distributions and models

Arthur Gretton, Dougal Sutherland, Wittawat Jitkrittum





Gatsby Unit UCL, TTI-Chicago→UBC, MPI for Intelligent Systems

NeurIPS, Vancouver, 2019

A motivation: comparing two samples

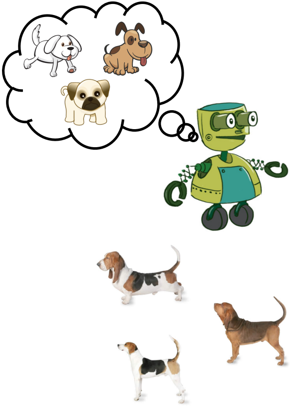
- **Given:** Samples from unknown distributions P and Q.
- **Goal:** do P and Q differ?

$\sim P$

$\sim Q$

A motivation: comparing a sample and a model

- **Given:** Sample from unknown Q, **model** P
- **Goal:** do P and Q differ?



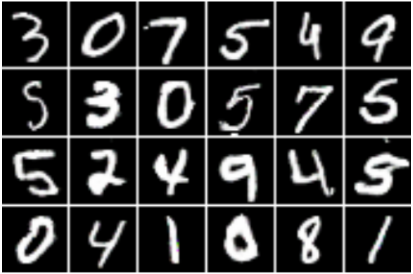
$\sim P$

$\sim Q$

A real-life example: two-sample tests

- **Have:** Two collections of samples X, Y from unknown distributions P and Q.
- **Goal:** do P and Q differ?

MNIST samples



Samples from a GAN

Significant difference in GAN and MNIST?

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, Xi Chen, NIPS 2016
Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.

Outline

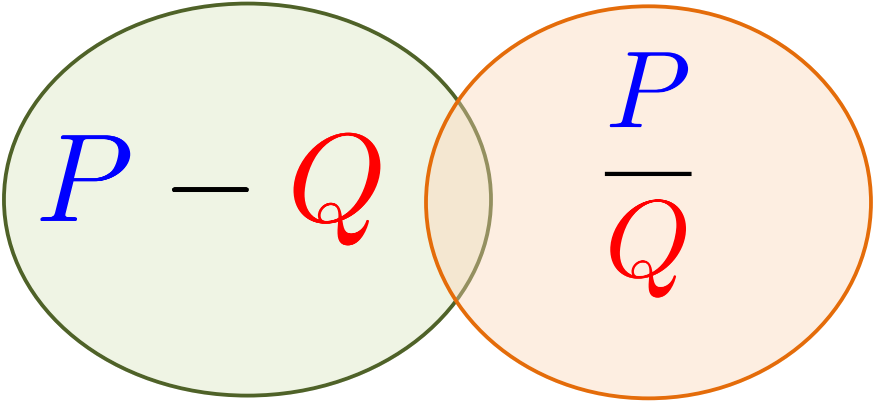
- Divergence measures
  - Integral probability metrics
  - ϕ-divergences (f-divergences)

- Statistical hypothesis testing
  - Using integral probability metrics
  - Learned features for powerful tests
  - Relation of testing and classification

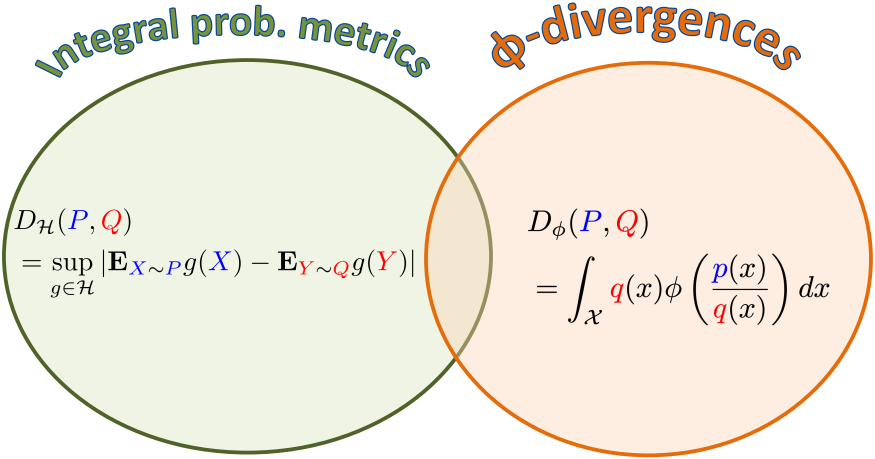
- Linear-time features and model criticism
  - Interpretable, linear time features for testing
  - Stein's method for model evaluation

Divergence measures

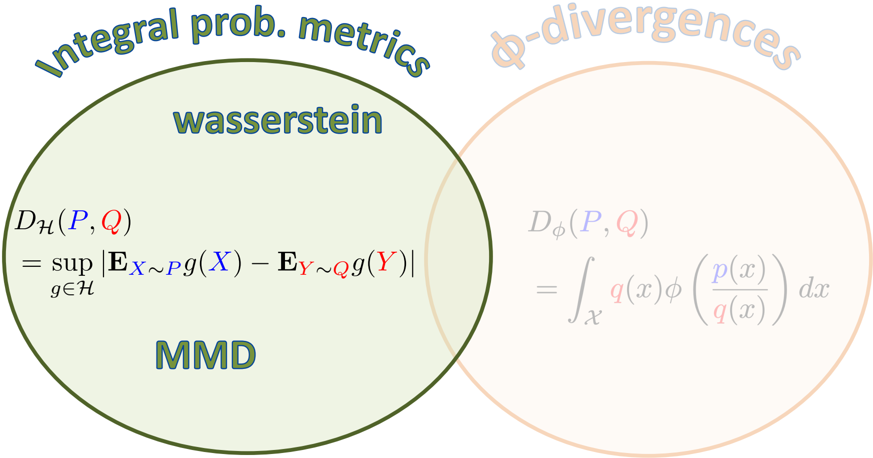
Divergences



Divergences

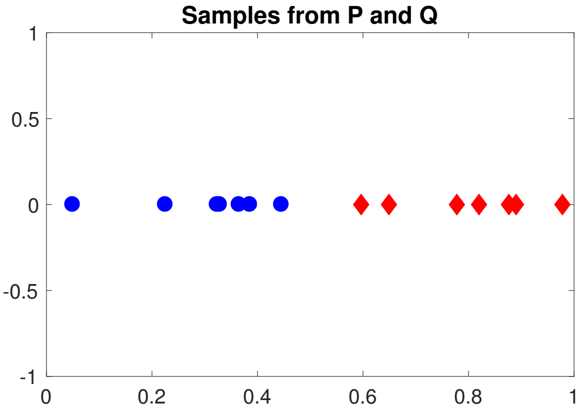


Divergences: integral probability metrics



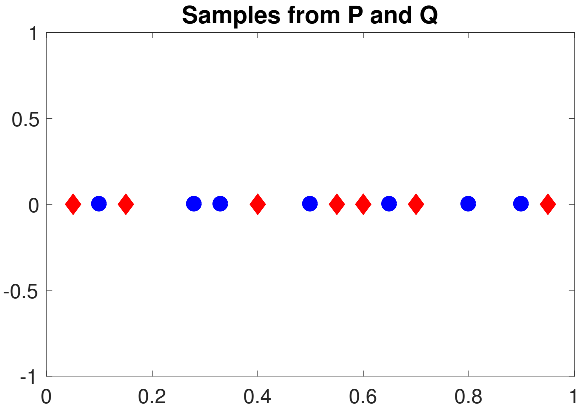
Integral probability metrics

Are P and Q different?



Integral probability metrics

Are P and Q different?

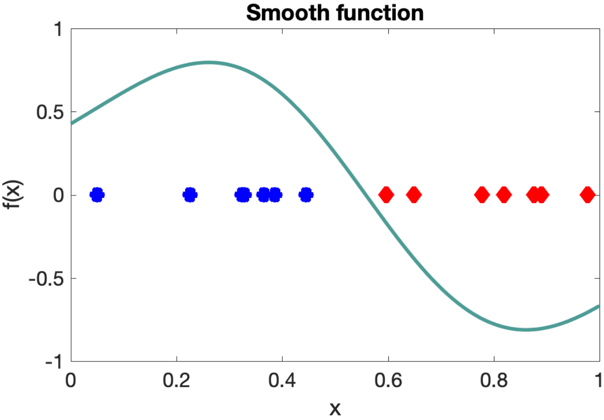


Integral probability metrics

Integral probability metric:

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

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

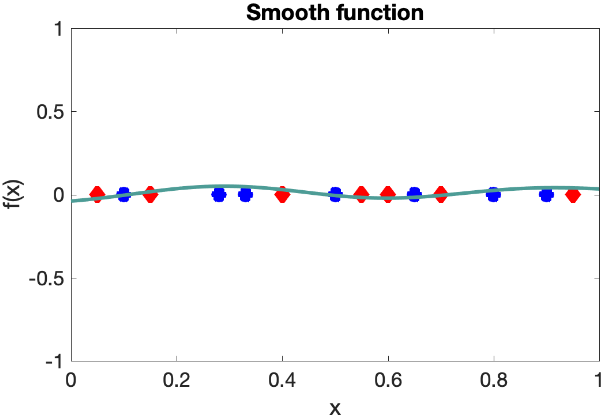


Integral probability metrics

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

The MMD: an integral probability metric

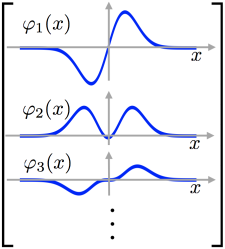
Maximum mean discrepancy: smooth function for P vs Q

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

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

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

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

For positive definite k,

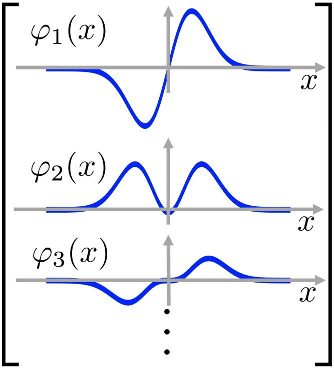
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

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

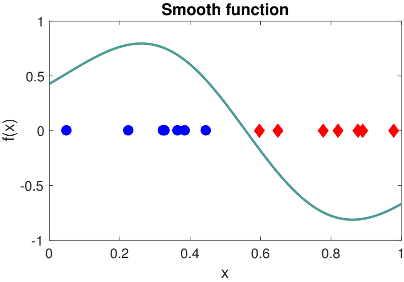
$$\mathbf{E}_P(f(X)) = \langle f, \mathbf{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

(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:

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

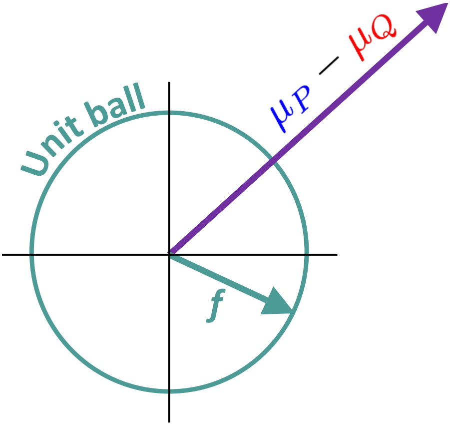
use

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

Integral prob. metric vs feature mean difference

The MMD:

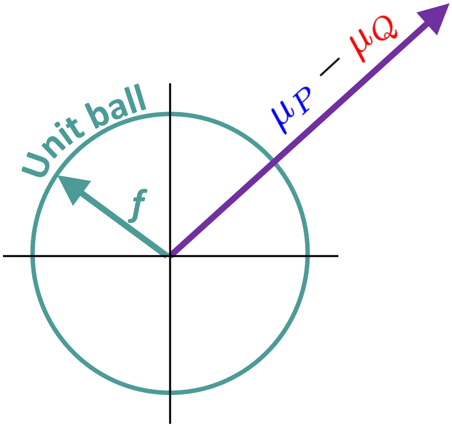
$$MMD(P, Q; F) = \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right] = \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature mean difference

The MMD:

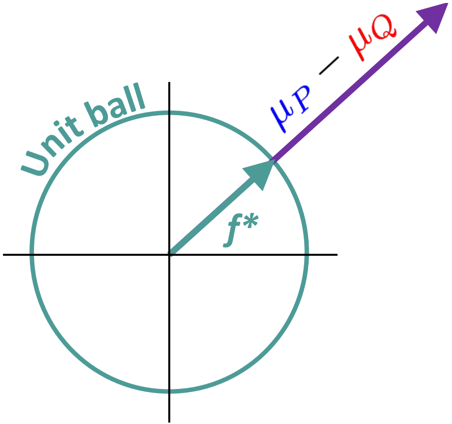
$$
\begin{aligned}
&MMD(P, Q; F) \\
&= \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right] \\
&= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\end{aligned}
$$



Integral prob. metric vs feature mean difference

The MMD:

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



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

Integral prob. metric vs feature mean difference

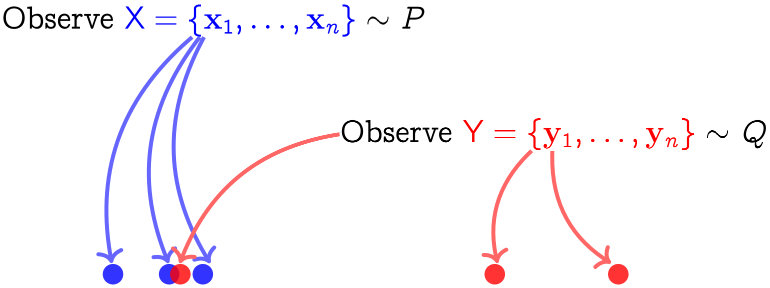
The MMD:

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

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

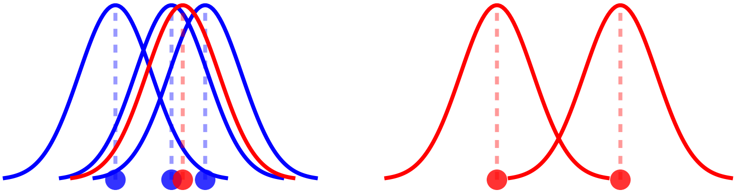
Construction of MMD witness

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



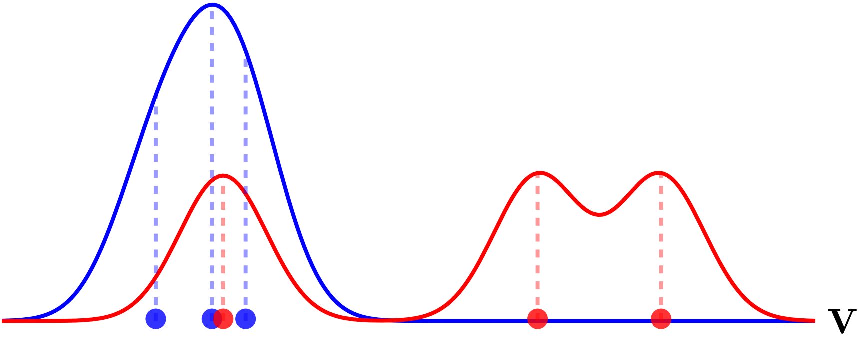
Construction of MMD witness

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



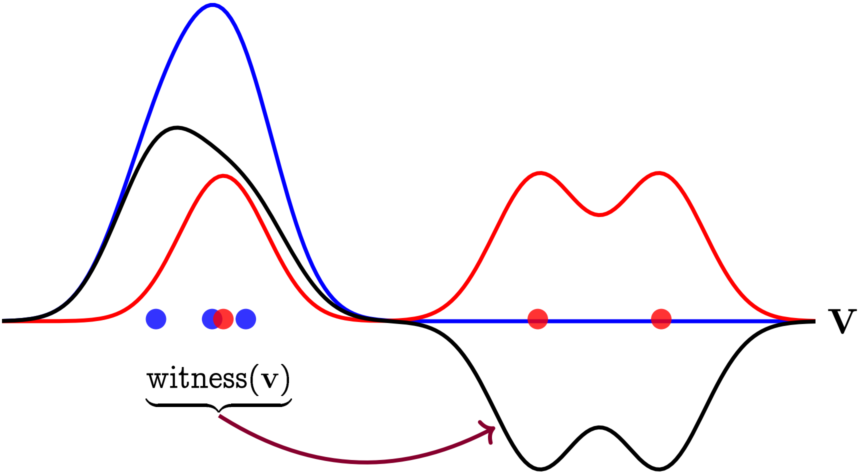
Construction of MMD witness

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

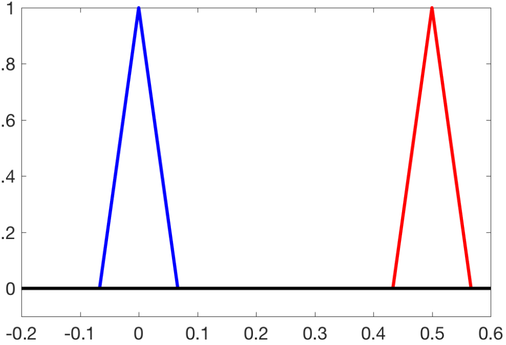
$$\begin{aligned} f^*(v) &= \langle f^*, \varphi(v) \rangle_{\mathcal{F}} \\ &\propto \langle \widehat{\mu}_P - \widehat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \\ &= \frac{1}{n} \sum_{i=1}^{n} k(x_i, v) - \frac{1}{n} \sum_{i=1}^{n} k(\mathrm{y}_i, v) \end{aligned}$$

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

IPMs in practice

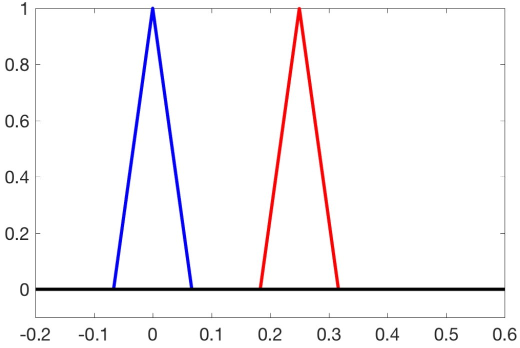
How do the IPMs behave?

- **A simple setting:** distributions with disjoint support, Q approaches P



How do the IPMs behave?

- **A simple setting:** distributions with disjoint support, Q approaches P

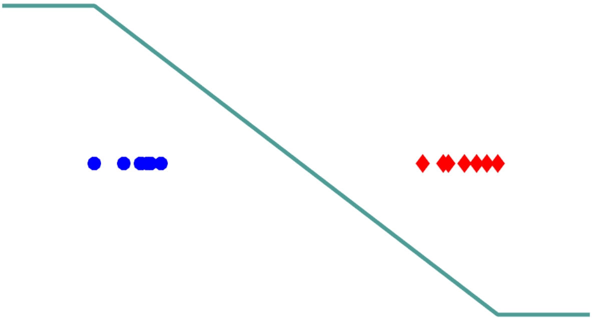


How does the Wasserstein-1 behave?

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

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

W_1=0.88



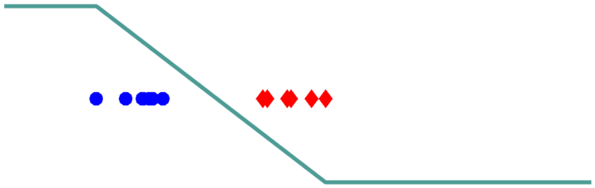
Santambrogio, Optimal Transport for Applied Mathematicians (2015, Section 5.4)
G Peyré, M Cuturi, Computational Optimal Transport (2019)
M. Cuturi, J. Solomon, NeurIPS tutorial (2017)

How does the Wasserstein-1 behave?

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

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

W_1=0.65



Santambrogio, Optimal Transport for Applied Mathematicians (2015, Section 5.4)

G Peyré, M Cuturi, Computational Optimal Transport (2019)

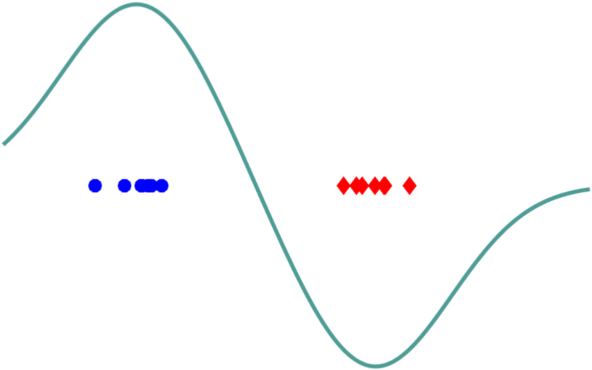
M. Cuturi, J. Solomon, NeurIPS tutorial (2017)

How does the MMD behave?

MMD with a broad kernel:

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

MMD=1.8

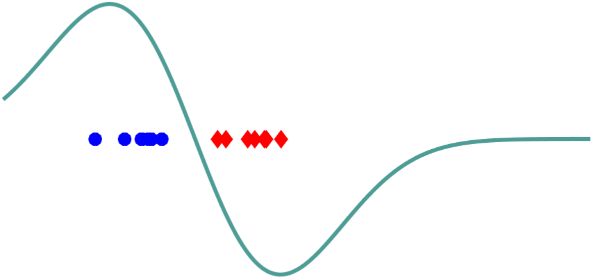


How does the MMD behave?

MMD with a broad kernel::

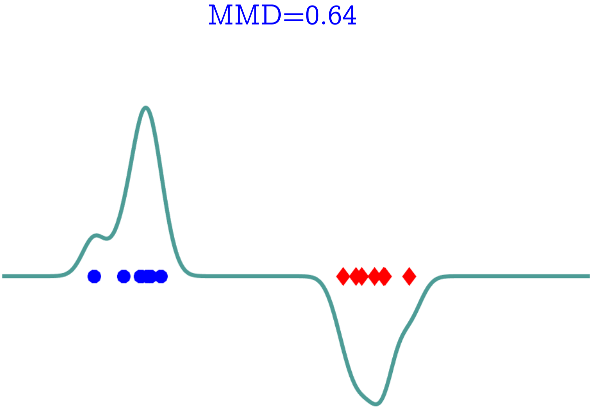
$MMD(P, Q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} E_P f(X) - E_Q f(Y)$

MMD=1.1



How does the MMD behave?

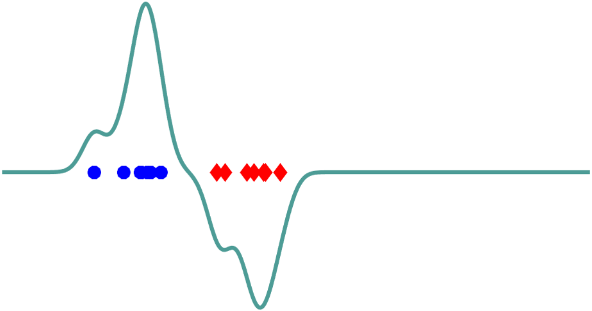
$MMD(P, Q)$ with a narrow kernel.



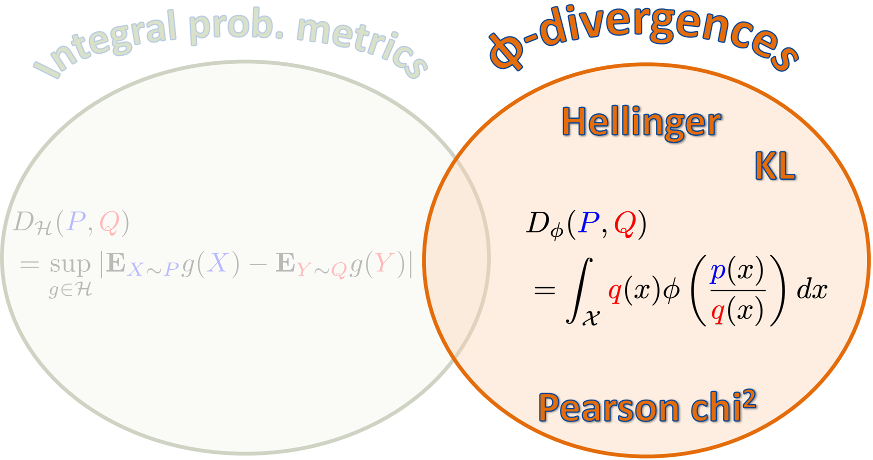
How does the MMD behave?

$MMD(P, Q)$ with a narrow kernel.

MMD=0.64



The ϕ-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,

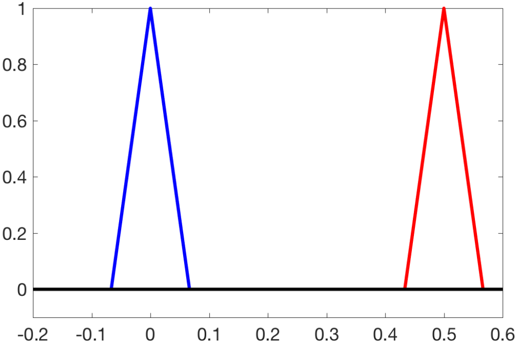
$$D_{KL}(Q, P) = \int \log\left(\frac{q(x)}{p(x)}\right) q(x) dx$$

How do ϕ-divergences behave?

Simple example: disjoint support, revisited.

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

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

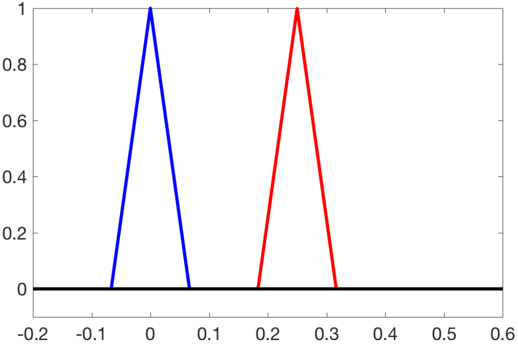


How do ϕ-divergences behave?

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

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

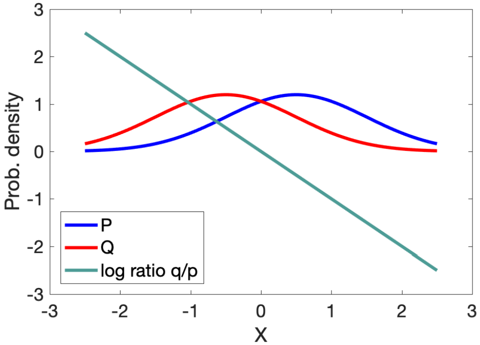
Case of the reverse KL

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

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

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

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

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

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

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?

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

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

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

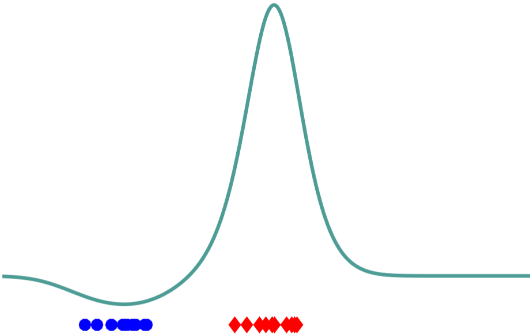
How does the KALE divergence behave?

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

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

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

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



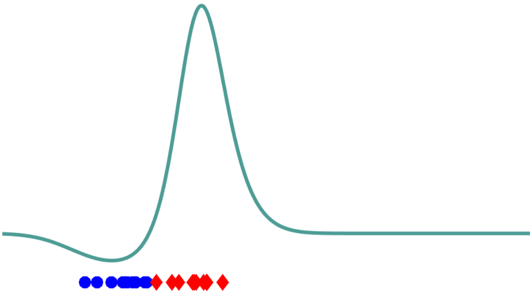
How does the KALE divergence behave?

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

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

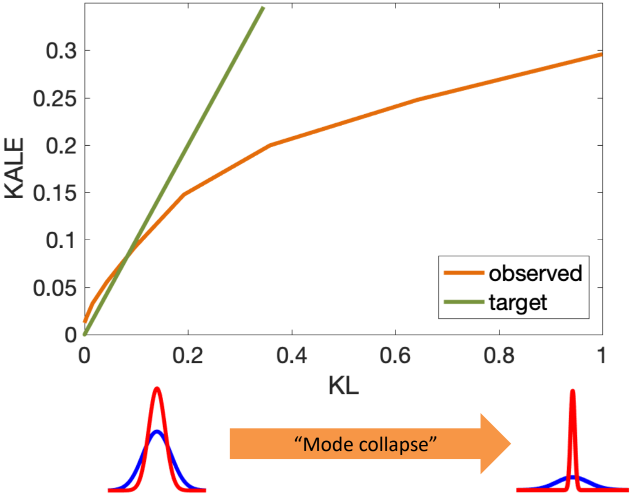
$$\|w\|^2_{\mathcal{F}} \quad \text{penalized : KALE smoothie}$$

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



The KALE smoothie and "mode collapse"

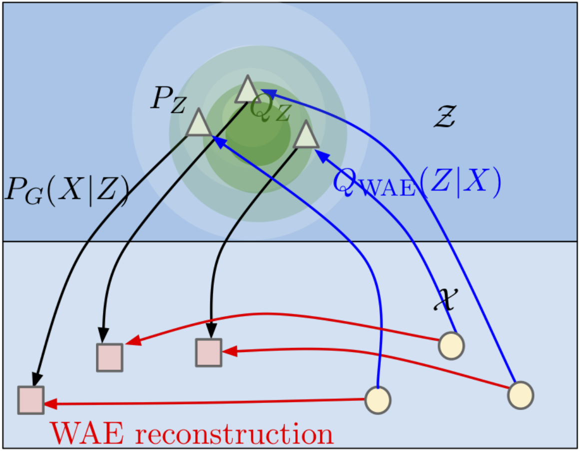
- Two Gaussians with same means, different variance



Example thanks to M. Arbel and M. Rosca

WAE-GAN Kale and WAE-MMD

The Wasserstein Autoencoder:



Tolstikhin, Bousquet, Gelly, Schölkopf (2018). New version with parameter sweep from 2019: see arxiv.

WAE-GAN Kale and WAE-MMD

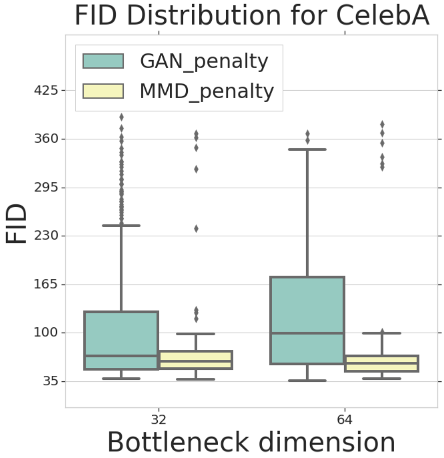
The Wasserstein Autoencoder:

Celeb-A performance (FID):

- WAE-MMD: 37
- WAE-GAN: 35
- Variational autoencoder: 45

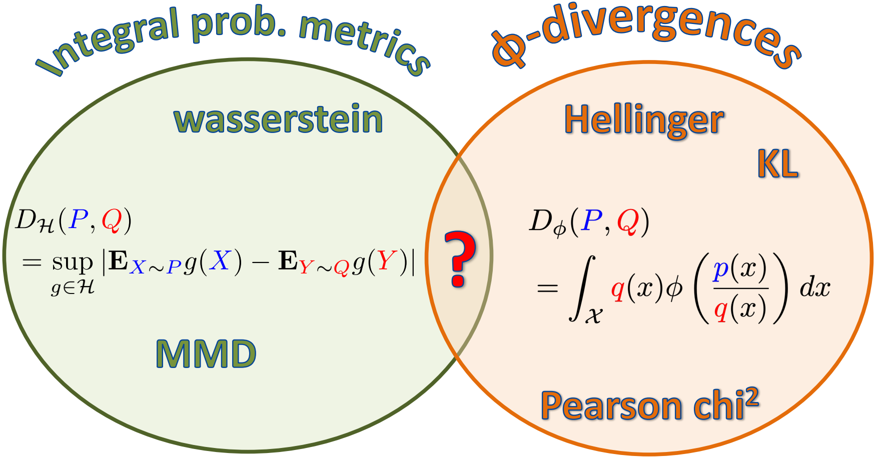
WAE-GAN Kale and WAE-MMD

The Wasserstein Autoencoder:

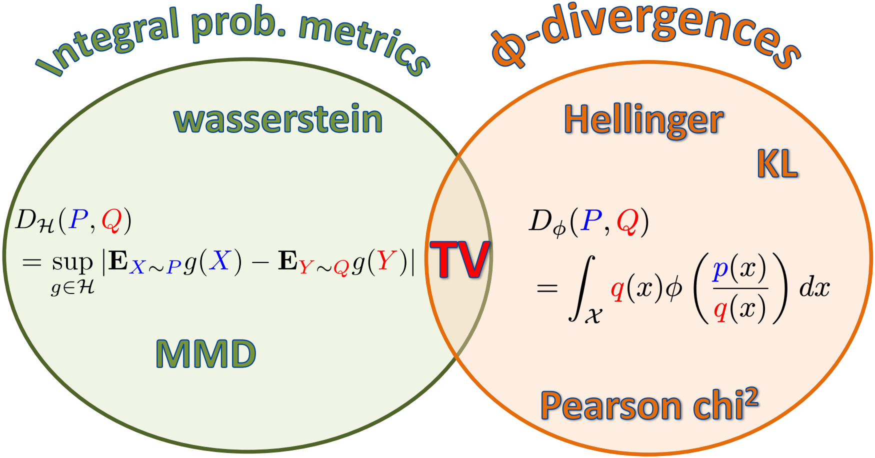


Sweep over: architectures of the Encoder and Decoder (DCGAN or ResNet50v2), regularization coefficient, learning rates, kernel width,...Parameters in both in WAE-MMD and WAE-GAN (i.e. AE learning rate, regularization coeff, etc) had the same ranges for both.

Divergences



Divergences



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

References and further reading

- Wasserstein distances:
  - Peyré, Cuturi. Computational Optimal Transport (2019)
  - Santambrogio. Optimal Transport for Applied Mathematicians (2015)
- The Maximum Mean Discrepancy:
  - Gretton, Borgwardt, Rasch. Schölkopf, Smola. A kernel two-sample test. (2012)
  - Arbel, Sutherland, Binkowski, Gretton. Gradient regularization for MMD GANS (2018)
- Variational estimates of ϕ-divergences:
  - Nguyen, Wainwright, Jordan. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization (2010)
  - Nowozin, Cseke, Tomioka. F-GAN: Training Generative Neural Samplers using Variational Divergence Minimization (2016)
- Divergences and generative models:
  - Arora, Ge, Liang, Ma, Zhang. Generalization and Equilibrium in Generative Adversarial Nets (GANs) (2017)
  - Tolstikhin, Bousquet, Gelly, Schölkopf. Wasserstein Auto-encoders (2019 version)
  - Huang, Berard, Touati, Gidel, Vincent, Lacoste-Julien. Parametric Adversarial Divergences are Good Task Losses for Generative Modeling (2018)
  - Bottou, Arjovsky, Lopez-Paz, Oquab. Geometrical Insights for Implicit Generative Modeling (2018)

Bound for Jensen-shannon

Case of the Jensen Shannon divergence

$$
\begin{aligned}
&D_{JS}(Q, P) \\
&= \frac{1}{2}\int p(z)\log\left(\frac{2p(z)}{p(z)+q(z)}\right)dz + \frac{1}{2}\int q(z)\log\left(\frac{2q(z)}{p(z)+q(z)}\right)dz
\end{aligned}
$$

Bound for Jensen-shannon

Case of the Jensen Shannon divergence

$$
\begin{aligned}
&D_{JS}(Q, P) \\
&= \frac{1}{2}\int p(z)\log\left(\frac{2p(z)}{p(z)+q(z)}\right)dz + \frac{1}{2}\int q(z)\log\left(\frac{2q(z)}{p(z)+q(z)}\right)dz \\
&\geq \sup_{f<0,\, f\in\mathcal{H}} \Bigg\{ \mathbf{E}_P f(X) \\
&\quad - \mathbf{E}_Q \underbrace{\left[-(f(Y)+1)\log\left(\frac{f(Y)+1}{2}\right) + f(Y)\log f(Y)\right]}_{\phi^*(f(Y))} \Bigg\}
\end{aligned}
$$

Bound for Jensen-shannon

Case of the Jensen Shannon divergence

$$
\begin{aligned}
& D_{JS}(Q, P) \\
&= \frac{1}{2} \int p(z) \log\left(\frac{2p(z)}{p(z)+q(z)}\right) dz + \frac{1}{2} \int q(z) \log\left(\frac{2q(z)}{p(z)+q(z)}\right) dz \\
&\geq \sup_{f<0, f\in\mathcal{H}} \Bigg\{ \mathbf{E}_P f(X) \\
&\quad - \mathbf{E}_Q \underbrace{\left[-(f(Y)+1)\log\left(\frac{f(Y)+1}{2}\right) + f(Y)\log f(Y)\right]}_{\phi^*(f(Y))} \Bigg\}
\end{aligned}
$$

Bound tight when:

$$
f^\diamond(z) = \log\left(\frac{2p(x)}{p(x)+q(x)}\right)
$$