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?
A motivation: comparing a sample and a model

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

Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.
Outline

■ Divergence measures
  - Integral probability metrics
  - $\phi$-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

\[ P \quad Q \]

\[ \frac{P}{Q} \]
Divergences

Integral prob. metrics

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

\[ \phi\text{-divergences} \]

\[ D_{\phi}(P, Q) = \int_{\mathcal{X}} q(x) \phi \left( \frac{p(x)}{q(x)} \right) dx \]
Divergences: integral probability metrics

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

\[ D_{\phi}(P, Q) = \int_{\mathcal{X}} q(x) \phi \left( \frac{p(x)}{q(x)} \right) \, dx \]
Integral probability metrics

Are $P$ and $Q$ different?

Samples from $P$ and $Q$
Integral probability metrics

Are $P$ and $Q$ different?

Samples from $P$ and $Q$
Integral probability metrics

Integral probability metric:
Find a "well behaved function" $f(x)$ to maximize

$$E_P f(X) - E_Q f(Y)$$
Integral probability metrics

Integral probability metric:
Find a "well behaved function" $f(x)$ to maximize

$$E_P f(X) - E_Q f(Y)$$
Maximum mean discrepancy: smooth function for $P$ vs $Q$

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

($F = \text{unit ball in RKHS } \mathcal{F}$)
The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for $P$ vs $Q$

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

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

Functions are linear combinations of features:

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

$$\|f\|^2_{\mathcal{F}} := \sum_{i=1}^{\infty} f_i^2 \leq 1$$
**Infinitely many features using kernels**

**Kernels: dot products of features**

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

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in l_2$$

For positive definite $k$,

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

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

Kernels: dot products of features

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

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

For positive definite $k$,

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

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

Exponentiated quadratic kernel

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

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 $Q$

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

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

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

Other choices for witness function class:

- Bounded continuous [Dudley, 2002]
- Bounded variation 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 $Q$

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

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

Expectations of functions are linear combinations of expected features

\[
\mathbb{E}_P(f(X)) = \langle f, \mathbb{E}_P \varphi(X) \rangle_{\mathcal{F}} = \langle f, \mu_P \rangle_{\mathcal{F}}
\]

(always true if kernel is bounded)
The MMD:

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

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

use

\[
\mathbb{E}_P f(X) = \langle \mu_P, f \rangle_F
\]

\[
= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_F
\]
The MMD:

\[
MMD(P, Q; F) = \sup_{\|f\| \leq 1} [E_P f(X) - E_Q f(Y)] \\
= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_F
\]
The MMD:

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

\[
= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}
\]
The MMD:

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

\[
= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_F
\]

\[
f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}
\]
Integral prob. metric vs feature mean difference

The MMD:

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

$$= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_F$$

$$= \|\mu_P - \mu_Q\|$$

IPM view equivalent to feature mean difference (kernel case only)
Construction of MMD witness

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

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

Observe $Y = \{y_1, \ldots, y_n\} \sim Q$
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 \textit{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 \( v \)

\[ f^*(v) = \langle f^*, \varphi(v) \rangle_F \]
Derivation of empirical witness function

Recall the witness function expression

\[ f^* \propto \mu_P - \mu_Q \]

The empirical feature mean for \( P \)

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

The empirical witness function at \( v \)

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

\[ \propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_{\mathcal{F}} \]
Derivation of empirical witness function

Recall the witness function expression

\[ f^* \propto \mu_P - \mu_Q \]

The empirical feature mean for \( P \)

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

The empirical witness function at \( v \)

\[ f^*(v) = \langle f^*, \varphi(v) \rangle_F \]
\[ \propto \langle \hat{\mu}_P - \hat{\mu}_Q, \varphi(v) \rangle_F \]
\[ = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, v) \]

Don’t need explicit feature coefficients \( f^* := \begin{bmatrix} f_1^* & f_2^* & \cdots \end{bmatrix} \)
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, $\mathcal{Q}$ approaches $\mathcal{P}$
How does the Wasserstein-1 behave?

\[
W_1(P, Q) = \sup_{\|f\|_{L^1} \leq 1} E_P f(X) - E_Q f(Y).
\]

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

\[
W_1 = 0.88
\]
How does the Wasserstein-1 behave?

\[ W_1(P, Q) = \sup_{\|f\|_{L^1} \leq 1} E_P f(X) - E_Q f(Y). \]

\[ \|f\|_L := \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|} \]

\[ W_1 = 0.65 \]

Santambrogio, Optimal Transport for Applied Mathematicians (2015, Section 5.4)
G Peyré, M Cuturi, Computational Optimal Transport (2019)
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.

\[ MMD = 0.64 \]
How does the MMD behave?

\[ MMD(P, Q) \] with a narrow kernel.

\[ MMD=0.64 \]
The $\phi$-divergences

Integral prob. metrics

$$D_\mathcal{H}(P, Q) = \sup_{g \in \mathcal{H}} |E_{\sim P} g(X) - E_{\sim Q} g(Y)|$$

$\phi$-divergences

Hellinger

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

KL

Pearson chi$^2$
### The $\phi$-divergences

Define the $\phi$-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 $\phi$ 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 $\phi$-divergences

Define the $\phi$-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 $\phi$ 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 $\phi$-divergences behave?

**Simple example:** disjoint support, revisited.

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

\[ D_{KL}(Q, P) = \infty \quad D_{JS}(P, Q) = \log 2 \]
How do \(\phi\)-divergences behave?

Simple example: disjoint support, revisited.

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

\[
D_{KL}(Q, P) = \infty \quad D_{JS}(P, Q) = \log 2
\]
**φ-divergences in practice**

Case of the reverse KL

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

ϕ-divergences in practice

Case of the reverse KL

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

\[ \geq \sup_{f < 0, f \in \mathcal{H}} \mathbb{E}_P f(X) + \mathbb{E}_Q \log (-f(Y)) + 1 - \phi^*(f(Y)) \]

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)
\( \phi \)-divergences in practice

Case of the reverse KL

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

\[
\geq \sup_{f < 0, f \in \mathcal{H}} E_P f(X) + E_Q \log (-f(Y)) + 1
\]

Bound tight when:

\[
f^\circ(z) = -\frac{q(z)}{p(z)}
\]

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)
\(\phi\)-divergences in practice

Case of the reverse KL

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

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

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

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)
**ϕ-divergences in practice**

Case of the reverse KL

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

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

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

This is a **KL Approximate Lower-bound Estimator.**
\(\phi\)-divergences in practice

Case of the reverse KL

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

\[
\geq \sup_{f<0, f \in \mathcal{H}} E_P f(X) + E_Q \log (-f(Y)) + 1
\]

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

This is a

K

A

L

E
**φ-divergences in practice**

**Case of the reverse KL**

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

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

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

---

**The KALE divergence**

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)
How does the KALE divergence behave?

\[
KALE(Q, P) = \sup_{f < 0, f \in \mathcal{H}} \mathbb{E}_P f(X) + \mathbb{E}_Q \log(-f(Y)) + 1
\]

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

\[
\|w\|_{\mathcal{F}}^2 \quad \text{penalized:}
\]
How does the KALE divergence behave?

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

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

$$\|w\|_{\mathcal{F}}^2 \quad \text{penalized: KALE smoothie}$$
How does the KALE divergence behave?

\[
KALE(Q, P) = \sup_{f \leq 0, f \in \mathcal{H}} EPf(X) + EQ \log(-f(Y)) + 1
\]

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

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

\[
KALE(Q, P) = 0.18
\]
How does the KALE divergence behave?

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

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

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

\[
KALE(Q, P) = 0.12
\]
The KALE smoothie and “mode collapse”

- Two Gaussians with same means, different variance

Example thanks to M. Arbel and M. Rosca
WAE-GAN Kale and WAE-MMD

The Wasserstein Autoencoder:

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

**Integral prob. metrics**

**Wasserstein**

\[ D_H(P, Q) = \sup_{g \in \mathcal{H}} |\mathbb{E}_{X \sim P} g(X) - \mathbb{E}_{Y \sim Q} g(Y)| \]

**MMD**

**\( \Phi \)-divergences**

**Hellinger**

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

**KL**

**Pearson chi\(^2\)**
Divergences

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

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

Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet, EJS (2012)
References and further reading

- **Wasserstein distances:**
  - Peyré, Cuturi. Computational Optimal Transport (2019)

- **The Maximum Mean Discrepancy:**
  - Arbel, Sutherland, Binkowski, Gretton. Gradient regularization for MMD GANS (2018)

- **Variational estimates of $\phi$-divergences:**

- **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)
Bound for Jensen-shannon

Case of the Jensen Shannon divergence

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

Case of the Jensen Shannon divergence

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

\[ \geq \sup_{f < 0, f \in \mathcal{H}} \left\{ E_P f(X) \right\} \]

\[ - E_Q \left[ - (f(Y) + 1) \log \left( \frac{f(Y) + 1}{2} \right) + f(Y) \log f(Y) \right] \]

\[ \phi^*(f(Y)) \]
Bound for Jensen-shannon

Case of the Jensen Shannon divergence

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

\[ \geq \sup_{f < 0, f \in \mathcal{H}} \left\{ E_P f(X) \right\} \]

\[ - E_Q \left[ - (f(Y) + 1) \log \left( \frac{f(Y) + 1}{2} \right) + f(Y) \log f(Y) \right] \]

\[ \phi^*(f(Y)) \]

Bound tight when:

\[ f^*(z) = \log \left( \frac{2p(x)}{p(x) + q(x)} \right) \]