UNIVERSITY
OF OSLO



IN4310/3310 Distribution Shifts

ali@uio.no



Key Learning Goals

Distribution shift

Covariate shift and label shift

Importance weights

Comparing distributions and discrepancy measures

Outline

1. Distribution Shift

2. Learning under Distribution Shift

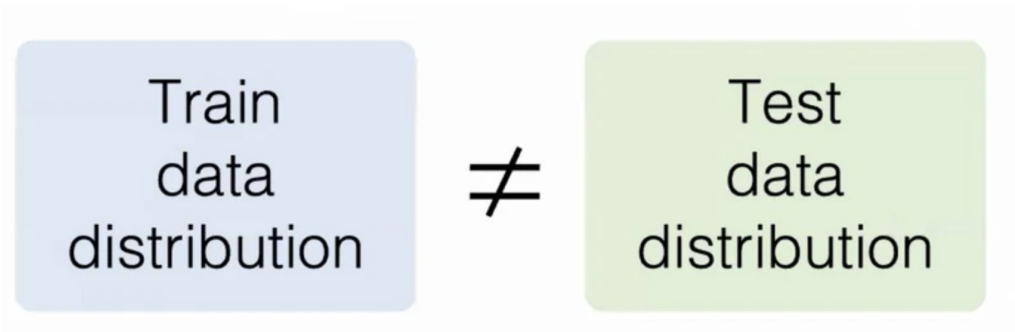
3. Comparing Distributions and Discrepancy Measures

Background

Training data: $\{(\mathbf{x}_i^{\mathrm{tr}}, \mathbf{y}_i^{\mathrm{tr}})\}_{i=1}^{n^{\mathrm{tr}}} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y})$

Learn predictor $y = f(\mathbf{x})$ minimizing the test risk $R(f)$

$\min_f R(f) \quad R(f) = \mathbb{E}_{p^{\mathrm{te}}(\mathbf{x},\mathbf{y})}[\ell(f(\mathbf{x}), \mathbf{y})] \quad \ell$: loss

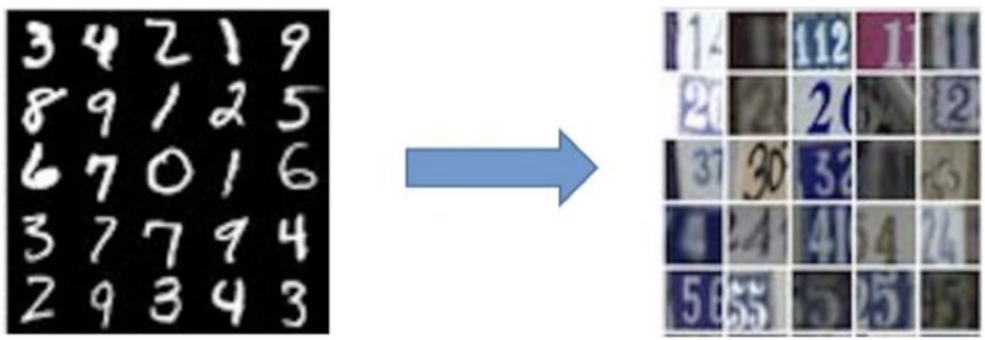


Example

Challenge: $p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y}) \neq p^{\mathrm{te}}(\mathbf{x}, \mathbf{y})$

Example: Dynamic environment, infection rates during a pandemic

Example: Domain adaptation



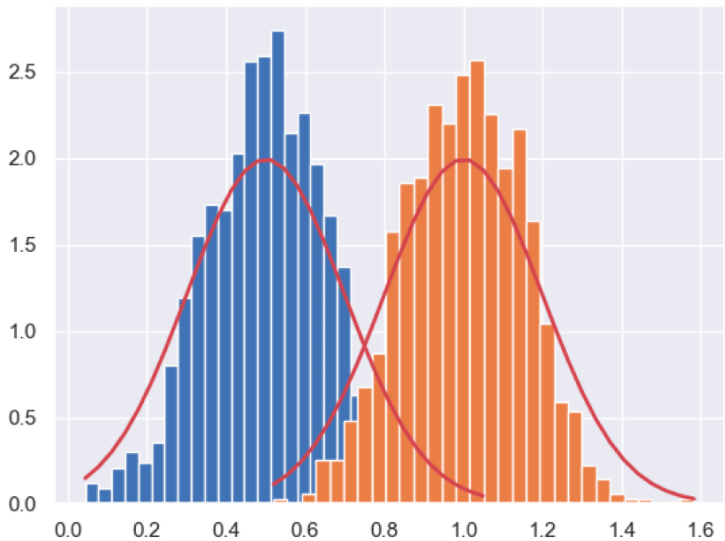
Source: MNIST

Target: SVHN

Distributions Shifts in Images (Data Augmentation Lecture)

Training images and test images are from different distributions:

- Different **resolution** (high res. SLR vs. low res. surveillance)
- Different **scale**, **perspective** and **pose** (closeup of face vs. full body)
- Different **lighting** and **colours** (indoors colour vs. outdoor IR/BW)
- Different **label distribution** (other cats in image, multiple cats in image)

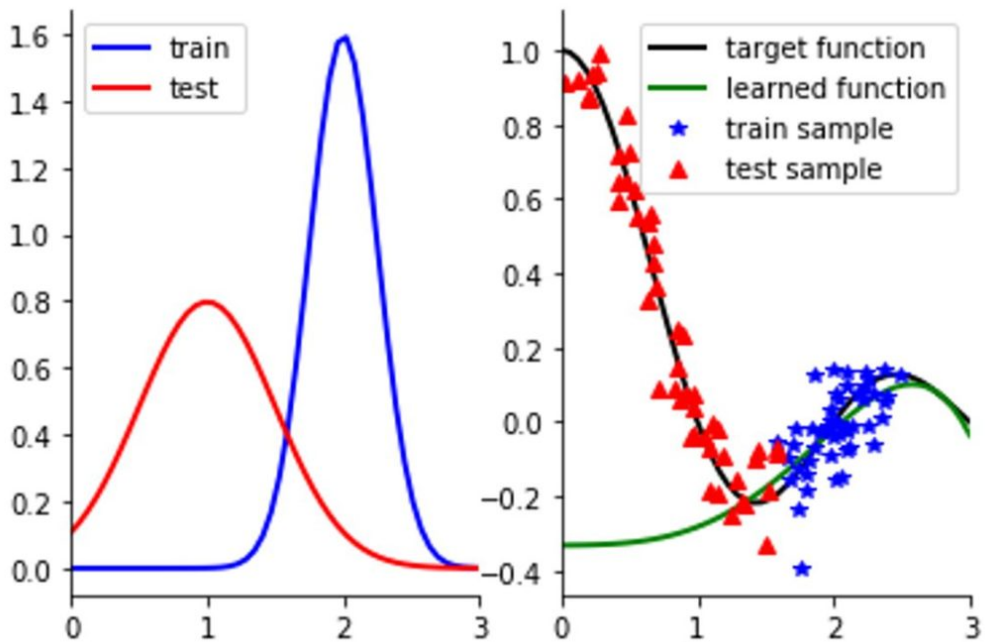


What is the Problem?

Distribution shifts can cause models to fail

Training and test input follow different distributions

Underlying target function remains unchanged



Types of Distribution Shift

General joint distribution $p(\mathbf{x}, \mathbf{y})$ factorization

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$$

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})p(\mathbf{x}|\mathbf{y})$$

Types of Distribution Shift

Covariate shift $p^{\mathrm{tr}}(\mathbf{x}) \neq p^{\mathrm{te}}(\mathbf{x})$

Label/target shift $p^{\mathrm{tr}}(\mathbf{y}) \neq p^{\mathrm{te}}(\mathbf{y})$

Output noise shift/concept shift $p^{\mathrm{tr}}(\mathbf{y}|\mathbf{x}) \neq p^{\mathrm{te}}(\mathbf{y}|\mathbf{x})$

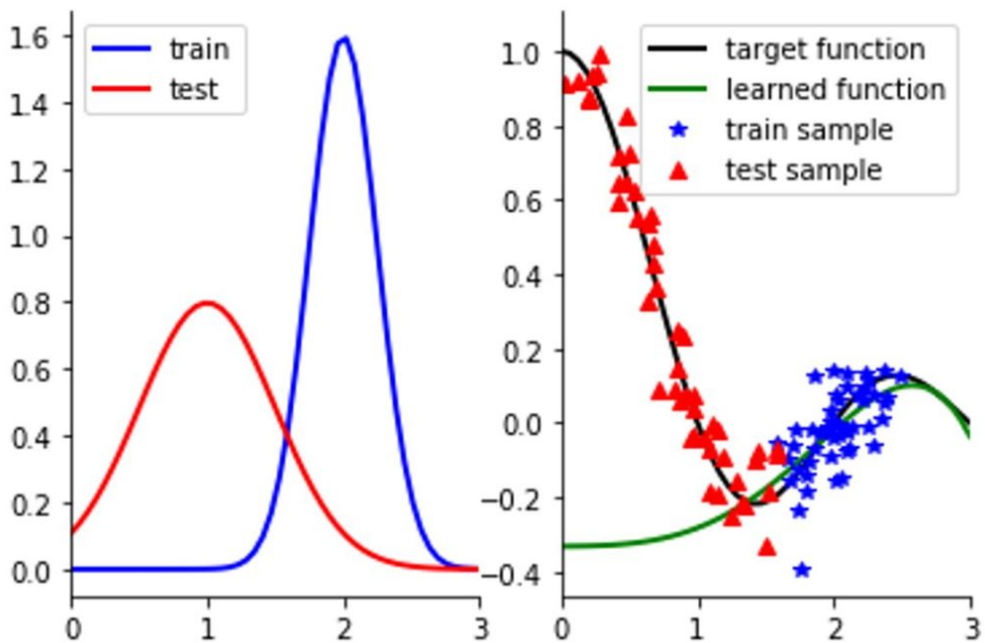
Class-conditional/concept shift $p^{\mathrm{tr}}(\mathbf{x}|\mathbf{y}) \neq p^{\mathrm{te}}(\mathbf{x}|\mathbf{y})$

Full distribution shift $p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y}) \neq p^{\mathrm{te}}(\mathbf{x}, \mathbf{y})$

Covariate Shift

Marginal training and test input distributions are different $p^{\mathrm{tr}}(\mathbf{x}) \neq p^{\mathrm{te}}(\mathbf{x})$

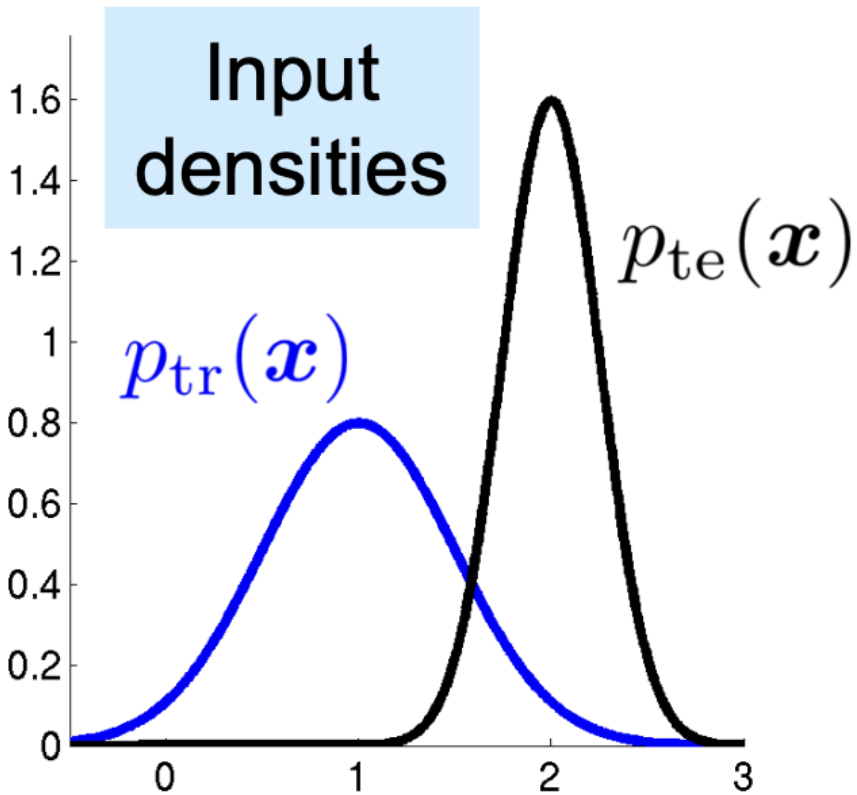
Conditional distribution is unchanged: $p^{\mathrm{tr}}(\mathbf{y}|\mathbf{x}) = p^{\mathrm{te}}(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$



Covariate Shift: Settings

Labeled training data: $\{(\mathbf{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n^{\mathrm{tr}}} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y})$

Unlabeled test data: $\{\mathbf{x}_j^{\mathrm{te}}\}_{j=1}^{n^{\mathrm{te}}} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{te}}(\mathbf{x})$

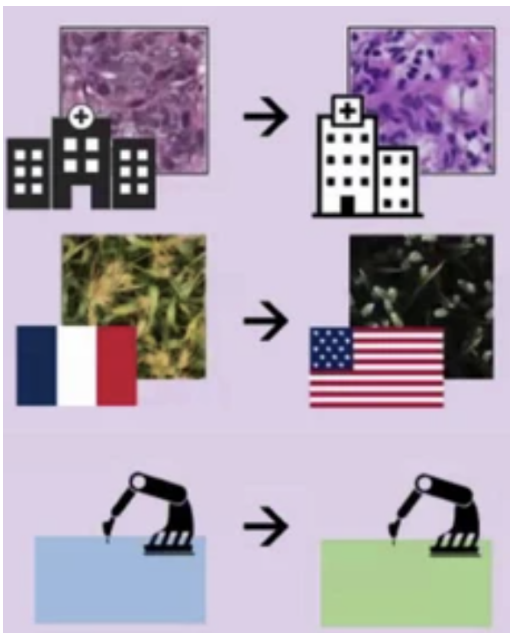


Covariate Shift: Examples

Medical imaging: various resolutions and equipment types

Agriculture: diverse weather conditions and soil types

Robotics: forest vs desert and sunny vs rainy scenarios



Label Shift: Assumptions and Example

Training and test label marginal distributions are different

$$p^{\mathrm{tr}}(\mathbf{y}) \neq p^{\mathrm{te}}(\mathbf{y})$$

But the conditional distribution is unchanged

$$p^{\mathrm{tr}}(\mathbf{x}|\mathbf{y}) = p^{\mathrm{te}}(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}|\mathbf{y})$$

COVID-19: varying infection rates across countries over time

Conditional symptom distribution: $p(\mathbf{x}|\mathbf{y} = \text{infected})$

Interactive Time-out

Join at menti.com | use code 8382 6709 (Quiz 1)

Interactive Time-out

Join at menti.com | use code 8382 6709 (Quiz 2)

Label Shift: Limitations

$p(\mathbf{x}|\mathbf{y} = \text{infected})$ may change with COVID variants such as Omicron

General shifts in conditional distribution $p(\mathbf{x}|\mathbf{y})$ due to diverse factors

Full distribution shift $p^{\text{tr}}(\mathbf{x}, \mathbf{y}) \neq p^{\text{te}}(\mathbf{x}, \mathbf{y})$

Generalized (Relaxed) Label Shift

Training and test label marginal distributions are different

$$p^{\mathrm{tr}}(\mathbf{y}) \neq p^{\mathrm{te}}(\mathbf{y})$$

Conditional distribution slightly changes

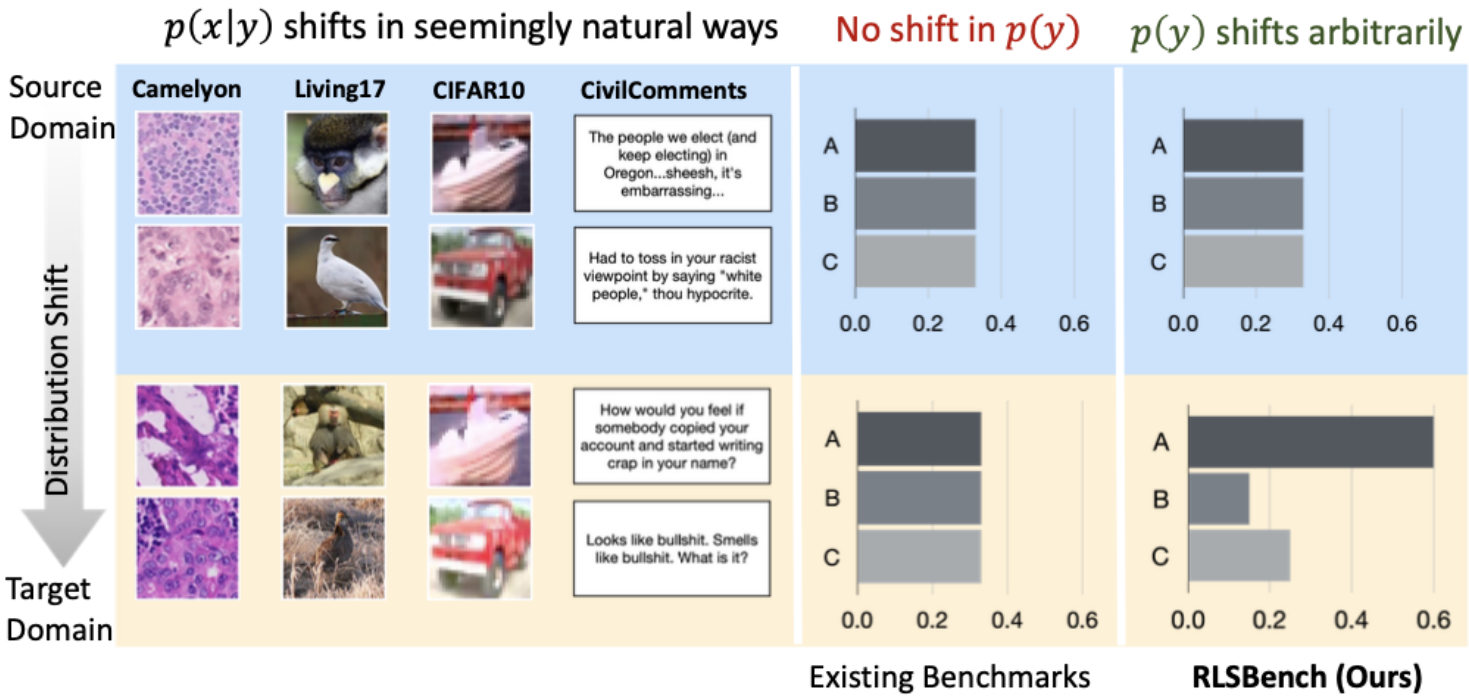
$$p^{\mathrm{tr}}(\mathbf{x}|\mathbf{y}) \neq p^{\mathrm{te}}(\mathbf{x}|\mathbf{y})$$

$D(p^{\mathrm{tr}}(\mathbf{x}|\mathbf{y}), p^{\mathrm{te}}(\mathbf{x}|\mathbf{y}))$ is bounded where D measures discrepancy of two distributions such as KL divergence

Unlike a distance (metric), divergence does not need to be symmetric or satisfy the triangle inequality

Generalized (Relaxed) Label Shift: Examples

Covariate shift and class imbalance



[1] https://arxiv.org/abs/2302.03020

Outline

1. Distribution Shift

2. Learning under Distribution Shift

3. Comparing Distributions and Discrepancy Measures

Recap on Empirical Risk Minimization (ERM)

Let $f_{\mathbf{w}} : \mathcal{X} \to \mathcal{Y}$ denote a predictor parameterized by $\mathbf{w} \in \mathbb{R}^d$. ERM:

$$\min_{\mathbf{w}} \left\{ R_n(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\mathbf{w}}(\mathbf{x}_i), \mathbf{y}_i) \right\}$$

where $\ell(f_{\mathbf{w}}(\mathbf{x}_i), \mathbf{y}_i)$ is the loss on $(\mathbf{x}_i, \mathbf{y}_i) \overset{\text{i.i.d.}}{\sim} p^{\text{tr}}(\mathbf{x}, \mathbf{y})$

ERM is over $p^{\text{tr}}(\mathbf{x}, \mathbf{y})$

$f_{\mathbf{w}^*}$ is not guaranteed to generalize well on $p^{\text{te}}(\mathbf{x}, \mathbf{y})$

Distributionally Robust Optimization (DRO) [3]

Let $f_{\mathbf{w}} : \mathcal{X} \to \mathcal{Y}$ denote a predictor parameterized by $\mathbf{w} \in \mathbb{R}^d$. DRO:

$$\min_{\mathbf{w}} \left\{ R(\mathbf{w}) := \sup_{Q \in \mathcal{Q}} \hat{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim Q}[\ell(f_{\mathbf{w}}(\mathbf{x}), \mathbf{y}))] \right\}$$

$\hat{\mathbb{E}}$ denotes empirical average

Minimize the worst-case loss over an uncertainty set of distributions $\mathcal{Q}$

$\mathcal{Q}$ includes the possible test distributions

DRO can lead to overly pessimistic models focusing on unrealistic worst-case distributions

Importance-weighted ERM (IWERM)

Let $f_{\mathbf{w}} : \mathcal{X} \to \mathcal{Y}$ denote a predictor parameterized by $\mathbf{w} \in \mathbb{R}^d$. IWERM:

$$\min_{\mathbf{w}} \left\{ \hat{\mathbb{E}}_{p^{\mathrm{tr}}(\mathbf{x},\mathbf{y})} \left[\frac{p^{\mathrm{te}}(\mathbf{x},\mathbf{y})}{p^{\mathrm{tr}}(\mathbf{x},\mathbf{y})} \ell(f_{\mathbf{w}}(\mathbf{x}),\mathbf{y}) \right] \right\}$$

Importance (density) ratio $r(\mathbf{x},\mathbf{y}) = \frac{p^{\mathrm{te}}(\mathbf{x},\mathbf{y})}{p^{\mathrm{tr}}(\mathbf{x},\mathbf{y})}$

Importance weighting is consistent

Learned function converges in probability to optimal test error minimizer

Why Does Importance Weighting (IWERM) Work? (won't be on the exam)

$$\frac{1}{n^{\mathrm{tr}}} \sum_{i=1}^{n^{\mathrm{tr}}} \frac{p^{\mathrm{te}}(\mathbf{x}_i^{\mathrm{tr}}, \mathbf{y}_i^{\mathrm{tr}})}{p^{\mathrm{tr}}(\mathbf{x}_i^{\mathrm{tr}}, \mathbf{y}_i^{\mathrm{tr}})} \ell(f_{\mathbf{w}}(\mathbf{x}_i^{\mathrm{tr}}), \mathbf{y}_i^{\mathrm{tr}})$$

$$\xrightarrow{n^{\mathrm{tr}} \to \infty} \mathbb{E}_{p^{\mathrm{tr}}(\mathbf{x},\mathbf{y})} \left[\frac{p^{\mathrm{te}}(\mathbf{x}, \mathbf{y})}{p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y})} \ell(f_{\mathbf{w}}(\mathbf{x}), \mathbf{y}) \right]$$

$$= \int_{\mathcal{Y}} \int_{\mathcal{X}} p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y}) \frac{p^{\mathrm{te}}(\mathbf{x}, \mathbf{y})}{p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y})} \ell(f_{\mathbf{w}}(\mathbf{x}), \mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}$$

$$= \int_{\mathcal{Y}} \int_{\mathcal{X}} p^{\mathrm{te}}(\mathbf{x}, \mathbf{y}) \ell(f_{\mathbf{w}}(\mathbf{x}), \mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}$$

$$= \mathbb{E}_{p^{\mathrm{te}}(\mathbf{x},\mathbf{y})} \left[\ell(f_{\mathbf{w}}(\mathbf{x}), \mathbf{y}) \right]$$

$$= \text{True risk of } f_{\mathbf{w}}$$

First line holds due to "law of large numbers"

Convergence in probability is established using the law of large numbers

IWERM under Covariate and Label Shift

Under covariate shift $p^{\mathrm{tr}}(\mathbf{y}|\mathbf{x}) = p^{\mathrm{te}}(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$, the importance weight (density ratio) becomes:

$$r(\mathbf{x}) = \frac{p^{\mathrm{te}}(\mathbf{x})}{p^{\mathrm{tr}}(\mathbf{x})}$$

Under label shift $p^{\mathrm{tr}}(\mathbf{x}|\mathbf{y}) = p^{\mathrm{te}}(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}|\mathbf{y})$, the importance weight (probability ratio) becomes:

$$r(\mathbf{y}) = \frac{p^{\mathrm{te}}(\mathbf{y})}{p^{\mathrm{tr}}(\mathbf{y})}$$

Importance Weight (Probability Ratio) under Label Shift

Class	$p^{\mathrm{tr}}(y)$	$p^{\mathrm{te}}(y)$	Test-to-Train Ratio
A	0.6	0.3	$\frac{0.3}{0.6} = 0.5$
B	0.3	0.5	$\frac{0.5}{0.3} \approx 1.67$
C	0.1	0.2	$\frac{0.2}{0.1} = 2.0$

Train and test class distributions with test-to-train probability ratios.

Importance Weight (Probability Ratio) under Label Shift

Class	$p^{\mathrm{tr}}(y)$	$p^{\mathrm{te}}(y)$	Test-to-Train Ratio
0	0.20	0.10	$\frac{0.10}{0.20} = 0.5$
1	0.15	0.20	$\frac{0.20}{0.15} \approx 1.33$
2	0.10	0.05	$\frac{0.05}{0.10} = 0.5$
3	0.10	0.15	$\frac{0.15}{0.10} = 1.5$
4	0.10	0.05	$\frac{0.05}{0.10} = 0.5$
5	0.10	0.10	$\frac{0.10}{0.10} = 1.0$
6	0.05	0.10	$\frac{0.10}{0.05} = 2.0$
7	0.05	0.05	$\frac{0.05}{0.05} = 1.0$
8	0.10	0.15	$\frac{0.15}{0.10} = 1.5$
9	0.05	0.05	$\frac{0.05}{0.05} = 1.0$

Class 0, 2, and 4 are overrepresented in training data

Class 1, 3, 6, 8 are underrepresented in training data

Class 5, 7, 9 have balanced distributions

Interactive Time-out

Join at menti.com | use code 8382 6709 (Quiz 3)

Density Ratio Estimation for IWERM

For IWERM, individual densities are actually not necessary. We need an estimate of density ratio:

$$r(\mathbf{x}, \mathbf{y}) = \frac{p^{\mathrm{te}}(\mathbf{x}, \mathbf{y})}{p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y})}$$

Density ratio estimation is challenging specially for high-dimensional data

Estimating the ratio is in general more effective than estimating densities in the numerator and denominator separately

A Simple Case: Label Shift

Under label shift $p^{\mathrm{tr}}(\mathbf{x}|\mathbf{y}) = p^{\mathrm{te}}(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}|\mathbf{y})$, the importance weight (probability ratio) becomes:

$$r(\mathbf{y}) = \frac{p^{\mathrm{te}}(\mathbf{y})}{p^{\mathrm{tr}}(\mathbf{y})}$$

$\mathbf{y}$ is usually a low-dimensional discrete vector

Ratio estimation can be written as a convex optimization problem

Basic settings: Label Shift

Labeled training data: $\{(\mathbf{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n^{\mathrm{tr}}} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y})$

Unlabeled test data: $\{\mathbf{x}_j^{\mathrm{te}}\}_{j=1}^{n^{\mathrm{te}}} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{te}}(\mathbf{x})$

Probability ratio estimation using the collected samples:

$$\hat{r}(\mathbf{y}) \approx \frac{p^{\mathrm{te}}(\mathbf{y})}{p^{\mathrm{tr}}(\mathbf{y})}$$

Maximum Likelihood Estimation: Discrepancy Minimization (won't be on the exam)

Given the mixture form of the density:

$$p_r(\mathbf{x}) = \sum_{i=1}^{|\mathcal{Y}|} p(\mathbf{x}|\mathbf{y}_i)p^{\text{tr}}(\mathbf{y}_i)r(\mathbf{y}_i) = \sum_{i=1}^{|\mathcal{Y}|} p^{\text{tr}}(\mathbf{x}, \mathbf{y}_i)r(\mathbf{y}_i)$$

With the true ratio $r = r^*$, we have for all $\mathbf{x}$:

$$\sum_{i=1}^{|\mathcal{Y}|} p^{\text{tr}}(\mathbf{x}, \mathbf{y}_i)r^*(\mathbf{y}_i) = \sum_{i=1}^{|\mathcal{Y}|} p^{\text{te}}(\mathbf{x}, \mathbf{y}_i) = p^{\text{te}}(\mathbf{x})$$

Estimate r by minimizing the KL divergence $D_{KL}(p^{\text{te}}(\mathbf{x}), p_r(\mathbf{x}))$:

$$\hat{r} = \arg\max_r \mathbb{E}_{\text{te}} \left[\log \sum_{i=1}^{|\mathcal{Y}|} p^{\text{tr}}(\mathbf{y}_i|\mathbf{x})r_i \right]$$

Maximum Likelihood Estimation: Discrepancy Minimization (won't be on the exam)

With a black-box predictor $f(\mathbf{x}) \approx p^{\text{tr}}(\mathbf{y}|\mathbf{x})$ as in [2], the convex problem to estimate density ratio is:

$$r_f = \arg\max_r \mathbb{E}_{\text{te}} \left[\log(f(\mathbf{x})^\top r)\right]$$

The optimization is efficient with convex solver

Predictor calibration plays a significant role

Model calibration: predicted probabilities reflect true likelihoods

If a model predicts 100 images have a 95% probability of being cats, ~ 95 images are indeed cats

Interactive Time-out

Join at menti.com | use code 8382 6709 (Quiz 4)

Outline

1. Distribution Shift

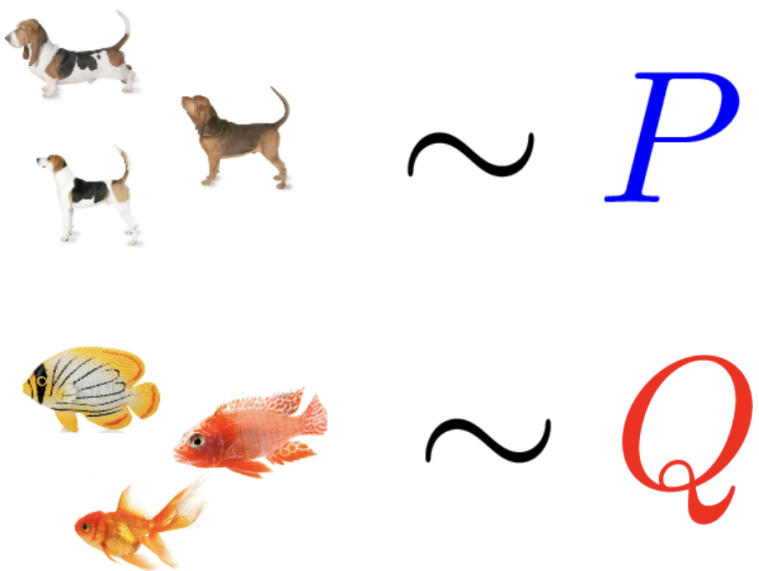
2. Learning under Distribution Shift

3. Comparing Distributions and Discrepancy Measures

Comparing Two Distributions

Given: Samples from unknown distributions P and Q

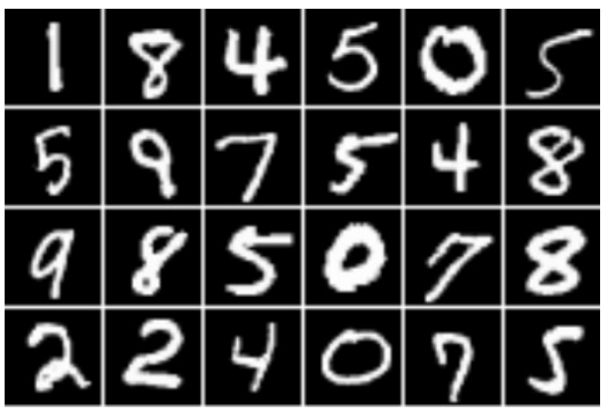
Goal: Do P and Q differ?



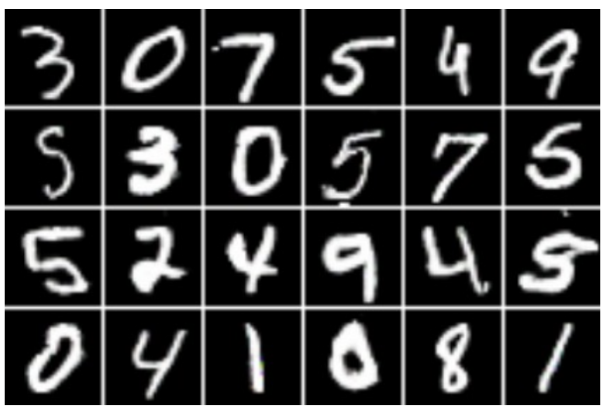
An Example in ML

Two collections of samples $(\mathbf{x}, \mathbf{y})$ from unknown distributions P and Q

Do P and Q differ?



MNIST samples

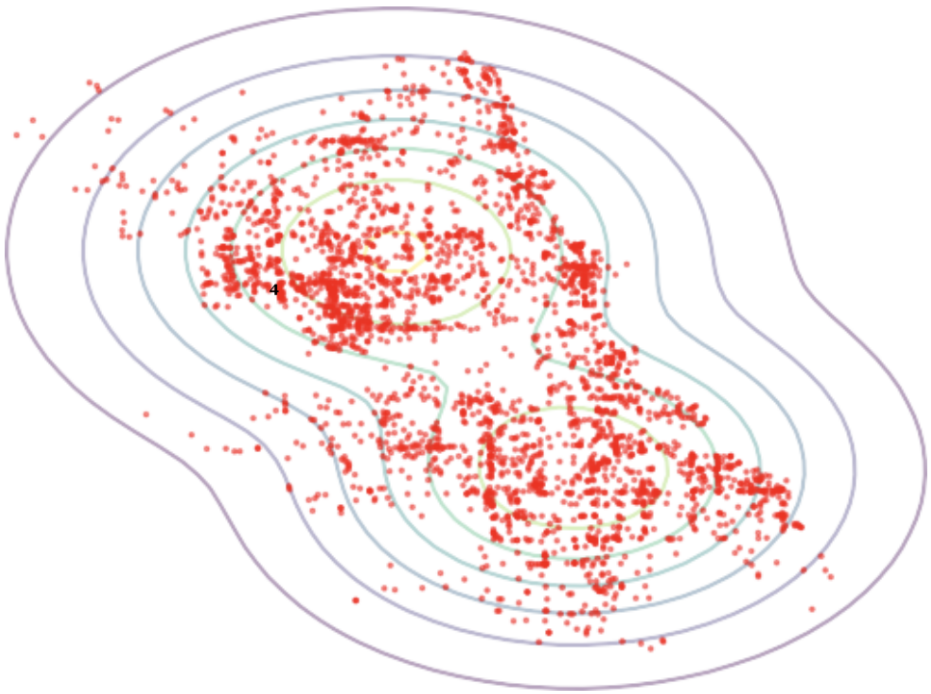


Samples from a GAN

Performance Measure of Fit

Given a Gaussian mixture model f with two components and samples from true underlying Q

Is f a good fit for Q?



Chicago crime data

Testing Independence

Given: samples from a distribution $P_{\mathbf{X},\mathbf{Y}}$

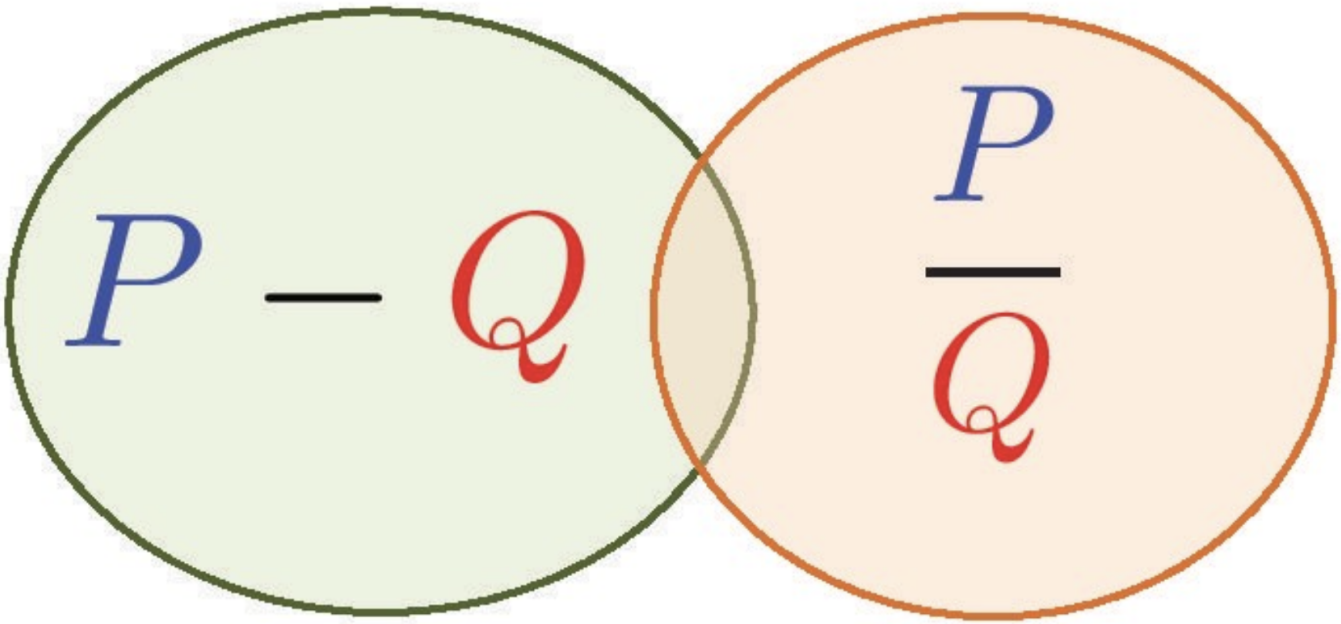
Are $\mathbf{X}$ and $\mathbf{Y}$ independent?

X	Y
	A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.
	Their noses guide them through life, and they're never happier than when following an interesting scent.
	A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

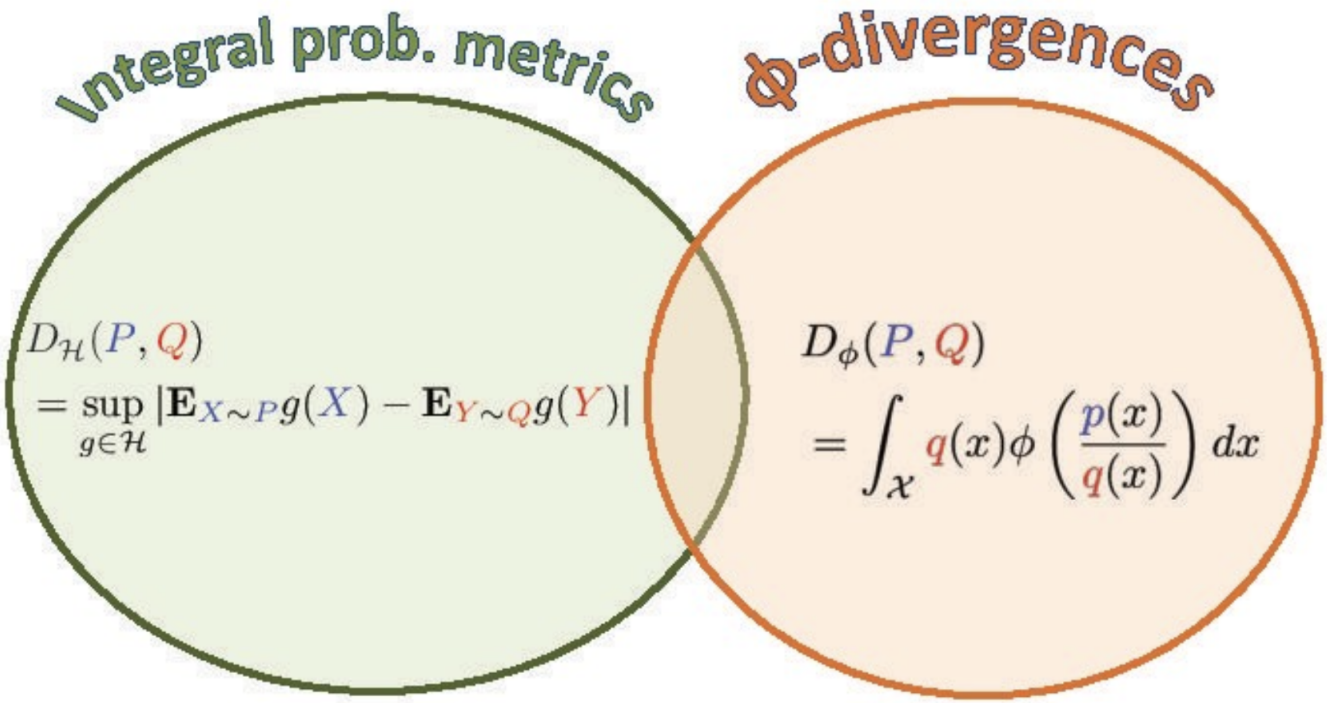
Text from dogtime.com and petfinder.com

Discrepancy Measures

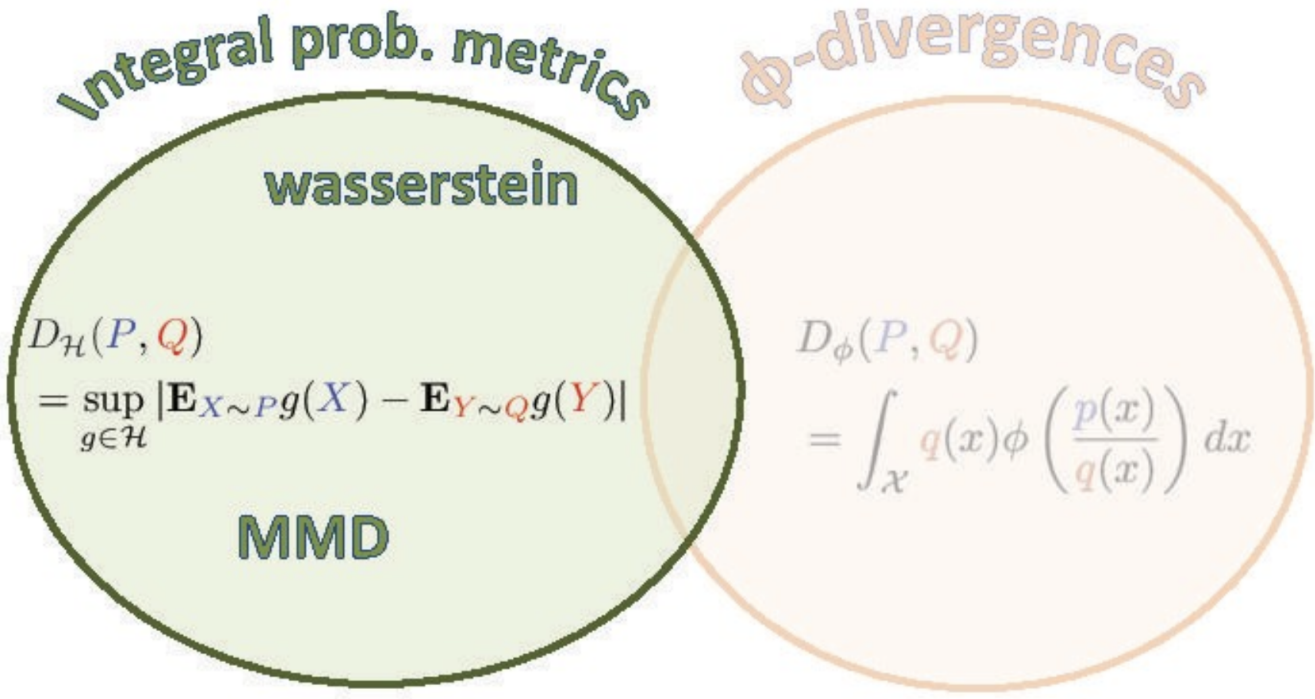
The following slides won't be on the exam



Discrepancy Measures



Discrepancy Measures



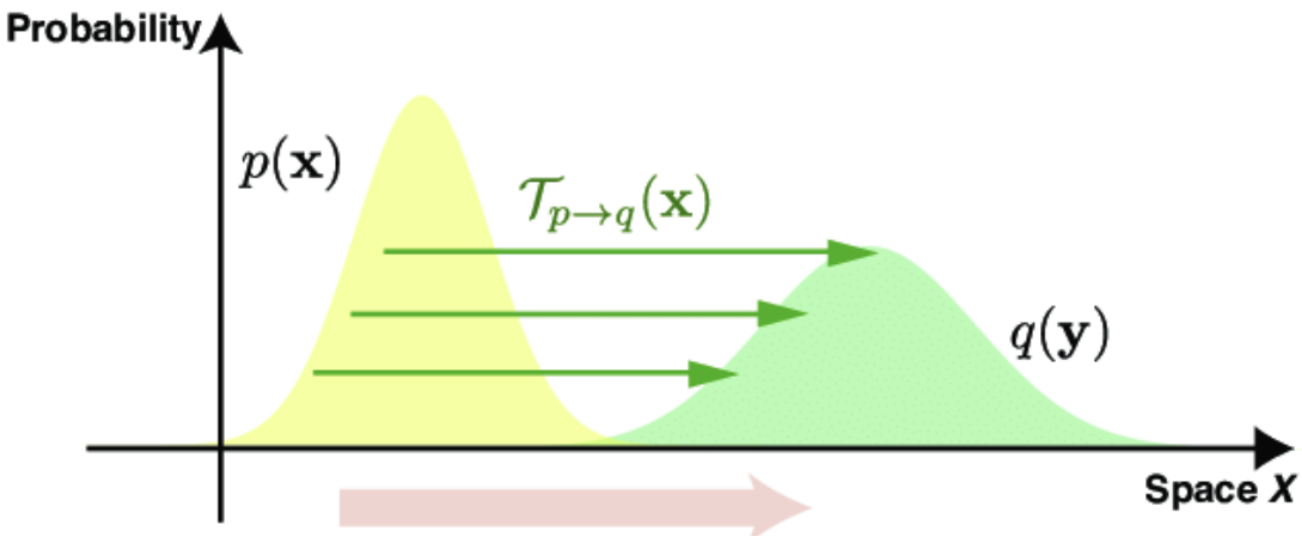
Wasserstein Distance

Earth Mover distance or Wasserstein-1:

$$W(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

$\Pi(P, Q)$ is the set of all couplings of P and Q, which are all joint distributions on $\mathcal{X}^2$ with marginals P and Q

Optimal transport:



Kantorovich-Rubinstein Duality

Kantorovich-Rubinstein duality:

$$W(P, Q) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{X \sim Q}[f(X)]$$

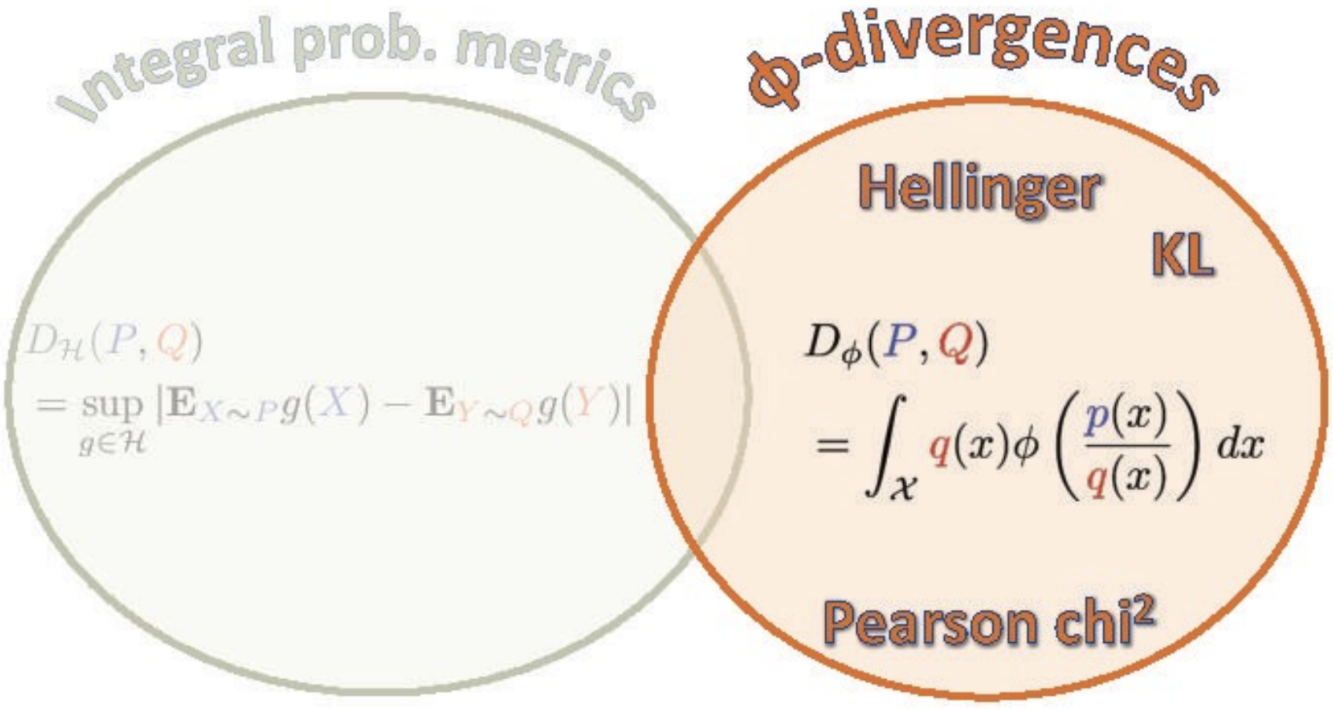
where $f : \mathcal{X} \to \mathbb{R}$ and is 1-Lipschitz.

A function $f : X \to Y$ is K-Lipschitz if there exists a $K \geq 0$ s.t.

$$\|f(x_1) - f(x_2)\| \leq K \|x_1 - x_2\|$$

for all $x_1, x_2 \in X$

ϕ-divergences



The ϕ-divergences

Define the ϕ-divergence (f-divergence):

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

where ϕ is convex with $\phi(1) = 0$.

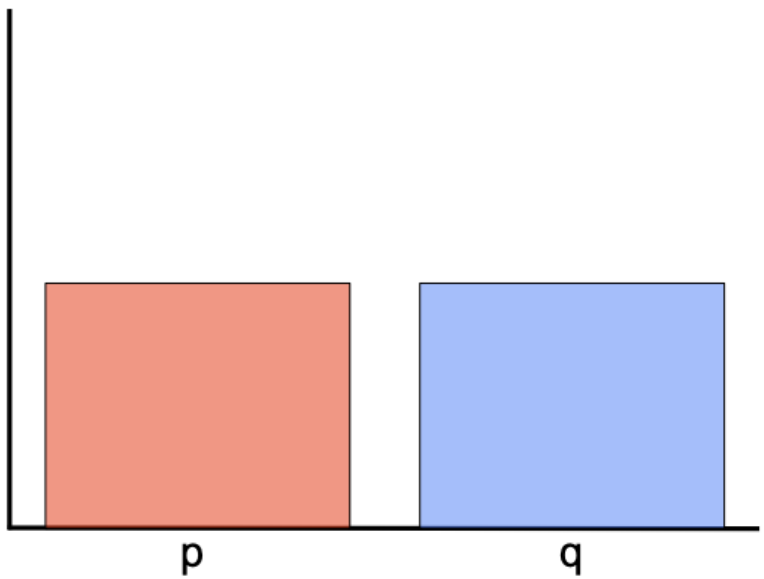
Example: $\phi(x) = x\log(x)$ gives KL divergence:

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

The limitation of ϕ-divergences

Support(p) and Support(q) are disjoint

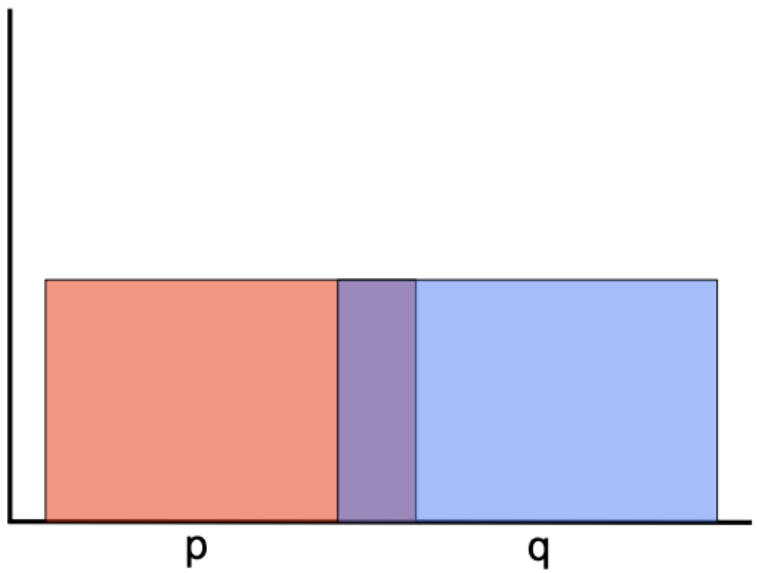
$D_{KL}(p, q) = \infty$ and $D_{KL}(q, p) = \infty$



The limitation of ϕ-divergences

Supports overlap, but neither is a subset of the other

$D_{KL}(p, q) = \infty$, $D_{KL}(q, p) = \infty$

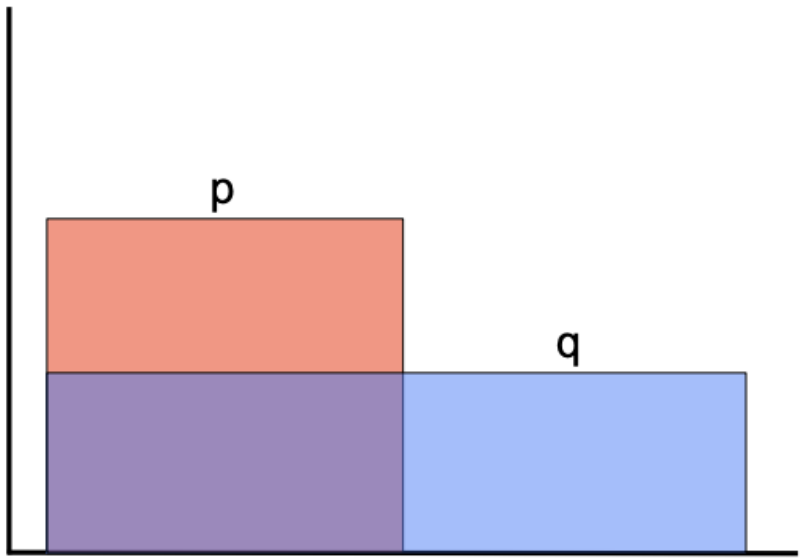


The limitation of ϕ-divergences

Support(p) $\subseteq$ Support(q)

$D_{KL}(p, q)$ is finite, $D_{KL}(q, p) = \infty$

KL divergence is asymmetric



Wasserstein distance is symmetric and defined for non-overlapping distributions

References

Masashi Sugiyama - Lecture on Density Ratio Estimation, ICLR 2023: https://www.ms.k.u-tokyo.ac.jp/sugi/slide/20230502_ICLR.pdf

Venkatesan Guruswami - Introduction to Information Theory: https://www.cs.cmu.edu/~venkatg/teaching/ITCS-spr2013/notes/lect-jan22.pdf

[1] Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Lipton. Rlsbench: Domain adaptation under relaxed label shift, 02 2023.

[2] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C. Lipton. A unified view of label shift estimation, 2020.

[3] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020.