Adversarial Support Alignment

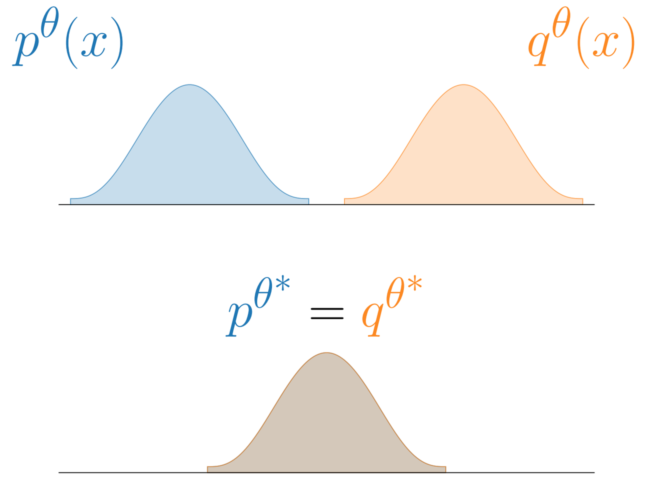
Shangyuan Tong[1] Timur Garipov[1] Yang Zhang[2] Shiyu Chang[3] Tommi Jaakkola[1]

[1]MIT CSAIL [2]MIT-IBM Watson AI Lab [3]UC Santa Barbara

Background: distribution alignment 1

Given $\mathcal{P} = \{p^\theta \mid \theta \in \Theta\}$ $\mathcal{Q} = \{q^\theta \mid \theta \in \Theta\}$

Find $\theta^* : p^{\theta^*} = q^{\theta^*}$



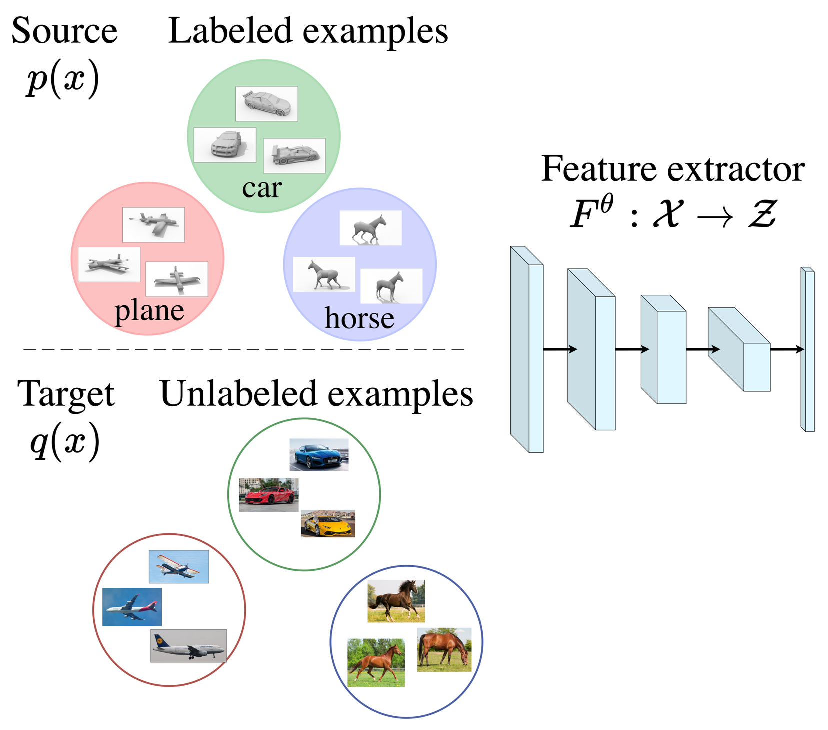
- **Generative Models** (GAN, Goodfellow et al. 2014)
  - alignment of generated and data distributions.
- **Domain Adaptation** (DANN, Ganin et al. 2016)
  - alignment of representations across domains.

Domain adaptation

Goal: learn classifier on target domain using
- labeled dataset in source domain.
- unlabeled dataset in target domain.

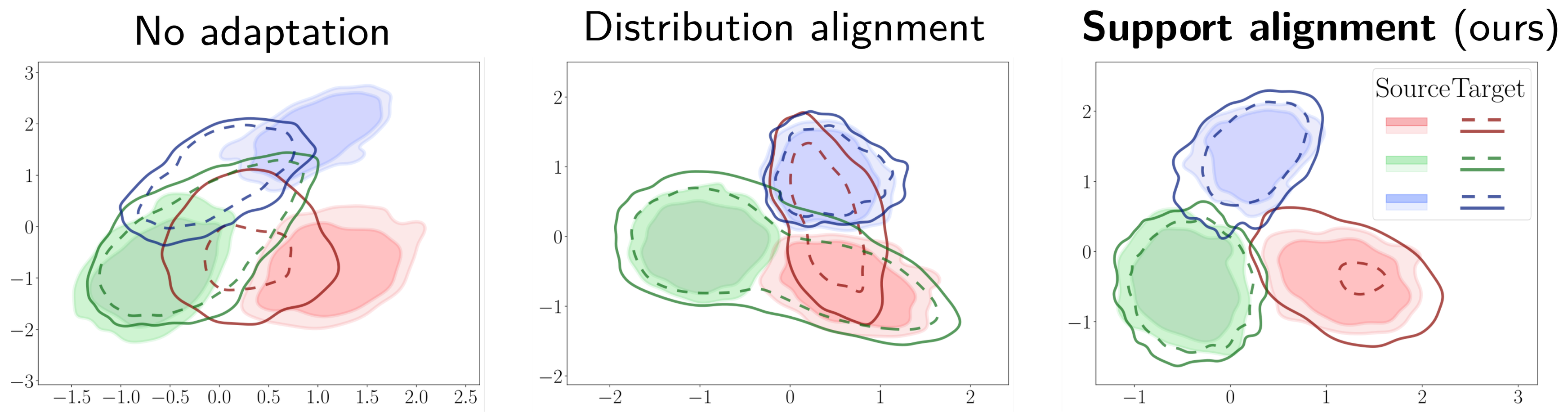
Distribution alignment approach:
Feature extractor network $F^\theta : \mathcal{X} \to \mathcal{Z}$

Training:
- separate classes in source.
- align distributions of source/target features.



Motivation 2

- Distribution alignment is not always desired.
- Example: domain adaptation under **label distribution shift.**
- We propose to align only the **supports** of distributions.



Target accuracy: 63% Target accuracy: 75% Target accuracy: 94%

Source class distribution [33%, 33%, 33%]

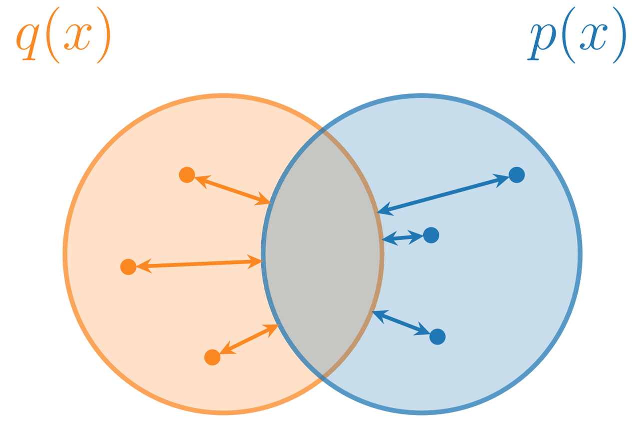
Target class distribution [23%, 65%, 12%]

Support divergence 3

Symmetric Support Difference (SSD) divergence

$$\mathcal{D}_\triangle(p,q) = \mathop{\mathbb{E}}_{x^q\sim q}\Big[d(x^q, \mathrm{supp}(p))\Big] + \mathop{\mathbb{E}}_{x^p\sim p}\Big[d(x^p, \mathrm{supp}(q))\Big]$$

$$d(x^q, \mathrm{supp}(p)) = \inf_{x^p \in \mathrm{supp}(p)} d(x^q, x^p) \qquad d(x^p, \mathrm{supp}(q)) = \inf_{x^q \in \mathrm{supp}(q)} d(x^p, x^q)$$



$$1)\ \mathcal{D}_\triangle(p,q) \ge 0\ \forall p, q; \qquad 2)\ \mathcal{D}_\triangle(p,q) = 0 \iff \mathrm{supp}(p) = \mathrm{supp}(q)$$

Support alignment via log-loss discriminator 4

$$\sup_{f:\mathcal{X}\to[0,1]} \mathop{\mathbb{E}}_{x\sim p}[\log f(x)] + \mathop{\mathbb{E}}_{y\sim q}[\log(1-f(y))]$$

Optimal solution: $f^*(x) = \dfrac{p(x)}{p(x)+q(x)}$

Theorem

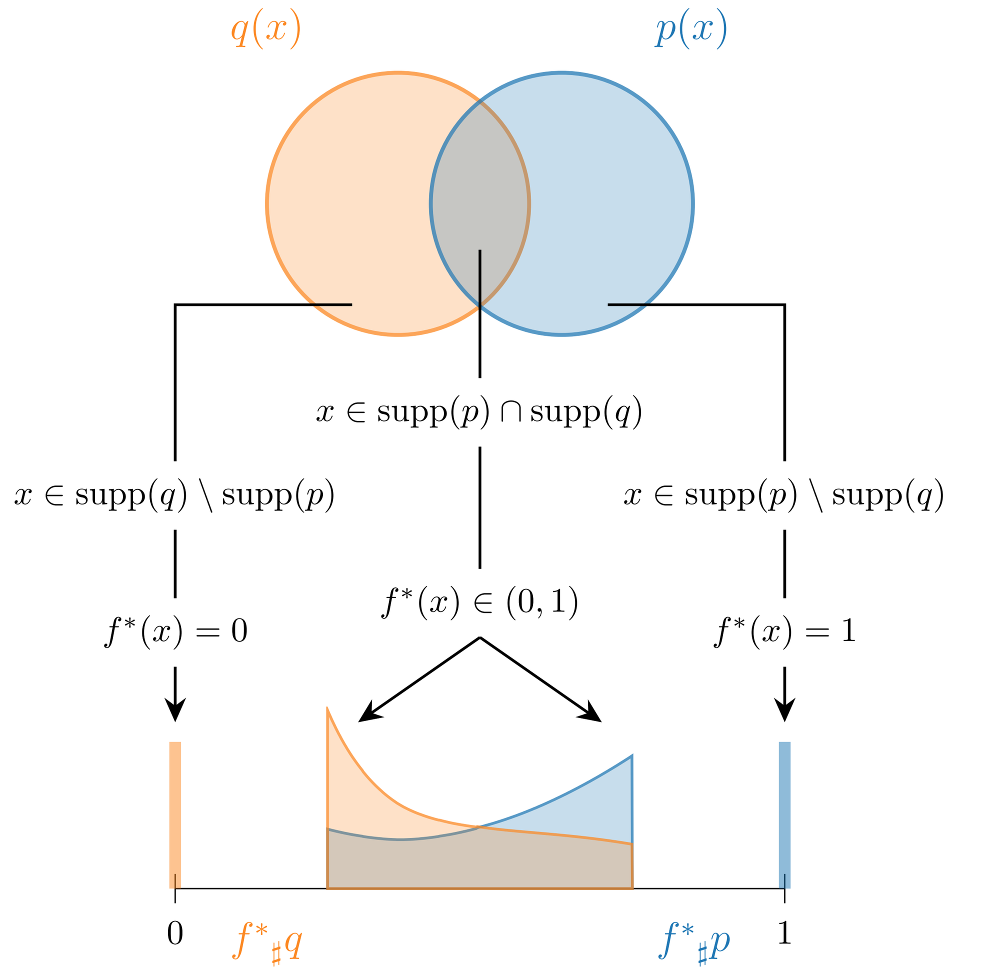
The mapping $f^* : \mathcal{X} \to [0,1]$ realized by the optimal discriminator preserves support discrepancy

$$\mathcal{D}_\triangle(p,q) = 0 \iff \mathcal{D}_\triangle(f^*_\sharp p, f^*_\sharp q) = 0$$

$f^*_\sharp p,\ f^*_\sharp q$ — pushforward distributions

Remark: the result holds for $g : \mathcal{X} \to \mathbb{R}$

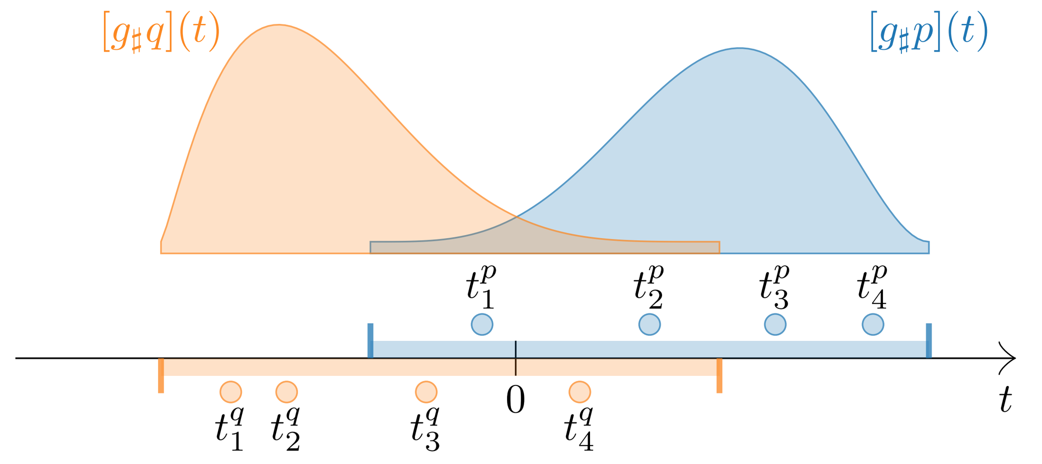
$$g(x) : f(x) = \mathrm{sigmoid}(g(x))$$



Method: Adversarial Support Alignment (ASA) 5

Discriminator's objective: $\min_{g:\mathcal{X}\to\mathbb{R}} \mathcal{L}_D(\theta, g)$

$$\mathcal{L}_D(\theta, g) = \mathop{\mathbb{E}}_{x\sim p^\theta}\left[\log\left(1+e^{-g(x)}\right)\right] + \mathop{\mathbb{E}}_{x\sim q^\theta}\left[\log\left(1+e^{g(x)}\right)\right]$$

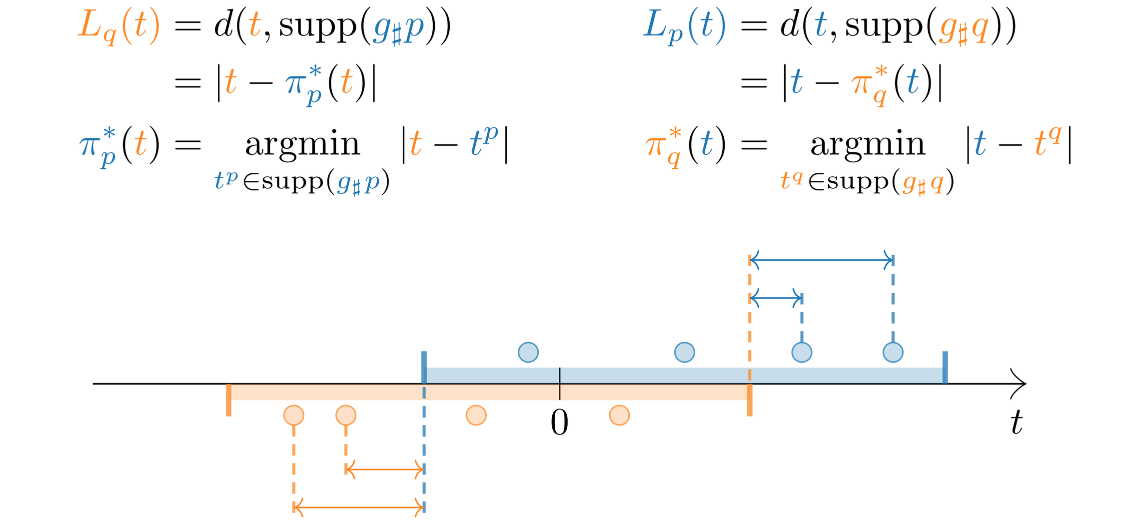


Alignment objective: $\min_\theta \mathcal{L}_A(\theta, g)$

$$\mathcal{L}_A(\theta, g) = \mathop{\mathbb{E}}_{x\sim p^\theta}[d(g(x), \mathrm{supp}(g_\sharp q))] + \mathop{\mathbb{E}}_{x\sim q^\theta}[d(g(x), \mathrm{supp}(g_\sharp p))]$$

$$L_q(t) = d(t, \mathrm{supp}(g_\sharp p)) = |t - \pi_p^*(t)| \qquad L_p(t) = d(t, \mathrm{supp}(g_\sharp q)) = |t - \pi_q^*(t)|$$

$$\pi_p^*(t) = \mathop{\mathrm{argmin}}_{t^p \in \mathrm{supp}(g_\sharp p)} |t - t^p| \qquad \pi_q^*(t) = \mathop{\mathrm{argmin}}_{t^q \in \mathrm{supp}(g_\sharp q)} |t - t^q|$$



Results: domain adaptation under label distribution shift 6

Average and minimum class accuracy (%) on USPS→MNIST, LeNet across different levels of shifts in label distributions (α).

Algorithm	$\alpha = 0.0$ no shift average	min	$\alpha = 1.0$ average	min	$\alpha = 1.5$ average	min	$\alpha = 2.0$ severe shift average	min
No DA	71.9	20.3	72.9	25.8	71.3	27.5	71.3	16.6
DANN	97.8	96.0	83.5	25.1	70.0	01.1	57.8	00.9
VADA	**98.0**	96.2	88.2	48.9	78.2	06.6	61.9	01.4
IWDAN	97.5	95.7	95.7	81.3	86.5	15.2	74.4	07.3
IWCDAN	**98.0**	**96.6**	**96.7**	85.1	91.3	66.5	77.5	22.2
sDANN-4	87.4	05.6	94.9	**85.7**	86.8	21.6	81.5	39.3
ASA-sq (ours)	93.7	89.2	92.3	83.5	90.9	69.9	87.2	62.5
ASA-abs (ours)	94.1	88.9	92.8	78.9	**92.5**	**82.4**	**90.4**	**68.4**

STL→CIFAR, DeepCNN

Algorithm	$\alpha = 0.0$ no shift average	min	$\alpha = 2.0$ severe shift average	min
No DA	69.9	49.8	65.8	43.7
DANN	75.3	54.6	63.3	27.0
VADA	**76.7**	**56.9**	63.2	25.5
IWDAN	69.9	50.5	64.4	36.8
IWCDAN	70.1	47.8	64.5	37.0
sDANN-4	71.8	52.1	66.4	39.0
ASA-sq (ours)	71.7	52.9	**68.1**	**44.7**
ASA-abs (ours)	71.6	49.0	67.8	40.9

VisDA-17, ResNet-50

Algorithm	$\alpha = 0.0$ no shift average	min	$\alpha = 2.0$ severe shift average	min
No DA	49.5	22.2	45.3	19.5
DANN	**75.4**	36.7	43.1	03.6
VADA	75.3	40.5	43.9	08.5
IWDAN	73.2	31.7	45.1	04.6
IWCDAN	71.6	27.6	38.3	00.6
sDANN-4	72.4	37.8	50.7	18.6
ASA-sq (ours)	64.9	35.7	51.9	18.3
ASA-abs (ours)	64.8	**40.6**	**52.5**	**19.7**

Takeaways 7

- **Support alignment** as extreme relaxation of **distribution alignment**.
- **Adversarial Support Alignment (ASA):** method to align distribution supports.
- Evaluation on domain adaptation datasets under **label distribution shift**.
- Bonus: spectrum of **relaxed alignment** approaches based on **optimal transport**.

Wasserstein distance	**β-Wasserstein distance** (Wu et al. 2019)	**SSD divergence**
$\mathcal{D}_W(p,q) = \inf_\gamma \mathop{\mathbb{E}}_{(x,y)\sim\gamma}[d(x,y)]$, s.t. $\int \gamma(x,y)dy = p(x)$, $\int \gamma(x,y)dx = q(y)$	$\mathcal{D}_W^{\beta,\beta}(p,q) = \mathcal{D}_W^\beta(p,q) + \mathcal{D}_W^\beta(q,p)$; $\mathcal{D}_W^\beta(p,q) = \inf_\gamma \mathop{\mathbb{E}}_{(x,y)\sim\gamma}[d(x,y)]$, s.t. $\int \gamma(x,y)dy = p(x)$, $\int \gamma(x,y)dx \le (1+\beta)q(y)$	$\mathcal{D}_\triangle(p,q) = \mathcal{D}_W^\infty(p,q) + \mathcal{D}_W^\infty(q,p)$; $\mathcal{D}_W^\infty(p,q) = \mathop{\mathbb{E}}_{x\sim p(x)}\left[d(x, \mathrm{supp}(q))\right]$
$\mathcal{D}_W(p,q) = 0 \Updownarrow p = q$	$\mathcal{D}_W^{\beta,\beta}(p,q) = 0 \Updownarrow \frac{1}{1+\beta} \le \frac{p}{q} \le 1+\beta$	$\mathcal{D}_\triangle(p,q) = 0 \Updownarrow \mathrm{supp}(p) = \mathrm{supp}(q)$