Adversarial Support Alignment

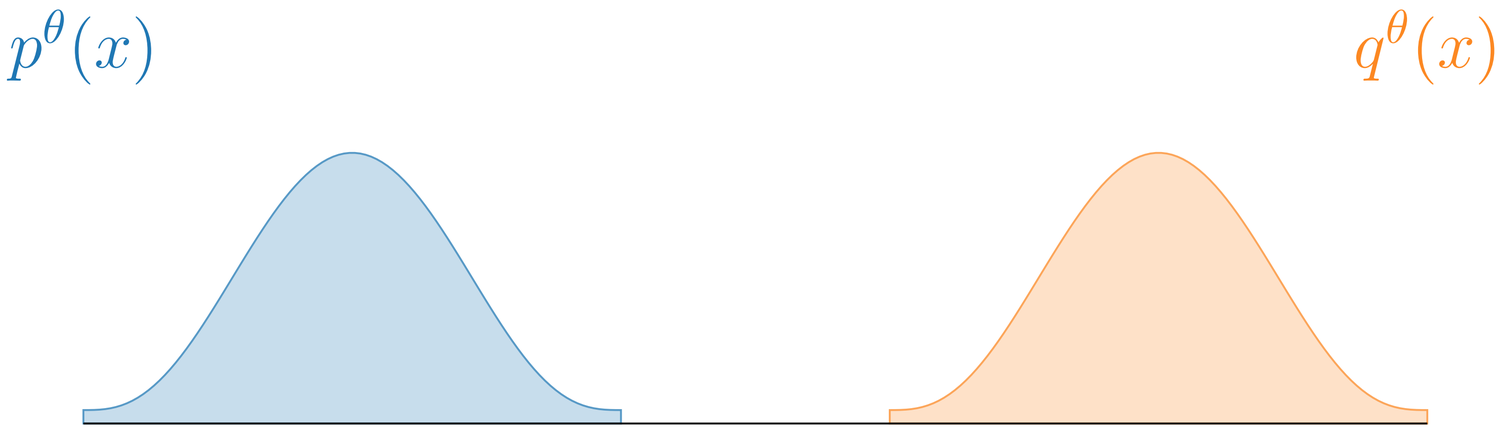
Shangyuan Tong[1] Timur Garipov[1]

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

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

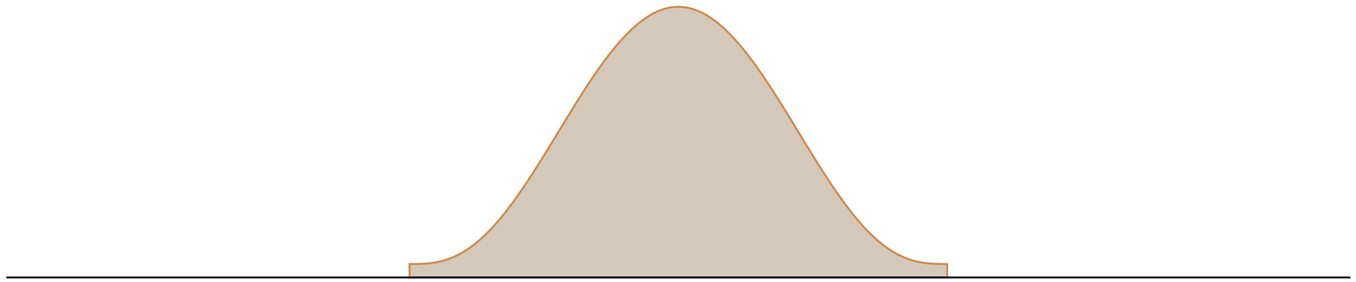


Background: distribution alignment

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

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

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



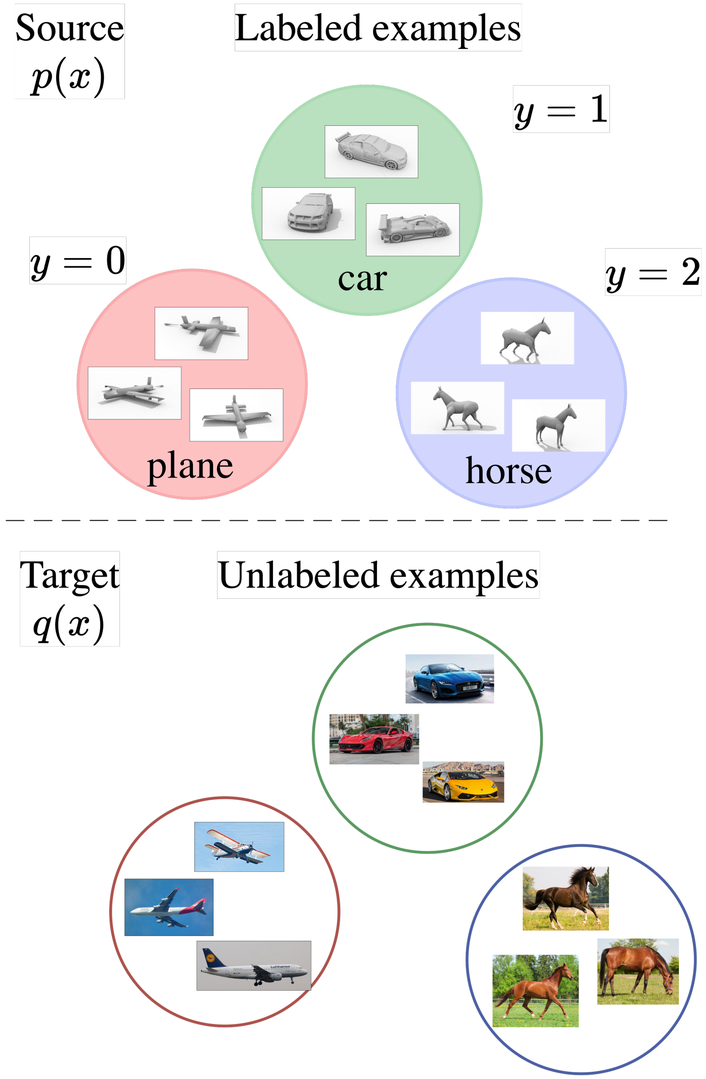
Background: distribution alignment

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

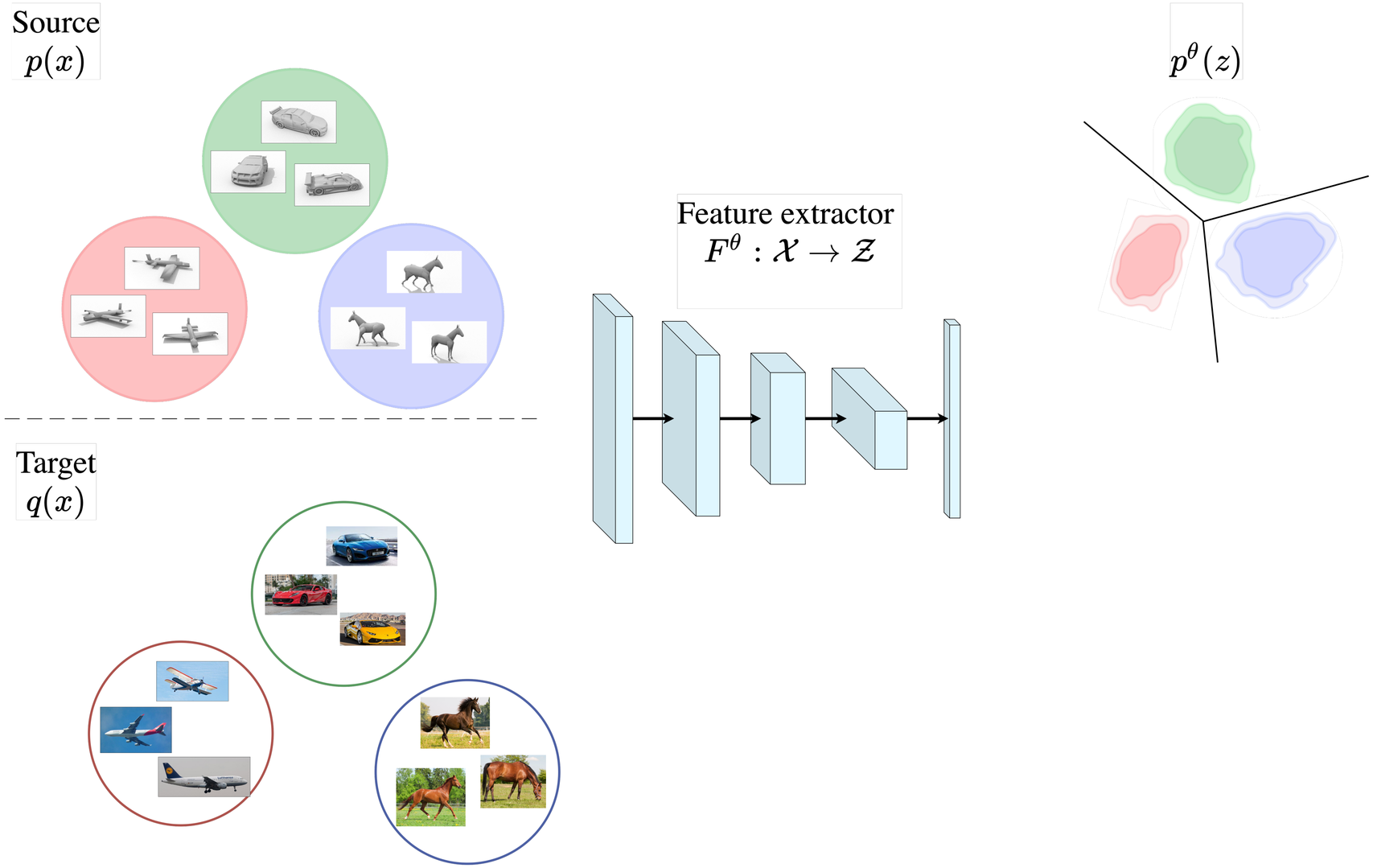
Find $\theta^* : \; p^{\theta^*} = q^{\theta^*}$

- **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

Background: distribution alignment in domain adaptation

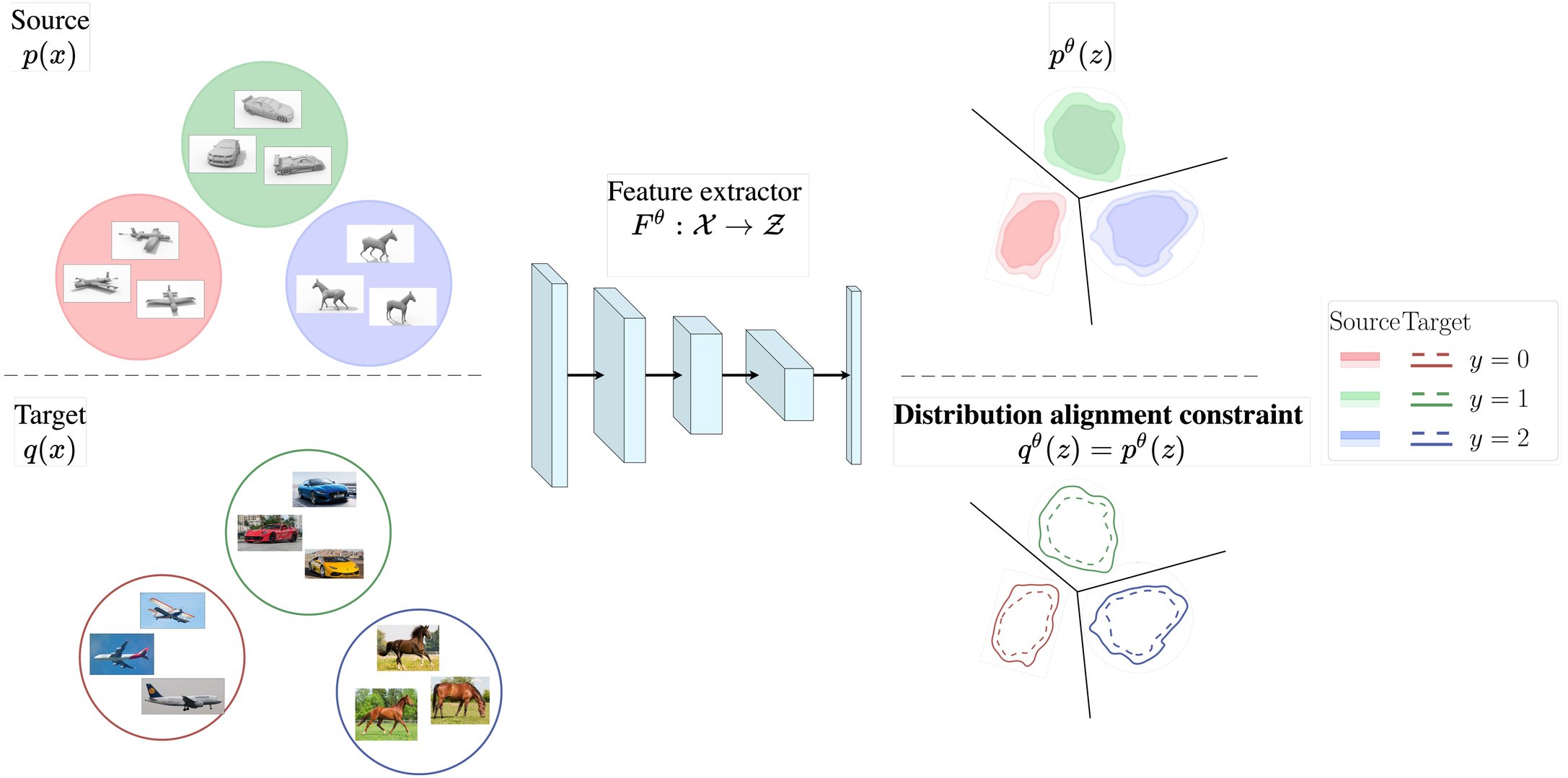


Background: distribution alignment in domain adaptation



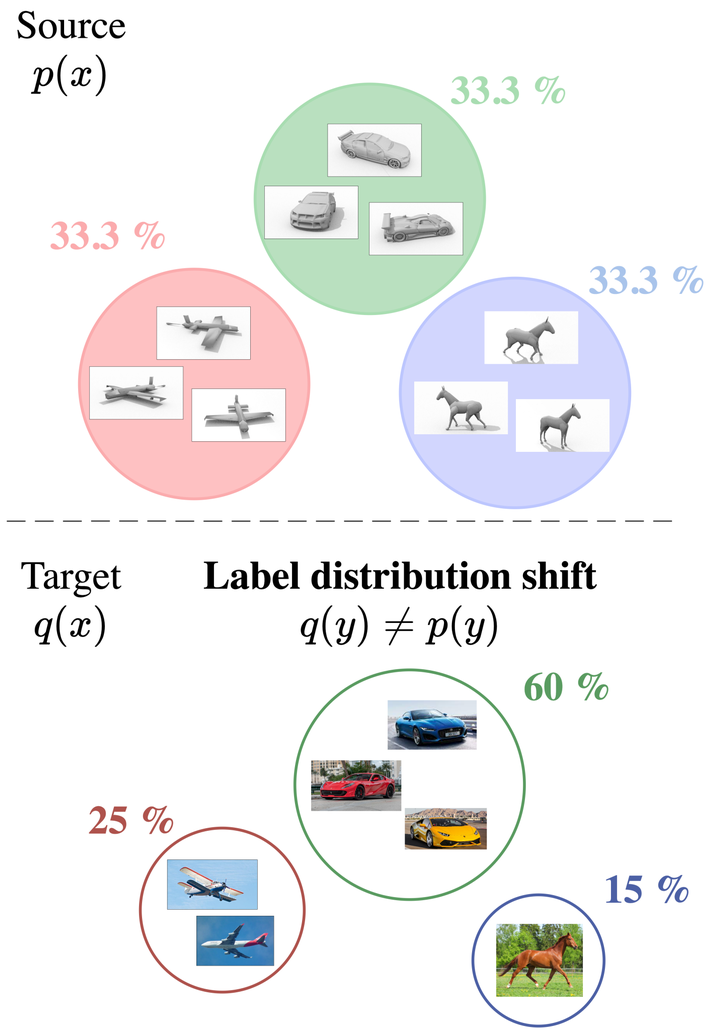
DANN: domain adversarial neural networks [Ganin et al., 2016]

Background: distribution alignment in domain adaptation



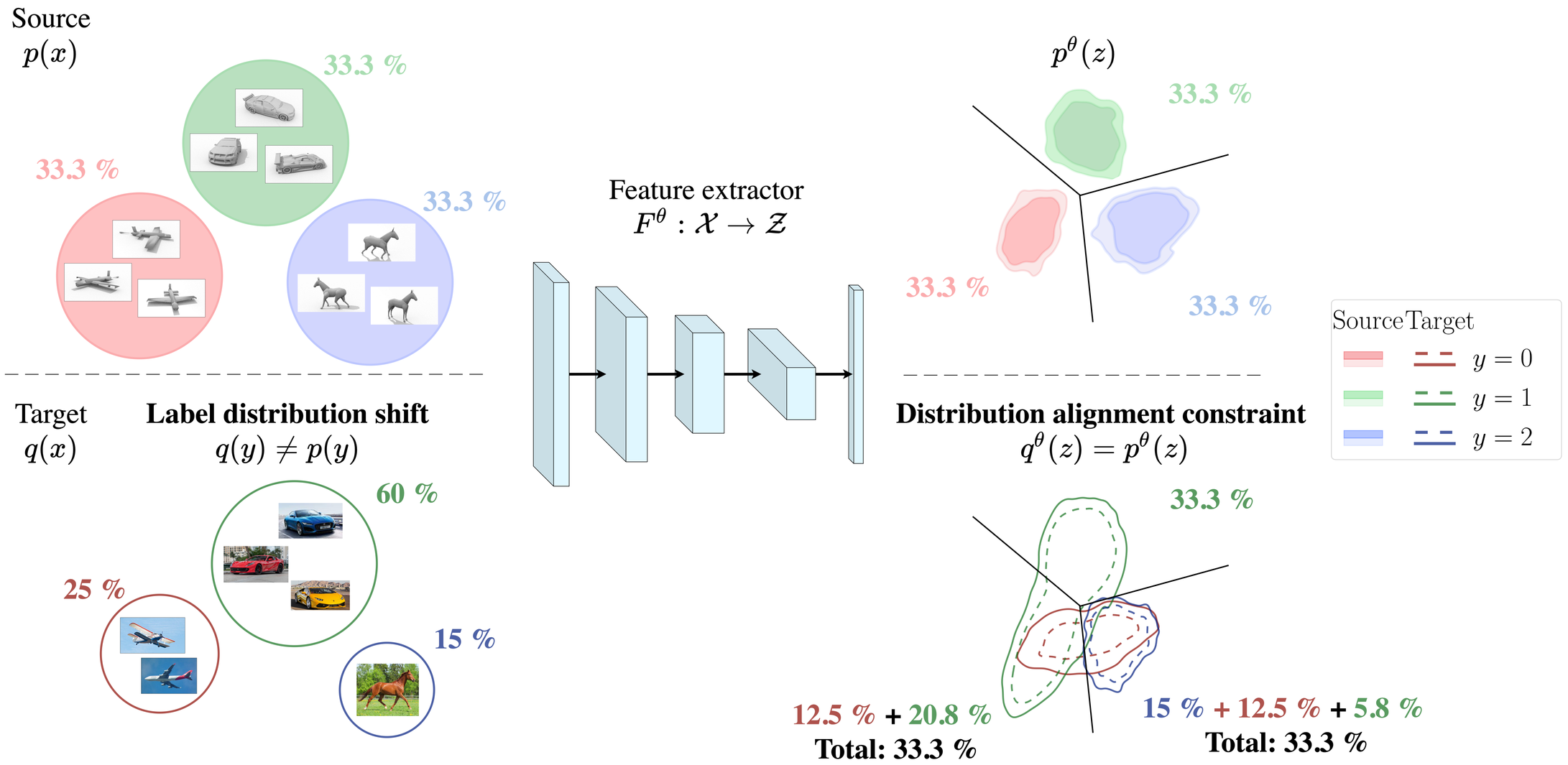
DANN: domain adversarial neural networks [Ganin et al., 2016]

Problem: label distribution shift



[Zhao et al., 2019; Wu et al., 2019; Tachet des Combes et al, 2020, ...]

Problem: label distribution shift

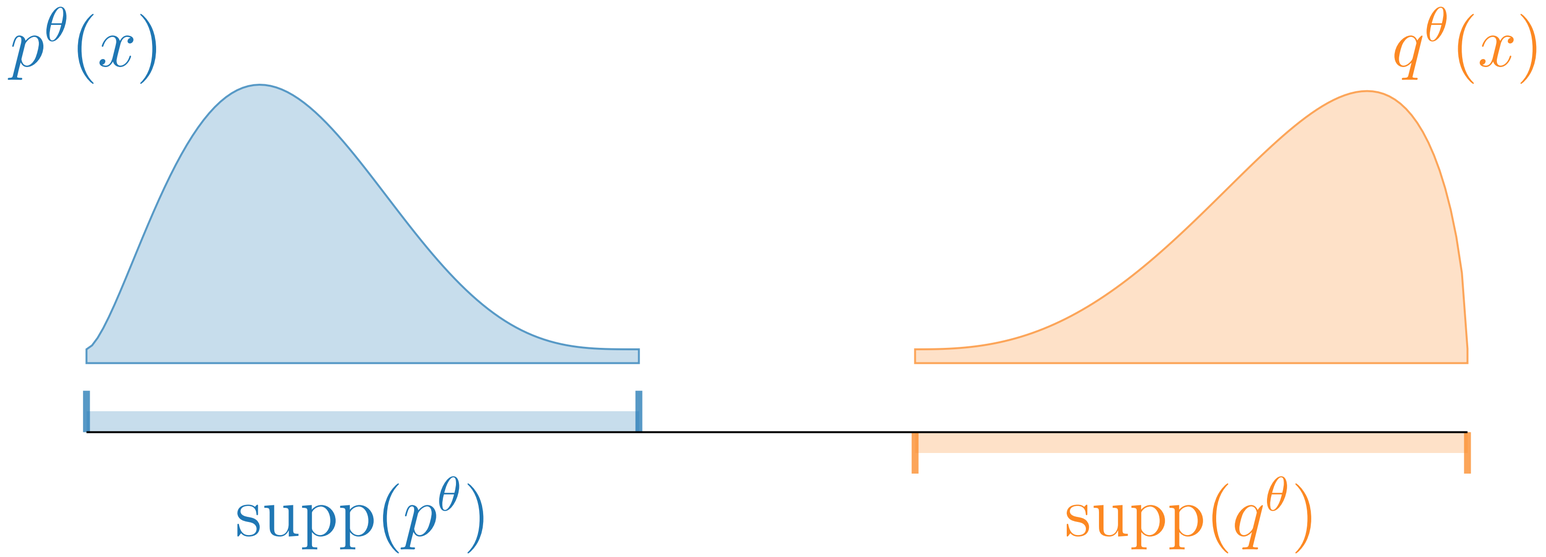


[Zhao et al., 2019; Wu et al., 2019; Tachet des Combes et al, 2020, ...]

Support alignment

We study the problem of aligning the supports of distributions

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

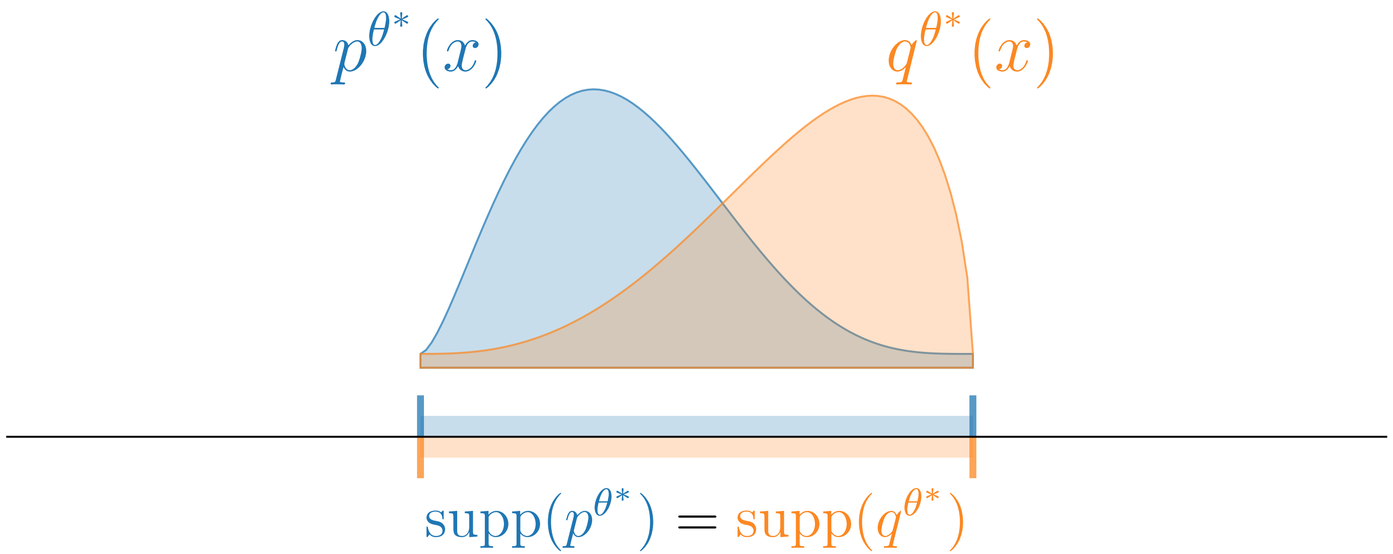


Support alignment

We study the problem of aligning the supports of distributions

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

Find $\theta^* : \ \mathrm{supp}(p^{\theta^*}) = \mathrm{supp}(q^{\theta^*})$



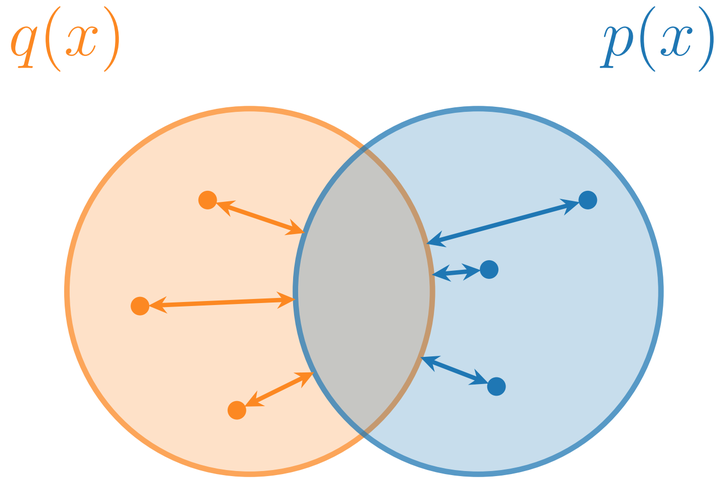
Support divergence

Symmetric Support Difference (SSD) divergence

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

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

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



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

Support alignment via log-loss discriminator

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

$$f^*(x) = \frac{p(x)}{p(x)+q(x)}$$

Support alignment via log-loss discriminator

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

$$f^*(x) = \frac{p(x)}{p(x)+q(x)}$$

Theorem

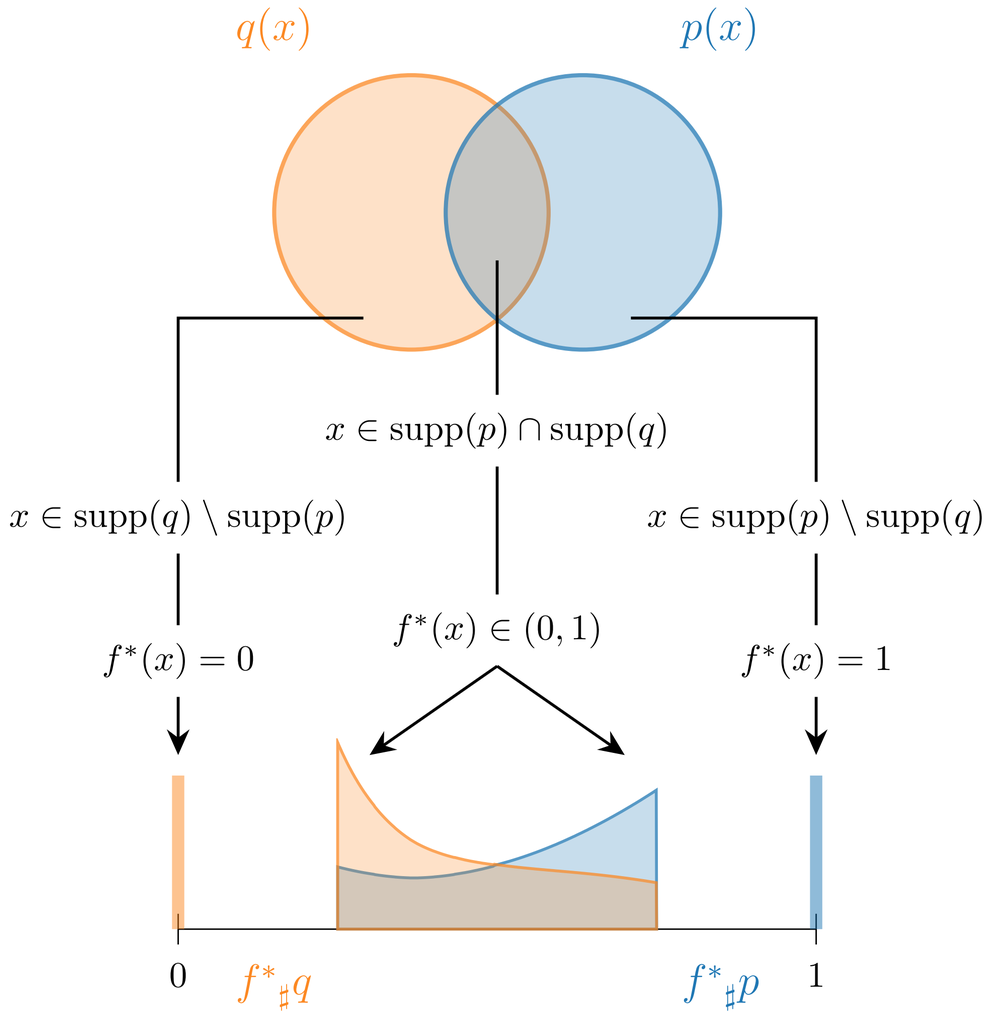
The mapping $f^* : \mathcal{X} \to [0,1]$

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



Support alignment via log-loss discriminator

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

$$f^*(x) = \frac{p(x)}{p(x)+q(x)}$$

Theorem

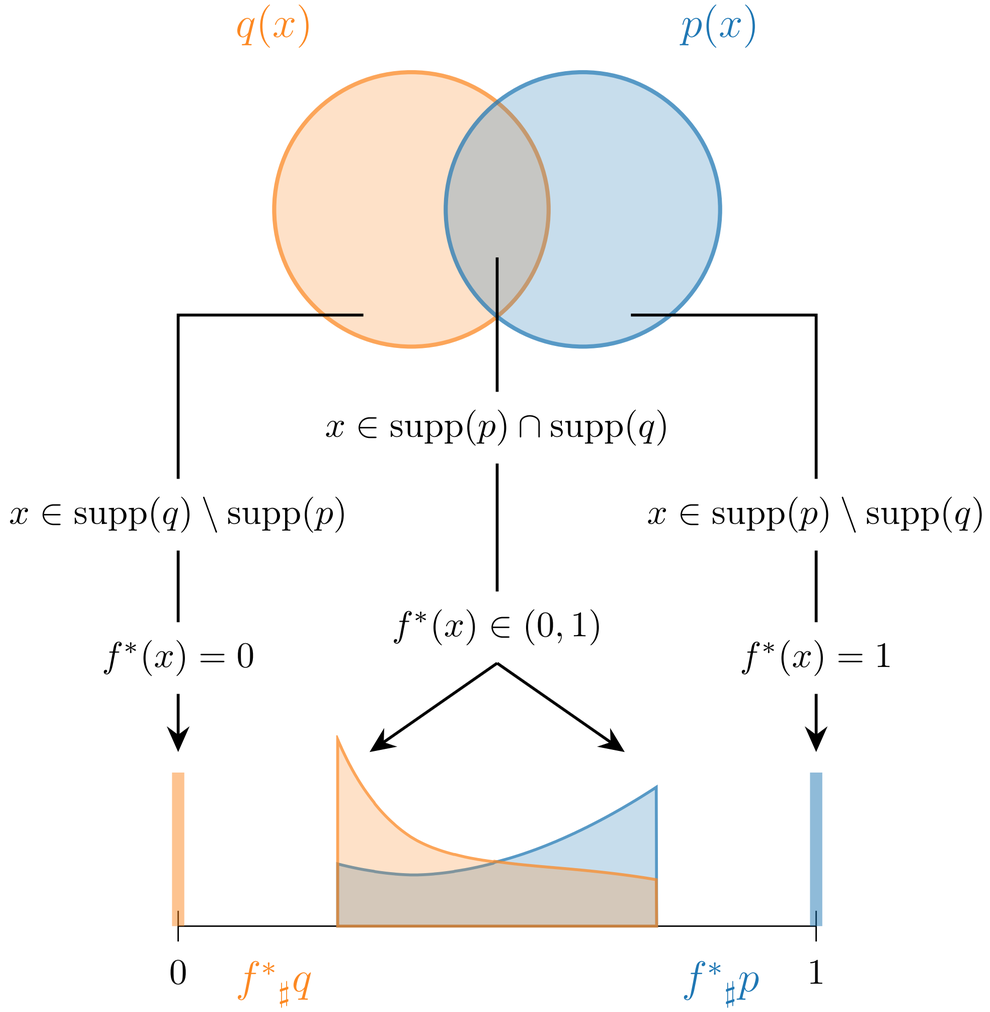
The mapping $f^* : \mathcal{X} \to [0,1]$

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 $g : \mathcal{X} \to \mathbb{R}$

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

Adversarial Distribution Alignment (GAN/DANN)	**Adversarial Support Alignment (ASA)**
Goal: find $\theta : p^\theta = q^\theta$	**Goal**: find $\theta : \mathrm{supp}(p^\theta) = \mathrm{supp}(q^\theta)$

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

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

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

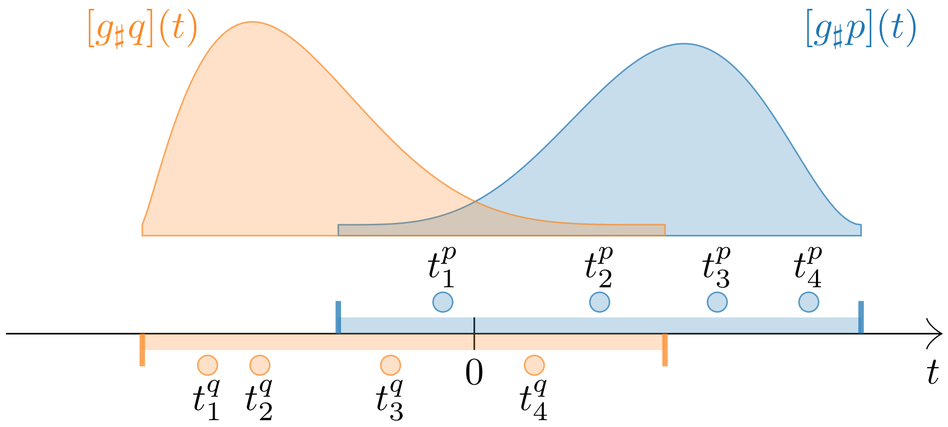
Adversarial Distribution Alignment (GAN/DANN)	**Adversarial Support Alignment (ASA)**
Goal: find $\theta : p^\theta = q^\theta$	**Goal**: find $\theta : \mathrm{supp}(p^\theta) = \mathrm{supp}(q^\theta)$

1. Sample examples

$$x_1^p, x_2^p, \ldots, x_N^p \sim p^\theta(x)$$
$$x_1^q, x_2^q, \ldots, x_N^q \sim q^\theta(x)$$

2. Apply discriminator mapping

$$p^\theta(x) \xrightarrow{g} [g_\sharp p^\theta](t) : \quad t_i^p = g(x_i^p)$$
$$q^\theta(x) \xrightarrow{g} [g_\sharp q^\theta](t) : \quad t_i^q = g(x_i^q)$$



Adversarial Distribution Alignment (GAN/DANN)

Goal: find $\theta : p^\theta = q^\theta$

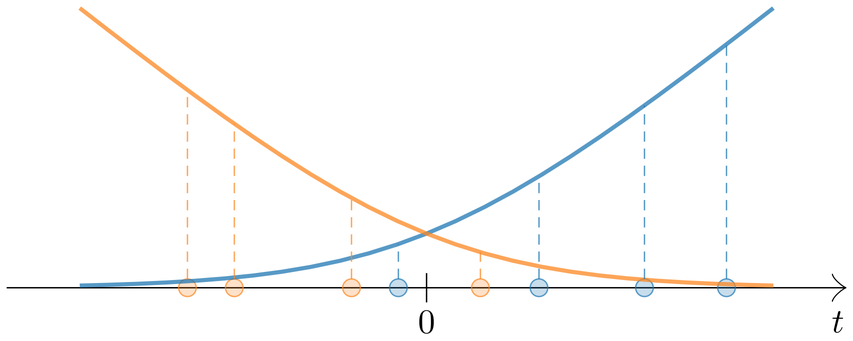
Adversarial Support Alignment (ASA)

Goal: find $\theta : \text{supp}(p^\theta) = \text{supp}(q^\theta)$

3. Compute loss function

$$\mathcal{L}_A = \frac{1}{N}\sum_i \underbrace{\log\left(1+e^{-t_i^q}\right)}_{L_q(t_i^q)} + \frac{1}{N}\sum_i \underbrace{\log\left(1+e^{t_i^p}\right)}_{L_p(t_i^p)}$$

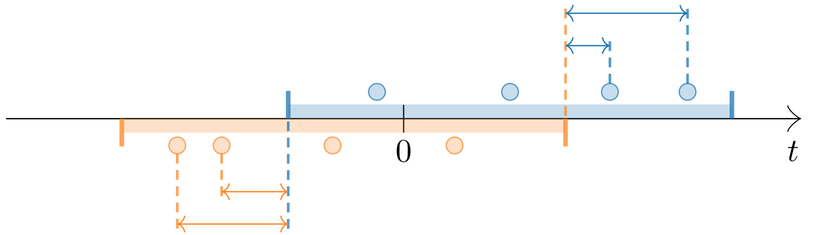
$L_q(t) = \log\left(1+e^{-t}\right)$ $\quad$ $L_p(t) = \log\left(1+e^{t}\right)$



$$\mathcal{L}_A = \frac{1}{N}\sum_i \underbrace{d(t_i^q, \text{supp}(g_\sharp p))}_{L_q(t_i^q)} + \frac{1}{N}\sum_i \underbrace{d(t_i^p, \text{supp}(g_\sharp q))}_{L_p(t_i^p)}$$

$$L_q(t) = d(t, \text{supp}(g_\sharp p)) = |t - \pi_p^*(t)|$$
$$\pi_p^*(t) = \underset{t^p \in \text{supp}(g_\sharp p)}{\text{argmin}} |t - t^p|$$

$$L_p(t) = d(t, \text{supp}(g_\sharp q)) = |t - \pi_q^*(t)|$$
$$\pi_q^*(t) = \underset{t^q \in \text{supp}(g_\sharp q)}{\text{argmin}} |t - t^q|$$



Adversarial Distribution Alignment (GAN/DANN)

Goal: find $\theta : p^\theta = q^\theta$

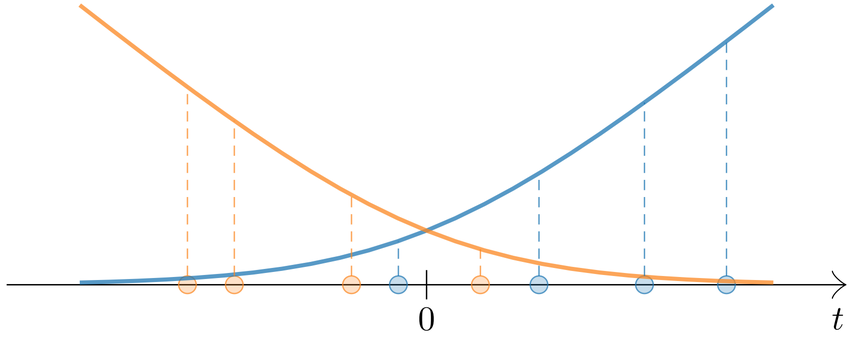
Adversarial Support Alignment (ASA)

Goal: find $\theta : \text{supp}(p^\theta) = \text{supp}(q^\theta)$

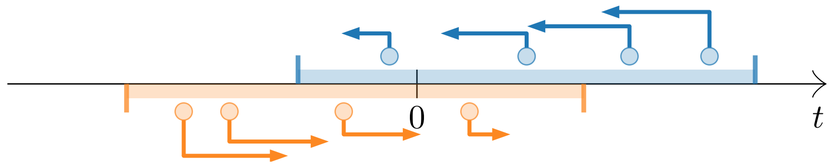
4. Compute and propagate gradients

$$\mathcal{L}_A = \frac{1}{N}\sum_i \underbrace{\log\left(1+e^{-t_i^q}\right)}_{L_q(t_i^q)} + \frac{1}{N}\sum_i \underbrace{\log\left(1+e^{t_i^p}\right)}_{L_p(t_i^p)}$$

$L_q(t) = \log\left(1+e^{-t}\right)$ $\qquad$ $L_p(t) = \log\left(1+e^{t}\right)$



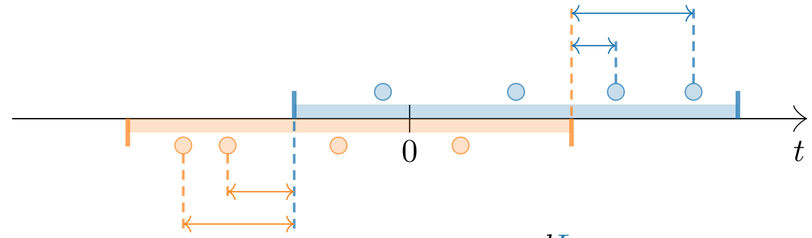
$-\frac{dL_q}{dt}(t) = -(1+e^t)^{-1}$ $\qquad$ $-\frac{dL_p}{dt}(t) = -(1+e^{-t})^{-1}$



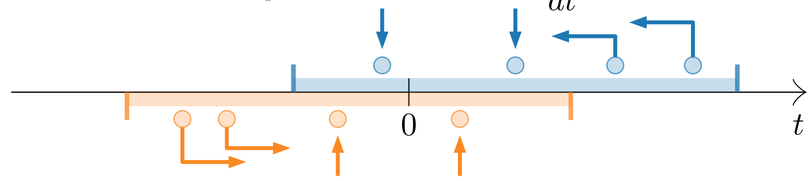
$$\mathcal{L}_A = \frac{1}{N}\sum_i \underbrace{d(t_i^q, \text{supp}(g_\sharp p))}_{L_q(t_i^q)} + \frac{1}{N}\sum_i \underbrace{d(t_i^p, \text{supp}(g_\sharp q))}_{L_p(t_i^p)}$$

$L_q(t) = d(t, \text{supp}(g_\sharp p))$
$= |t - \pi_p^*(t)|$
$\pi_p^*(t) = \underset{t^p \in \text{supp}(g_\sharp p)}{\text{argmin}} |t - t^p|$

$L_p(t) = d(t, \text{supp}(g_\sharp q))$
$= |t - \pi_q^*(t)|$
$\pi_q^*(t) = \underset{t^q \in \text{supp}(g_\sharp q)}{\text{argmin}} |t - t^q|$



$-\frac{dL_q}{dt}(t) = -\,\text{sign}\left(t - \pi_p^*(t)\right)$ $\qquad$ $-\frac{dL_p}{dt}(t) = -\,\text{sign}\left(t - \pi_q^*(t)\right)$

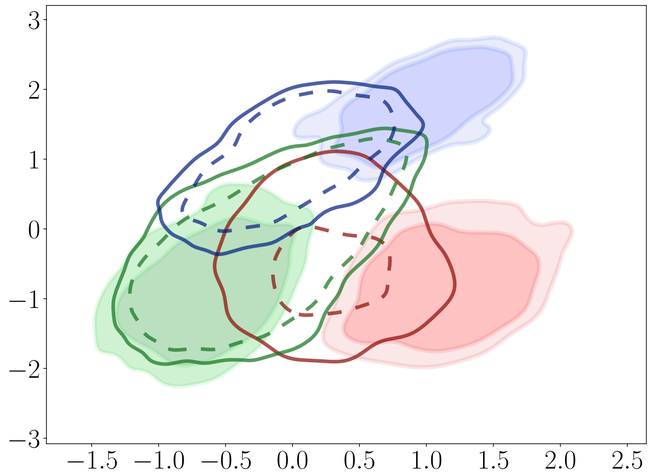


Domain adaptation results: 2D embeddings visualization

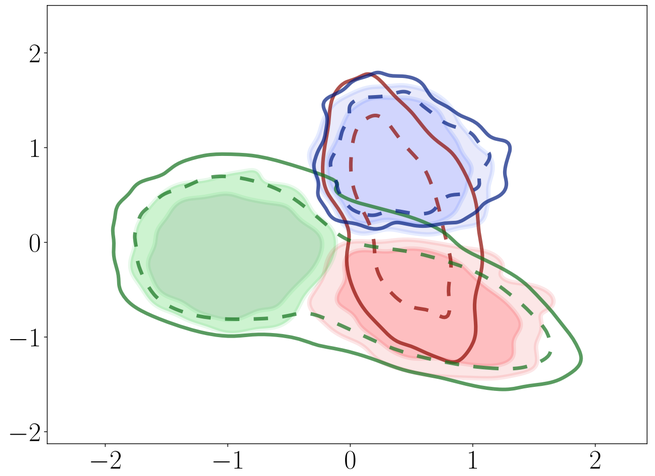
USPS➜MNIST, 3 classes

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

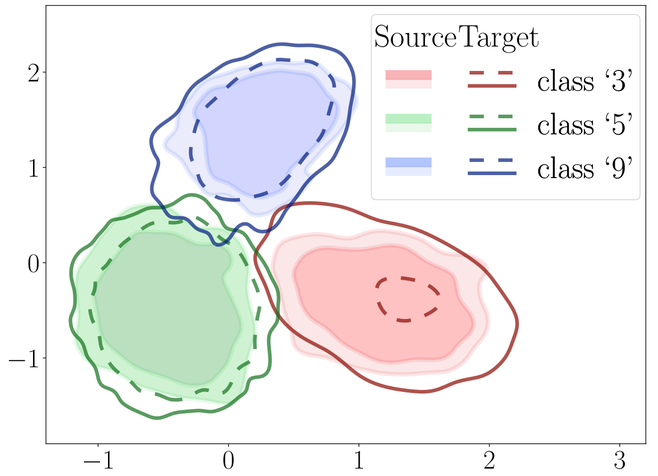
Target class distribution
[**23%**, **65%**, **12%**]



(a) No DA (avg acc: 63%)
$\mathcal{D}_W(p_Z^\theta, q_Z^\theta) = 0.78$
$\mathcal{D}_\triangle(p_Z^\theta, q_Z^\theta) = 0.10$



(b) DANN (avg acc: 75%)
$\mathcal{D}_W(p_Z^\theta, q_Z^\theta) = 0.07$
$\mathcal{D}_\triangle(p_Z^\theta, q_Z^\theta) = 0.02$



(c) ASA-abs (avg acc: 94%)
$\mathcal{D}_W(p_Z^\theta, q_Z^\theta) = 0.59$
$\mathcal{D}_\triangle(p_Z^\theta, q_Z^\theta) = 0.03$

Domain adaptation results: USPS➜MNIST, LeNet

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

	$\alpha = 0.0$ no shift		$\alpha = 1.0$		$\alpha = 1.5$		$\alpha = 2.0$ severe shift	
Algorithm	average	min	average	min	average	min	average	min
No DA	71.9	20.3	72.9	25.8	71.3	27.5	71.3	16.6
Distribution Alignment: DANN	97.8	96.0	83.5	25.1	70.0	01.1	57.8	00.9
Distribution Alignment: VADA	**98.0**	96.2	88.2	48.9	78.2	06.6	61.9	01.4

Domain adaptation results: USPS➜MNIST, LeNet

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

		$\alpha = 0.0$ no shift		$\alpha = 1.0$		$\alpha = 1.5$		$\alpha = 2.0$ severe shift	
	Algorithm	average	min	average	min	average	min	average	min
	No DA	71.9	20.3	72.9	25.8	71.3	27.5	71.3	16.6
Distribution Alignment	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
Relaxed Distribution Alignment	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

Domain adaptation results: USPS→MNIST, LeNet

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

		$\alpha = 0.0$ no shift		$\alpha = 1.0$		$\alpha = 1.5$		$\alpha = 2.0$ severe shift	
	Algorithm	average	min	average	min	average	min	average	min
	No DA	71.9	20.3	72.9	25.8	71.3	27.5	71.3	16.6
Distribution Alignment	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
Relaxed Distribution Alignment	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
Support Alignment (ours)	ASA-sq	93.7	89.2	92.3	83.5	90.9	69.9	87.2	62.5
	ASA-abs	94.1	88.9	92.8	78.9	**92.5**	**82.4**	**90.4**	**68.4**

Domain adaptation results

STL➜CIFAR10

Deep CNN

	$\alpha = 0.0$ no shift		$\alpha = 2.0$ severe shift	
Algorithm	average	min	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	71.7	52.9	**68.1**	**44.7**
ASA-abs	71.6	49.0	67.8	40.9

VisDA-17

ResNet50

	$\alpha = 0.0$ no shift		$\alpha = 2.0$ severe shift	
Algorithm	average	min	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	64.9	35.7	51.9	18.3
ASA-abs	64.8	**40.6**	**52.5**	**19.7**

Summary

- **Support alignment** as extreme relaxation of **distribution alignment**
- **Adversarial Support Alignment (ASA):** method to align distribution supports
- **Optimal Transport perspective:** spectrum of **relaxed alignment** approaches
- Evaluation on domain adaptation datasets under **label distribution shift**