Lecture 13 Distribution Shift

IEMS 402 Statistical Learning

Northwestern

References

https://hsnamkoong.github.io/assets/html/b9145/index.html

Distribution Shift

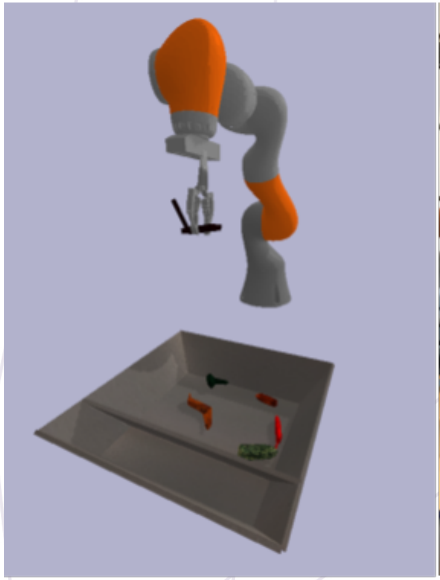
Reconsider the ML Theory…

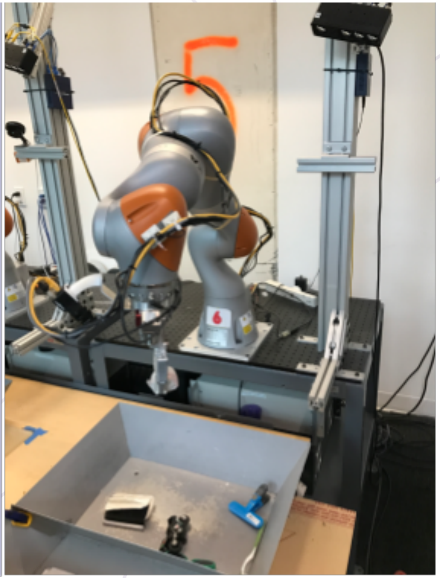
$$\left| \mathbb{E}_{\hat{\mathbb{P}}} f - \mathbb{E}_{\mathbb{P}} f \right| \leq \cdots$$

$\hat{\mathbb{P}}$ is i.i.d. from $\mathbb{P}$.

What if the test distribution $\mathbb{P}_{test}$ is different??

However…



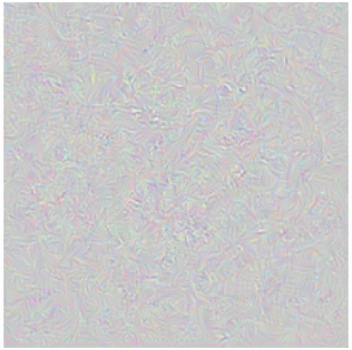


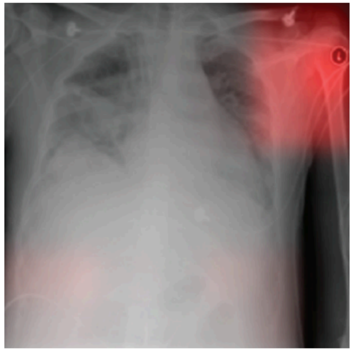
抖音

TikTok

Elephant or Cat

Shortcut learning





Article: Super Bowl 50

Paragraph: "*Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.*"

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"

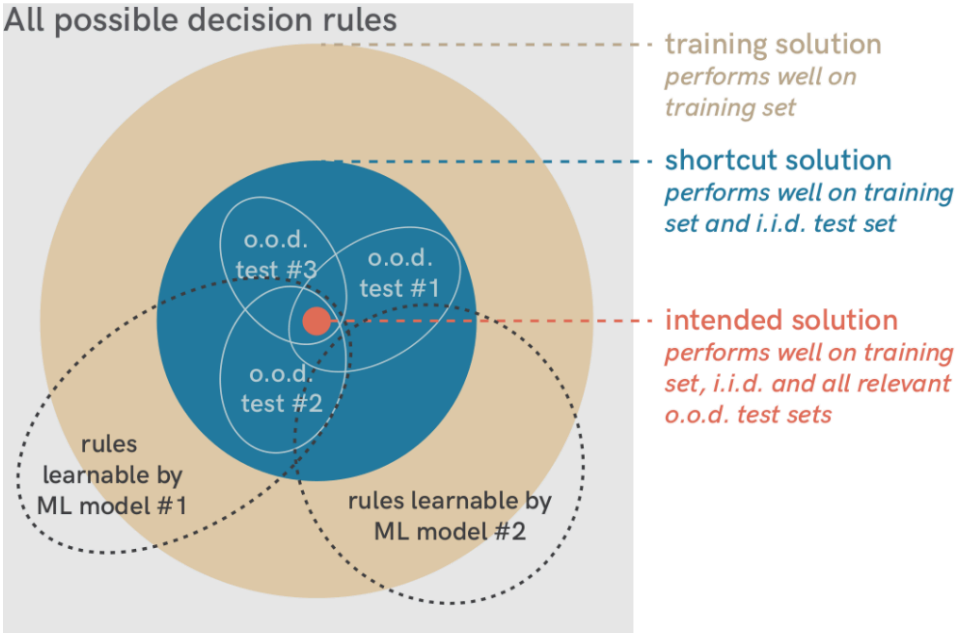
Original Prediction: John Elway

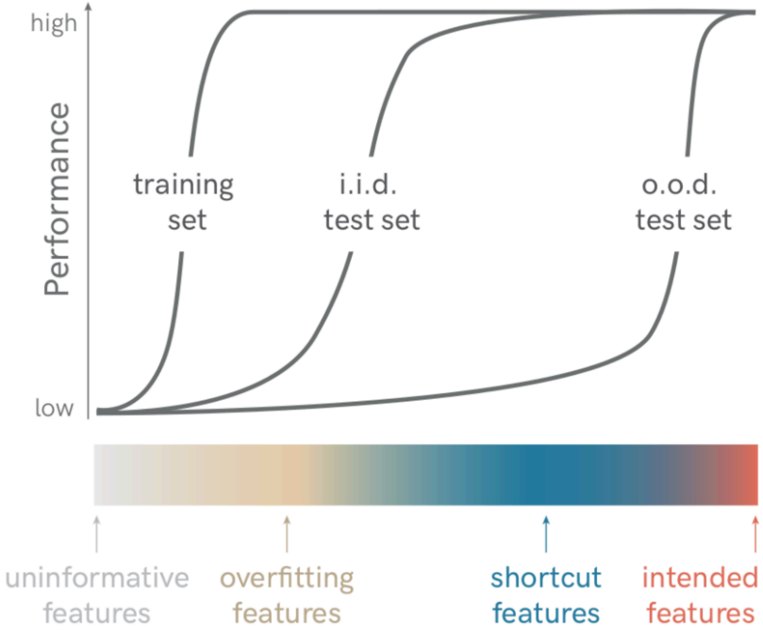
Prediction under adversary: Jeff Dean

Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

spurious correlation

From i.i.d to o.o.d





Importance Weighting

Importance Weighting

How do we deal with covariate / label shifts?

reweighting data1 data2

① 5 loss(data1) + loss(data1)

What we have

$E_{p_{train}}[\ell(z;\theta)]$

What we want

$E_{p_{test}}[\ell(z;\theta)]$ resample

weighted loss.

② data1 data1 ... data1 data2

Most basic approach: reweight the loss

$$E_{p_{train}}\left[\frac{p_{test}(z)}{p_{train}(z)}\ell(z;\theta)\right] = E_{p_{test}}[\ell(z;\theta)]$$

Weighted loss over the training distribution

in expectation, they are same

but these variance are different

(also possible: resample the dataset)

Importance weighting

I don't know $\frac{p_{test}}{p_{train}}$

An alternative algorithm: use a classifier that separates p_{train} and p_{test}

1. Estimate a classifier $f(x) \approx \frac{p_{train}(x)}{p_{test}(x)+p_{train}(x)}$

   (collect another dataset. (X_{train}, 1) (X_{test}, 0))

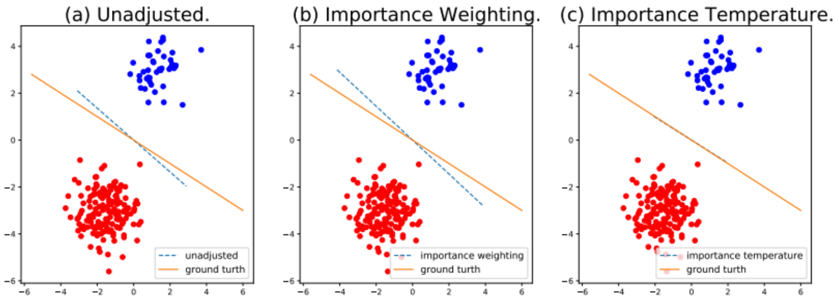
2. Reweight by $h(x) = \frac{1}{f(x)} - 1$

3. Fit a model by minimizing the loss $h(x)\ell(x, y; \theta)$

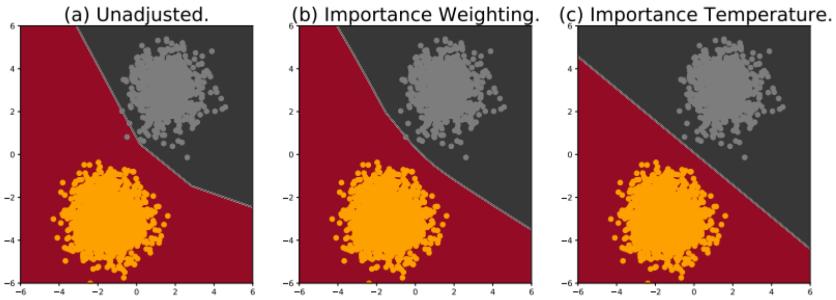
Discriminative Learning for Differing Training and Test Distributions

Steffen Bickel BICKEL@MPI-INF.MPG.DE
Michael Brückner BRUM@MPI-INF.MPG.DE
Tobias Scheffer SCHEFFER@MPI-INF.MPG.DE
Max Planck Institute for Computer Science, Saarbrücken, Germany

Not Working for Over-parameterized Model



(a) Linear Model for Separable Data



(b) Multilayer Perceptron with two hidden layers of size 200

Byrd J, Lipton Z. What is the effect of importance weighting in deep learning? International conference on machine learning. PMLR, 2019: 872-881.

IPM

Background material: integral probability measures

$\mathbb{E}_{P_{test}} \ell - \mathbb{E}_{P_{train}} \ell$

To state this clearly, we need to first go into some background.

Definition (IPM):

For two probability distributions p and q, the integral probability metric (IPM) for a family of functions $\mathcal{F}$ is defined as

uniform

$$d_{\mathcal{F}}(p,q) = \sup_{f \in \mathcal{F}} |E_p[f(x)] - E_q[f(x)]|$$

distance between distribution P_{train} / P_{test}

Intuition: $\mathcal{F}$ are 'test functions' that can distinguish p and q

If two have the same function value for all $\mathcal{F}$, then they are similar

IPM and distribution shift

What we want — What we have — Domain distance

$$E_{p_{test}}[\ell(x,y,\theta)] = E_{p_{train}}[\ell(x,y,\theta)] + \Delta$$

$E_{P_{test}}\, \ell \leq E_{P_{train}}\, \ell + d_{\mathcal{F}}(P_{train}, P_{test})$

From the trivial restatement

$$\Delta = E_{p_{test}}[\ell(x,y,\theta)] - E_{p_{train}}[\ell(x,y,\theta)]$$

This looks like an IPM! (if $\ell(x,y,\theta) \in \mathcal{F}$ for all θ)

$$\Delta \leq \sup_{f \in \mathcal{F}} E_{p_{test}}[f(x,y)] - E_{p_{train}}[f(x,y)] = d_{\mathcal{F}}(p_{train}, p_{test})$$

Takeaway: IPMs bound excess error under transfer

Example: L1 distance

$d_{\mathcal{F}}(p, q) = \sup_{\{f \mid -1 \le f \le 1\}} \left| \sum f(x)\,(p(x) - q(x)) \right| = \sum_x |p(x) - q(x)|$

We can now bound test performance in terms of IPMs

$\mathcal{F} := \{f \mid -1 \le f(x) \le 1,\ \forall x\}$, $d_{\mathcal{F}}(p, q) = \sum_x |p(x) - q(x)|$

For $0 \le \ell(x, y, \theta) \le 1$ and under covariate shift,

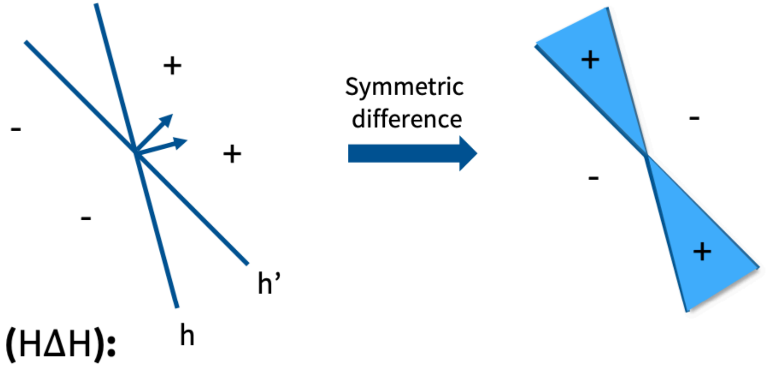
$$E_{p_{test}}[\ell(x, y, \theta)] \le E_{p_{train}}[\ell(x, y, \theta)] + ||p_{train}(x) - p_{test}(x)||_1$$

$||p_{train}(x) - p_{test}(x)||$

	Reweighting	IPM
Goals	Correct train-test mismatch	Estimate train-test mismatch (can also correct)
Assumptions	Overlap $\frac{P_{test}}{P_{train}}$ ⇒ if $P_{train}(x)=0$ then $P_{test}(x)=0$	Boundedness $d_{\mathcal{F}}(P, \hat{P})$
Training	Weighted/modified loss	No change (you may use $E_{P_{test}} \ell + d_{\mathcal{F}}(P, \hat{P})$ as loss function)
Costs	More samples (variance)	Inaccurate models (bias)
		Curse of dimensionality (next lecture)

Defining HΔH (disagreement)

For a hypothesis class $\mathcal{H}$, the HΔH set is defined as the symmetric difference



Definition (HΔH):

For a hypothesis class $\mathcal{H}$, the symmetric difference set $H\Delta H$ is defined as

$$\text{H}\Delta\text{H} \coloneqq \{g\colon g(x) = \text{XOR}\big(h(x), h'(x)\big) \text{ and } h, h' \in \mathcal{H}\}$$

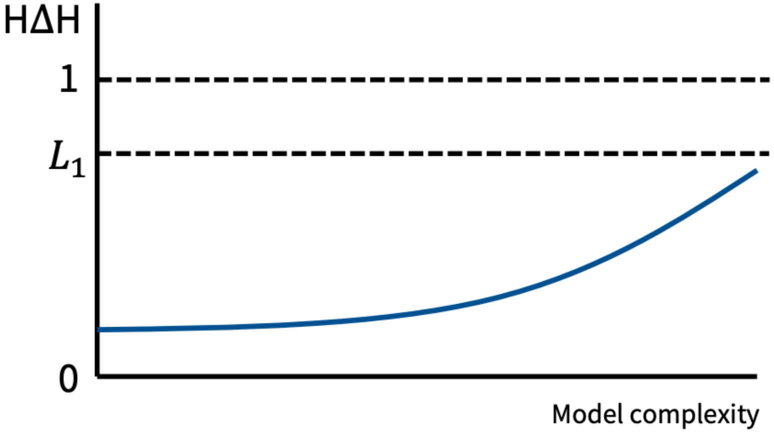
Dependency on Hypothesis Space

For a hypothesis class $\mathcal{H}$, the HΔH-divergence is

$$d_{H\Delta H}(p_{train}, p_{test}) = 2 \sup_{g \in H\Delta H} \left| E_{p_{train}}[g(x)] - E_{p_{test}}[g(x)] \right|$$

F

HΔH: $\frac{1}{2} d_{H\Delta H}(p_{train}, p_{test})$



$d_{H\Delta H}$ is upper bounded by the L_1 distance

$d_{H\Delta H}$ increases monotonically with model complexity. If $H \subset H'$,

$$d_{H\Delta H} \leq d_{H'\Delta H'}$$

Another trade-off

Let's walk through the main bound.

$$E_{p_{test}}[\ell(x,y,h)]$$
$$\leq E_{p_{train}}[\ell(x,y,h)] + \frac{1}{2} d_{H\Delta H}(p_{train}, p_{test}) + \lambda$$

Different answer on two domains

same answer but
Both are wrong

Training domain error

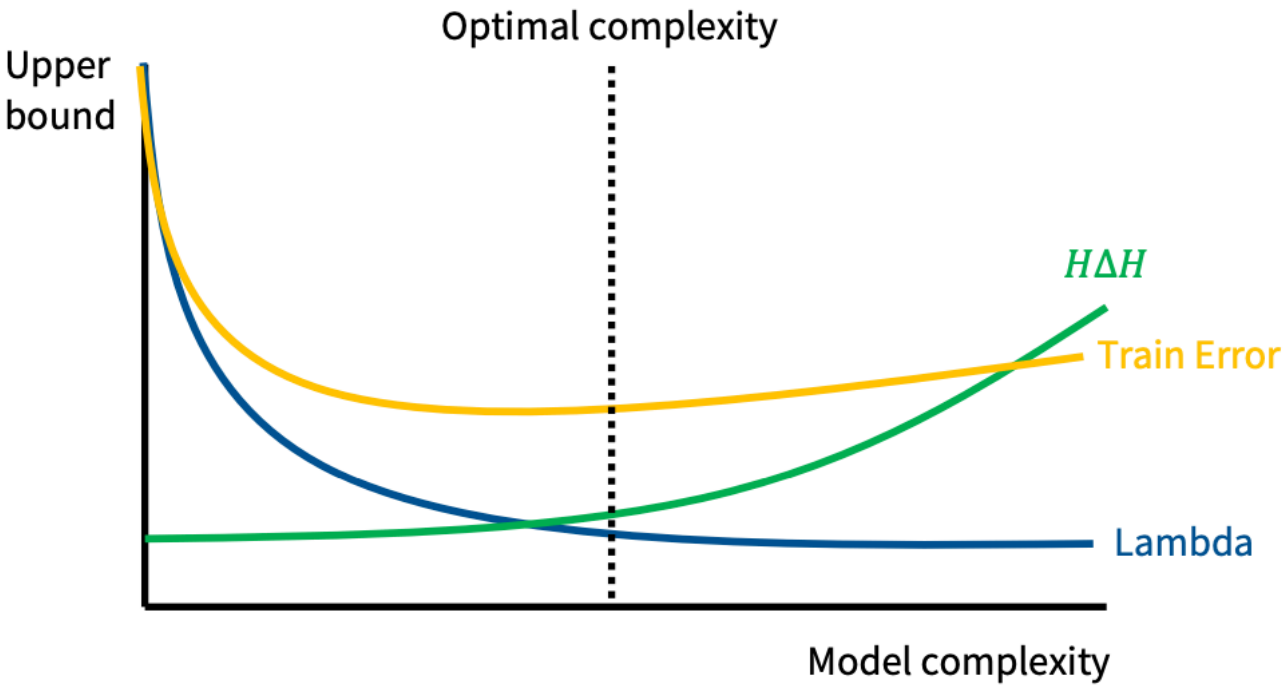
Domain distinguishability

Minimal error of a classifier on both domains

$$\lambda = \inf_{h \in \mathcal{H}} p_{train}\big(y \neq h(x)\big) + p_{test}(y \neq h(x))$$

HΔH claim: Low training domain error + low $H\Delta H$ divergence + rich $\mathcal{H}$
= good generalization to target domain

Another tradeoff



Distributionally Robust Optimization

F-divergence

f-divergence. $D_f(Q \| P) := \int f\left(\frac{dQ}{dP}\right) dP$.

$f(1) = 0$

f is a convex function

$f(t) = t \log t$, then KL-divergence

$f(t) = |t-1|$ → TV distance (Total Variation)

$f(t) = (t-1)^2$ χ^2 divergence

Distributionally Robust Optimization

$dist(P_{train}, P_{test}) \le \rho$

Empirical Risk Minimization

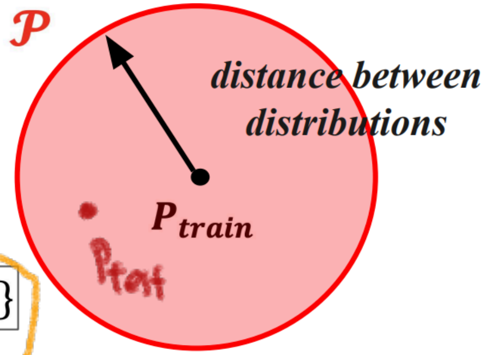
$$\min_{\theta \in \Theta} \mathbb{E}_{Z \sim P_{train}}[\ell(\theta; Z)]$$

$\mathbb{E}_{Z \sim P_{test}}[\ell(\theta; Z)]$

DRO

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

$$\mathcal{P} = \{Q : Dist(Q, P_{train}) \le \rho\}$$



Instead of minimizing loss over training distribution, minimize loss over distributions ***near*** it

Generalization of DRO

automatically can be built.

$$\sup_{d(Q,P)\le\rho} \mathbb{E}_{z\sim Q}[\ell(\theta;z)] \geq \mathbb{E}_{z\sim P_{test}}[\ell(\theta;z)]$$

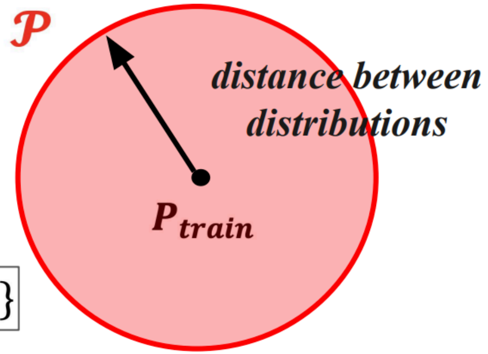
Empirical Risk Minimization

$$\min_{\theta\in\Theta} \mathbb{E}_{Z\sim P_{train}}[\ell(\theta;Z)]$$

DRO

$$\min_{\theta\in\Theta} \sup_{Q\in\mathcal{P}} \mathbb{E}_{Z\sim Q}[\ell(\theta;Z)]$$

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \le \rho\}$$



Instead of minimizing loss over training distribution, minimize loss over distributions ***near*** it

Duality of DRO

$R_f(\theta;P) := \sup_{D_f(Q\|P)\le\rho} \mathbb{E}_Q[\ell(\theta;Z)]$

$$R_f(\theta;P) = \inf_{\lambda\ge 0,\eta\in\mathbb{R}} \left\{\lambda \mathbb{E}_P\left[f^*\left(\frac{\ell(\theta;Z)-\eta}{\lambda}\right)\right] + \lambda\rho + \eta\right\}$$

$$f^*(s) := \sup_t\{st - f(t)\}.$$

$\rho - D_f(Q\|P) \ge 0$

Q is a distribution.

$$= \sup_{L\ge 0} \inf_{\lambda\ge 0,\eta\in\mathbb{R}} \{\mathbb{E}_P[L(Z)\ell(\theta;Z)] + \lambda(\rho - \mathbb{E}_P[f(L(Z))] - \eta(\mathbb{E}_P[L(Z)] - 1)\}$$

$L(Z) = \frac{dQ}{dP}$

$\mathbb{E}_P \frac{dQ}{dP} = \mathbb{E}_Q 1 = 1$

Next step: Solve the sup over L

Duality of DRO

$$R_f(\theta;P) = \inf_{\lambda\geq 0,\eta\in\mathbb{R}} \left\{ \lambda \mathbb{E}_P\left[f^*\left(\frac{\ell(\theta;Z)-\eta}{\lambda} \right) \right] + \lambda\rho + \eta \right\}$$

$$f^*(s) := \sup_t \{st - f(t)\}.$$

$$= \sup_{L\geq 0} \inf_{\lambda\geq 0,\eta\in\mathbb{R}} \{\mathbb{E}_P[L(Z)\ell(\theta;Z)] + \lambda(\rho - \mathbb{E}_P[f(L(Z))] - \eta(\mathbb{E}_P[L(Z)] - 1)\}$$

$$= \inf_{\lambda\geq 0,\eta\in\mathbb{R}} \sup_{L\geq 0} \left\{ \lambda \mathbb{E}_P\left[\frac{L(Z)(\ell(\theta;Z)-\eta)}{\lambda} - f(L(Z)) \right] \right\} + \lambda\rho + \eta.$$

$$= \mathbb{E}_P\left[f^*\left(\frac{\ell(\theta;Z)-\eta}{\lambda} \right) \right].$$

→ reference

↳ The loss function : f^*

f^*(relation)

Duality of f-divergence DRO, is changing loss function e.

Variance Regularization

χ^2 divergence. $f = (t-1)^2$

$$\inf_{\substack{\lambda \geq 0 \\ \eta \in \mathbb{R}}} \mathbb{E}_{\mathbb{P}} \left(\frac{\ell(\theta; z) - \eta}{\lambda} \right)^2 + \ell(\theta; z).$$

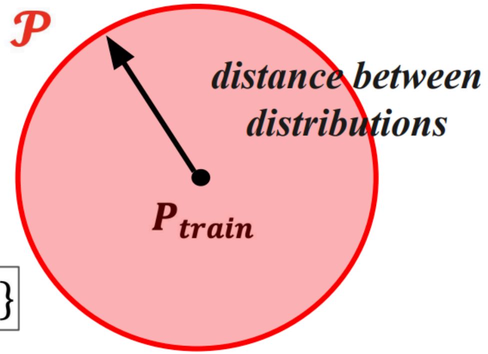
$\downarrow$ $\eta = \mathbb{E}_{\mathbb{P}} \ell$

$$\lambda \mathrm{Var}(\ell) + \mathbb{E}_{\mathbb{P}} \ell$$

Generalization of DRO

Empirical Risk Minimization	$\min_{\theta \in \Theta} \mathbb{E}_{Z \sim P_{train}}[\ell(\theta; Z)]$
DRO	$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$

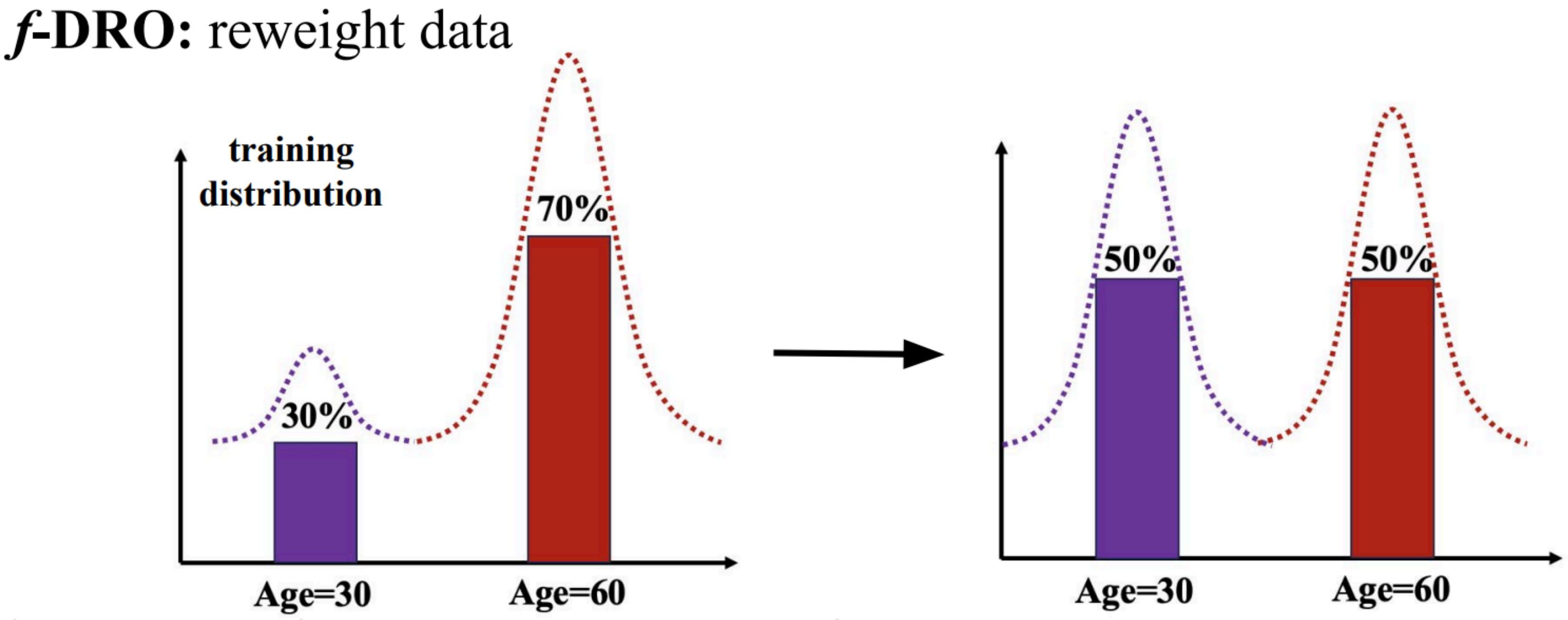
$$\mathcal{P} = \{Q : Dist(Q, P_{train}) \leq \rho\}$$



Instead of minimizing loss over training distribution, minimize loss over distributions ***near*** it

Is DRO Working?

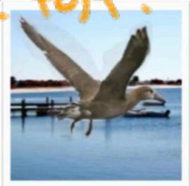
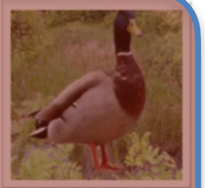
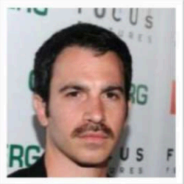
F-divergence DRO only reweighting



spurious correlation

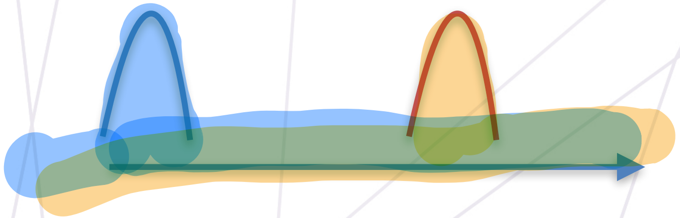
f-DRO	IW
weight more on the data have higher loss	weight more on the data appearing in the test.

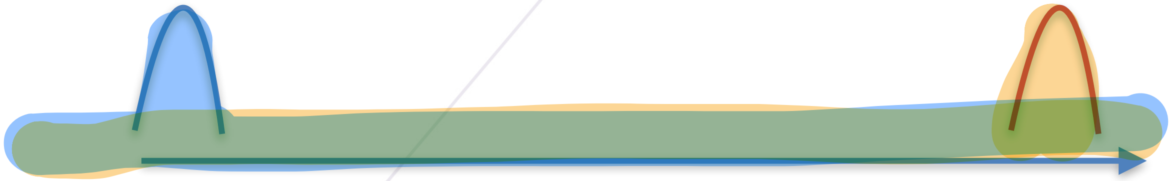
Weights more on rare data!

	Common training examples				Test examples	
Waterbirds	y: waterbird a: water background		y: landbird a: land background		y: waterbird a: land background	
CelebA	y: blond hair a: female		y: dark hair a: male		y: blond hair a: male	
MultiNLI	y: contradiction a: has negation (P) The economy could be still better. (H) The economy has never been better.		y: entailment a: no negation (P) Read for Slate's take on Jackson's findings. (H) Slate had an opinion on Jackson's findings.		y: entailment a: has negation (P) There was silence for a moment. (H) There was a short period of time where no one spoke.	

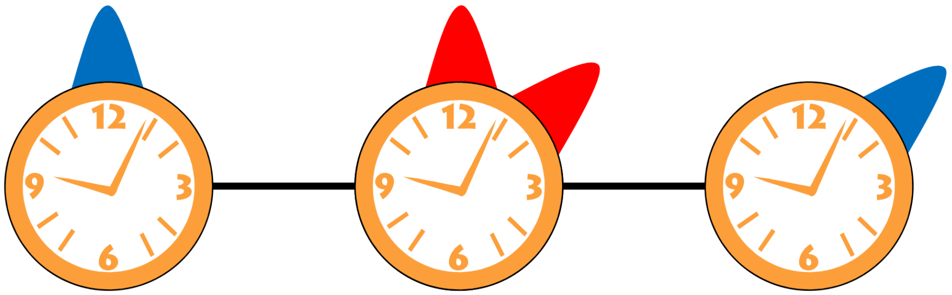
What's wrong about f-divergence

$$D_f(P \| Q) = \mathbb{E}_P f\left(\frac{dQ}{dP}\right)$$

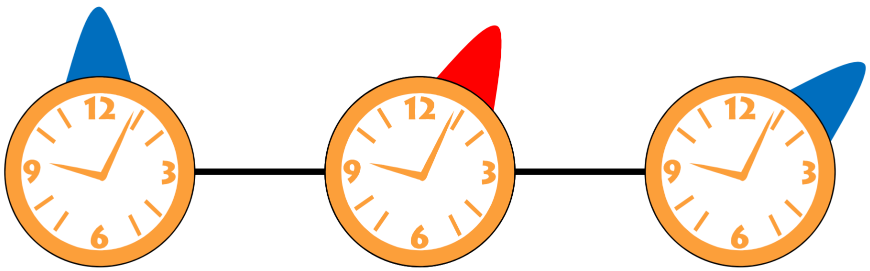




What’s wrong about f-divergence



Or



Over-parameterization?

			Average Accuracy		Worst-Group Accuracy	
			ERM	DRO	ERM	DRO
Standard Regularization	Waterbirds	Train	100.0	100.0	100.0	100.0
		Test	97.3	97.4	60.0	76.9
	CelebA	Train	100.0	100.0	99.9	100.0
		Test	94.8	94.7	41.1	41.1
	MultiNLI	Train	99.9	99.3	99.9	99.0
		Test	82.5	82.0	65.7	66.4
Strong ℓ_2 Penalty	Waterbirds	Train	97.6	99.1	35.7	97.5
		Test	95.7	96.6	21.3	84.6
	CelebA	Train	95.7	95.0	40.4	93.4
		Test	95.8	93.5	37.8	86.7

No improvement

model class is small.

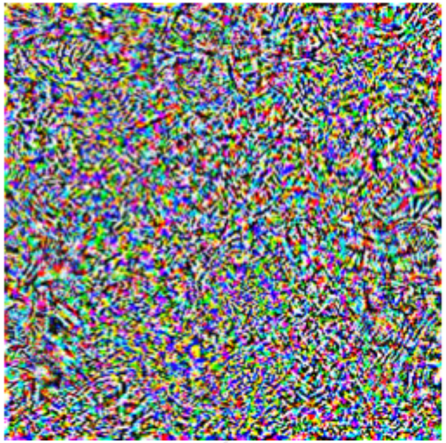
Northwestern

Adversarial Learning

adversarial training

“pig”

+ 0.005 x

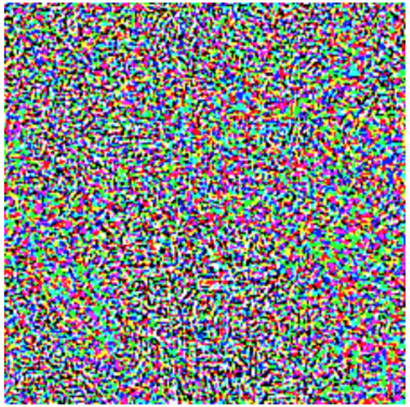


=

“airliner”

How to find Adversarial Examples?

$+ .007 \times$



$=$

$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Optimization that maximize the loss

Adversarial Training

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)\right].$$

https://arxiv.org/pdf/1706.06083

Adversarial Training Can Hurt Generalization

	Standard training	Adversarial training
Robust test	3.5%	45.8%
Robust train	-	100%
Standard test	95.2%	87.3%
Standard train	100%	100%

Real World?

Lots of progress on ImageNet over the past 10 years, but models are still not robust.

Evaluation: **new test sets**

ImageNetV2

[Recht, Roelofs, Schmidt, Shankar '19]

ObjectNet

[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz '19]

ImageNet-Sketch

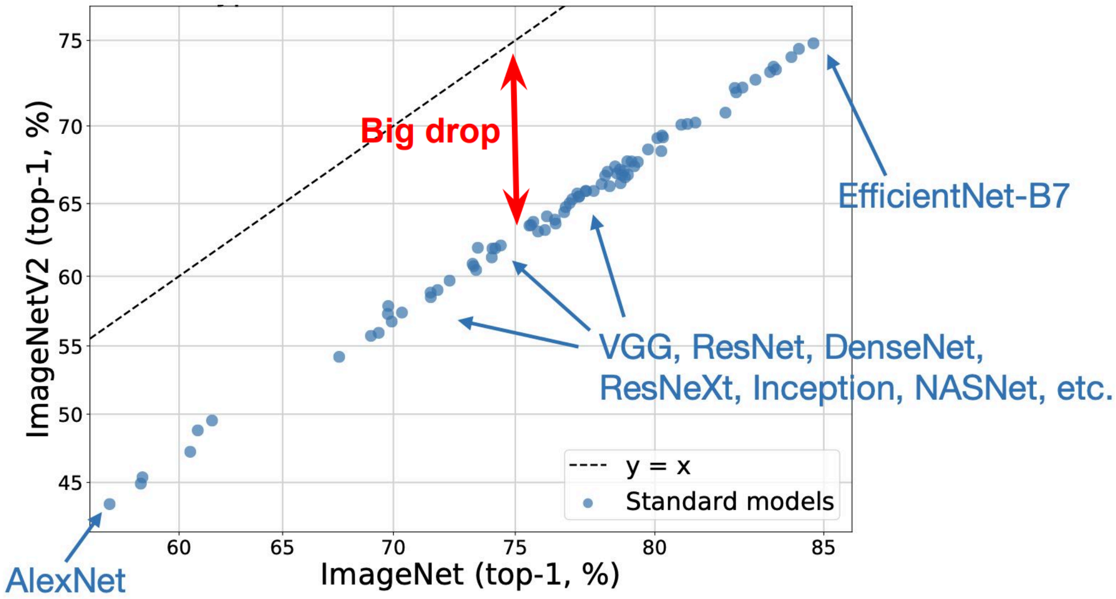
[Wang, Ge, Lipton, Xing '19]

ImageNet-R

[Hendrycks, Basart, Mu, Kadavath, Wang, Dorundo, Desai, Zhu, Parajuli, Guo, Song, Steinhardt, Gilmer '20]

Agree on the line!

Recht B, Roelofs R, Schmidt L, et al. Do imagenet classifiers generalize to imagenet?[C]// International conference on machine learning. PMLR, 2019: 5389-5400.



[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]

Why?