Learning Representations using Causal Invariance

LÉON BOTTOU

FACEBOOK AI RESEARCH & NEW YORK UNIVERSITY

Joint work with

Martin Arjovsky
New York University

Ishaan Gulrajani
Google

David Lopez-Paz
Facebook AI Research

Summary

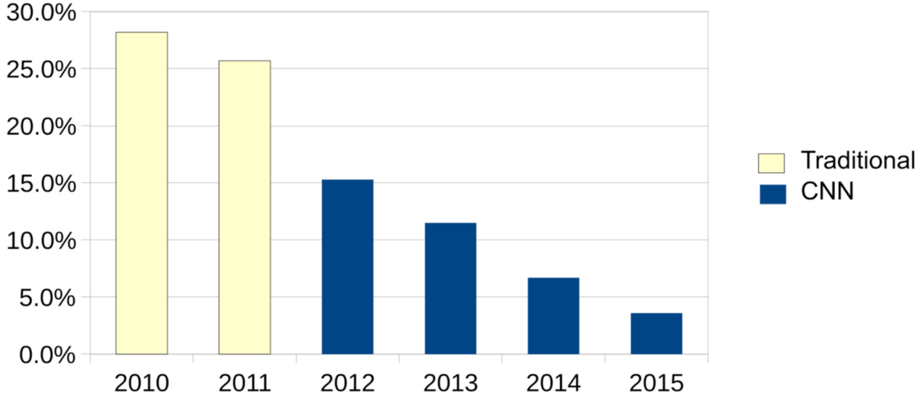
1. The statistical problem is only a proxy
2. Nature does not shuffle the examples. We do!
3. From interpolation to extrapolation
4. Related work
5. Linear invariant regression
6. Invariant regularization and nonlinear models
7. Aiming for zero training errors makes sense

1

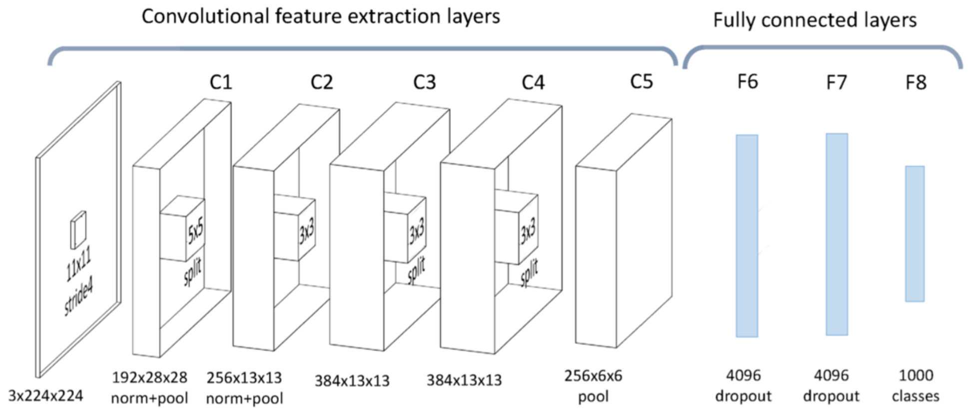
Why this work?

THE STATISTICAL PROBLEM IS ONLY A PROXY FOR THE REAL TASK

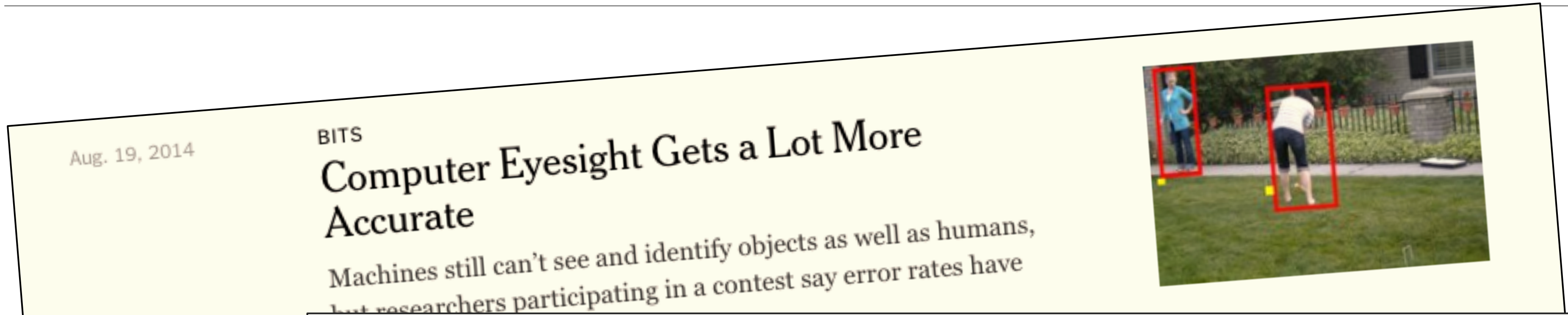
The AlexNet moment (2012)



Top5 error rate of the annual winner of the ImageNet image classification challenge. CNNs break through in 2012.



The AlexNet moment (2014-present)



Aug. 19, 2014

BITS

Computer Eyesight Gets a Lot More Accurate

Machines still can't see and identify objects as well as humans, but researchers participating in a contest say error rates have

July 17, 2016

Artificial Intelligence Swarms Silicon Valley on Wings and Wheels

The valley has found its next shiny new thing in A.I., and financiers and entrepreneurs are digging in with remarkable exuberance.

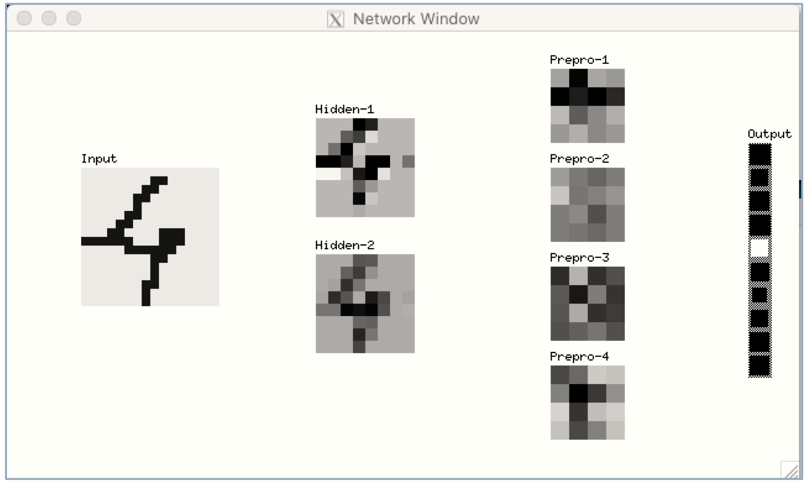
March 25, 2016

The Race Is On to Control Artificial Intelligence, and Tech's Future

Amazon, Google, IBM and Microsoft are using high salaries and games pitting humans against computers to try to claim the

A couple decades earlier (1988)

320+160
mouse-written
characters,
and a
convolutional
network



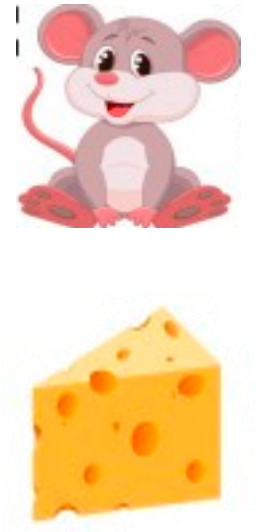
The same code was later used for the 1989 “LeNet” paper,
with a whopping 9000 training examples and 2000 testing examples

+ three decades of Moore’s law…

Why machine learning?

Absent a formal specification of what makes an image represent a mouse or a piece of cheese, we must

- either formulate heuristic specifications,
  and write a program that targets them.
- or rely on data, formulate a statistical proxy problem,
  and use a learning algorithm.



Why machine learning?

Absent a f
mouse or

- either fo
  and write
- or rely on
  and use a

Big data and big computing power

When data and computation increase
- defining heuristic specifications becomes harder.
- training learning systems becomes more effective.

Why machine learning?

Absent a f
mouse or

- either fo
  and write
- or rely on
  and use a

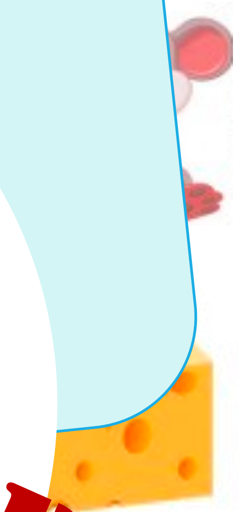
Big data and big computing power

When data and computatio
- defining heuristic specif
- training learning syster

Why machine learning?

Absent a f
mouse or

- either fo
  and write
- or rely on
  and use a



Big data and big comput ower

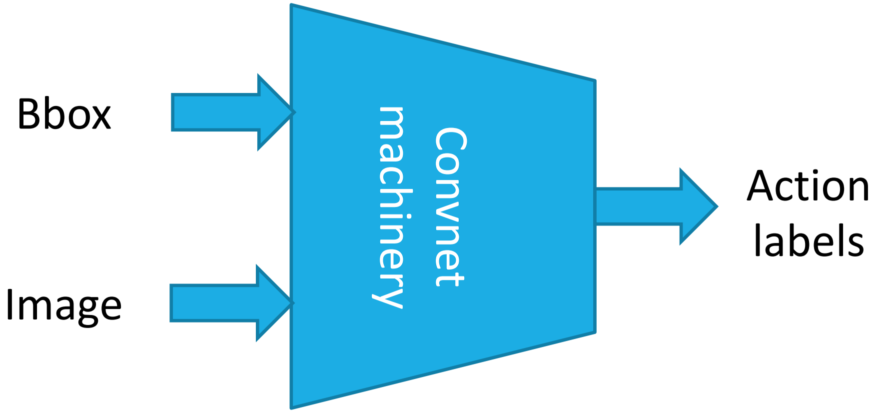
When data and computat
- defining heuristic specif
- training learning syster

WORLD DOMINATION

POSTPONED

The statistical problem is only a proxy

Example: detection of the action “*giving a phone call*”



Bbox

Image

Convnet machinery

Action labels

(Oquab et al., CVPR 2014)
~70% correct (SOTA in 2014)

The statistical problem is only a proxy

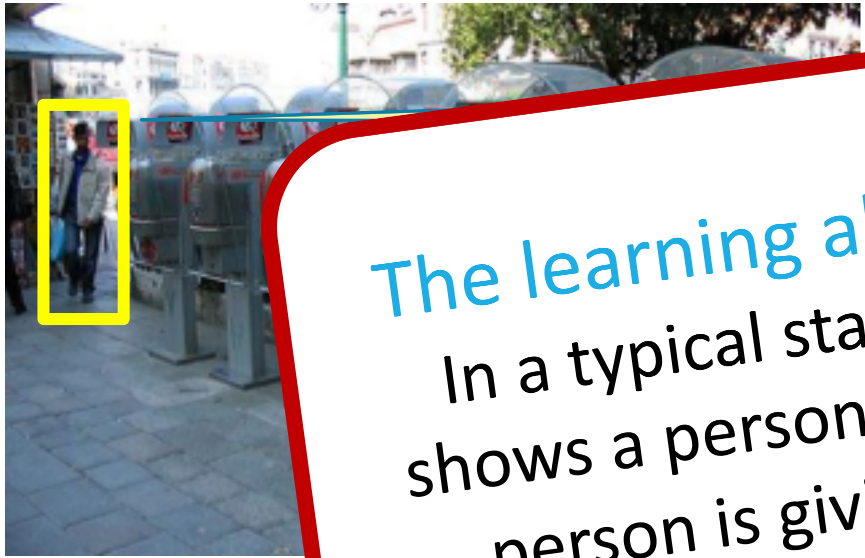
Example: detection of the action “*giving a phone call*”

Not giving a phone call.

Giving a phone call ????

The statistical problem is only a proxy

Example: detection of the action “*giving a phone call*”



The learning algorithm is statistically correct!

In a typical static image dataset, when an image shows a person near a phone, chances are that the person is giving a phone call (a selection bias.)

The statistical problem is only a proxy

Example: detection of the action *“giving a phone call”*

The learning algorithm is statistically correct!

al static image dataset, when an image

hances are that the

The learning algorithm is <u>statistically correct</u>
and is also <u>missing the point</u>!

Dataset curation and biases

Machine learning in the 1990s

- Training set carefully curated to cover all the cases of interest.
- Actual deployments (e.g. ATT-Lucent-NCR check reading machines with CNNs.)

Machine learning in the 2010s

- Datasets are too big to be carefully curated
- Data collection biases, confounding biases, feedback loops, …
- Machine learning algorithms recklessly take advantage of spurious correlations.

The statistical problem is only a proxy

Unbiased Look at Dataset Bias

Torralba & Efros

… With the focus on beating the latest benchmark numbers

ataset, have we perhaps lost sight of the

e?

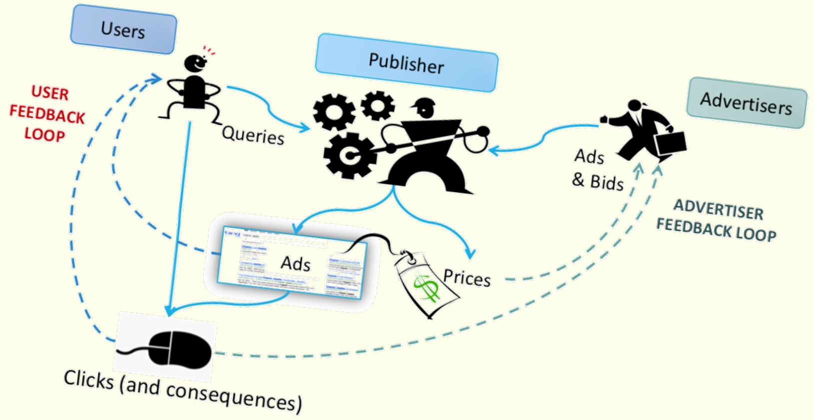
Revisiting Visual Question Answering Baselines

Jabri, Joulin & van der Maaten

… Overall, our results suggest that the performance of current VQA systems is not significantly better than that of systems designed to exploit dataset biases …

Counterfactual Reasoning and Learning Systems

Bottou, Peters, et al.



Adversarial Examples Are Not Bugs, They Are Features

Logan Engstrom, Andrew Ilyas, Aleksander Madry, Shibani Santurkar, Brandon Tran, Dimitris Tsipras · May 6, 2019

11 minute read

2

Multiple environments

NATURE DOES NOT SHUFFLE THE DATA. WE DO!

Spurious correlations

Spurious correlations are correlations that we do not expect to hold in future use cases

What informs such an expectation?

- Substantive knowledge
- Past observations ← Finding stable properties

Past observations

We do not expect spurious correlations to hold in the future.

We know this because they did not always hold in the past.

But these spurious correlations precisely appear in the data we have collected in the past!

Nature does not shuffle the data. We do!

We collect data

- at different points in time and in space
- in different experimental settings
- with different biases

environments

Then we shuffle the records and pretend that they are independent and identically distributed

Nature does not shuffle the data. We do!

We collect data

- at different points in time and in space
- in different experimental setti
- with different bi

Then we shuf nd that
they are nd identically distributed

Shuffling the data is a loss of information.

Multiple environments

Following Peters et al. (2016), we consider that data from each environment e comes with a different distribution P_e.

$$P_e = P(X_e, Y_e) \text{ for } e = 1,2,3 \ldots$$

- Training sets $D_e = \{(x_i^e, y_i^e) \sim P_e\}$ are provided for <u>some</u> e.
- We want a predictor $f(x) \approx y$ that works for <u>many</u> e.

3

From robustness to invariance

FROM INTERPOLATION TO EXTRAPOLATION

The robust approach

A very classic move in statistics

Minimize the largest error across training environments

$$f^* \;\in\; \operatorname*{ArgMin}_{f\in\mathcal{F}} \left\{ \max_{e} \; \mathbb{E}[(Y_e - f(X_e))^2] \;-\; r_e \right\}$$

Training
environments

Squared loss
or some other loss..

Per-environment
baseline

The robust approach demystified

After rewriting as a constrained optimization problem,

$$\operatorname*{ArgMin}_{f\in\mathcal{F}} M \quad \text{subject to } \ \forall e \ \ M \geq \mathbb{E}[(Y_e - f(X_e)] - r_e$$

Proposition *Subject to the Karush-Kuhn-Tucker differentiability and qualification conditions, there exist coefficients* $\lambda_e \geq 0$ *such that the robust regression* f^* *is a first order stationary point of the weighted square error*

$$C(f) = \sum_e \lambda_e \, \mathbb{E}[(Y_e - f(X_e))^2]$$

The robust approach demystified

After rewriting as a constrained optimization problem

$$\operatorname*{ArgMin}_{f\in\mathcal{F}} M \quad \text{subject to } \forall$$

Prop *ty*

and q *ch*

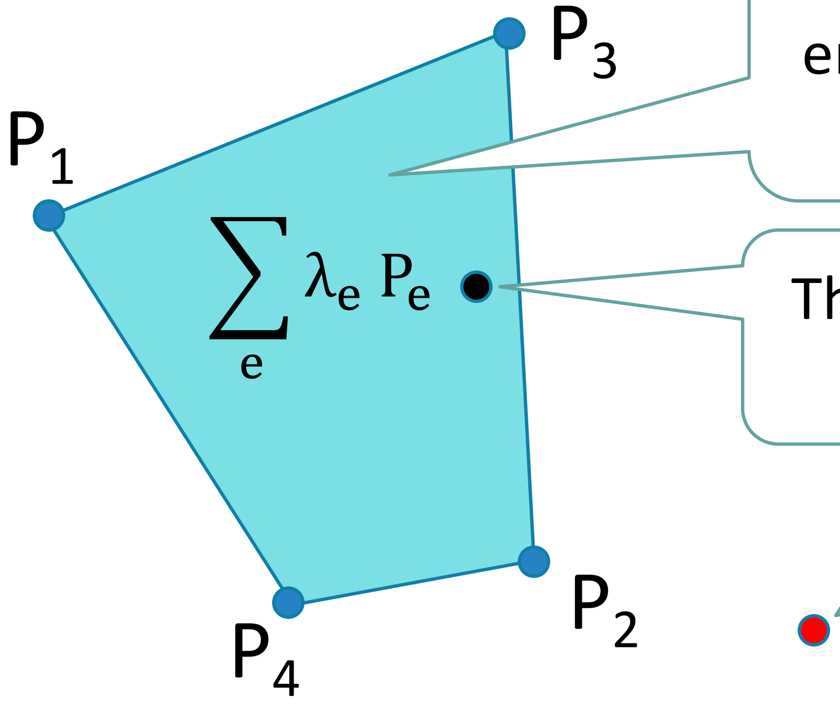
that th *point of the*

weighte

$$) = \sum_{e} \lambda_e \, \mathbb{E}[(Y_e - f(X_e))^2]$$

The robust approach
means mixing the environments
with the correct proportions.

The robust approach demystified



The robust approach guarantees a maximal error for any distribution in this convex hull, that is, a mixture with positive weights.

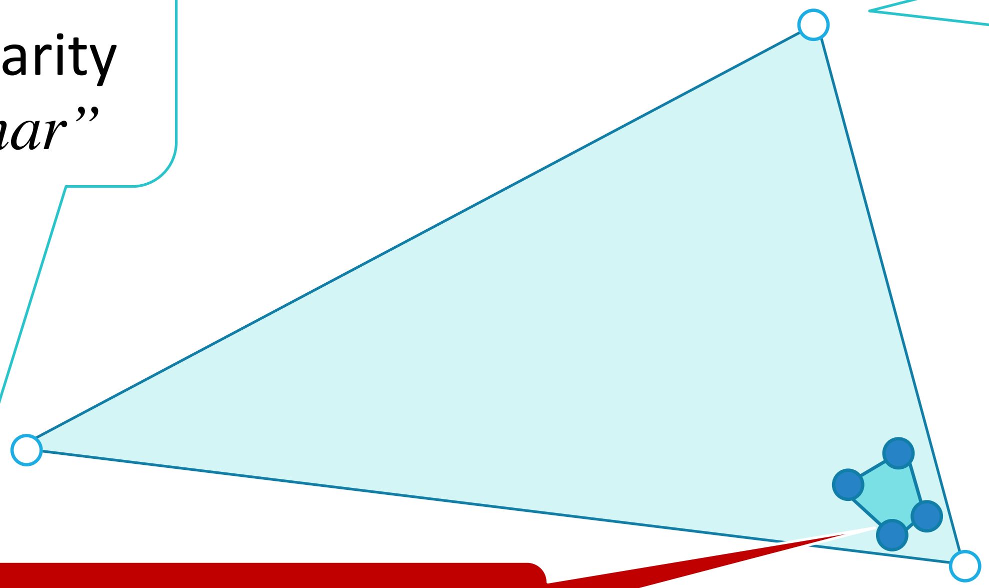
This is attained by minimizing the error for a specific mixture with positive weights.

Although valid distributions maybe reachable with negative weights, they come with no error guarantees

Negative mixtures matter!

“Easter bu

Consider a search engine query classification problem.
Let the X_e be search engine queries observed on day e=1,2,3,4.



Queries with
increasing popularity
e.g. *“ECE seminar”*

Queries with
decreasing popularity
e.g. *“Back to school*"

Queries with
constant popularity
e.g. *“Orange juice*"

Training environments

Negative mixtures matter!

“*Easter bu*

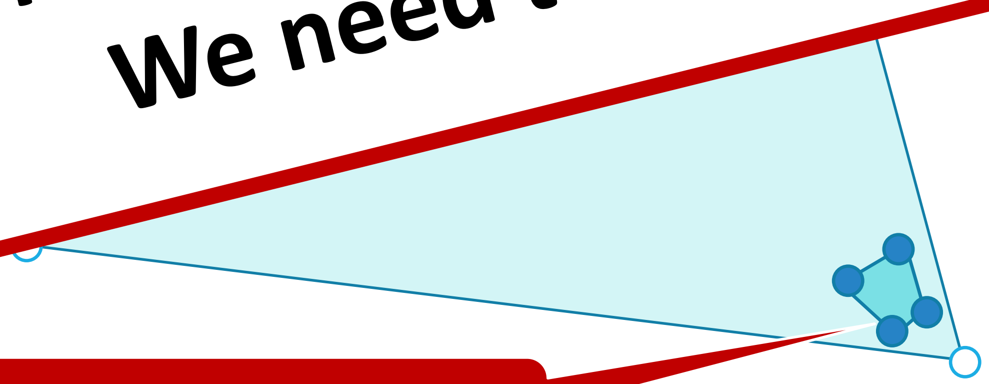
Consider a search engine query classification problem

Let the X_e be search engine queries observ

Queries with
increasing po

arity

Interpolating is not enough.
We need to extrapolate.



Queries with
constant popularity
e.g. “*Orange juice*"

Training environments

Learning stable properties

When the environments tell different stories…

- Selection bias favors pictures of a call.
- Spurious positive correlation between *person-near-phone* and *person calling*.

- Same selection bias for movies.
- But there are frames showing the person right before or right after the call.

Learning stable properties

When the environments tell different stories…

- Selectio

movies.

there are frames showing the person right before or right after the call.

gg91272144 GoGraph ©

Both environments exhibit the spurious correlation but with different strengths.
→ different regressions

Learning stable properties

When the environments tell different stories…

- Selecti

son

gg91:

Both environments exhibit the spurious correlation but with different strengths.
→ different regressions

Can we learn only what remains *invariant across environments?*

Invariant regression

A strong requirement

Simultaneously minimize the error in each training environment.

$$\forall e \quad f^* \in \operatorname*{ArgMin}_{f\in\mathcal{F}} \left\{ \mathbb{E}[(Y_e - f(X_e))^2] \right\}$$

All training environments

Squared loss or some other loss.

- Not necessarily possible without a bit of help.
- What does this mean in terms of mixture coefficients?

Invariance buys extrapolation powers

f^* is a stationary point of

$$\mathbb{E}[(Y_e - f(X_e))^2]$$

for all e.

$$\Longrightarrow$$

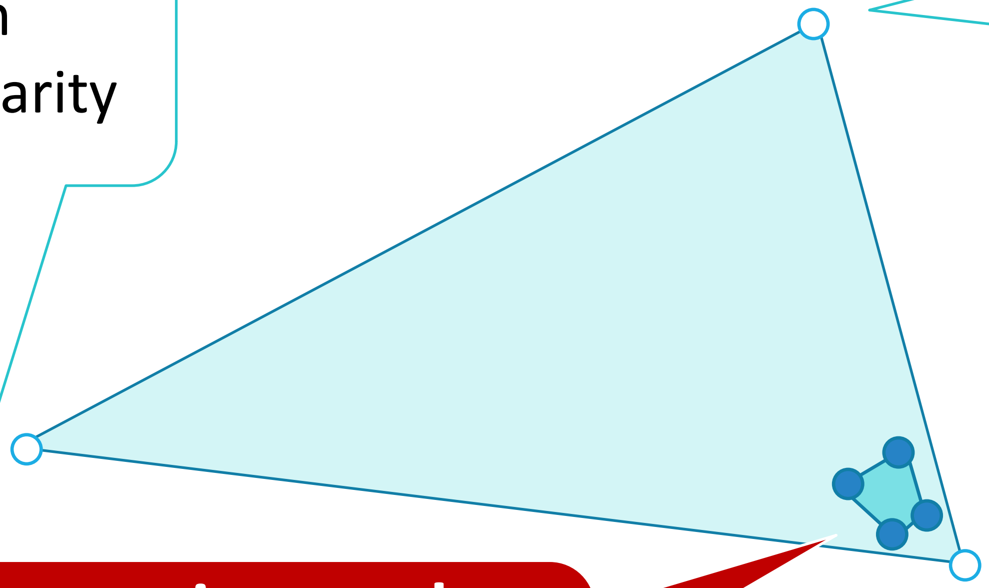
f^* is a stationary point of

$$\textstyle\sum_e \lambda_e \ \mathbb{E}[(Y_e - f(X_e))^2]$$

for all $\lambda_1 \ldots \lambda_m \in \mathbb{R}$.

These “mixture” coefficients can now be negative!

Invariance buys extrapolation powers



Queries with increasing popularity

Queries with decreasing popularity

Queries with constant popularity
e.g. *“Orange juice*"

An invariant regression on the training environments is optimal far beyond their convex hull.

Trivial existence cases

$$\forall e \quad f^* \in \operatorname*{ArgMin}_{f\in\mathcal{F}} \left\{ \mathbb{E}[(Y_e - f(X_e))^2] \right\}$$

Two cases where the invariant regression trivially exist.

- The noiseless case

  There is $f^* \in \mathcal{F}$ such that $f^*(X) = Y$ for all X.

- The realizable case

  There is $f^* \in \mathcal{F}$ such that $f^*(X_e) = \mathbb{E}[Y_e|X_e]$ whenever $\mathbb{P}(X_e) > 0$.

Trivial existence cases

$$\forall e \quad f^* \in \operatorname*{ArgMin}_{f\in\mathcal{F}} \left\{ \mathbb{E}[(Y_e - f(X_e))^2] \right\}$$

Two cases where the invariant regression exist.

- The noiseless case:
- The

whenever $\mathbb{P}(X_e) > 0$.

Things get interesting
when f^* does not exist a priori.

Playing with the function family

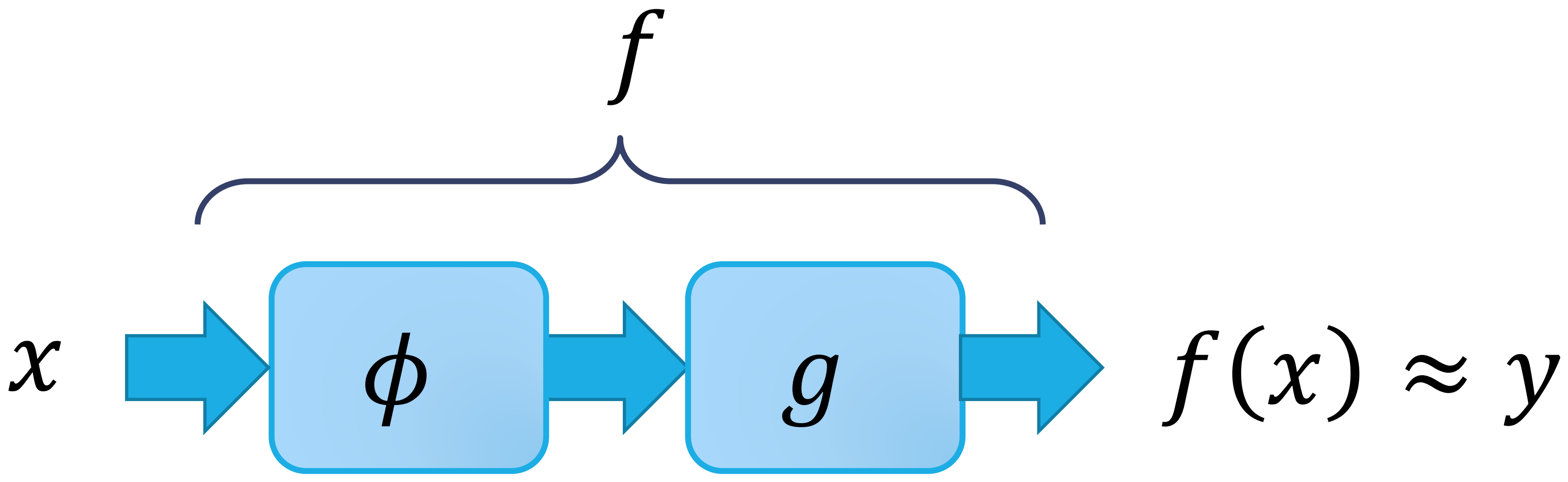
The invariant regression may not exist
when the environments tell different stories

$$\forall e \quad f^* \in \underset{f \in \mathcal{F}}{\mathrm{ArgMin}} \left\{ \mathbb{E}[(Y_e - f(X_e))^2] \right\}$$

but we can play with the family $\mathcal{F}$.

- Recall substantive modeling
- Making $\mathcal{F}$ insensitive to the spurious correlations

Invariant representation



Find a representation $\phi(x)$

Such that the regression from $\phi(x)$ to y is invariant across environments

Finding the relevant variables

> *"If we find a representation in which*
> *all falling objects obey the same laws,*
> *then we possibly understand something useful."*

4

Related work

INSPIRATIONS

1- Invariance and causation

Invariance as the main element for causal inference

To predict the outcome of an intervention, we rely on

- the properties of our intervention
- the properties assumed invariant after the intervention

Pearl's *do-calculus on causal graphs* is a framework that tells which conditionals remain invariant after an intervention. Rubin's *ignorability* assumption plays the same role.

(Cartwright, 2003; Woodward, 2005, chap 6)

1- Invariance for causal inference

Discovering invariant properties is easier than discovering causal graphs

- Finding the structure is hard, orienting the arrows is harder.
- Maybe easier with multiple environments (Bengio 2019)
- Sometimes causal graphs do not exist at all (equilibria.)

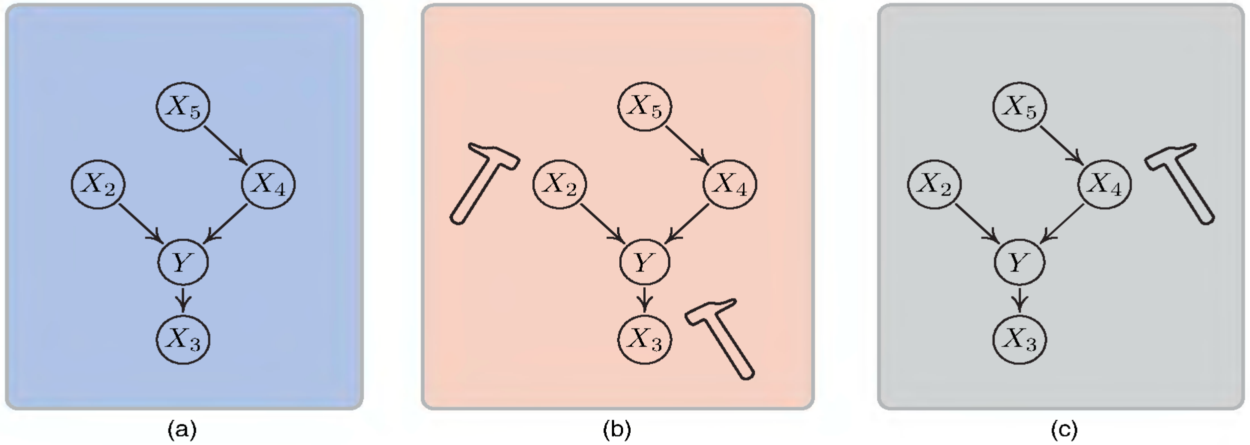
Using invariant properties is also easy

- If they’re invariant, they’re ready to use!

2- Invariant causal prediction

Inspiration

- Environments result from interventions on a causal graph.



(a) (b) (c)

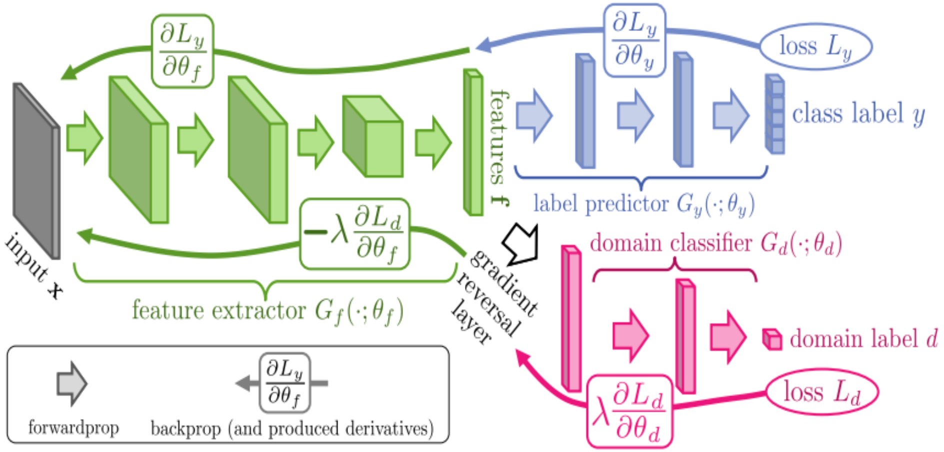
- The set of variables in the graph is assumed known.
- Representations ϕ merely select a subset of the variables.

If we find an invariant representation,
we have recovered the direct causes of Y.

(Peters, Bühlman, Meinshausen, 2016; Heinze-Deml, Peters, Meinshausen 2018)

3- Adversarial Domain Adaptation

- The goal is to learn a classifier that does not depend on the environment.
- An adversarial term makes it hard to recover the environment label e from the representation $\phi(x)$.



- This implies that $\mathbb{P}\big(\phi(X_e)\big)$ does not depend on e. Therefore $\mathbb{P}\{f(X_e)\}$ does not depend on e either. But $\mathbb{P}\{Y_e\}$ might..
- Conditional ADA stratifies on Y to achieve $\mathbb{P}(\phi(X_e)|Y_e)\perp\!\!\!\perp e$. Hence $\mathbb{E}(\phi(X_e)|Y_e)\perp\!\!\!\perp e$ instead of $\mathbb{E}\big(Y_e|\phi(X_e)\big)\perp\!\!\!\perp e$.

(Ganin et al., 2015; Edwards & Storkey, 2015; Louppe et al., 2016; Li et al,

4- Robust supervised learning

$$\underset{f\in\mathcal{F}}{\mathrm{ArgMin}}\left\{\ \sup_{Q\in\mathcal{D}_P}\ \mathbb{E}_{(X,Y)\sim Q}[\ell(Y,f(X))]\ \right\}$$

Max error for all distributions in a predefined neighborhood of the training data distribution.
e.g., $\mathcal{D}_P = \{\, Q \,:\, D(Q||P) \leq \delta \,\}$

In contrast:

- We use multiple environments to define $\mathcal{D}_P$.
- We invoke invariance to achieve robustness well beyond the convex hull of the training environment distributions (long distance.)

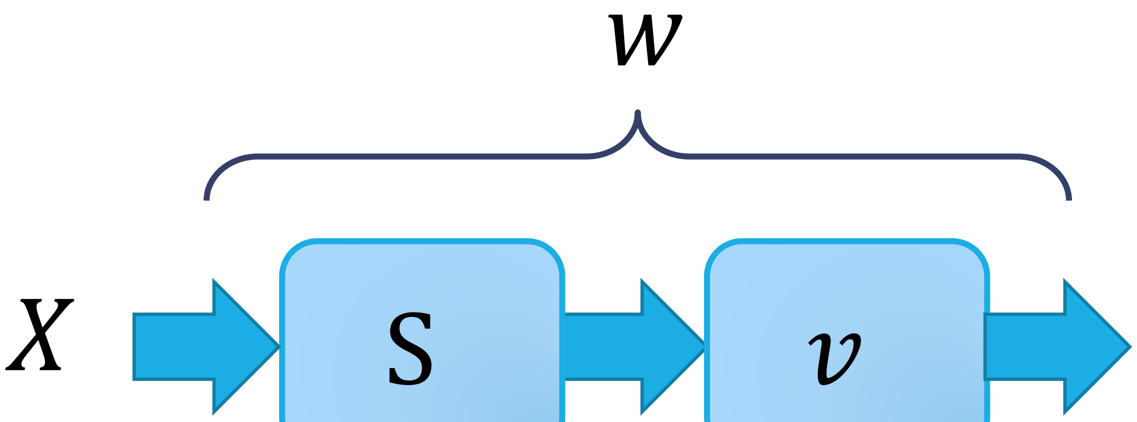
(Bagnell 2005; Duchi et al, 2015; and others)

5

The linear case

SOLVING LINEAR INVARIANT REGRESSION

The linear least square case



$$X\,\mathrm{S}^{\top}\,v = X\,w \approx Y$$

Find a matrix S

Such that v simultaneously minimizes $||Y_e - X_e\,\mathrm{S}^{\top}\,v||^2$ for all e

Issues

Find S, v such that v minimizes $||Y_e - X_e\ \mathrm{S}^{\top}\ v||^2$ for all e.

- Lots of uninteresting solutions such as $S = 0$.
- What matters is the nullspace of S: the censored information.
- Noncontinuity: an infinitesimal change in S can change its rank.

Characterization of the solutions

Let $C^e : \mathbb{R}^d \to \mathbb{R}$ be convex differentiable cost functions.

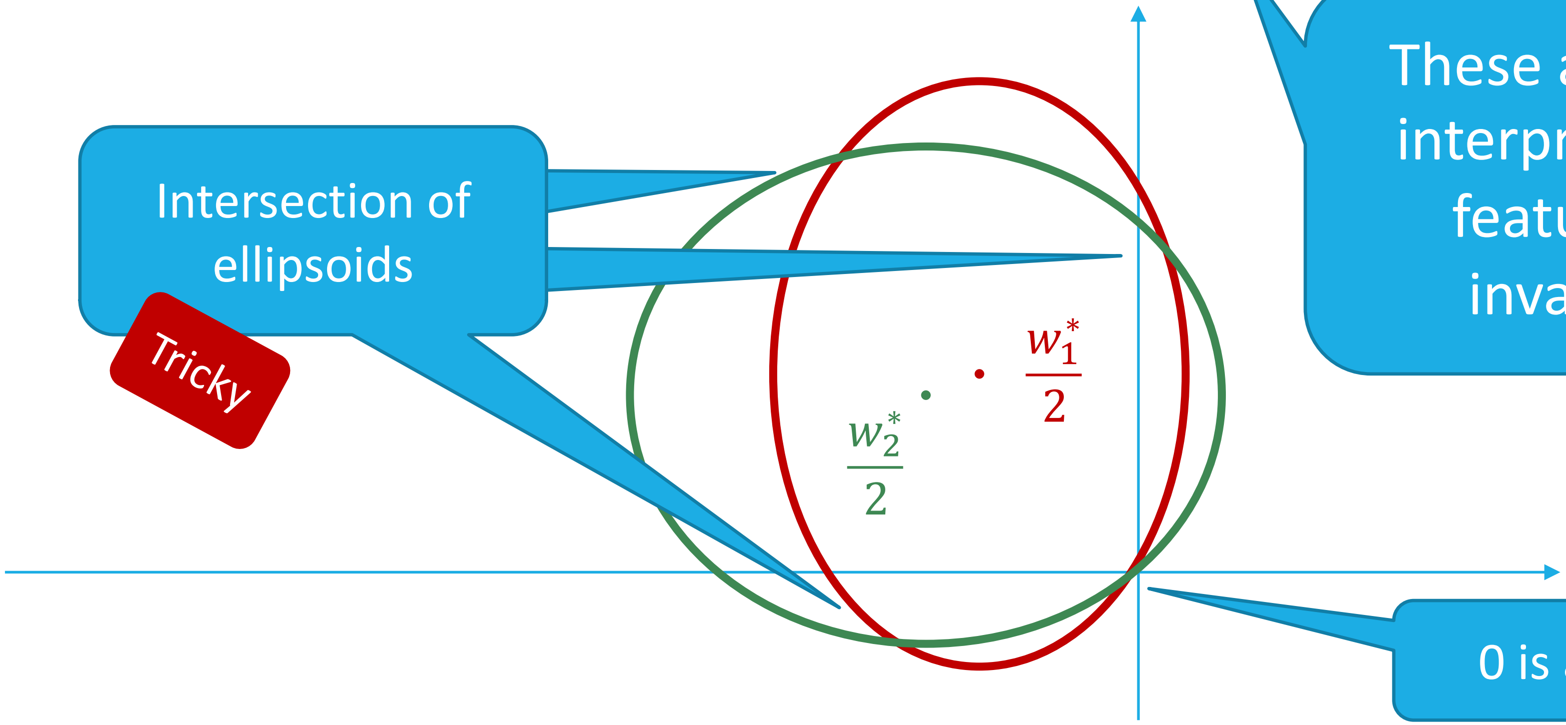
Theorem 4. *A vector* $w \in \mathbb{R}^d$ *can be written* $w = S^\top v$, *where* $S \in \mathbb{R}^{p \times d}$, *and where* $v \in \mathbb{R}^p$ *simultaneously minimize* $C^e(S^\top v)$ *for all* $e \in \mathcal{E}$, *if and only if* $w^\top \nabla C^e(w) = 0$ *for all* $e \in \mathcal{E}$.

Furthermore, the matrices S *for which such a decomposition exists are the matrices whose nullspace* $\mathrm{Ker}(S)$ *is orthogonal to* w *and contains all the* $\nabla C^e(w)$.

Where are the solutions?

$$w^\top \nabla C_e(w) = w^\top X_e^\top X_e w - w^\top X_e^\top Y_e = 0 \quad \text{for all } e.$$

These are the w than can be interpreted as products of a feature matrix S and an invariant regression v.



Intersection of ellipsoids

Tricky

0 is a solution

Rank of the feature matrix S

From the theorem:

The nullspace of S must contain all the gradients $\nabla C_e(w)$

When the gradients $\nabla C_e(w)$ are independent, $\mathrm{rank}(S) \leq d - m$.

Is it always the case?

High rank solutions

High rank solutions exist when the $\nabla C_e(w)$ are linearly dependent.

→ There are coefficients λ_e, not all zero, such that

$$\sum_e \lambda_e \nabla C_e(w) = 0$$

Dimension counting says that such w form a discrete set

→ Therefore

w is a stationary point of $\sum_e \lambda_e C_e(w)$

Potentially negative mixture coefficients again!

Exact recovery of high rank solutions

Two set of environments

- $\mathcal{E}_{all}$: the many environments we could encounter in the future.

  Assume there is a unique invariant solution on $\mathcal{E}_{all}$ with rank r.

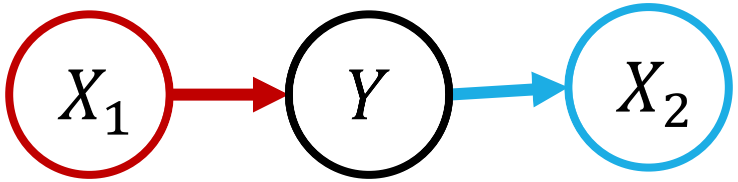
- $\mathcal{E}_{tr}$: the m environments known at training time.

  Assume that $m > d - r$ and the environments are in generic positions.
  The only invariant solution on $\mathcal{E}_{tr}$ of rank greater than r form a discrete set.
  The solution on $\mathcal{E}_{all}$ is one of them.

  In fact the solution is **uniquely** identified when $m > d - r + \frac{d}{r}$!

A minimal example

$$\begin{aligned} X_1 &= e \times \texttt{randn()} \\ Y &= X_1 + e \times \texttt{randn()} \\ X_2 &= Y + \texttt{randn()} \end{aligned}$$



Analytical derivation of the invariant representation

$$\mathbb{E}\left[Y|X_1, X_2\right] = \frac{1}{1+e^2} X_1 + \frac{e^2}{1+e^2} X_2$$

$Z = cX_1 + sX_2$

$$\mathbb{E}\left[Y|Z\right] = \frac{(c+2s)}{(c+s)^2+s^2(1+e^2)e^{-2}} Z$$

$(c, s) = (1, 0)$

$$\mathbb{E}\left[Y|X_1\right] = X_1$$

Invariant solution

Enumerating all the high rank solutions

Setup

- Two environments $e = 1$ and $e = 0.5$ with 10,000 examples each.

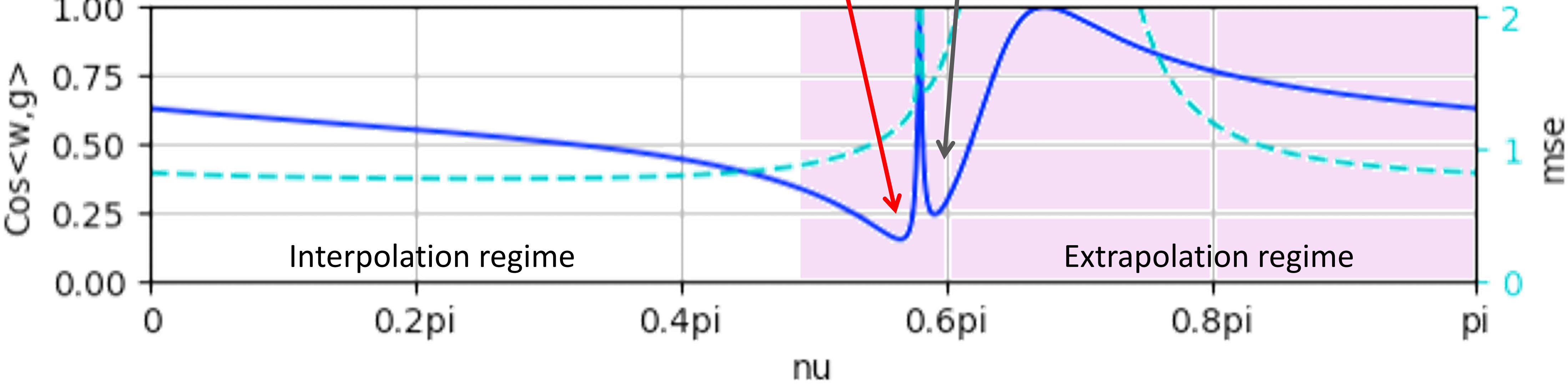
Method

- For $\nu \in [0, \pi]$, solve $\cos(\nu)\,\nabla C_1(w) + \sin(\nu)\,\nabla C_{0.5}(w) = 0$
- Plot the cosine of the angle between w and $\{\nabla C_e(w)\}$ against ν.
- This cosine is zero when invariance is exactly achieved

Enumerating the high rank solutions

$\nu = 0.567\pi \quad w = [0.9181, 0.0321]$

$\nu = 0.591\pi \quad w = [1.1908, -0.0871]$

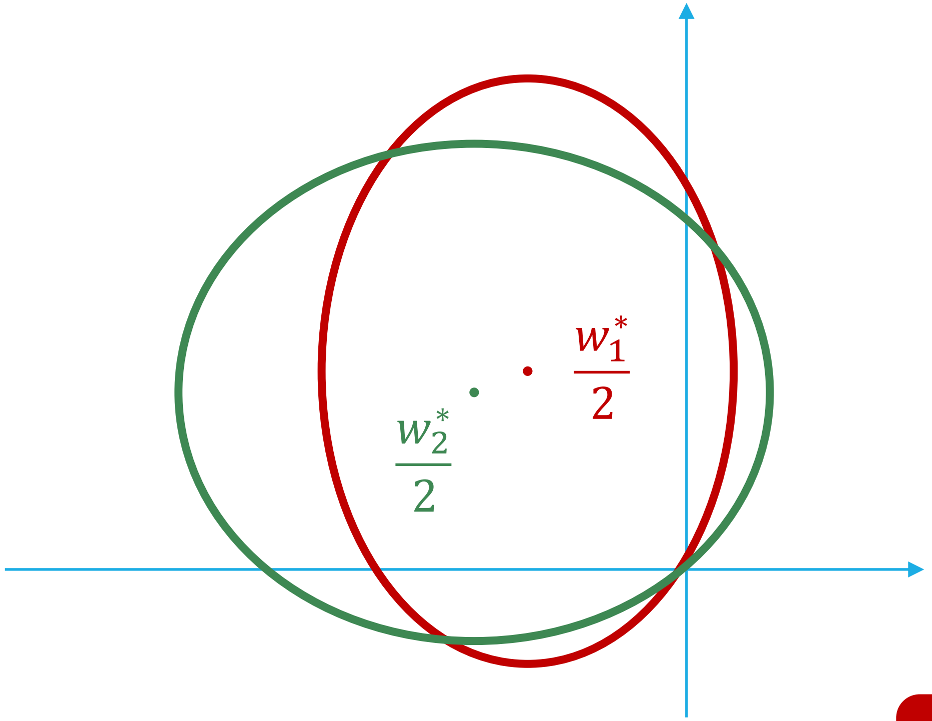


6

Invariant regularization

EXPANDING TO NONLINEAR MODELS

Favor solutions near the ellipsoids



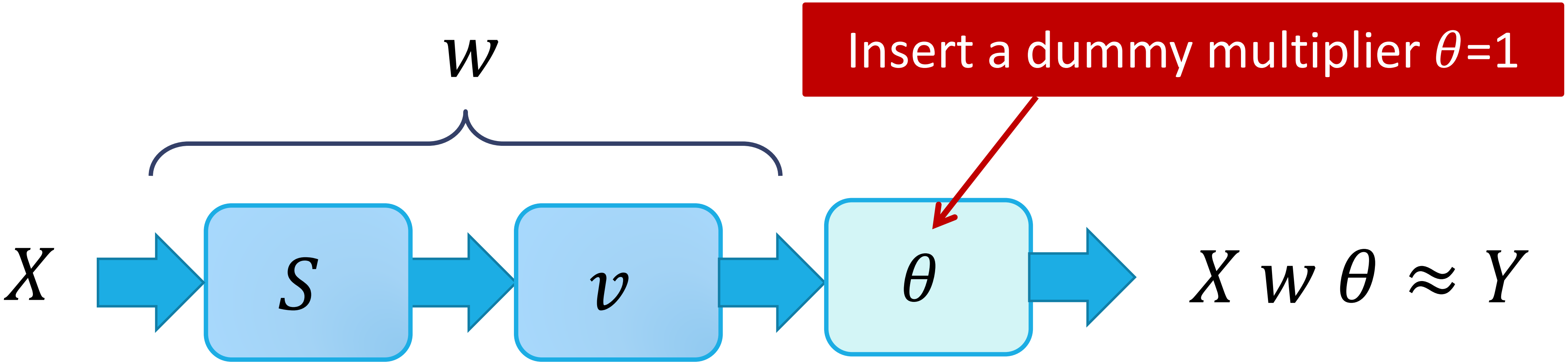
Minimize a regularized cost

$$\sum_e \overbrace{\frac{1}{n_e}\|Y_e - X_e w\|^2}^{C_e(w)}$$

$$+\kappa \sum_e \Omega_e(w)$$

Something like $\Omega_e(w) = \left(w^\top \nabla C_e(w)\right)^2$

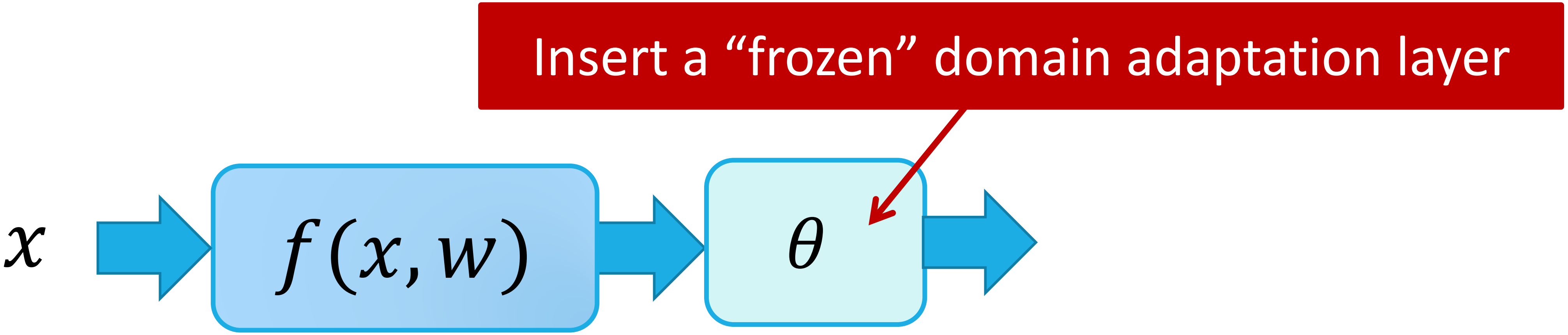
A more general perspective



Insert a dummy multiplier θ=1

$$\Omega_e(w) = \left(\frac{\partial C_e}{\partial \theta}\Big|_{\theta=1}\right)^2 = \left(w^\top \nabla C_e(w)\right)^2$$

Nonlinear version



Insert a “frozen” domain adaptation layer

$$\Omega_e(w) = \left(\frac{\partial C_e}{\partial \theta} \Big|_{\theta=1} \right)^2$$

Regularization favors weights w such that no environments would benefit from $\theta \neq 1$.

Colored MNIST

Digits with misleading colors

	Y=0	Y=1
{0,1,2,3,4}	0.75	0.25
{5,6,7,8,9}	0.25	0.75

	Red	**Green**
Y=0	$1-e$	e
Y=1	e	$1-e$

The optimal classification rate on the basis of the shape only is 75%.

Random guess is 50%.

During the training $e \in \{\, 0.1, 0.2 \,\}$. The color is a better indicator than the shape, but not a stable one.

Then we test with $e = 0.9$.

Colored MNIST

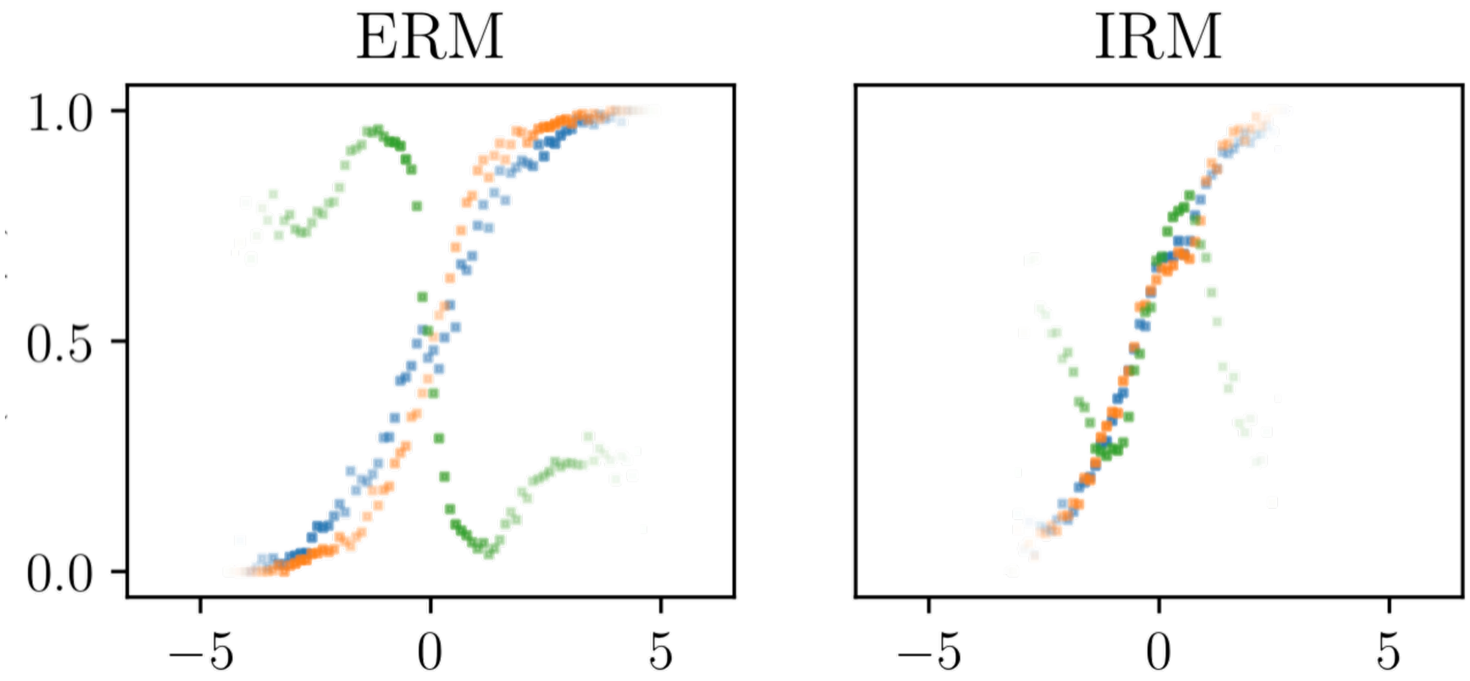
Training with $e \in \{0.1, 0.2\}$	Testing with $e \in \{0.1, 0.2\}$	Testing with $e = 0.9$
Minimize empirical risk after mixing data from both environments	84.3%	10.1%
Minimize empirical risk with invariant regularization	70.8%	66.9%

- Network is a MLP with 256 hidden units on 14x14 images.
- Invariant regularization tuned high : regularization term is nearly zero.

Colored MNIST

How invariant the representation?

$\mathbb{P}(Y|H)$ where H is the state just before the frozen adaptation layer.



🟩 $e = 0.9$

🟦 $e = 0.1$

🟨 $e = 0.2$

Scaling up
invariant regularization

Issue #1 : Numerical issues

- The regularization term is very nonconvex.

Issue #2 : Realizable problems are different…

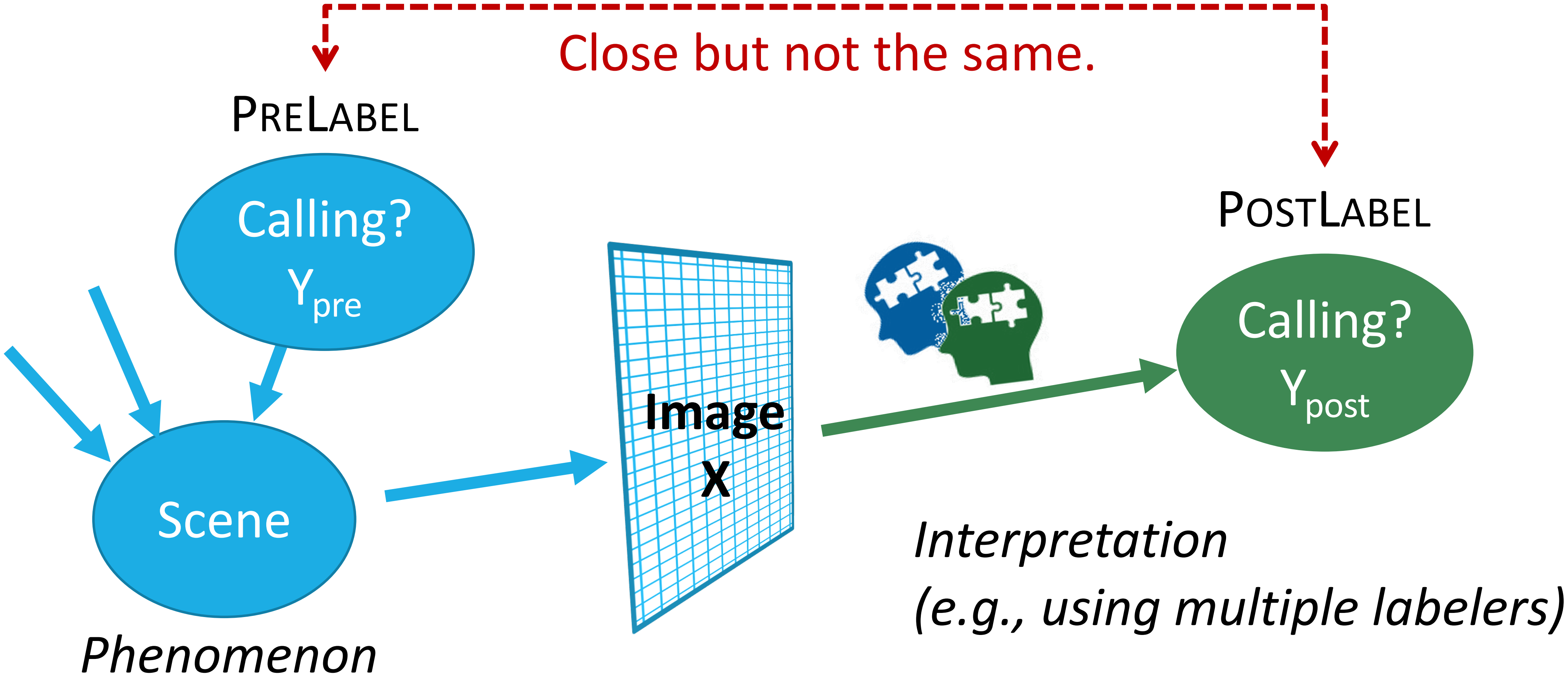
- Both the minimal example and the colored MNIST example are non-realizable: more data does not fix the problem. Many real problems are not like that…

7

Back to the realizable case

AIMING FOR ZERO TRAINING ERRORS MAKE SENSE

Phenomenon and interpretation



Close but not the same.

PRELABEL

Calling?
Y_{pre}

Scene

Phenomenon

Image
X

POSTLABEL

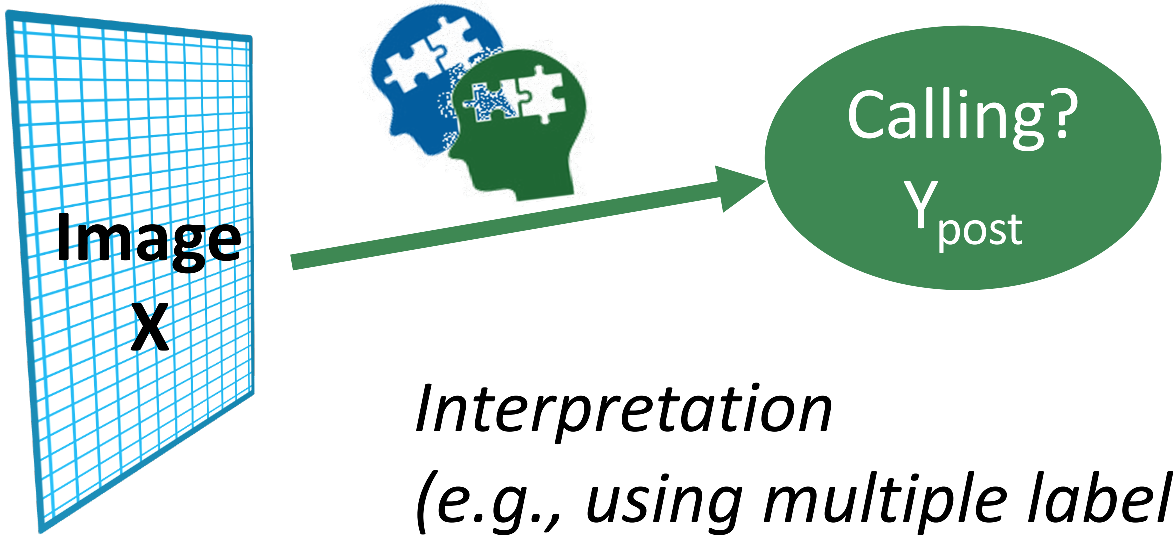
Calling?
Y_{post}

Interpretation
(e.g., using multiple labelers)

Supervised learning

The labeling process often is designed to be as deterministic as possible:

$$Y_{\text{post}} = f^*(X)$$



Image

X

POSTLABEL

Calling?

Y_{post}

Interpretation
(e.g., using multiple labelers)

Supervised learning

The lab

often is

be as d

as p

$Y_{\text{post}} = f$ (

If such an invariant function f^* exists in $\mathcal{F}$, training on the merged training sets asymptotically yields f^*

on th nion of the supports of the training environment distributions.

- The good : we gain invariance for free.
  --- up to finite training set issues.
- The bad : the idea of invariant representations is vacuous.
  -- they're all invariant!

8

Conclusions

Main points

- The statistical problem is only a proxy.
- Nature does not shuffle the examples. We shouldn’t.
- Invariance across environments buys extrapolation powers ☺
- Invariance across environments is related to causation ☺
- Invariant representations enable invariance ☺
- We need something else for the (frequent) realizable cases ☹
- This is far from cooked ☹