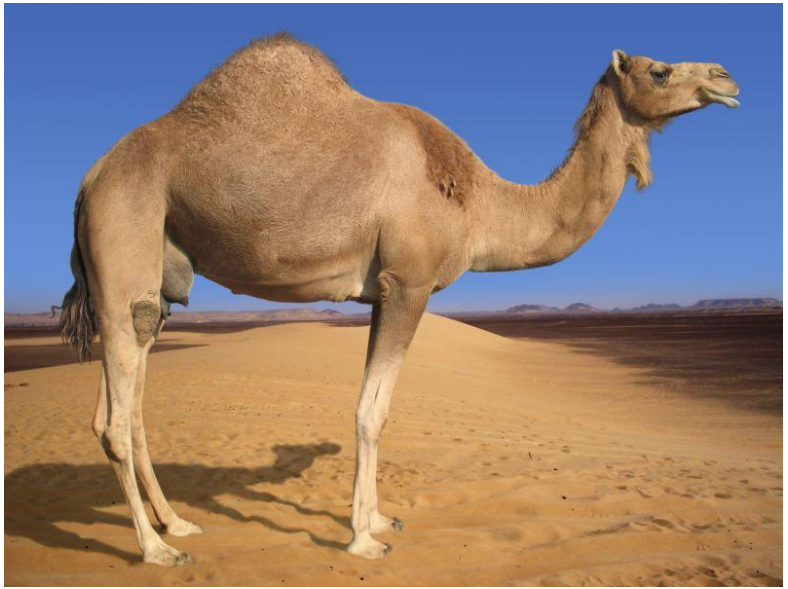
Invariant Risk Minimization

by Martin Arjovsky

https://arxiv.org/abs/1907.02893

Classification: cows vs. camels



A camel?!

Sandy background

5 Unusual Facts About Camels - YouTube
youtube.com

Camel Facts
thoughtco.com

Army Bring Camels to Texas ...
texashillcountry.com

Where Do Camels Live? - WorldAtlas.com
worldatlas.com

Camel | San Diego Zoo Animals & Plants
animals.sandiegozoo.org

Camels' Humps Are Not Filled Wi...
todayifoundout.com

E-chips to be installed on camels ...
egypttoday.com

When and Where Wer...
thoughtco.com

Camels - All About Camels Facts ...
animalcorner.co.uk

Grassy background

cow | Description & Facts | Britan…
britannica.com

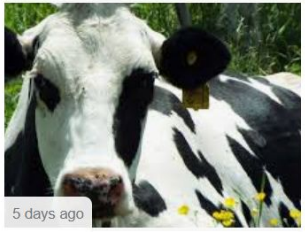
From Two Bulls, 9 Million Dairy Cows
undark.org

Dairy cattle - Wikipedia
en.wikipedia.org

Stomach Could Help Your Health ...
ucdavis.edu

Cow toilets' in Netherlands aim to cut ...
phys.org

Isis rigs cows with explosives in ...
independent.co.uk



videos will make you want to hug a c…
wbaltv.com

Escaped Rodeo Cow Is On The Lam In …
huffpost.com

Atypical BSE Confirmed in Fl…
dairyherd.com

Meet Knickers, the 1,400kg cow from ...
youtube.com

Correlation-versus-causation dilemma

Minimizing training error leads machines into recklessly absorbing **all** the correlations found in training data.

However, *spurious* correlations stemming from data biases are unrelated to the *causal* explanation of interest.

Problem: identify which properties of the training data describe spurious correlations (landscapes and contexts), and which properties represent the phenomenon of interest (animal shapes).

Causation ⇒ invariance

Spurious correlations **do not** appear to be *stable* properties.

Yet, there exists an intimate link between *invariance* and *causation* useful for generalization.

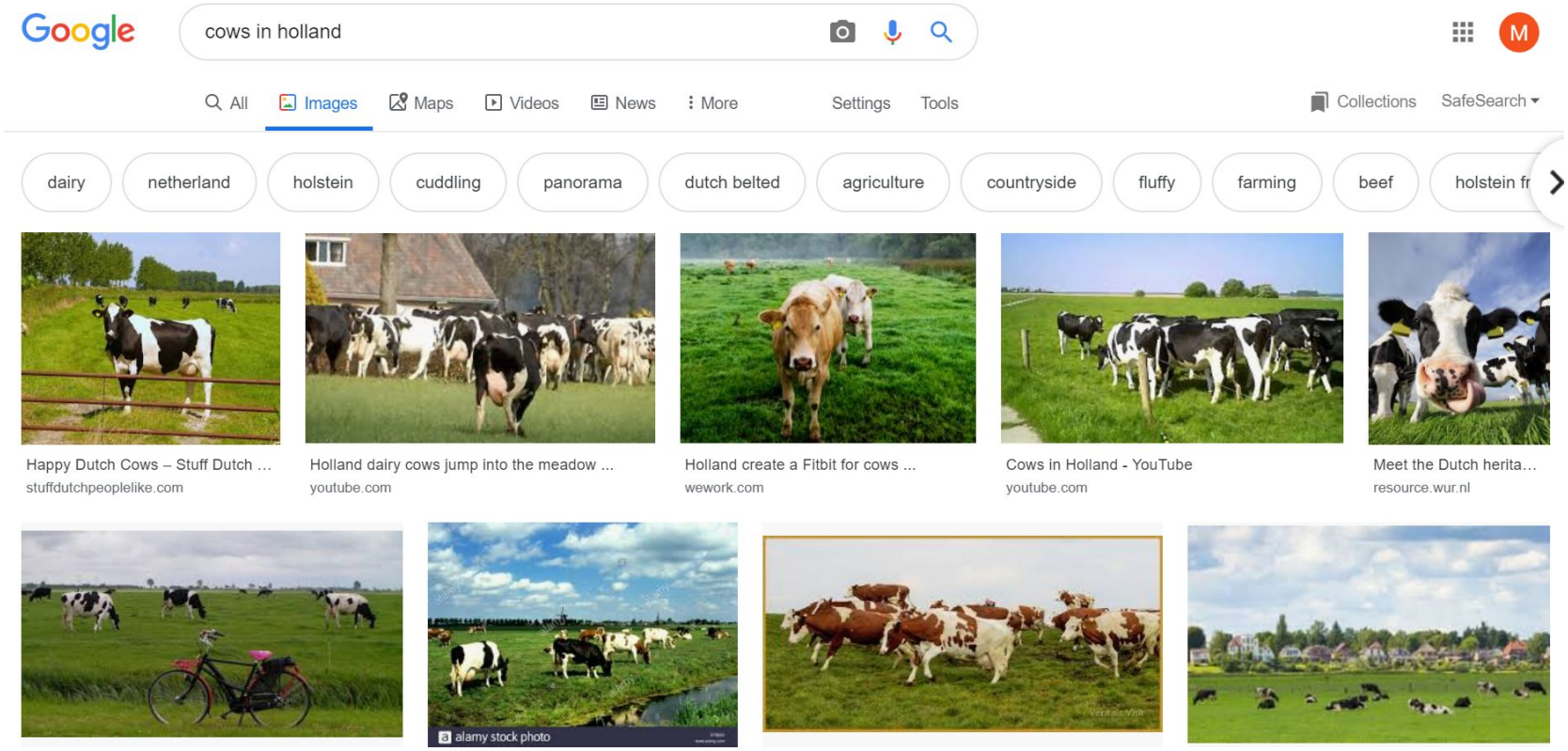
(That's why shuffling is actually not that OK: we destroy information about how the data distribution changes ⇒ no way to know which properties are **stable**.)

Strategy

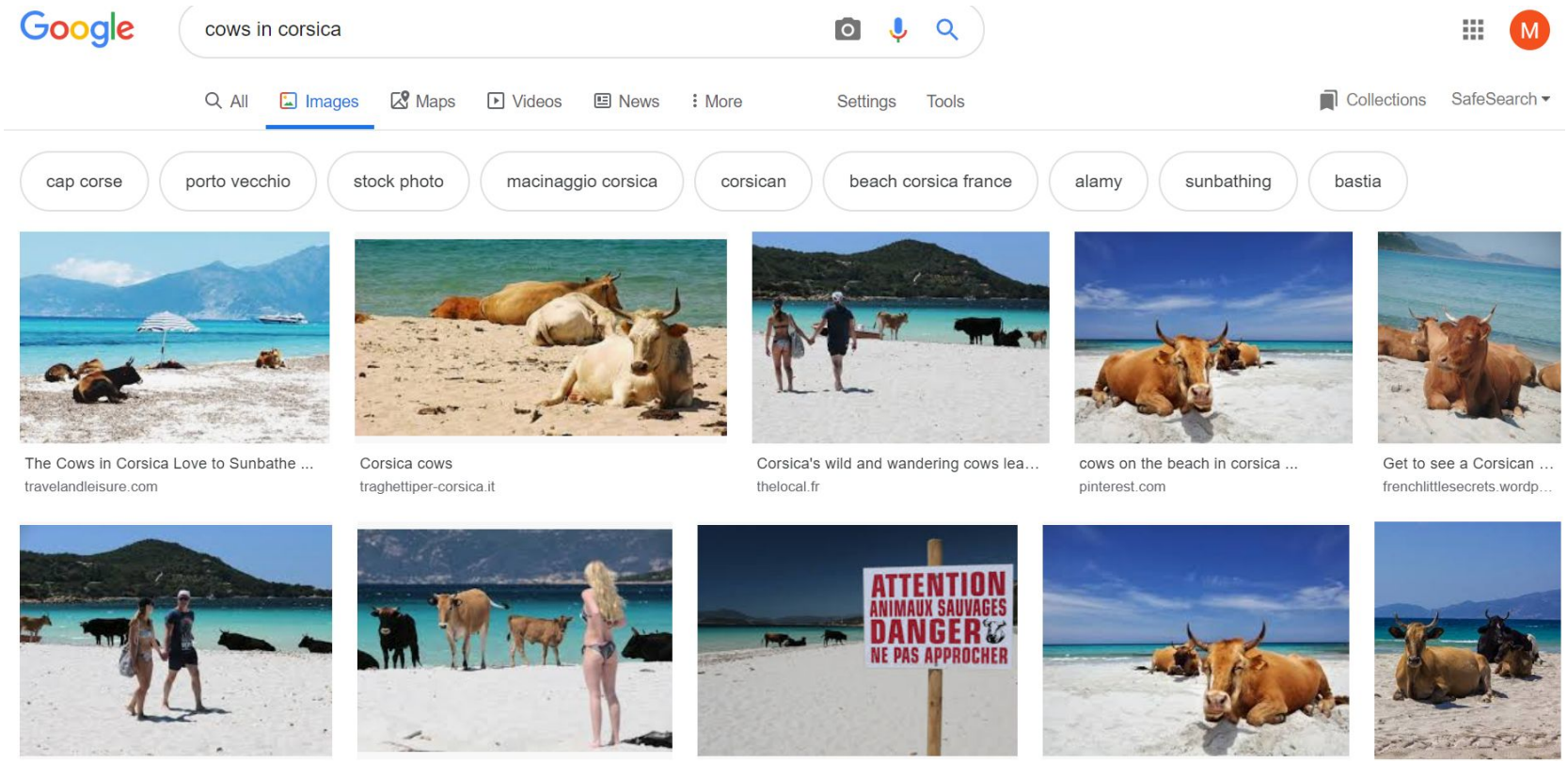
Assume that the training data is collected into distinct, separate *environments*.

We promote learning correlations that are **stable** across training environments, as these should also hold in novel testing environments.

Different cows data environments



Different cows data environments



Invariant Risk Minimization (IRM) principle

To learn invariances across environments, find a data representation such that the optimal classifier on top of that representation matches for all environments.

Basic formulation of the problem

Consider datasets $D_e := \{(x_i^e, y_i^e)\}_{i=1}^{n_e} \quad e \in \mathcal{E}_{\text{tr}} \quad \mathcal{E}_{\text{all}} \supset \mathcal{E}_{\text{tr}}$

Our **goal** is to learn a predictor $Y \approx f(X)$

We wish to minimize $R^{\text{OOD}}(f) = \max_{e \in \mathcal{E}_{\text{all}}} R^e(f)$

where $R^e(f) := \mathbb{E}_{X^e, Y^e}[\ell(f(X^e), Y^e)]$ is the risk under environment e.

Example

$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$

Let $$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$

$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

and $$\mathcal{E}_{\text{tr}} = \{\text{replace } \sigma^2 \text{ by } 10, \text{ replace } \sigma^2 \text{ by } 20\}$$

Consider $\hat{Y}^e = X_1^e \hat{\alpha}_1 + X_2^e \hat{\alpha}_2$ ⇒ we can:

- regress from X_1^e, to obtain $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0$,
- regress from X_2^e, to obtain $\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 = \sigma(e)/(\sigma(e) + \frac{1}{2})$,
- regress from (X_1^e, X_2^e), to obtain $\hat{\alpha}_1 = 1/(\sigma(e) + 1)$ and $\hat{\alpha}_2 = \sigma(e)/(\sigma(e) + 1)$.

Analysis

The regression using X_1 is our first example of an *invariant correlation*: this is **the only** regression whose coefficients do not depend on the environment.

Conversely, the second and third regressions exhibit coefficients that *vary* from environment to environment.

The *invariant* rule $\hat{Y} = 1 \cdot X_1 + 0 \cdot X_2$ is **the only** predictor with finite R^{OOD} and this actually **is** the *causal* explanation.

The many faces of generalization (prior work)

1. *Empirical Risk Minimization (ERM)*
2. *Robust learning*: minimize $R^{\mathrm{rob}}(f) = \max_{e \in \mathcal{E}_{\mathrm{tr}}} R^e(f) - r_e$
   Turns out to be equivalent to weighted ERM.
3. *Domain adaptation*: estimate a data representation that follows the same distribution for all environments.
   Sometimes attempts to enforce the wrong type of invariance.
4. *Invariant causal prediction (ICP)*: search for the subset of variables which produce equally distributed regression residuals.

IRM formulation

Definition 3. *We say that a data representation $\Phi : \mathcal{X} \to \mathcal{H}$ elicits an invariant predictor $w \circ \Phi$ across environments $\mathcal{E}$ if there is a classifier $w : \mathcal{H} \to \mathcal{Y}$ simultaneously optimal for all environments, that is, $w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w} \circ \Phi)$ for all $e \in \mathcal{E}$.*

This concept of invariance clarifies common induction methods in science!

$$\begin{aligned} \min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ w:\mathcal{H}\to\mathcal{Y}}} \quad & \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w \circ \Phi) \\ \text{subject to} \quad & w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w} \circ \Phi), \text{ for all } e \in \mathcal{E}_{\text{tr}}. \end{aligned} \tag{IRM}$$

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0} R^e(w \cdot \Phi)\|^2, \tag{IRMv1}$$

From (IRM) to (IRMv1)

1. Phrasing the constraints as a penalty

$$\begin{aligned} &\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} && \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w\circ\Phi) \\ &\text{subject to} && w \in \operatorname*{arg\,min}_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{\text{tr}}. \end{aligned} \tag{IRM}$$

becomes

$$L_{\text{IRM}}(\Phi, w) = \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w\circ\Phi) + \lambda\cdot\mathbb{D}(w,\Phi,e)$$

2. Choosing a penalty for linear classifiers

Consider linear-least squares regression case ⇒ then the optimum is:

$$w_{\Phi}^{e} = \mathbb{E}_{X^e}\left[\Phi(X^e)^{\top}\Phi(X^e)\right]^{-1}\mathbb{E}_{X^e,Y^e}\left[\Phi(X^e)^{\top}Y^e\right]$$

And hence we can consider two similar types of penalty:

$$\mathbb{D}_{\text{dist}}(w,\Phi,e) = \|w - w_{\Phi}^{e}\|^2$$

and

$$\mathbb{D}_{\text{lin}}(w,\Phi,e) = \left\|\mathbb{E}_{X^e}\left[\Phi(X^e)^{\top}\Phi(X^e)\right]w - \mathbb{E}_{X^e,Y^e}\left[\Phi(X^e)^{\top}Y^e\right]\right\|^2$$

2. Choosing a penalty for linear classifiers

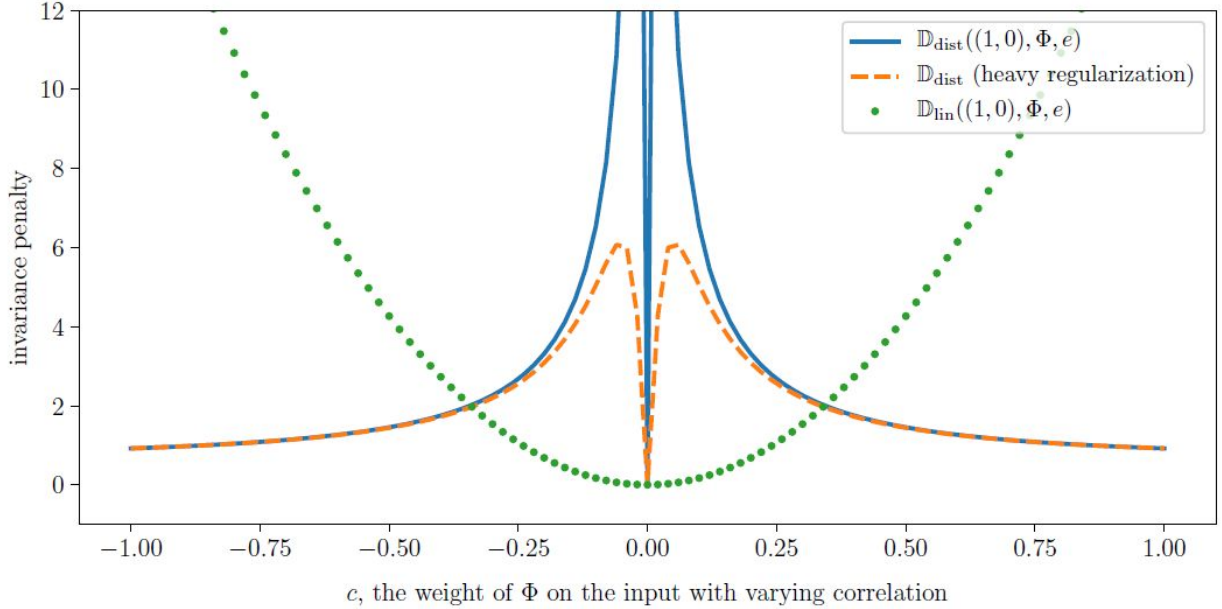


Figure 1: Different measures of invariance lead to different optimization landscapes in our Example 1. The naïve approach of measuring the distance between optimal classifiers $\mathbb{D}_{\text{dist}}$ leads to a discontinuous penalty (solid blue unregularized, dashed orange regularized). In contrast, the penalty $\mathbb{D}_{\text{lin}}$ does not exhibit these problems.

3. Fixing the linear classifier

The problem is over-parameterized: $w \circ \Phi = \underbrace{\left(w \circ \Psi^{-1}\right)}_{\tilde{w}} \circ \underbrace{\left(\Psi \circ \Phi\right)}_{\tilde{\Phi}}.$

Let's restrict our search! *We will fix the non-zero classifier $\tilde{w}$ and find a data representation such that the optimal classifier, on top of that data representation, is $\tilde{w}$ for all environments*:

$$L_{\mathrm{IRM},w=\tilde{w}}(\Phi) = \sum_{e \in \mathcal{E}_{\mathrm{tr}}} R^e(\tilde{w} \circ \Phi) + \lambda \cdot \mathbb{D}_{\mathrm{lin}}(\tilde{w}, \Phi, e).$$

4. Scalar classifiers are sufficient to monitor invariance

It turns out that $\tilde{w} = (1, 0, \ldots, 0)$ is sufficient! So:

$$L_{\text{IRM},w=1.0}(\Phi^\top) = \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(\Phi^\top) + \lambda \cdot \mathbb{D}_{\text{lin}}(1.0, \Phi^\top, e).$$

Theorem 4. *For all $e \in \mathcal{E}$, let $R^e : \mathbb{R}^d \to \mathbb{R}$ be convex differentiable cost functions. A vector $v \in \mathbb{R}^d$ can be written $v = \Phi^\top w$, where $\Phi^\top \in \mathbb{R}^{p \times d}$, and where $w \in \mathbb{R}^p$ simultaneously minimize $R^e(w \circ \Phi)$ for all $e \in \mathcal{E}$, if and only if $v^\top \nabla R^e(v) = 0$ for all $e \in \mathcal{E}$. Furthermore, the matrices Φ for which such a decomposition exists are the matrices whose nullspace $\text{Ker}(\Phi)$ is orthogonal to v and contains all the $\nabla R^e(v)$.*

Theorem 4 illustration

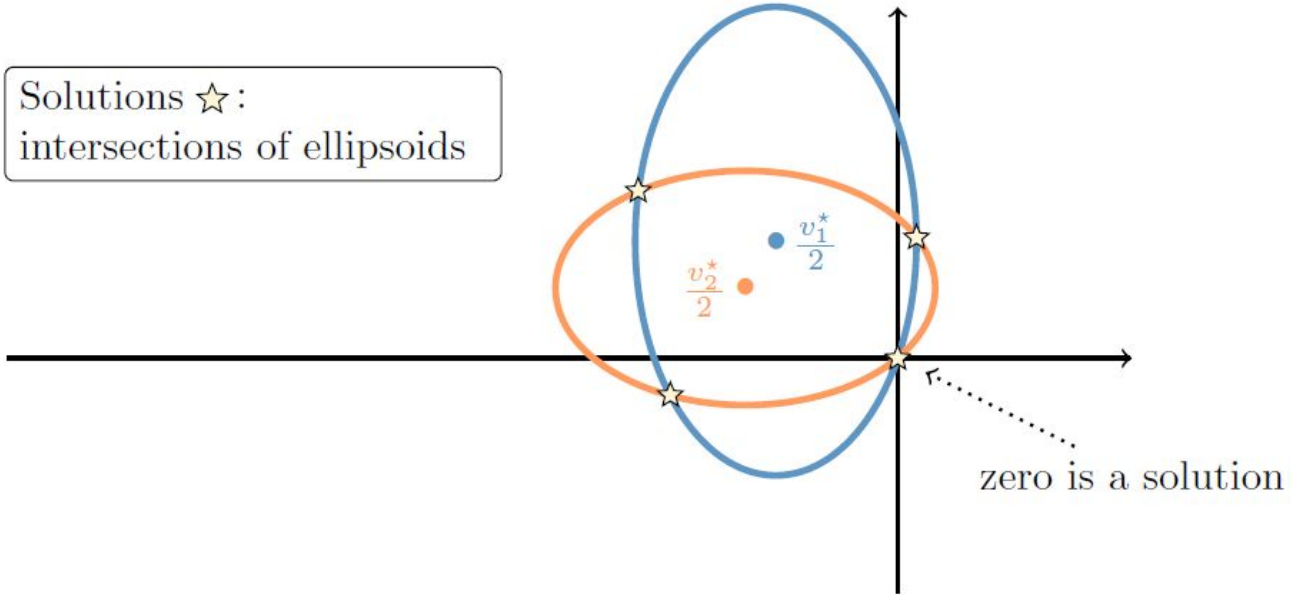


Figure 2: The solutions of the invariant linear predictors $v = \Phi^\top w$ *coincide with the intersection of the ellipsoids representing the orthogonality condition* $v^\top \nabla R^e(v) = 0$.

5. Extending to general losses and multivariate outputs

Recall $\mathbb{D}_{\text{lin}}(w, \Phi, e) = \left\| \mathbb{E}_{X^e}\left[\Phi(X^e)^\top \Phi(X^e)\right] w - \mathbb{E}_{X^e, Y^e}\left[\Phi(X^e)^\top Y^e\right] \right\|^2$

We can rewrite this linear penalty: $\mathbb{D}(1.0, \Phi, e) = \|\nabla_{w|w=1.0} R^e(w \cdot \Phi)\|^2$

If the target space $\mathcal{Y}$ returned by Φ has multiple outputs, we multiply all of them by the fixed scalar classifier $w = 1.0$.

$$\min_{\Phi: \mathcal{X} \to \mathcal{Y}} \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0} R^e(w \cdot \Phi)\|^2, \qquad \text{(IRMv1)}$$

Invariance, causality and generalization

Fundamental questions

IRM promotes low error and invariance across **training environments**.

*When do these conditions imply **invariance** across **all** environments?*

*When do these conditions lead to **low error** across **all** environments? (Basically, OOD generalization)*

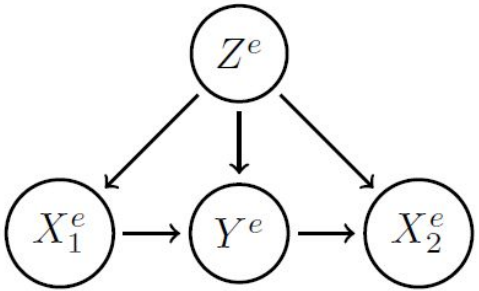
*How does statistical invariance and out-of-distribution generalization relate to concepts from the **theory of causation**?*

Please see the article for details...

Experiments

https://github.com/facebookresearch/InvariantRiskMinimization

Synthetic data



$$
\begin{aligned}
Z^e &\leftarrow \mathcal{N}(0, e^2) \\
X_1^e &\leftarrow Z^e \cdot W_{h\to 1} + \mathcal{N}(0, e^2) \\
Y^e &\leftarrow Z^e \cdot W_{h\to y} + X_1^e \cdot W_{1\to y} + \mathcal{N}(0, \sigma_y^2) \\
X_2^e &\leftarrow Z^e \cdot W_{h\to 2} + Y^e \cdot W_{y\to 2} + \mathcal{N}(0, \sigma_2^2)
\end{aligned}
$$

Figure 3: In our synthetic experiments, the task is to predict Y^e *from* $X^e = (X_1^e, X_2^e) \cdot S$ *across multiple environments* $e \in \mathbb{R}$.

- *Scrambled* (S) observations, where S is an orthogonal matrix, or *unscrambled* (U) observations, where $S = I$.
- *Fully-observed* (F) graphs, where $W_{h\to 1} = W_{h\to y} = W_{h\to 2} = 0$, or *partially-observed* (P) graphs, where $(W_{h\to 1}, W_{h\to y}, W_{h\to 2})$ are Gaussian.
- *Homoskedastic* (O) Y-noise, where $\sigma_y^2 = e^2$ and $\sigma_2^2 = 1$, or *heteroskedastic* (E) Y-noise, where $\sigma_y^2 = 1$ and $\sigma_2^2 = e^2$.

Synthetic data

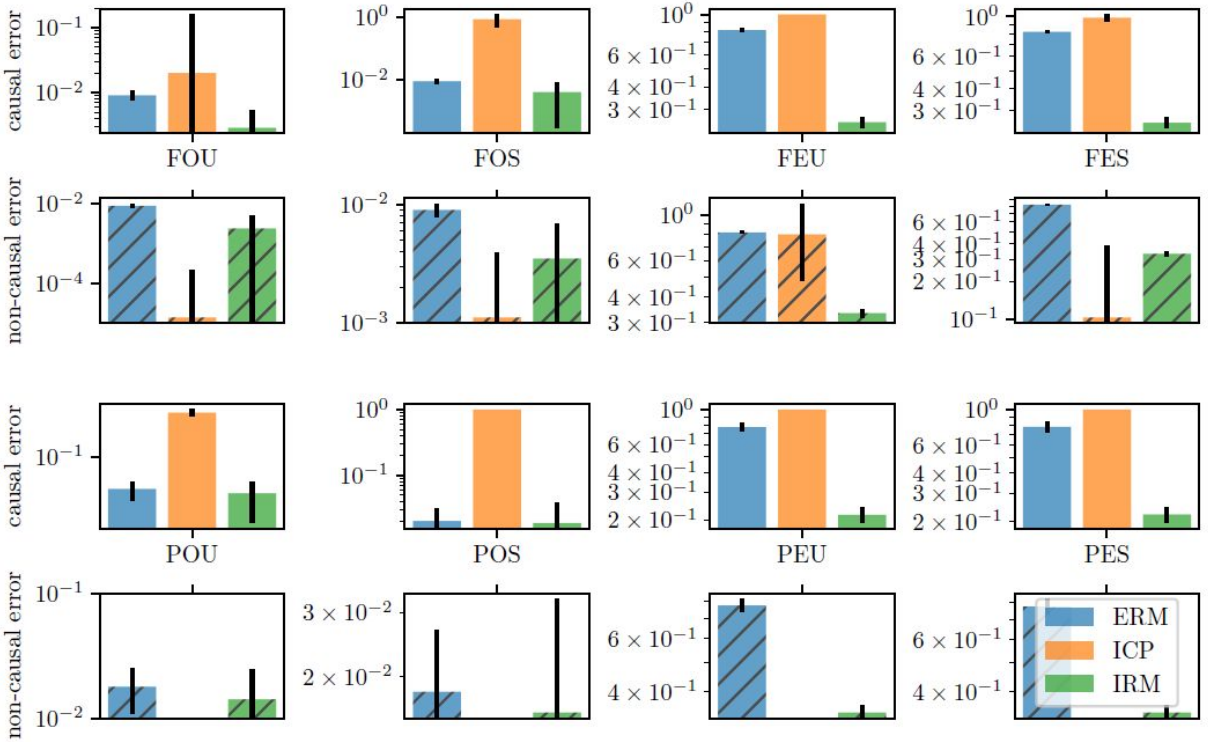


Figure 4: Average errors on causal (plain bars) and non-causal (striped bars) weights for our synthetic experiments. The y-axes are in log-scale. See main text for details.

Colored MNIST

Binary label: y = 0 for (0-4) / 1 for (5-9), then flip with prob. 0.25

Color: green(0) / red(1) by flipping y with prob. p_e

2 training environments: {p_e = 0.2, p_e = 0.1} and 1 testing: {p_e = 0.9}

Algorithm	Acc. train envs.	Acc. test env.
ERM	87.4 ± 0.2	17.1 ± 0.6
IRM (ours)	70.8 ± 0.9	$\mathbf{66.9 \pm 2.5}$
Random guessing (hypothetical)	50	50
Optimal invariant model (hypothetical)	75	75
ERM, grayscale model (oracle)	73.5 ± 0.2	73.0 ± 0.4

Table 1: Accuracy (%) of different algorithms on the Colored MNIST synthetic task. ERM fails in the test environment because it relies on spurious color correlations to classify digits. IRM detects that the color has a spurious correlation with the label and thus uses only the digit to predict, obtaining better generalization to the new unseen test environment.

Colored MNIST

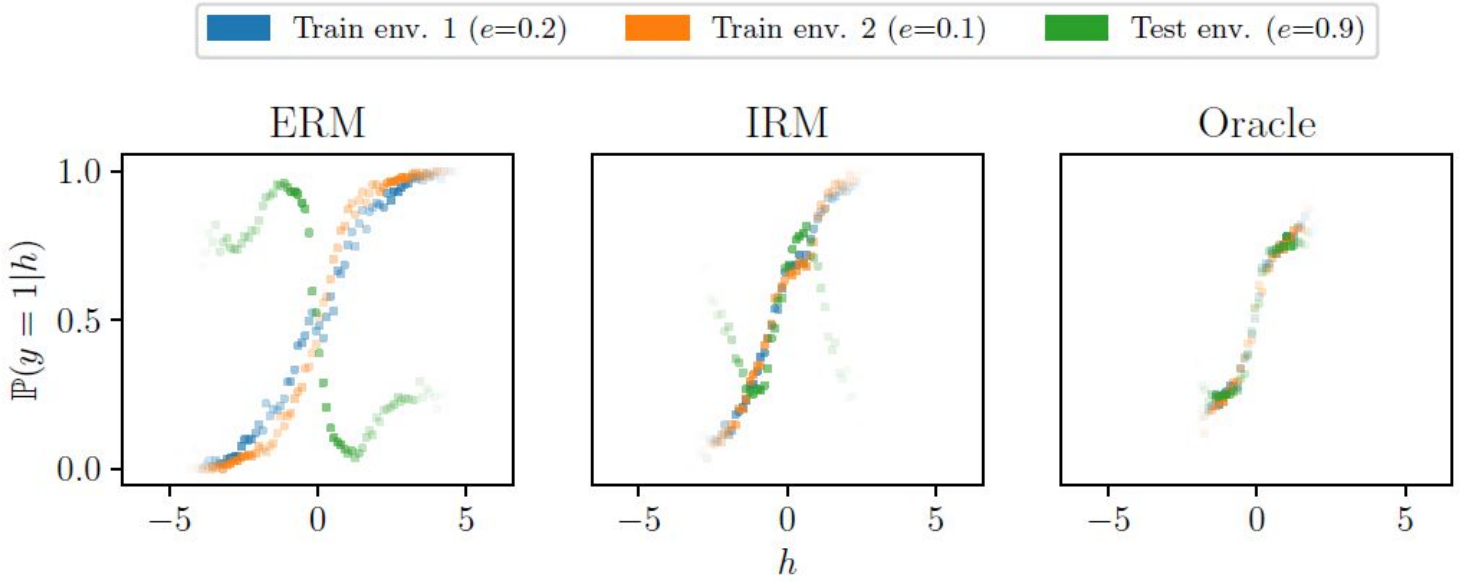
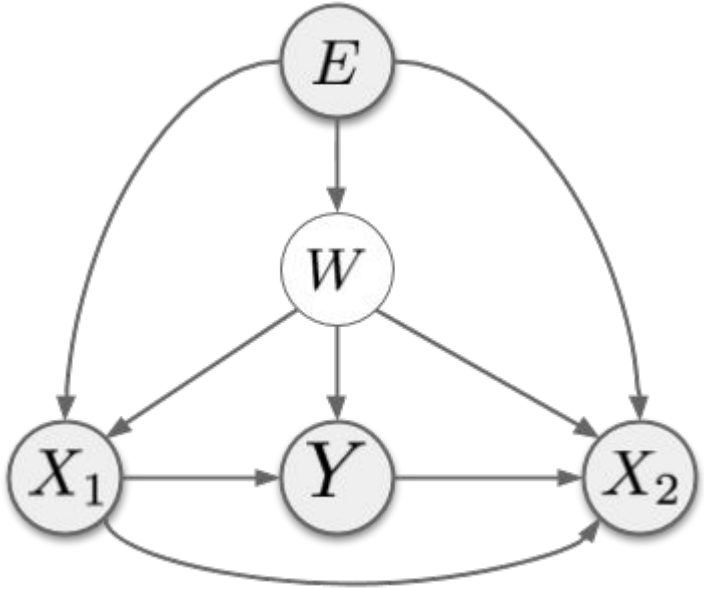


Figure 5: $P(y = 1|h)$ as a function of h for different models trained on Colored MNIST: (left) an ERM-trained model, (center) an IRM-trained model, and (right) an ERM-trained model which only sees grayscale images and therefore is perfectly invariant by construction. IRM learns approximate invariance from data alone and generalizes well to the test environment.

An information theoretic view

https://www.inference.vc/invariant-risk-minimization/

Information theoretic formulation of IRM



$\phi(X)$ such that:

- $Y \perp E|\phi(X)$, and
- ϕ is informative about y, i.e. we can predict y accurately from $\phi(x)$

Smells like information bottleneck...

$$\max_{\phi} \{I[Y, Z] - \beta I[X, Z]\}$$

$$\max_{\phi} \{I[Y, Z] - \beta I[Y, E|Z]\} \qquad Z = \phi(X)$$

$$L_{\text{IRM}}(\Phi, w) = \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(w \circ \Phi) + \lambda \cdot \mathbb{D}(w, \Phi, e)$$

$$\min_{\Phi: \mathcal{X} \to \mathcal{Y}} \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0} R^e(w \cdot \Phi)\|^2, \qquad \text{(IRMv1)}$$

From information to gradient penalties

$$I[Y, E|\phi(x)] \approx \min_{\theta} \mathbb{E}_e \mathcal{R}^e(f_\theta \circ \phi) - \mathbb{E}_e \min_{\theta_e} \mathcal{R}^e(f_{\theta_e} \circ \phi)$$

$$I[Y, E|\phi(x)] \geq \min_{\theta} \mathbb{E}_e \left\{ \mathcal{R}^e(f_\theta \circ \phi) - \min_{\|d_e\| \leq \epsilon} \mathcal{R}^e(f_{\theta + d_e} \circ \phi) \right\}$$

$$I[Y, E|\phi(x)] \geq \min_{\theta} \mathbb{E}_e \|\nabla_\theta \mathcal{R}^e(f_\theta \circ \phi)\|_2$$

$$\min_{\phi} \left\{ \min_{\theta} \mathbb{E}_e \mathcal{R}^e(f_\theta \circ \phi) + \lambda \min_{\theta} \mathbb{E}_e \|\nabla_\theta \mathcal{R}^e(f_\theta \circ \phi)\|_2 \right\}$$

$$\min_{\phi} \min_{\theta} \{ \mathbb{E}_e \mathcal{R}^e(f_\theta \circ \phi) + \lambda \mathbb{E}_e \|\nabla_\theta \mathcal{R}^e(f_\theta \circ \phi)\|_2 \}$$

Summing up

- We would like to learn **robust predictors** that are based on *causal* associations between variables, rather than *spurious* surface correlations that might be present in our data.
- *Invariance* and *causation* are quite related; we can leverage this connection by promoting **out-of-distribution generalization**.
- Assume that data are sampled from **different environments**.
- **IRM principle**: *find a representation of features, such that the optimal predictor is simultaneously optimal in all environments*.

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0}\, R^e(w\cdot\Phi)\|^2, \qquad \text{(IRMv1)}$$

Thank you!