Invariant Risk Minimization

Authors: Martin Arjovsky, L´eon Bottou, Ishaan Gulrajani, David Lopez-Paz

February 21, 2025

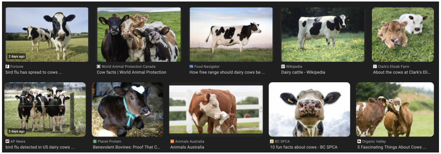
Motivation

ML training is done via minimizing some training loss

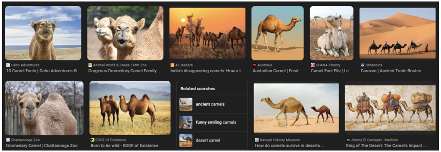
Figure: Task: Classification of cows vs camels

Motivation

The problem



(a) Grassy background



(b) Sandy background

Figure: Training data contains biases

Camel? 

The problem

Correlations-vs-causations Minimizing training error leads machines into recklessly absorbing all the correlations found in training data.

Spurious correlations (landscape, contexts) are unrelated to causal explanations of interest (animal shapes) **Causation** Correlations that are stable (invariant) across training 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.

Outline

1. IRM training objective to learn invariance features across different **training** environments
2. After achieving the desired invariance and a model with low error across training environments, we want to know:
   a. When do these conditions imply invariance across **all** environments
   b. When do these conditions lead to low error across **all** environments (basically, OOD generalization)
   c. Connect invariance and OOD generalization to **theory of causation**

Problem formulation

Datasets $D_e := \{(x_i^e, y_i^e\}_{i=1}^{n_e}$ under multiple environments $e \in \mathcal{E}_{\text{tr}}$

A large set of unseen but related environments $\mathcal{E}_{\text{all}} \supset \mathcal{E}_{\text{tr}}$

Intuitive goal: Learn predictor $Y \approx f(X)$ that performs well across $\mathcal{E}_{\text{all}}$

Denote

$$R^e(f) := \mathbb{E}_{X^e, Y^e}[\ell(f(X^e), Y^e)]$$

is risk under environment e

Problem formulation

Take $\ell =$ MSE or cross-entropy, then the *optimal predictors* can be written as conditional expectations.

Problem formulation

Take $\ell =$ MSE or cross-entropy, then the *optimal predictors* can be written as conditional expectations.
We say a data representation $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ **elicits** an invariant predictor across environment $\mathcal{E}$ if and only if

$$\mathbb{E}[Y^e \mid \Phi(X^e) = h] = \mathbb{E}[Y^{e'} \mid \Phi(X^{e'}) = h]$$

$\forall h \in \cap_{e \in \mathcal{E}} \operatorname{supp}(\Phi(X^e))$

Problem formulation

Take $\ell =$ MSE or cross-entropy, then the *optimal predictors* can be written as conditional expectations.
We say a data representation $\Phi : \mathcal{X} \to \mathcal{H}$ **elicits** an invariant predictor across environment $\mathcal{E}$ if and only if

$$\mathbb{E}[Y^e \mid \Phi(X^e) = h] = \mathbb{E}[Y^{e'} \mid \Phi(X^{e'}) = h]$$

$\forall h \in \cap_{e \in \mathcal{E}} \operatorname{supp}(\Phi(X^e))$

Formal Def Say data representation Φ elicits an invariant predictor $w \circ \Phi$ across $\mathcal{E}$ if there is a classifier $w : \mathcal{H} \to \mathcal{Y}$ simultaneously optimal $\forall e \in \mathcal{E}$:

$$w \in \arg\min_{\bar{w}} R^e(\bar{w} \circ \Phi) \qquad \text{(optimization constraint)}$$

IRM as optimization problem

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

Instantiate IRM into the practical version (derived in the paper):

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

$w = 1$ is a scalar and fixed "dummy" classifier, $\lambda\in[0,\infty)$ is a regularizer balancing between predictive power and the invariance of the predictor $1\cdot\Phi$

Implementing IRMv1

Estimate the objective IRMv1 using mini-batches for stochastic gradient descent (unbiased),

$$\sum_{k=1}^{b} \left[\nabla_{w|w=1.0} \ell(w \cdot \Phi(X_k^{e,i}), Y_k^{e,i}) \cdot \nabla_{w|w=1.0} \ell(w \cdot \Phi(X_k^{e,j}), Y_k^{e,j}) \right],$$

where $(X^{e,i}, Y^{e,i})$ and $(X^{e,j}, Y^{e,j})$ are two random mini-batches of size b from environment e.

Going from IRM to IRMv1

1. Phrasing the constraints as a penalty

$$L_{\text{IRM}}(\Phi, w) = \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(w \circ \Phi) + \lambda \cdot \mathbb{D}(w, \Phi, e) \tag{1}$$

$\mathbb{D}(w, \Phi, e)$ measures how close w is to minimizing $R^e(w \circ \Phi)$, and $\lambda \in [0, \infty)$ is a hyper-parameter balancing predictive power and invariance.

Going from IRM to IRMv1

2. Choosing a penalty $\mathbb{D}$ for linear classifiers w

Consider learning an invariant predictor $w \circ \Phi$, where w is a linear-least squares regression, and Φ is a nonlinear data representation.

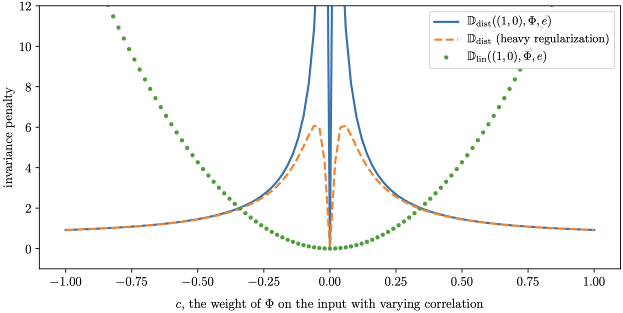


Figure: Different measures of invariance lead to different optimization landscapes. The naïve approach of measuring the distance between optimal classifiers $\mathbb{D}_{\text{dist}}$ leads to a discontinuous penalty (solid blue unregularized, dashed orange regularized). In contrast, the penalty $\mathbb{D}_{\text{lin}}$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

3. Fixing the linear classifier w

We recognize that when optimizing over (Φ, w) using $\mathbb{D}_{\text{lin}}$, a pair $(\gamma\Phi, \frac{1}{\gamma}w)$ can pick $\gamma \approx 0$ to drive $\mathbb{D}_{\text{lin}}$ towards zero without touching the risk term. Similarly, note:

$$w \circ \Phi = \underbrace{\left(w \circ \Psi^{-1}\right)}_{\tilde{w}} \circ \underbrace{\left(\Psi \circ \Phi\right)}_{\tilde{\Phi}}.$$

$\rightarrow$ Can always re-parametrize our invariant predictor w and restrict it to be some non-zero value $\tilde{w}$ of our choosing. This turns (1) into a relaxed version of IRM, where optimization only happens over Φ:

$$L_{\mathrm{IRM}, w=\tilde{w}}(\Phi) = \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(\tilde{w} \circ \Phi) + \lambda \cdot \mathbb{D}_{\text{lin}}(\tilde{w}, \Phi, e). \tag{2}$$

Going from IRM to IRMv1

Scalar fixed classifiers $\tilde{w}$ are sufficient to monitor invariance

Theorem

For all $e \in \mathcal{E}$, let $R^e : \mathbb{R}^d \to \mathcal{R}$ be convex differentiable cost functions. A vector $v \in \mathbb{R}^d$ can be written $v = \Phi^\top w$, where $\Phi \in \mathbb{R}^{p \times d}$, 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 $\mathrm{Ker}(\Phi)$ is orthogonal to v and contains all the $\nabla R^e(v)$.

$\rightarrow$ Any linear invariant predictor can be decomposed as linear data representations of different ranks.
$\rightarrow$ can restrict our search to matrices $\Phi \in \mathbb{R}^{1 \times d}$ and let $\tilde{w} \in \mathbb{R}^1$ be the fixed scalar 1.0. This translates (2) into:

$$L_{\mathrm{IRM},w=1.0}(\Phi^\top) = \sum_{e \in \mathcal{E}_{\mathrm{train}}} R^e(\Phi^\top) + \lambda \cdot \mathbb{D}_{\mathrm{lin}}(1.0, \Phi^\top, e). \tag{3}$$

When does IRM work?

IRM: promotes low error and invariance across **training** environments $\mathcal{E}_{\text{tr}}$

$\overset{?}{\rightarrow}$ Invariance + low error across $\mathcal{E}_{\text{all}}$

Invariance $\overset{?}{\leftrightarrow}$ causality $\overset{?}{\leftrightarrow}$ OOD generalization

When does IRM work?

1. Environments ◦ The data from all the environments share the same underlying Structural Equation Model $\mathcal{C} := (\mathcal{S}, N)$ over the feature and outcome vector $(X_1, \ldots, X_d, Y)$

$$\mathcal{S} : X_i \leftarrow f_i(\mathrm{PA}(X_i), N_i)$$

◦ Then $\mathcal{E}_{\text{all}}(\mathcal{C})$ indexes all the interventional distributions $P(X^e, Y^e)$ obtainable by valid interventions e

When does IRM work?

1. Environments ◦ The data from all the environments share the same underlying Structural Equation Model $\mathcal{C} := (\mathcal{S}, N)$ over the feature and outcome vector $(X_1, \ldots, X_d, Y)$

$$\mathcal{S} : X_i \leftarrow f_i(\text{PA}(X_i), N_i)$$

◦ Then $\mathcal{E}_{\text{all}}(\mathcal{C})$ indexes all the interventional distributions $P(X^e, Y^e)$ obtainable by valid interventions e
◦ Intervention e is valid if they "do not destroy too much information about the target variable Y":
The causal graph remains acyclic,
$\mathbb{E}[Y^e \mid \text{Pa}(Y)] = \mathbb{E}[Y \mid \text{Pa}(Y)]$,
$\mathbb{V}[Y^e \mid \text{Pa}(Y)]$ remains within a finite range.

When does IRM work?

Invariance $\leftrightarrow$ Causation: predictor $v : \mathcal{X} \rightarrow \mathcal{Y}$ is invariant on $\mathcal{E}_{\text{all}} \Leftrightarrow$ attains optimal $R^{\text{OOD}} \Leftrightarrow$ uses only the direct causal parents of Y to predict, $v(x) = \mathbb{E}_{N_Y}[f_Y(Pa(Y), N_Y)]$

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

When does IRM work?

◦ Diversity requirement: limits the extent to which the training environments are co-linear

Assumption

A set of training environments $\mathcal{E}_{tr}$ lie in linear general position of degree r if $|\mathcal{E}_{tr}| > d - r + \frac{d}{r}$ for some $r \in$, and for all non-zero $x \in^d$:

$$\dim\left(\mathrm{span}\left(\left\{_{X^e}\left[X^e X^{e\top}\right]x - _{X^e,\epsilon^e}\left[X^e \epsilon^e\right]\right\}_{e\in\mathcal{E}_{tr}}\right)\right) > d - r.$$

When does IRM work?

2. Invariant Causal Prediction (ICP) theory (Peters, 2015)

Theorem (Invariant Causal Prediction - ICP)

Consider a (linear) Gaussian SEM with interventions. Then given the identifiable causal predictors $S(\mathcal{E})$ under interventions $\mathcal{E}$, all causal predictors are identifiable, that is

$$S(\mathcal{E}) = Pa(Y)$$

if the interventions are do-interventions, noise interventions or simultaneous noise interventions

$\rightarrow$ IRM allows for non-Gaussian data, for linear transformation of the variables with stable and spurious correlations, does not require specific types of interventions or the existence of a causal graph

When does IRM work?

2. Invariant Causal Prediction (ICP) theory (Peters, 2015)
Theorem (roughly stated): If one finds a representation $\Phi \in \mathbb{R}^{d \times d}$ of rank r eliciting an invariant predictor $w \circ \Phi$ across $\mathcal{E}_{\text{tr}}$, and $\mathcal{E}_{\text{tr}}$ satisfying the diversity requirement, then $w \circ \Phi$ is invariant across $\mathcal{E}_{\text{all}}$.

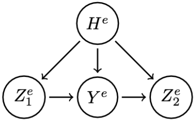
The setting in consideration:
$\circ$ $Y^e = Z_1^e \cdot \gamma + \epsilon^e, \quad Z_1^e \perp \epsilon^e, \quad \mathbb{E}[\epsilon^e] = 0$. Z_1: causal variables, Z_2: non-causal variables
$\circ$ $X^e = S(Z_1^e, Z_2^e)$, Z_1 component of S is invertible

3. OOD generalization (low error) across $\mathcal{E}_{\text{tr}}$ + invariance across $\mathcal{E}_{\text{all}}$ = OOD generalization across $\mathcal{E}_{\text{all}}$

$$\Rightarrow \text{ Invariance } \leftrightarrow \text{ OOD generalization}$$

Experiments results

Synthetic data generation process.



$$
\begin{aligned}
H^e &\leftarrow \mathcal{N}(0, e^2) \\
Z_1^e &\leftarrow \mathcal{N}(0, e^2) + W_{h \to 1} H^e \\
Y^e &\leftarrow Z_1^e \cdot W_{1 \to y} + \mathcal{N}(0, \sigma_y^2) + W_{h \to y} H^e \\
Z_2^e &\leftarrow W_{y \to 2} Y^e + \mathcal{N}(0, \sigma_2^2) + W_{h \to 2} H^e
\end{aligned}
$$

Figure 3: In our synthetic experiments, the task is to predict Y^e from $X^e = S(Z_1^e, Z_2^e)$.

Along with the following variations
◦ *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 $\sigma_2^2 = e^2$.
◦ The 3 training environments are $e \in \{0.2, 2, 5\}$ and we draw 1000 samples from each environment.

Experiments results

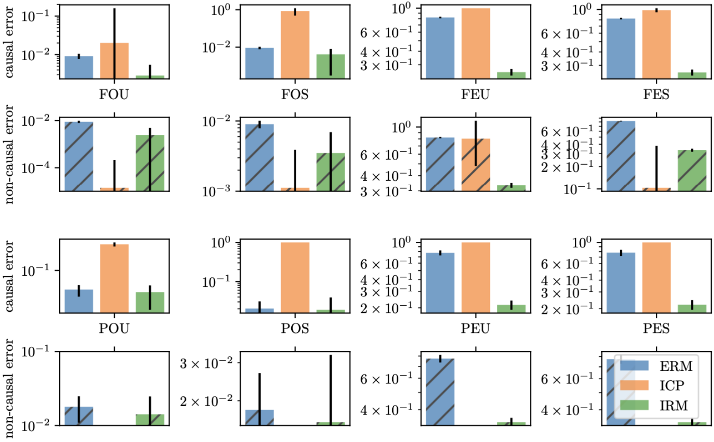


Figure 4: Average errors on causal (plain bars) and non-causal (striped bars) weights for our synthetic experiments. The y-axes are in log-scale. See main text for details.

Experiments results

Color each image in MNIST with either red or green in a way that correlates strongly (but spuriously) with the class label.

Three environments (two training, one test) formed by:

∘ Assign a preliminary binary label $\tilde{y}$ based on the digit: $\tilde{y} = 0$ for digits 0-4 and $\tilde{y} = 1$ for digits 5-9, then flip $\tilde{y}$ with probability 0.25 to get the final label y.

∘ Sample a color ID z by flipping y with probability p_e, which is 0.2 (first environment), 0.1 (second), or 0.9 (test).

∘ Color each image red if $z = 1$ or green if $z = 0$.

Experiments results

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.

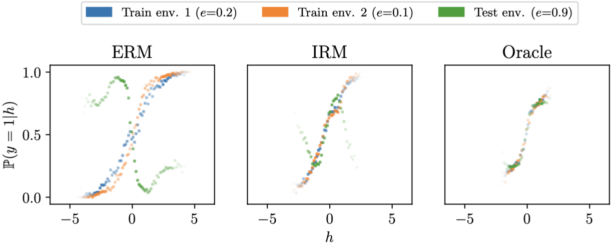


Figure 5: $P(y = 1|h)$ as a function of h for different models trained on Colored MNIST: (left) an ERM-trained model, (center) an IRM-trained model, and (right) an ERM-trained model which only sees grayscale images and therefore is perfectly invariant by construction. IRM learns approximate invariance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