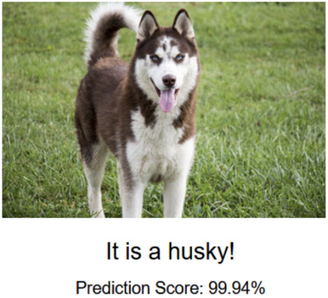
# Causal Optimization: Aligning Prediction and Causal Estimation in Machine Learning

Mingzhang Yin (University of Florida)
Joint work with Yixin Wang (U. Michigan) and David Blei (Columbia)

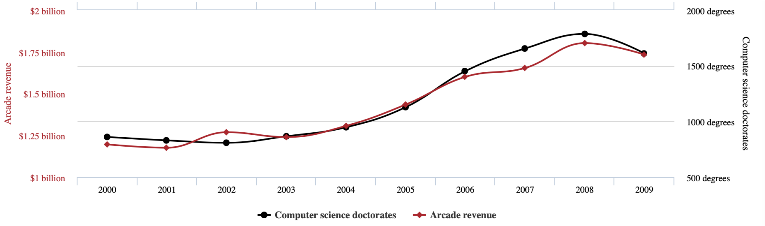
# Prediction and Causal Estimation

- One of the major successes of modern machine learning is their powerful predictive capability.

# Prediction and Causal Estimation

- However, accurate prediction does not guarantee accurate causal estimation.[1]




[1] Efron, B. (2020). Prediction, estimation, and attribution.

# Spurious association problem

- Some elements of the observed covariates $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)$ are predictive to the outcome $y$, but they are not the true causes.

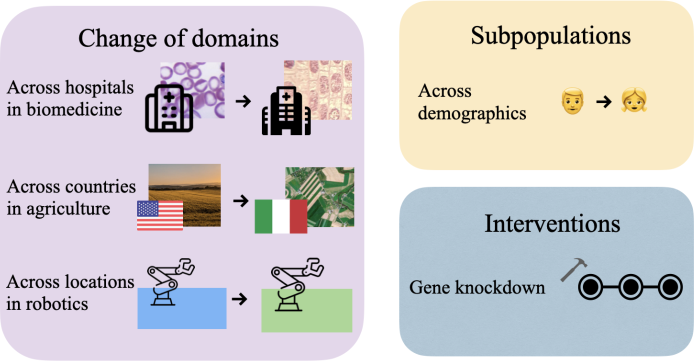
- Classical machine learning often relies on the empirical risk minimization (ERM)

$$\min_{\boldsymbol{\alpha}} \; R(\boldsymbol{\alpha}) = \mathbb{E}[L(\hat{y}(\boldsymbol{x}; \boldsymbol{\alpha}), y)].$$

- ERM leverages causal and non-causal information in $\boldsymbol{x}$.

- A parametric model $\hat{y}(\boldsymbol{x}; \boldsymbol{\alpha})$ learned by ERM
  1. is biased for causal estimation;
  2. cannot generalize its prediction under interventions.

# Environments




- We will leverage multi-environment data to distinguish causality.
- Each environment $e$ has distribution $p^e(X, Y)$.
- Observations per environment are $(X_i^e, Y_i^e) \sim p^e(X, Y),\ e \in \mathcal{E}$.

# Data generating process

Consider a linear structural equation model

$$y^e \leftarrow (\boldsymbol{\beta}^*)^\top \boldsymbol{x}^e + \varepsilon^e, \quad e \in \mathcal{E}$$

- $\mathcal{E}$: a collection of environments.
- $S \subset \{1, 2, \cdots, p\}$: the index set of direct causes.
- $\boldsymbol{x}$: observed covariates; $\boldsymbol{x}_S$ are the causes, $\boldsymbol{x}_{\setminus S}$ are the spurious covariates.
- $\boldsymbol{\beta}^*$: causal coefficients or direct causal effects; $\boldsymbol{\beta}^*_S \neq \mathbf{0}$, $\boldsymbol{\beta}^*_{\setminus S} = \mathbf{0}$.
- Goal: (i) estimate $S$ and $\boldsymbol{\beta}^*$; (ii) make predictions based on causes.

# Formalize spurious association

- Spurious association is an endogeneity problem

$$\boldsymbol{x}^e_{\setminus S} \not\perp \varepsilon^e, \text{ hence } \mathbb{E}[\varepsilon^e | \boldsymbol{x}^e] \neq 0$$

- Possible reasons

1. Unobserved confounding $y \leftarrow \epsilon \rightarrow \boldsymbol{x}_{\setminus S}$
2. Observing descendents $y \rightarrow \boldsymbol{x}_{\setminus S}$
3. Observing colliders $y \rightarrow x_1 \leftarrow x_2,\ x_1, x_2 \in \boldsymbol{x}_{\setminus S}$

# Assumptions

- (i) Linear DGP $y^e \leftarrow (\boldsymbol{\beta}^*)^\top \boldsymbol{x}^e + \varepsilon^e$; it will be relaxed to nonlinear models for methodology
- (ii) Moment conditions: $\mathbb{E}[\epsilon^e] = 0$, $\mathrm{Var}[\epsilon^e], \mathrm{Var}[x_j^e] < \infty$ for all $j \in \{1, 2, \cdots, p\}$
- (iii) Exogeneity of causes: the observed causes

$$\boldsymbol{x}_S^e \perp\!\!\!\perp \epsilon^e,$$

  which is weaker than standard assumption $\boldsymbol{x}^e \perp\!\!\!\perp \epsilon^e$.
- (iv) Invariance: across environments

$$\mathbb{E}[y^e | \mathrm{Pa}(y^e) = \mathbf{c}] = \mathbb{E}[y^{e'} | \mathrm{Pa}(y^{e'}) = \mathbf{c}], \text{ for all } e, e' \in \mathscr{E},$$

  while $p^e(\boldsymbol{x})$ changes.

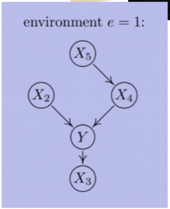
# Invariance of causality

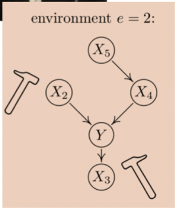
- Philosophy: constant conjunction (Hume, 1740); Econometrics: autonomy and modularity (Haavelmo, 1944, Hoover 2008); Computer Science: independent causal mechanism (Schölkopf, et al., 2021)

- Invariant Causal Prediction (Peters, Bühlmann and Meinshausen, 2016)

- Invariant Risk Minimization (Arjovsky et al., 2019)

A more comprehensive history is in Peters et al. (2017), Chapter 2.[1]


[1] Elements of causal inference: foundations and learning algorithms, 2017.

# Our main idea

1. Find an *idealized* optimization problem with the causal coefficients as *the* solution.
2. Relax it to be a *feasible* optimization problem with the causal coefficients as *a* solution.
3. Restore the identification using multi-environment data.

# Idealized optimization in an environment

- Consider a predictor $\hat{y}(\boldsymbol{x}, \boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \boldsymbol{x}$
- Throughout, $\boldsymbol{\alpha}$ denotes the model parameters and $\boldsymbol{\beta}^*$ denotes the unknown causal parameters.
- Direct ERM $\min_{\boldsymbol{\alpha}}\ R(\boldsymbol{\alpha}) = \mathbb{E}[(1/2)(\hat{y}(\boldsymbol{x}, \boldsymbol{\alpha}) - y)^2]$ produces biased estimate $\hat{\boldsymbol{\alpha}}$ due to spurious association.
- A simple constrained optimization provides causal optimality

$$\begin{aligned} &\min_{\boldsymbol{\alpha}}\ R(\boldsymbol{\alpha}) \\ &\text{s.t. } \alpha_j = 0, \quad j \notin S \text{ (the index set of causes).} \end{aligned}$$

Its solution $\hat{\boldsymbol{\alpha}} = \boldsymbol{\beta}^*$.

# First order condition

- We will turn the constrained optimization into an unconstrained optimization while keeping causal optimality.
- Derive the first order condition of constrained optimization by the directional derivative method.
- Directional derivative in direction $\boldsymbol{v}$ is

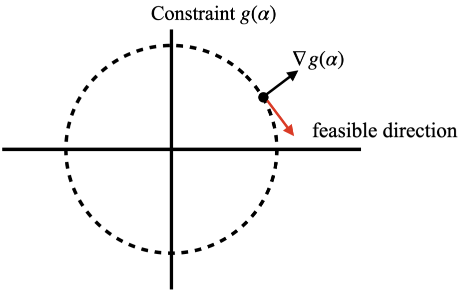
$$\mathbf{D}_v R(\boldsymbol{\alpha}) := \lim_{t\to 0}(R(\boldsymbol{\alpha}+t\boldsymbol{v})-R(\boldsymbol{\alpha}))/t = \langle \nabla R(\boldsymbol{\alpha}), \boldsymbol{v}\rangle$$

- Principle: the first-order condition for optimality is that the directional derivative in all feasible directions vanishes (Marban, 1969).

# Feasible directions

$$\min_{\boldsymbol{\alpha}} R(\boldsymbol{\alpha})$$
$$\text{s.t. } \alpha_j = 0, \quad j \notin S$$

- Feasible directions are where the optimizer can go without violating the constraints. They are tangent to the constraint surface in $\mathbb{R}^p$.
- Our constraints $g_j(\boldsymbol{\alpha}) = \alpha_j = 0$ for $j \notin S$
- The feasible directions form a linear space $\mathscr{U} = \text{span}\{\mathbf{e}_j : j \in S\}$ with basis vector $\mathbf{e}_j$.



# Single environment objective

- Given the feasible directions, the first order condition is

$$\mathbf{D}_{\mathbf{e}_j} R(\boldsymbol{\alpha}) = \langle \nabla R(\boldsymbol{\alpha}), \mathbf{e}_j \rangle = 0, \text{ for } j \in S,$$

  or equivalently written with Hadamard product $\circ$

$$\|\nabla R(\boldsymbol{\alpha}) \circ \boldsymbol{\beta}^*\|_2 = 0$$

- Relaxation: the causal coefficients $\boldsymbol{\beta}^*$ by construction is the optimum, which satisfy the first order condition as

$$\|\nabla R(\boldsymbol{\beta}^*) \circ \boldsymbol{\beta}^*\|_2 = 0.$$

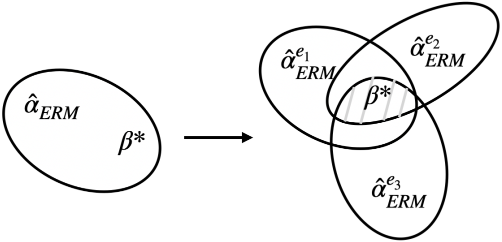
- In other words,

$$\boldsymbol{\beta}^* \in \arg\min_{\boldsymbol{\alpha}} \|\nabla R(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2 . \tag{1}$$

# No free lunch

- The objective $\min_{\boldsymbol{\alpha}} \|\nabla R(\boldsymbol{\alpha}; X, Y) \circ \boldsymbol{\alpha}\|_2$
  - Only depends on the observational data.
  - Unlike $R(\boldsymbol{\alpha}; X, Y)$, it has $\boldsymbol{\beta}^*$ as an optima.
  - It is simple and easy to compute.
- However, the optima is not unique, which can be $\boldsymbol{\beta}^*$, $\hat{\boldsymbol{\alpha}}_{\text{ERM}}$, $\mathbf{0}$, and others.

# Multi-environment objective




- Causal coefficients $\boldsymbol{\beta}^*$ is invariant and shared across environments.
- We aggregate single-environment objectives over multiple environments $\mathcal{E}$

$$\min_{\boldsymbol{\alpha}} f_{\mathcal{E}}(\boldsymbol{\alpha}) := \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left( \|\nabla R^e(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2 \right). \tag{2}$$

- Due to invariance assumption: (1) $\boldsymbol{\beta}^* \in \arg\min f_{\mathcal{E}}(\boldsymbol{\alpha})$, and (2) $\arg\min_{\boldsymbol{\alpha}} f_{\mathcal{E}}(\boldsymbol{\alpha}) = \bigcap_{e \in \mathcal{E}} \arg\min_{\boldsymbol{\alpha}} \|\nabla R^e(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2$ so $|\mathcal{E}| \uparrow$ helps.

# Last step

We need to remove the $\mathbf{0}$-vector from the minimizers if $\boldsymbol{\beta}^* \neq \mathbf{0}$

- If a set of variables C are known to be exogenous, i.e. $X_j \perp\!\!\!\perp \epsilon, j \in C$, we can safely regress over this set of variables (Approach 1).

- Modify the objective with $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \circ (\mathbf{1} - \mathbf{1}_C) + \mathbf{1}_C$,

$$\min_{\boldsymbol{\alpha}} f_{\mathscr{E}}(\boldsymbol{\alpha}) = \frac{1}{|\mathscr{E}|} \sum_{e \in \mathscr{E}} \|\nabla R^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}]\|_2 \tag{3}$$

- We can show $f_{\mathscr{E}}(\boldsymbol{\beta}^*) = 0$ while $f_{\mathscr{E}}(\mathbf{0}) > 0$ almost surely when $\boldsymbol{\beta}^*_C \neq \mathbf{0}$

- Alternatively, we can use the risk function as a regularization as $R^e(\mathbf{0}) \geq R^e(\boldsymbol{\beta}^*)$. It recovers ERM for one environment (Approach 2).

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{|\mathscr{E}|} \sum_{e \in \mathscr{E}} \left\{ \|\nabla R^e(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2 + \lambda_r R^e(\boldsymbol{\alpha}) \right\}, \ \ \lambda_r > 0. \tag{4}$$

# Algorithm

## Conditional Causal Optimization (CoCo)

---

**Algorithm 1** CoCo with known exogenous variables

---

**input :** Data $\mathbf{D}^e = \{\mathbf{Y}^e, \mathbf{X}^e\}$, $\mathbf{X}^e \in \mathbb{R}^{n^e \times p}$; the risk function $R^e$ for each environment $e \in \mathcal{E}$; the set of known non-descendant variables $\mathcal{C}$; the predictor $f(\cdot)$.

**output :** Coefficient estimation $\boldsymbol{\alpha}$ with causal interpretation.

Initialize $\boldsymbol{\alpha}$ randomly

**while** *not converged* **do**

  **for** *$e$ in $\mathcal{E}$* **do**

    Compute the gradient of the empirical risk:

$$\boldsymbol{g}^e(\boldsymbol{\alpha}) = \frac{1}{n_e} \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_{i=1}^{n_e} R^e(\boldsymbol{\alpha}; y_i^e, \hat{y}_i^e), \;\; \hat{y}_i^e = f(\boldsymbol{x}_i^e; \boldsymbol{\alpha})$$

    Set $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \circ (\mathbf{1} - \mathbf{1}_{\mathcal{C}}) + \mathbf{1}_{\mathcal{C}}$

    Compute the optimization objective:

$$\mathcal{L}^e(\boldsymbol{\alpha}) = \|\boldsymbol{g}^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}\|_2$$

  **end**

  Update $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_{e \in \mathcal{E}} \mathcal{L}^e(\boldsymbol{\alpha})$ with step size $\eta$

**end**

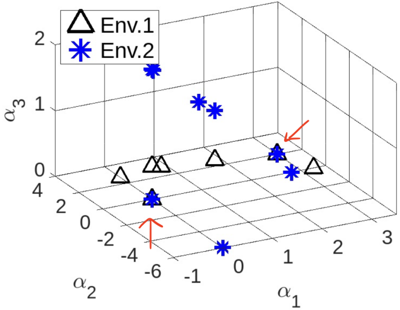
---

# Example

- The data generation follows

$$x_2^e \leftarrow \mathcal{N}(m_2^e, (\gamma^e)^2)$$
$$x_1^e \leftarrow \mathcal{N}(m_1^e, (\gamma^e)^2)$$
$$y^e \leftarrow 3x_1^e + 2x_2^e + \mathcal{N}(0,1)$$
$$x_3^e \leftarrow \gamma^e y^e + \mathcal{N}(0, (\gamma^e)^2)$$

- The two environments correspond to parameters $(m_1^{(1)}, m_2^{(1)}, \gamma^{(1)}) = (2, 0.5, 2)$, $(m_1^{(2)}, m_2^{(2)}, \gamma^{(2)}) = (3, -1, 0.5)$, and $\boldsymbol{\beta}^* = (3, 2, 0)$.




CoCo optima (Two envs.)

# Analytic connections with IRM

- Invariant Risk Minimization is a popular approach for causal representation learning under spurious association (Arjovsky et al., 2019) by solving

$$\min_{\boldsymbol{\alpha}} \sum_{e \in \mathcal{E}} \Big[ \underbrace{R^e(\boldsymbol{\alpha}; f(\boldsymbol{x}_i^e; \boldsymbol{\alpha}))}_{\text{Empirical risk}} + \lambda \underbrace{\left( \nabla_{w|w=1.0} R^e(\boldsymbol{\alpha}; w \cdot f(\boldsymbol{x}_i^e; \boldsymbol{\alpha})) \right)^2}_{\text{IRM regularization}} \Big].$$

- For Linear-Gaussian and Linear-Bernoulli outcome models, IRM regularization is a directional derivative

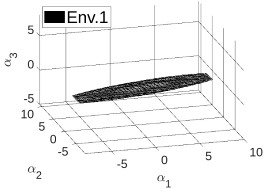
$$\left( \nabla_{w|w=1.0} R^e(\boldsymbol{\alpha}; w \boldsymbol{\alpha}^\top \boldsymbol{x}^e) \right)^2 = (\langle \nabla R^e(\boldsymbol{\alpha}), \boldsymbol{\alpha} \rangle)^2$$

- It explains the success of IRM because $\boldsymbol{\beta}^* \in \arg\min_{\boldsymbol{\alpha}} (\langle \nabla R^e(\boldsymbol{\alpha}), \boldsymbol{\alpha} \rangle)^2$

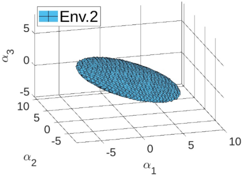
- It suggests IRM regularization could be a loose lower bound because $(\langle \nabla R(\boldsymbol{\alpha}), \boldsymbol{\alpha} \rangle)^2 \leq p \, \|\nabla R(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2^2$

# Geometric connections with IRM

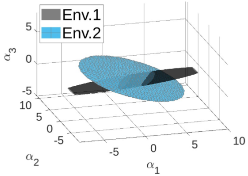
Back to the toy example, CoCo solutions are always less than that by IRM regularization



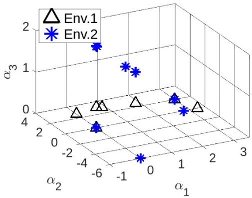

(a) IRM optima (Env.#1)



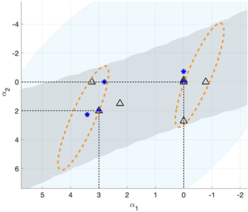

(b) IRM optima (Env.#2)




(c) IRM optima (Two envs.)




(d) CoCo optima (Two envs.)




(e) IRM & CoCo optima (Two envs.)

# Identification

- The goal is to find sufficient conditions for the uniqueness of the solutions for $\min_{\boldsymbol{\alpha}} f_{\mathcal{E}}(\boldsymbol{\alpha}) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|\nabla R^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}]\|_2$

- For each $\hat{\boldsymbol{\alpha}} \in \arg\min_{\boldsymbol{\alpha}} f_{\mathcal{E}}(\boldsymbol{\alpha})$, there exists $H \subset \{1, 2, \cdots, p\}$ such that $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_H, \hat{\boldsymbol{\alpha}}_{\setminus H} = \mathbf{0})^\top$ and

$$\nabla \mathbb{E}[(y - \hat{\boldsymbol{\alpha}}_H^\top \boldsymbol{x}_H^e)^2] = 0.$$

- We call $H$ an invariant set if regression on $\boldsymbol{x}_H^e, \boldsymbol{x}_H^{e'}$ for any environments $e, e'$ produces the same $\hat{\boldsymbol{\alpha}}_H^e = \hat{\boldsymbol{\alpha}}_H^{e'}$.

# Sufficient condition for identification

**Theorem.** Under Assumptions (i-iv) and (v) Effective interventions: there is only one invariant sets $H$, $C \subset H \subset \{1, 2, \cdots, p\}$. Then

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\alpha}} \frac{1}{|\mathscr{E}|} \sum_{e \in \mathscr{E}} \|\nabla R^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}]\|_2\,,$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \circ (\mathbf{1} - \mathbf{1}_C) + \mathbf{1}_C$.

- The effectiveness can be checked from data, though it can be computationally expensive.
- It guarantees the identification of the whole vector $\boldsymbol{\beta}^*$.
- We also provide a simple to check sufficient condition based on the rank of Gram matrix. It guarantees identification of $\boldsymbol{\beta}^*_C$ for the effects of exogenous treatment variables in C.

# Generalize to nonlinear models

- Consider the nonlinear data generation and predictor:

$$y^e \leftarrow f(\mathbf{B}^* x_S^e; \boldsymbol{\gamma}^*) + \varepsilon^e, \quad \hat{y}^e = f(\mathbf{A} x^e; \boldsymbol{\gamma}).$$

- The optimality of the causal model still holds for the constrained optimization: $\min_{\boldsymbol{\alpha}} \; R(\boldsymbol{\alpha})$ s.t. $\alpha_j = 0, j \notin S$

- The same objective can be derived using the directional derivative similarly to the linear settings.

- This nonlinear model contains the fully-connected neural net as a special case.

# Robust prediction

- The fitted model has local optimality when applied to a new environment.
- **Proposition.** Suppose $\boldsymbol{\alpha}'$ minimizes CoCo objective with $f_{\mathcal{E}}(\boldsymbol{\alpha}') = 0$. Suppose a new environment $l$ satisfies

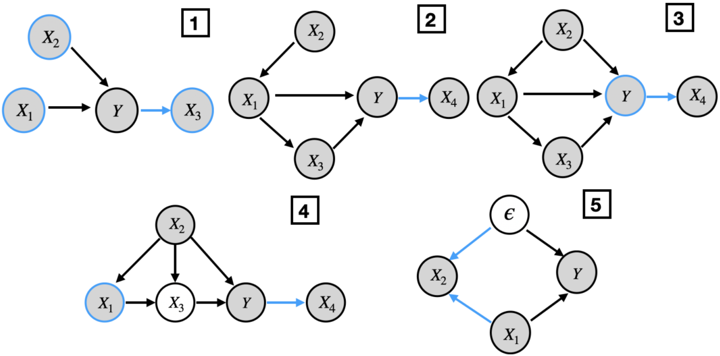
$$p^l(x,y) = \sum_{e\in\mathcal{E}} w_e p^e(x,y), \ \sum_{e\in\mathcal{E}} w_e = 1,$$

then $\frac{\partial}{\partial \alpha_\pi} R^l(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}'} = 0$, $\pi = \mathrm{supp}(\boldsymbol{\alpha}')$.

# Empirical studies

# Causal estimation

- Consider 5 *independent* cases; each case is represented by a graph below
- Data in each case are collected from two environments
- Suppose $X_1$ is known as an exogenous variable



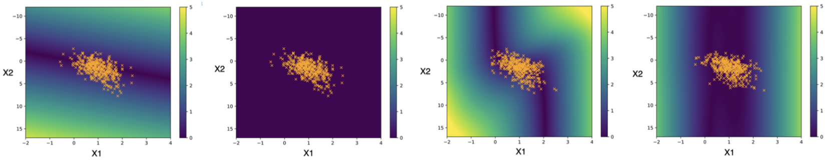
# Causal estimation

The mean absolute error of the $\boldsymbol{\beta}^*$ estimates

| Case | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| ERM | 0.31 (0.06) | 0.16 (0.00) | 0.32 (0.00) | 0.19 (0.03) | 0.38 (0.01) |
| V-REx | 0.16 (0.06) | 0.11 (0.01) | 0.44 (0.01) | 0.13 (0.04) | 0.06 (0.10) |
| RVP | 0.10 (0.04) | 0.10 (0.01) | 0.43 (0.01) | 0.11 (0.04) | 0.05 (0.04) |
| Dantzig | 0.54 (0.62) | 3.23 (2.64) | 4.95 (3.06) | 0.43 (0.05) | 0.20 (0.01) |
| IRMv1 | 2.12 (0.70) | 0.01 (0.00) | 0.02 (0.01) | 2.17 (0.65) | 0.72 (0.35) |
| CoCo | 0.01 (0.00) | 0.02 (0.01) | 0.01 (0.01) | 0.01 (0.01) | 0.01 (0.00) |

RVP, V-REx, Dantzig, IRM are related optimization methods.

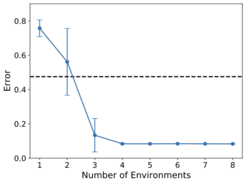
# Robust prediction: synthetic data



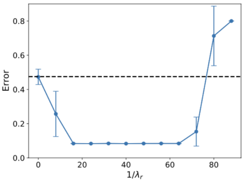

(a) ERM, linear (b) CoCo, linear (c) ERM, nonlinear (d) CoCo, nonlinear

- $x_1$ is a true cause, $x_2$ is spurious, the DGP is linear, the yellow points are data.
- Consider a linear predictor (correctly specified) and a nonlinear predictor (misspecified).
- Heatmap is the predictive error. Causal optimization better generalizes beyond the data region.

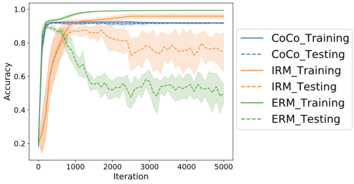
# A nonlinear, non-Gaussian case




# of environments




Regularization strength




- Data generation:

$$
\begin{aligned}
\boldsymbol{x}_1^e &\leftarrow \textstyle\sum_{k=1}^{K} \frac{1}{K} \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{I}) \\
y^e &\leftarrow \text{Categorical}(p_1, \cdots, p_K) \\
\boldsymbol{x}_2^e &\leftarrow (1-p^e)\delta_{\boldsymbol{u}^e_{y^e}} + p^e \delta_{\boldsymbol{u}^e_{k_1}},
\end{aligned}
$$

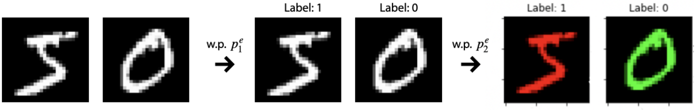
$$p_k = \mathcal{N}(\boldsymbol{x}_1^e; \boldsymbol{\mu}_k, \mathbf{I}) / \textstyle\sum_{k'=1}^{K} \mathcal{N}(\boldsymbol{x}_1^e; \boldsymbol{\mu}_{k'}, \mathbf{I}),\ k_1 \sim \text{Multinomial}(1/K, \cdots, 1/K).$$

- Test in a new environment with distribution shift.

# Robust prediction: semi-synthetic data

Colored-MNIST:

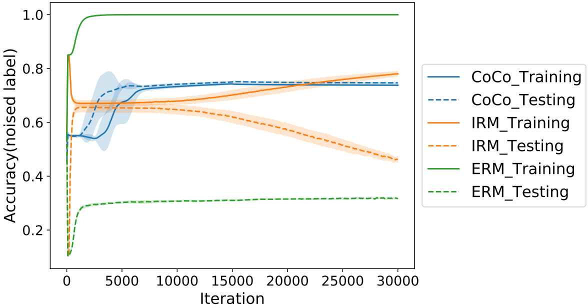
- Data generation: Even/odd digits $\to y_i^e \in \{0, 1\} \to \text{color} \in \{\text{green}, \text{red}\}$.




- Covariates are the colored digits $\boldsymbol{x}_i^e \in \mathbb{R}^{28\times 28\times 2}$
- Causal: shape$\to y_i^e$, Spurious: color$\to y_i^e$.
- Evaluate at a *new* environment with different label-color relationships.

# Predictive accuracy

Predictor is a fully connected neural network.




| Methods | ERM | IRM | V-REx | CoCo | Random guess | Oracle |
|---|---|---|---|---|---|---|
| Test env. accuracy | 31.1 (0.3) | 46.5 (4.1) | 31.8 (1.4) | **74.7** (0.2) | 50 | 74.8 |

IRM (M. Arjovsky et al., 2019), V-REx (D. Krueger et al., 2020)

# Robust prediction: real-world data

- Environments: camera locations.
- Classify coyotes or raccoons, $y_i^e \in \{0, 1\}$.

- Causal: animal shape $\to y_i^e$,  Spurious: physical factors $\to y_i^e$.
- Evaluate on the images taken at a *new* camera location.

# Prediction accuracy

Predictive accuracy is evaluated with images from a new camera location.

| | Wildlife | |
|---|---|---|
| | Training Environment | Testing Environment |
| ERM | 99.6 (0.2) | 58.4 (0.8) |
| IRM | 83.4 (0.7) | **84.9** (0.8) |
| V-REx | 96.2 (0.4) | 67.3 (1.6) |
| CoCo | 86.1 (0.3) | **85.2** (0.3) |
| Random guess | 50 | 50 |

# Takeaway

- Causal optimization enables accurate causal estimation and robust prediction when there is spurious association.
- Multiple environments and the invariance assumption help identify the causal model.
- It can potentially be applied to any differentiable model.
- Worth considering regularizations on the direction of derivatives, beyond the magnitude of parameters.
- Representation learning?

- Thank you!

- M. Yin, Y. Wang, and D.M. Blei
  Optimization-based Causal Estimation from Heterogeneous Environments
  *Journal of Machine Learning Research, 2024*

  m.yin@ufl.edu