The Risks of Invariant Risk Minimization

Elan Rosenfeld Pradeep Ravikumar Andrej Risteski

Machine Learning Department, Carnegie Mellon University

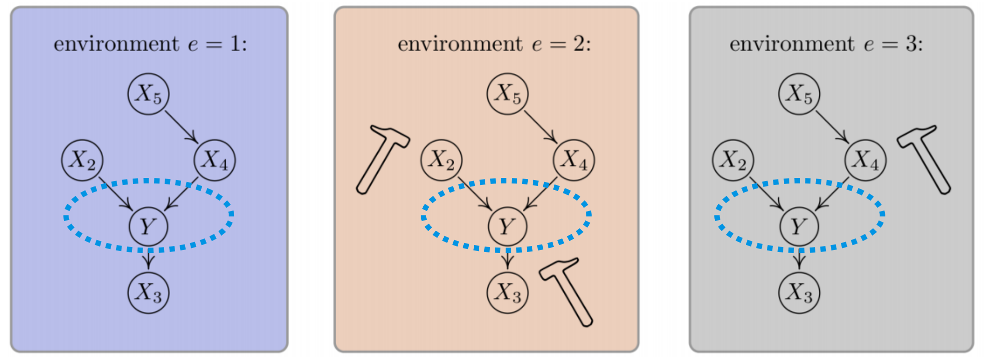
MACHINE LEARNING DEPARTMENT

You can skip this section if you're already familiar with IRM!

Invariant Causal Prediction

Q: How can we train a predictor to ignore **non-causal features** whose correlation with the target may not hold at test time?

Assume training data can be partitioned into distinct **environments**. Each environment represents an **intervention**, inducing a different joint distribution.



Across environments, the **causal mechanism** $P(Y \mid Parents(Y))$ remains fixed. Recovering precisely the parents of Y ensures that predictions are **invariant** and therefore **minimax** across all possible interventions.

Deep Invariant Feature Learning

Q: How can we accomplish this when the features are **latent**?

Learn a feature embedder Φ such that the resulting feature distribution $\Phi(X)$ induces **invariance**. Some recent suggestions:

(**IRM**): Invariant $\mathbb{E}[Y \mid \Phi(X)]$
Require optimal regression vector β on top of features to be invariant.

$$\min_{\Phi,\beta} \sum_{e\in E} R^e(\beta \circ \Phi) \quad s.t.\ \beta \in \operatorname*{argmin}_{\hat\beta} R^e(\hat\beta \circ \Phi), \forall e \in E$$

(**REx**): Invariant $\mathbb{V}[Y \mid \Phi(X)]$
Require equal risk across environments.

$$\min_{\Phi,\beta} \sum_{e\in E} R^e(\beta \circ \Phi) + \lambda \mathrm{Var}[R^e(\beta \circ \Phi)]$$

We prove that IRM and all proposed variants can rarely, if ever be expected to recover the correct invariant features. Thus they all fail under distribution shift just like ERM.

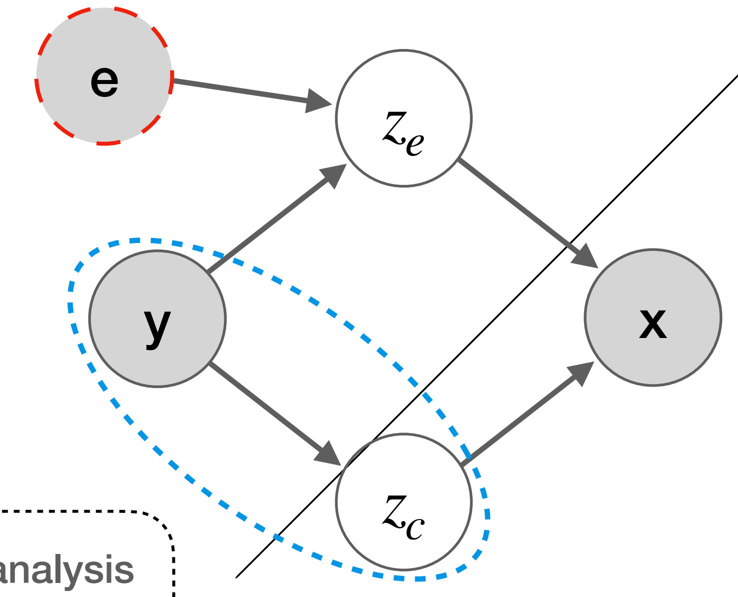
A Formal Model of Latent Invariant Features

$$y = \begin{cases} +1, & \text{w.p. } \eta, \\ -1 & \text{otherwise} \end{cases}$$

$$z = \begin{bmatrix} z_c \\ z_e \end{bmatrix} \begin{cases} z_c \sim \mathcal{N}(\mu_c(y), \sigma_c^2 I) \\ z_e \sim \mathcal{N}(\mu_e(y), \sigma_e^2 I) \end{cases}$$

$$x = f(z)$$

We present the *first* analysis under **nonlinear observations** of the latent variables



$p(y, z_c)$ is **invariant**, capturing the intuition behind invariant feature learning.

Under this model, our goal is to learn a feature embedder Φ^* which recovers just the **invariant features**: $\Phi^*(x) = z_c$. We would then also learn the regression vector $\beta^* = \operatorname*{argmin}_{\beta} R(\beta \circ \Phi^*)$.

We call the predictor $\beta^* \circ \Phi^*$ the **optimal invariant predictor (OIP)**.
We assume we observe E environments, with infinite samples.
Let d_e be the dimensionality of z_e. Typically expect $d_e \gg E$.

Linear Observations

For conditionally Gaussian features, optimal classifier is $\beta^* = \Sigma^{-1}(\mu_1 - \mu_0)$. So long as this vector is the same for all environments, the solution is feasible under the IRM objective.

If $\mathbf{E} \le \mathbf{d_e}$: We construct a **feasible linear** Φ which recovers z_c plus an additional set of features which depend on the non-invariant latents z_e. Provably has **lower training risk** than the OIP.

If $\mathbf{E} > \mathbf{d_e}$: We prove that any feasible linear Φ **can *only* depend on** z_c. Among such Φ, the OIP has lowest training risk.

Corollary: the optimal invariant predictor is the global minimum of the IRM objective if and only if $\mathbf{E} > \mathbf{d_e}$.

Nonlinear Observations

We study IRMv1, a regularized form of IRM used in practice which penalizes the environmental gradient norm of the classifier.

We construct a predictor $\beta \circ \Phi$ with the following properties:

- Penalty term is **exponentially small in dimension** d_e.
- *Exactly equivalent to the OIP* on all but an **exponentially small** fraction of the training distribution.
  — polynomial # of samples $\Longrightarrow$ **indistinguishable**!
- On any environment **slightly different** from the training environments, it is *exactly equivalent to the ERM-optimal solution* on all but an **exponentially small** fraction of the test data.

Implication: the solution behaves *just like ERM* at test time! Furthermore, results apply to *all recently proposed variants*.