CMU - 10732

Week-4

Domain Generalization

- Sameer Jain
- Pranav Mani

Causal Inference using Invariant Prediction Identification and Confidence intervals

Causal Model → Invariant conditional distributions

Goal: Invariant conditional distributions → Causal Model

* $\mathcal{E}$: set of environments. $e \in \mathcal{E}$ specifies a distribution over X, Y. Alternate denote X, Y by X^e, Y^e
* the distributions are valid interventions on the causal structure.
  - ? what's fair game?
  - Mess with anything except SEM(Y)
* $S \subseteq \{1 \dots p\}$, $p = |X|$, $X_S : \{X_i \mid i \in S\}$
* S^*: $\{i \mid X_i \in Pa(Y)\}$
* for allowed interventions: $P(Y \mid Pa(Y))$ is an invariant structure across $\mathcal{E}$

for linear SEM

ASSUMPTION - I :

$\exists\ \gamma^*$ with support S^* s.t $\forall\ e \in \mathcal{E}$

$$Y = X^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon, \quad \varepsilon^e \perp\!\!\!\perp X_{S^*}$$

* By definition assumption-I is satisfied by $Pa(Y)$
* But many S can satisfy assumption (eg) X, Y: jointly Gaussian
* Frame as hypothesis

$$H_{0,\gamma,S}(\mathcal{E}) \triangleq \begin{cases} \gamma_k = 0 \text{ if } k \neq S \\ y^e = x_e \gamma + \varepsilon^e \text{ where } \varepsilon^e \sim F_\varepsilon \ \forall e \in \mathcal{E} \end{cases}$$

* if $\exists\ \gamma \in \mathbb{R}^p$ such that $H_{0,\gamma,S}(\mathcal{E})$ is true

then S: plausible causal predictors

* Define $S(\mathcal{E}) = \bigcap_{S,\ H_{0,S} \text{ is true}} S$

$S(\mathcal{E}) \triangleq$ Identifiable causal predictors

Easy to notice:

① since $H_{0,S^*,\gamma}(\mathcal{E})$ is true.

$$\bigcap_{S, H_{0,\gamma,S}} S \subseteq S^*$$

$$\therefore \quad S(\mathcal{E}) \subseteq S^*$$

② If $\mathcal{E}_1 \subseteq \mathcal{E}_2$, $S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2) \subseteq S^*$

$\Longrightarrow$ closer to identifying S^*

* $\Gamma_S(\mathcal{E}) \triangleq \{\gamma \in R^p \mid H_{0,\gamma,S} \text{ is true}\}$

* $\Gamma(\mathcal{E}) \triangleq \bigcup_S \Gamma_S(\mathcal{E})$

Easy to notice:

if $\mathcal{E}_1 \subseteq \mathcal{E}_2$, $\Gamma(\mathcal{E}_2) \subseteq \Gamma(\mathcal{E}_1)$

Alternate definition of null hypothesis:

$\exists\, \beta \in R^p$, such that $\forall\, e \in \mathcal{E}$, $\beta = \beta^{pred,e}(S)$

$$\beta^{pred,e}(S) = \underset{\beta,\ \beta_k = 0,\ k \notin S}{\operatorname{argmin}} \left\{ E(Y^e - \beta X^e)^2 \right\}$$

Identifiability Results:

formalizing interventions: A_e: contains indices of variables that we intervene on

$A_1 = \phi$

$$\beta^e_{jik} = \begin{cases} \beta^1_{jik} & \text{if } j \notin A_e \\ 0 & \text{if } j \in A_e \end{cases}$$

$$\varepsilon^e_j = \begin{cases} \varepsilon^1_j & \text{if } j \notin A_e \\ a^e_j & \text{if } j \in A_e \end{cases}$$

→ Do interventions

$$\beta^e_{jk} = \beta^1_{jk} \quad \forall j$$

$$\varepsilon^e_j = \begin{cases} \varepsilon^1_j & \text{if } j \notin A_e \\ A^e_j \varepsilon^1_j & \text{if } j \in A_e \\ (\text{or}) \; C^e_j + \varepsilon^1_j & \text{if } j \in A_e \end{cases}$$

→ Noise intervention

Simultaneous: pool all intervention environments into $e : 2$ + noise intervention as here

Identifiability Result:

relaxed further below. Proofs can be found in Appendix F.

Theorem 2 *Consider a (linear) Gaussian SEM as in (19) and (20) with interventions. Then, with $S(\mathcal{E})$ as in (6), all causal predictors are identifiable, that is*

$$S(\mathcal{E}) = \mathbf{PA}(Y) = \mathbf{PA}(1) \tag{22}$$

if one of the following three assumptions is satisfied:

i) The interventions are **do-interventions** *(Section 4.2.1) with $a_j^e \neq E(X_j^1)$ and there is at least one single intervention on each variable other than Y, that is for each $j \in \{2, \dots, p+1\}$ there is an experiment e with $\mathcal{A}^e = \{j\}$.*

ii) The interventions are **noise interventions** *(Section 4.2.2) with $1 \neq E(A_j^e)^2 < \infty$, and again, there is at least one single intervention on each variable other than Y. If the interventions act additively rather than multiplicatively, we require $EC_j^e \neq 0$ or $0 < \operatorname{Var} C_j^e < \infty$.*

iii) The interventions are **simultaneous noise interventions** *(Section 4.2.3). This result still holds if we allow changing linear coefficients $\beta_{j,k}^{e=2} \neq \beta_{j,k}^{e=1}$ in (21) with (possibly random) coefficients $\beta_{j,k}^{e=2}$.*

The statements remain correct if we replace the null hypothesis (10) with its weaker version (16).

These are examples for sufficient conditions for identifiability but there may be many more. For example, one may also consider random coefficients or changing graph structures (only the parents of Y must remain the same).

Extensions :

Some cases these strong conditions can be relaxed

Analysis of Estimation

Assume we have level-α test available

by defn: $\sup_P P\left(H_{0,s}(\mathcal{E}) \text{ rejected} \mid H_{0,s}(\mathcal{E}) \text{ is true}\right) \leq \alpha$

$\therefore S(\mathcal{E}) \subseteq S^*$ with $P \geq 1-\alpha$

$\gamma^* \in \hat{\Gamma}(\mathcal{E})$ with $P \geq 1-2\alpha$

proof:

$P(S^* \text{ rejected}) \leq \alpha$

$\Rightarrow P(S^* \text{ accepted}) \geq 1-\alpha \Rightarrow P\left(S(\mathcal{E}) \subseteq S^*\right) \geq 1-\alpha$

note: $\hat{\Gamma}(\mathcal{E}) = \bigcup \hat{C}(S)$, S makes null hypothesis true

$\hat{C}(S)$: $1-\alpha$ confidence set for γ or $g^{pred}(S)$

$P(\gamma^* \notin \hat{\Gamma}(\mathcal{E})) = P(S^* \text{ rejected}) + P(\gamma^* \notin \hat{C}(S))$

($P(S^* \text{ rejected}) \leq \alpha$; $P(\gamma^* \notin \hat{C}(S)) \leq \alpha$)

$P\left(\gamma^* \notin \hat{\Gamma}(\mathcal{E})\right) \leq 2\alpha$

$P\left(\gamma^* \in \hat{\Gamma}(\mathcal{E})\right) \geq 1-2\alpha$

Generic method for invariant prediction

1) For each set $S \subseteq \{1, \dots, p\}$, test whether $H_{0,S}(\mathcal{E})$ holds at level α (we will discuss later concrete examples).

2) Set $\hat{S}(\mathcal{E})$ as

$$\hat{S}(\mathcal{E}) := \bigcap_{S: H_{0,S}(\mathcal{E}) \text{ not rejected}} S. \tag{12}$$

3) For the confidence sets, define

$$\hat{\Gamma}(\mathcal{E}) := \bigcup_{S \subseteq \{1,\dots,p\}} \hat{\Gamma}_S(\mathcal{E}), \tag{13}$$

where

$$\hat{\Gamma}_S(\mathcal{E}) := \begin{cases} \emptyset & H_{0,S}(\mathcal{E}) \text{ can be rejected at level } \alpha \\ \hat{C}(S) & \text{otherwise.} \end{cases} \tag{14}$$

Here, $\hat{C}(S)$ is a $(1-\alpha)$-confidence set for the regression vector $\beta^{\text{pred}}(S)$ that is obtained by pooling the data.

Last piece: Coming up with a hypothesis test at level alpha.

Method II: Invariant prediction using fast(er) approximate test on residuals

1) For each $S \subseteq \{1, \dots, p\}$ and $e \in \mathcal{E}$:
   (a) Fit a linear regression model on all data to get an estimate $\hat{\beta}^{\text{pred}}(S)$ of the optimal coefficients using set S of variables for linear prediction in regression. Let $R = Y - X\hat{\beta}^{\text{pred}}(S)$.
   (b) Test the null hypothesis that the mean of R is identical for each set I_e and $e \in \mathcal{E}$, using a two-sample t-test for residuals in I_e against residuals in I_{-e} and combing via Bonferroni correction across all $e \in \mathcal{E}$. Furthermore, test whether the variances of R are identical in I_e and I_{-e}, using an F-test, and combine again via Bonferroni correction for all $e \in \mathcal{E}$. Combine the two p-values of equal variance and equal mean by taking twice the smaller of the two values. If the p-value for the set S is smaller than α, we reject the set S.
2) As in the generic algorithm, using (12).
3) If we do reject a set S we set $\hat{\Gamma}_S(\mathcal{E}) = \emptyset$. Otherwise, we set $\hat{\Gamma}_S(\mathcal{E})$ to be the conventional $(1-\alpha)$-confidence region for $\beta^{\text{pred}}(S)$ when using all data simultaneously. For simplicity, we will use rectangular confidence regions, exactly as in step 3 of Method I.

2 things not covered:

Extension to non-linear models
Extension to include hidden variables
} Basic extension allowed

Model misspecification: what happens if we use linear model on non-linear SEM?

→ If S^* is accepted: then confidence statements hold as satisfies hypothesis

→ If nothing satisfies: $\hat{S}(\varepsilon) = \phi$ max power to reject false hypothesis

→ If S^* does not, but something else: Under mild assump. $\hat{S}(\varepsilon) = AN(\gamma)$ as opposed to $PA(\gamma)$

Invariant Risk Minimization:

Invariant Representations of objects relate to the causal explanation of the object itself.

Canonical Example: Cows on grass

Goal: Enable OOD Generalisation by learning invariant predictors

Approach: Learn a data representation such that it elicits an optimal predictor on top of it that is invariant

Notation Setup:

R: risk, R^e: risk when expectation is considered over $e \in \mathcal{E}$

$\mathcal{E}_{all}$: all environments in a modal realism sense

what's your game? SEM(Y): Can only be varied by varying noise model's variance in a finite range

SEM (anything else): mess arbitrarily

note: ICP requires $P(Y|Pa(Y))$ invariant
IRM requires $E[Y|Pa(Y)]$ invariant

$$R^{OOD}(f) \triangleq \max_{e \in \mathcal{E}_{all}} R^e(f) \longrightarrow \text{Goal: } f = \arg\min_f R^{OOD}(f)$$

$$R^{rob}(f) \triangleq \max_{e \in \mathcal{E}_{tr}} R^e(f)$$

$\mathcal{E}_{tr}$: Set of training environments
[we have access to labeled samples]

Some candidate approaches & their shortcomings

1) ERM : obvious shortcomings discussed & motivated ERM (pooling data is ERM)

2) min $R^{rob}(f)$: can be shown is weighted ERM under some regularity conditions

3) Domain Adaptation strategy : we are okay with $P(X_1), P(X_2), P(Y)$, etc to change.

not justified to enforce $P(X_1, X_2)$ with invariant distribution.

4) ICP : allow noise model to change variance in finite range

IRM :

To learn $f : \mathcal{X} \to \mathcal{Y}$, learn $\Phi : \mathcal{X} \to H$ and $w : H \to \mathcal{Y}$, $f : w \circ \Phi$.

Key idea : w on top of Φ should be simultaneously optimal for all $e \in \mathcal{E}_{all}$

why ? Common risks elicit optimal predictors of the form $E[Y^e / X^e]$

$\therefore$ Invariance of $w \Rightarrow E[Y^e / \Phi(X^e) = x] = E[Y^{e'} / \Phi(X^{e'}) = x]$

formally :

$\Phi : \mathcal{X} \to H$, $w : H \to \mathcal{Y}$

objective : $\min_{\Phi, w} \sum_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi)$. such that $w \in \operatorname{argmin}_{\bar{w} : H \to \mathcal{Y}} R^e(\bar{w} \circ \Phi) \quad \forall e \in \mathcal{E}$

IRM :

IRM in its original form: Complex bi-level optimisation problem

Several relaxations:

a) Make constraint soft

$$\min_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi) + \lambda D(w, \Phi, e)$$

choice of D:

Say $w \circ \Phi$ linear (w is LS Regression on $\Phi(x_e)$)

$$w_\Phi^e = \underset{x^e}{E}\left[\Phi(x^e)\Phi(x^e)^T\right]^{-1} \underset{x^e, y^e}{E}\left[\Phi(x^e) y^e\right]$$

$$D_{dist}(w, \Phi, e) = \|w - w_\Phi^e\|^2$$

poor: as $\Phi \longrightarrow \begin{bmatrix} 1 & 0 \\ 0 & c \end{bmatrix} \cdot$ as $c \to 0$. $\Phi \to \Phi^*$

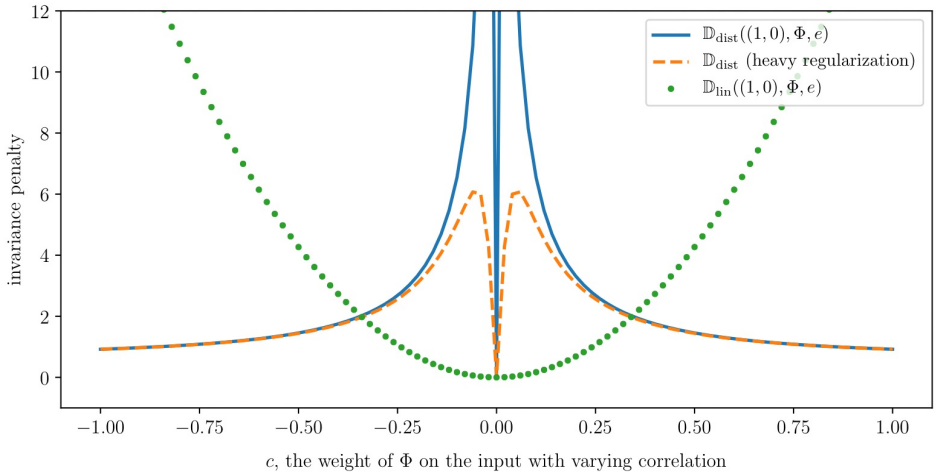
but as $c \to 0$ $D_{dist} \to \infty$

similarly as $c \to \infty$, $D_{dist} \to 0$

Root of this problem: Inverse

fix it: $D_{Lin}(w, \Phi, e)$

$$= \left\| \underset{x^e}{E}\left[\Phi(x^e)\Phi(x^e)^T\right] w - \underset{x^e, y^e}{E}\left[\Phi(x^e) y^e\right] \right\|$$



New Issue: Consider same mapping $w \circ \Phi$

but $w \circ \Phi = w \circ \psi^{-1} \circ \psi \circ \Phi$

$\tilde{w} \circ \tilde{\Phi}$

ERM part is the same

but say $\psi = \gamma$, $\gamma \in (0,1)$

then $D'_{Lin} = \gamma D_{Lin} \to 0$ as $\gamma \to 0$

Final fix: fixing the linear classifier

Rephrase IRM: find a data representation Φ such that optimal linear classifier ontop of that is $\tilde{w}$.

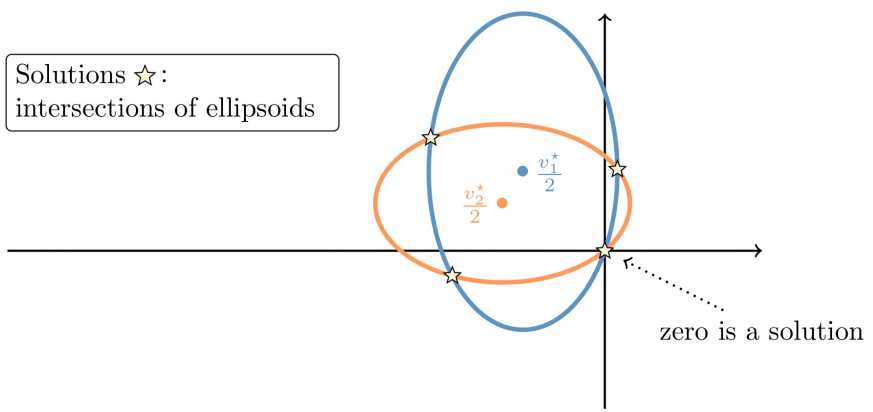
So optimization only over Φ

Why ?

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 $\Phi \in \mathbb{R}^{p \times d}$, 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 $\nabla R^e(v)$.*

So, any linear invariant predictor can be decomposed as linear data representations of different ranks. In particular, we 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 (5) into:

$$L_{\mathrm{IRM},w=1.0}(\Phi^\top) = \sum_{e \in \mathcal{E}_{\mathrm{tr}}} R^e(\Phi^\top) + \lambda \cdot \mathbb{D}_{\mathrm{lin}}(1.0, \Phi^\top, e). \tag{6}$$



Finally :

IRM v1 : $\min_{\Phi} \sum_{e \in \mathcal{E}_{tr}} R^e(\Phi) + \lambda \|\nabla_{w, w=1.0} R^e(w\Phi)\|^2$

Generalization theory for IRM.

When does invariance on $\mathcal{E}_{tr}$ translate to invariance on $\mathcal{E}_{all}$?

Linear General position of degree r:

$|\mathcal{E}_{tr}| > d - r + \frac{d}{r}$ for some $r \in \mathbb{N}$ and

$\forall\, x \neq 0 \in \mathbb{R}^d$

$$\dim\left(\text{span}\left(\left\{ \mathop{E}_{x^e}\left[x^e x^{eT}\right] x - \mathop{E}_{x^e, \epsilon^e}\left[x^e \epsilon^e\right] \right\}\right)\right) > d - r$$

then

if $y^e = z_1^e v + \epsilon^e$, $z_1^e \perp \epsilon^e$, $E[\epsilon^e] = 0$

$x^e = S(z_1, z_2)$, S is invertible in z_1 part

(ii) $\tilde{S}(S(z_1, z_2)) = z_1$

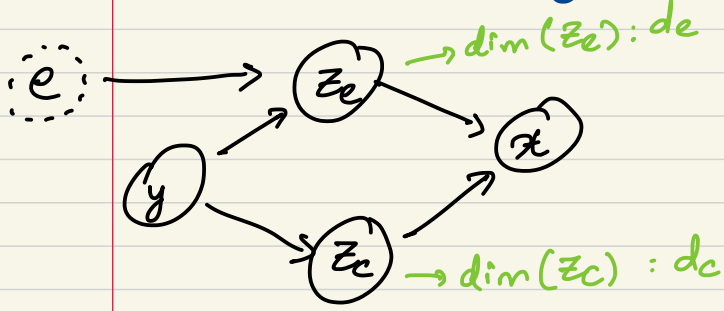
then if $\Phi : R^{d \times d}$, rank $r > 0$.

if $|\mathcal{E}_{tr}| > d - r + \frac{d}{r}$ and lie in L.G.P of degree r, $\Phi \mathop{E}_{x^e}\left[x^e x^{eT}\right] \Phi^T w - \Phi \mathop{E}_{x^e y^e}\left[x^e y^e\right] = 0$ on $\mathcal{E}_{tr}$

if and only if $\Phi^T w$ is invariant on $\mathcal{E}_{all}$

The Risks of Invariant Risk Minimization

Informal results based on the following pedagogical model (which can be relaxed in ways)



$$R^e(\Phi, \hat{\beta}) = \mathop{E}_{(x,y)\sim p^e}\left[\ell(\sigma(\hat{\beta}^T \Phi(x)), y)\right]$$

(I) linear f $\rightarrow$ $|\mathcal{E}| > d_e$: IRM is dope

$\rightarrow$ $|\mathcal{E}| \leq d_e$: IRM solutions $\Phi, \hat{\beta}$ relies upon non-invariant features

(II) linear f : $\exists$ linear predictor $\Phi, \hat{\beta}$ which uses only environmental features yet achieves lower risk than optimal invariant predictor

(III) Non-linear f : $\exists$ non-linear $\Phi, \hat{\beta}$, on train: nearly identical to OIP

on test : Equivalent to IRM

if test correlations are reversed compared to train : this achieves almost 0 accuracy

Very rough pic

IRM $\overset{\text{train}}{\equiv}$ Middle guy $\overset{\text{test}}{\equiv}$ ERM

if test highly opposite to train

IRM gets destroyed!

TABLE-1: A snapshot of results from Gulrajani et al, showing AVERAGE OOD accuracies for IRM and ERM across the seven datasets considered with three model selection criteria.

Model selection method: training domain validation set								
Algorithm	**CMNIST**	**RMNIST**	**VLCS**	**PACS**	**Office-Home**	**TerraInc**	**DomainNet**	**Avg**
ERM	52.0 ± 0.1	98.0 ± 0.0	77.4 ± 0.3	85.7 ± 0.5	67.5 ± 0.5	47.2 ± 0.4	41.2 ± 0.2	67.0
IRM	51.8 ± 0.1	97.9 ± 0.0	78.1 ± 0.0	84.4 ± 1.1	66.6 ± 1.0	47.9 ± 0.7	35.7 ± 1.9	66.0

Model selection method: Leave-one-domain-out cross-validation								
Algorithm	**CMNIST**	**RMNIST**	**VLCS**	**PACS**	**Office-Home**	**TerraInc**	**DomainNet**	**Avg**
ERM	34.2 ± 1.2	98.0 ± 0.0	76.8 ± 1.0	83.3 ± 0.6	67.3 ± 0.3	46.2 ± 0.2	40.8 ± 0.2	63.8
IRM	36.3 ± 0.4	97.7 ± 0.1	77.2 ± 0.3	82.9 ± 0.6	66.7 ± 0.7	44.0 ± 0.7	35.3 ± 1.5	62.9

Model selection method: Test-domain validation set *(oracle)*								
Algorithm	**CMNIST**	**RMNIST**	**VLCS**	**PACS**	**Office-Home**	**TerraInc**	**DomainNet**	**Avg**
ERM	58.5 ± 0.3	98.1 ± 0.1	77.8 ± 0.3	87.1 ± 0.3	67.1 ± 0.5	52.7 ± 0.2	41.6 ± 0.1	68.9
IRM	70.2 ± 0.2	97.9 ± 0.0	77.1 ± 0.2	84.6 ± 0.5	67.2 ± 0.8	50.9 ± 0.4	36.0 ± 1.6	69.2

TABLE-2: Domain generalization accuracies on CMNIST per domain.

Model selection method: training domain validation set			
Algorithm	**0.1**	**0.2**	**0.9**
ERM	72.7 ± 0.2	73.2 ± 0.3	10.0 ± 0.0
IRM	72.0 ± 0.3	73.2 ± 0.0	10.1 ± 0.2

Model selection method: leave-one-domain-out cross-validation			
Algorithm	**0.1**	**0.2**	**0.9**
ERM	46.0 ± 3.4	46.6 ± 3.8	10.0 ± 0.1
IRM	49.3 ± 0.9	49.5 ± 0.2	10.0 ± 0.2

Model selection method: test-domain validation set *(oracle)*			
Algorithm	**0.1**	**0.2**	**0.9**
ERM	72.3 ± 0.6	73.1 ± 0.3	30.0 ± 0.3
IRM	72.7 ± 0.1	72.8 ± 0.3	65.2 ± 0.8

→ The CMNIST Dataset

- Assign a preliminary binary label $\tilde{y}$ to the image based on the digit: $\tilde{y} = 0$ for digits 0-4 and $\tilde{y} = 1$ for 5-9.

- Obtain the final label y by flipping $\tilde{y}$ with probability 0.25
- Sample the colour ID z by flipping y with probability p^e, where $p^e = 0.2$ in the first environment, 0.1 in the second, and 0.9 in the third.
- Colour the image red if $z=1$ or green if $z=0$.

IRM vs ERM : A SAMPLE COMPLEXITY PERSPECTIVE.

→ Setup :

- Dataset $D = \{D^e\}_{e \in \mathcal{E}_{tr}}$, which is a collection of datasets $D^e = \{(x_i^e, y_i^e, e)_{i=1}^{n_e}\}$ obtained from a set of training environments $\mathcal{E}_{tr}$.
- Probability distribution $\{\pi^e\}_{e \in \mathcal{E}_{tr}}$, where π^e is the probability that a training data point is from environment e.
- Define a predictor $f : \mathcal{X} \to \mathbb{R}$ and the space $\mathcal{F}$ of all possible maps from $\mathcal{X} \to \mathbb{R}$
- Define risk achieved by f in environment e as $R^e(f) = \mathbb{E}^e\left[\ell(f(x^e), y^e)\right]$, where ℓ is the loss. Overall expected risk across the training environments is $R(f) = \sum_{e \in \mathcal{E}_{tr}} \pi^e R^e(f)$.

→ OOD generalization Problem :

$$\min_{f \in \mathcal{F}} \max_{e \in \mathcal{E}_{all}} R^e(f). \quad \rightarrow (1)$$

→ Assumption-1 : Invariance Condition.

There exists a representation Φ^* that transforms X^e to $Z^e = \Phi^*(X^e)$ and $\forall e, o \in \mathcal{E}_{all}$, $\forall z \in \Phi^*(\mathcal{X})$ satisfies $\mathbb{E}^e[Y^e | Z^e = z] = \mathbb{E}^o[Y^o | Z^o = z]$. Also, $\forall e \in \mathcal{E}_{all}$, $\forall z \in \Phi^*(\mathcal{X})$, $\text{Var}^e[Y^e | Z^e = z] = \xi^2$, where Var^e is the conditional variance.

→ Using assumption-1 define an invariant map $m : \Phi^*(\mathcal{X}) \to \mathbb{R}$ as :

$$\forall z \in \Phi^*(\mathcal{X}),\ m(z) = \mathbb{E}^e[Y^e | Z^e = z], \quad \rightarrow (2)$$

where $Z^e = \Phi^*(X^e)$.

→ Assumption-2 : There exists an environment $e \in \mathcal{E}_{all}$ s.t.

$$Y^e \perp X^e | Z^e$$

i.e, in environment e, the information that X^e has about Y^e is also contained in Z^e.

→ Define composition $m \circ \Phi^*$, $\forall x \in \mathcal{X}$, $m \circ \Phi^*(x) = \mathbb{E}^e[Y^e | Z^e = \Phi^*(x)]$

→ Proposition-1 : If l is the square loss, and assumptions 1 and 2 hold, then $m \circ \Phi^*$ solves the OOD problem.

→ Recall the IRM objective :

- A data representation Φ elicits an invariant predictor $w \circ \Phi$ across environments $e \in \mathcal{E}_{tr}$ if there is a classifier w that achieves minimum risk simultaneously for all environments.

- IRM selects the invariant predictor with the least sum of risk across all training environments:

$$\min_{\substack{\phi \in \mathcal{H}_\phi \\ \omega \in \mathcal{H}_\omega}} R(\omega \circ \phi) = \sum_{e \in \mathcal{E}_{tr}} \pi^e R^e(\omega \circ \phi) \longrightarrow (3)$$

$$\text{s.t.} \quad \omega \in \underset{\bar{\omega} \in \mathcal{H}_\omega}{\operatorname{argmin}} R^e(\bar{\omega} \circ \Phi), \ \forall e \in \mathcal{E}_{tr}.$$

→ $\omega \circ \Phi^*$ is a feasible solution (3) and is also the ideal solution as it solves (1).

→ Gradient constrained formulation of IRM:

$$\min_{\Phi \in \mathcal{H}_\phi} R(\Phi) \longrightarrow (4)$$

$$\text{s.t.} \ \| \nabla_{\omega | \omega = 1.0} R^e(\omega \cdot \Phi) \| = 0, \ \forall e \in \mathcal{E}_{tr}.$$

→ <u>ϵ- approximation of the IRM objective and its empirical version.</u>

- Define $R'(\Phi) = \sum_{e \in \mathcal{E}_{tr}} \pi^e \| \nabla_{\omega | \omega = 1.0} R^e(\omega \cdot \Phi) \|^2$ and set

$$S^{IV}(\epsilon) = \{ \Phi \mid R'(\Phi) \leq \epsilon, \Phi \in \mathcal{H}_\Phi \}$$

- The ϵ-approximation of (4) is:

$$\min_{\Phi \in S^{IV}(\epsilon)} R(\Phi) \longrightarrow (5)$$

- An empirical version of (5) replaces R and R' with empirical estimators $\hat{R}$ and $\hat{R}'$ respectively:

$$\min_{\Phi \in \hat{S}^{IV}(\epsilon)} \hat{R}(\Phi) \longrightarrow (6)$$

$$\text{where } \hat{S}^{IV}(\epsilon) = \{ \Phi \mid \hat{R}'(\Phi) \leq \epsilon, \Phi \in \mathcal{H}_\Phi \}$$

$\hat{R}$ comes from the sample mean of loss across samples from D.
$\hat{R}'$ is obtained in supplement 7.3.1 (pages 23 and 24).

→ Proposition-2 (highlight):

Sample complexity grows as $O\left(\max\left\{\frac{1}{\epsilon^2}, \frac{1}{\nu^2}\right\}\right)$ to ensure that every solution to empirical IRM (6) is a ν-approximation of IRM with probability at least $1-\delta$

→ Proposition-3 (Shalev-Shwartz and Ben-David, 2014):

Sample complexity grows as $O\left(\frac{1}{\nu^2}\right)$ to ensure that every solution to ERM is a ν-approximation of expected risk minimization with probability at least $1-\delta$.

→ OOD PERFORMANCE (ERM vs IRM) : COVARIATE SHIFT.

→ Assumption-4: Invariance w.r.t all features.

$\forall e, o \in \mathcal{E}_{all}$ and $\forall x \in \mathcal{X}$, $\mathbb{E}\left[Y^e \mid X^e = x\right] = \mathbb{E}\left[Y^o \mid X^o = x\right]$

$\forall e \in \mathcal{E}_{all}$, $X^e \sim \mathbb{P}^e_{Xe}$ and the support of $\mathbb{P}^e_{Xe}$ is equal to $\mathcal{X}$.

→ The first part of the assumption is equivalent to setting $\Phi^* = I$ in assumption-1. If $\Phi^* = I$, then m in equation (2) simplifies to $m(x) = \mathbb{E}\left[Y^e \mid X^e = x\right]$ and solves the OOD problem (1)

→ Proposition-4: Let l be the square loss, and

$\tilde{\alpha} = \min_{\Phi_1, \Phi_2 \in \mathcal{H}_\Phi, \Phi_1 \neq \Phi_2} |R(\Phi_1) - R(\Phi_2)|$ i.e, the minimum separation between the risks of any two distinct hypotheses in $\mathcal{H}_\Phi$.

For every $\nu > 0$, $\epsilon > 0$, and $\delta \in (0,1)$ if $\mathcal{H}_\Phi$ is a finite hypothesis class, $m \in \mathcal{H}_\Phi$, assumptions 3 and 4 hold, then:

- If the number of samples $|D|$ is greater than $\max\left\{\frac{8L^2}{\nu^2}\log\left(\frac{4|\mathcal{H}_\Phi|}{\delta}\right), \frac{16L'^4}{\epsilon^2}\log\frac{2}{\delta}\right\}$, then with probability at least $1-\delta$, every solution $\hat{\Phi}$ to ERM (equation ⑥) satisfies $R(m) \le R(\hat{\Phi}) \le R(m)+\nu$. If $\nu < \tilde{\kappa}$, then $\hat{\Phi} = m$.

- If the number of samples $|D|$ is greater than $\frac{8L^2}{\nu^2}\log\left(\frac{2|\mathcal{H}_\Phi|}{\delta}\right)$ then with probability at least $1-\delta$, every solution $\Phi^\dagger$ to ERM satisfies $R(m) \le R(\Phi^\dagger) \le R(m)+\nu$. If $\nu < \tilde{\kappa}$, then $\Phi^\dagger = m$.

→ OOD GENERALIZATION (ERM VS IRM): CONFOUNDERS OR/AND ANTI-CAUSAL VARIABLES.

→ Recall the linear model from Arjovsky et al. (2019):

$$e \sim \text{Categorical}(\{\pi^o\}_{o\in\mathcal{E}_{tr}}) \quad \forall o \in \mathcal{E}_{tr}, \pi^o > 0$$
$$Y^e \leftarrow \gamma^T(Z_1^e) + \varepsilon^e, \quad \varepsilon^e \perp Z_1^e, \quad \mathbb{E}[\varepsilon^e]=0, \quad \mathbb{E}[(\varepsilon^e)^2]=\sigma^2$$
$$X^e \leftarrow S(Z_1^e, Z_2^e).$$

→ Assumption-5

Z_1 component of S is invertible, i.e. $\exists\, \tilde{S}$ s.t. $\tilde{S}(S(Z_1, Z_2)) = Z_1$, and $\gamma \neq 0$.

Define $\Sigma^e = \mathbb{E}[X^e X^{e,T}]$. s.t. $\forall e \in \mathcal{E}_{tr}$, Σ^e is positive definite.

The support of Z^e and norm of S are bounded.

→ If Z_2^e is an effect of Y^e, then Z_2^e is an anti-causal variable.

→ If H^e causes both ε^e and Z_2^e, then H^e is a confounder.

→ Given the above, the linear model $\tilde{S}^T\gamma$ solves the OOD problem in equation ① ($\Phi^* = \tilde{S}$, $m = \gamma^T$ in proposition-1)

→ Assumption-6 : Linear general position (from IRM)

→ Proposition-5 (highlight):

If the slack on the IRM penalty (ϵ) is sufficiently small, and the data grows as $O(\frac{1}{\epsilon^2})$ then every solution $\hat{\Phi}$ to EIRM is in $\sqrt{\epsilon}$ radius of the OOD solution, i.e.,

$\|\hat{\Phi} - \tilde{S}^T r\| = O(\sqrt{\epsilon})$

Table 1: Summary of (empirical) IRM vs. ERM for finite hypothesis class $\mathcal{H}_\Phi$. ϵ: slack in IRM constraints, ν: approximation w.r.t optimal risk, δ: failure probability, $\mathcal{E}_{tr}$: set of training environments, n: data dimension, p: degree of the generative polynomial, L, L': bound on loss & its gradients.

Assumptions	Method	Sample complexity	OOD
Covariate shift case: $\mathbb{E}^e[Y^e \vert X^e]$ is invariant (Proposition 4)	ERM	$\frac{8L^2}{\nu^2}\log\left(\frac{2\vert\mathcal{H}_\Phi\vert}{\delta}\right)$	Yes
	IRM	$\max\left\{\frac{8L^2}{\nu^2}\log\left(\frac{4\vert\mathcal{H}_\Phi\vert}{\delta}\right), \frac{16L'^4}{\epsilon^2}\log\left(\frac{2}{\delta}\right)\right\}$	Yes
Confounder/Anti-causal variable case: $\mathbb{E}^e[Y^e \vert \Phi^*(X^e)]$ is invariant, Linear, Polynomial models, $\vert\mathcal{E}_{tr}\vert = \mathcal{O}(n^p)$ (Proposition 5, 6,17)	ERM	$\frac{8L^2}{\nu^2}\log\left(\frac{2\vert\mathcal{H}_\Phi\vert}{\delta}\right)$	No
	IRM	$\frac{16L'^4}{\epsilon^2}\log\left(\frac{2\vert\mathcal{H}_\Phi\vert}{\delta}\right)$	Yes