IRM研究背景

□ OOD泛化数学刻画：

Problem 1 (Supervised Learning). Given a set of n training samples of the form $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, which are drawn from training distribution $P_{tr}(X, Y)$, a supervised learning problem is to find an optimal model f_θ^* which can generalize best on data drawn from test distribution $P_{te}(X, Y)$:

$$f_\theta^* = \arg\min_{f_\theta} \mathbb{E}_{X,Y\sim P_{te}}[\ell(f_\theta(X), Y)]$$

$$\text{I.I.D:}\ P_{tr}(X,Y) = P_{te}(X,Y)$$

$$\text{OOD:}\ P_{tr}(X,Y) \neq P_{te}(X,Y)$$

$$\Downarrow$$

$$P_{tr}(Y|X)P_{tr}(X) \neq P_{te}(Y|X)P_{te}(X)$$

传统机器学习不能有效解决OOD泛化

(Cui Peng. Talk on causal learning and decision making, 2022)

IRM研究背景

□ 两类OOD泛化问题：

Shift Type	$\mathbb{P}(\mathbf{X})$	$\mathbb{P}(Y\|\mathbf{X})$	Training		Testing	
Marginal	$\mathbb{P}_{tr}(\mathbf{X}) \neq \mathbb{P}_{te}(\mathbf{X})$	$\mathbb{P}_{tr}(Y\|\mathbf{X}) = \mathbb{P}_{te}(Y\|\mathbf{X})$				
Conditional	$\mathrm{supp}(\mathbb{P}_{tr}(\mathbf{X})) \approx \mathrm{supp}(\mathbb{P}_{te}(\mathbf{X}))$	$\mathbb{P}_{tr}(Y\|\mathbf{X}_s) \neq \mathbb{P}_{te}(Y\|\mathbf{X}_s)$				

Causal Feature(Xc)		
Spurious Feature(Xs)		
Label(Y)	Cow	Camel

IRM学习依赖不变特征的模型，解决条件概率变化的OOD泛化

(Lin Yong, 2022)

不变风险最小化

□ I.I.D：

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w\circ\Phi)$$

经验风险最小化ERM

□ IRM问题定义：

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

不变风险最小化IRM

- **理解：** 增加的约束确保H为因果特征（跨环境不变特征）
- **缺点：** 双层优化问题，无法通过梯度下降求解

将优化目标转为梯度下降可以求解的形式

[Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization[J]. arXiv preprint arXiv:1907.02893, 2019]

From IRM to IRMv1

東南大學 SOUTHEAST UNIVERSITY

□ 第一步：将约束转为惩罚项

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

预测能力（ERM） **不变性**

- λ：$\lambda \in [0, \infty)$，平衡ERM与不变性
- $\mathbb{D}(w, \Phi, e)$：惩罚项衡量距离 w 与 $w^* \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min} R^e(\bar{w} \circ \Phi)$, for all $e \in \mathcal{E}_{\text{tr}}$
- **增加假设：** 考虑 w 为线性分类器

□ 第二步：给线性分类器选择惩罚项

考虑最小二乘回归:

$$Y^e = w \circ \Phi(X^e)$$

显式解为:

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

From IRM to IRMv1

两个分类器之间差距定义为：

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

因为 $w_{\Phi}^{e} = \mathbb{E}_{X^e}\left[\Phi(X^e)\Phi(X^e)^{\top}\right]^{-1}\mathbb{E}_{X^e,Y^e}[\Phi(X^e)Y^e]$，存在求**逆运算**，导致距离函数**不连续**，改为：

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

上式满足：$\mathbb{D}_{\text{lin}}(w, \Phi, e) = 0$ 当且仅当 $w \in \arg\min_{\bar{w}} R^e(\bar{w} \circ \Phi)$

◻ 第三步：固定线性分类器

当考虑 $\left(\gamma\Phi, \frac{1}{\gamma}w\right)$，随着 γ 趋于0，可使ERM项不变，但 $\mathbb{D}_{\text{lin}}(w, \Phi, e) = 0$

考虑对任何可逆映射Ψ，重写不变预测器：

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

From IRM to IRMv1

将所有环境最优分类器固定在 $\tilde{w}$，IRM定义改写为：

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

IRM relaxed version

更一般地，固定 $\tilde{w}=1$

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

考虑最小二乘损失：

$$R^e(w\cdot\Phi) = \frac{1}{2}(w\cdot\Phi(X^e)-Y^e)^T(w\cdot\Phi(X^e)-Y^e)$$

$$\|\nabla_{w|w=1.0}R^e(w\cdot\Phi)\|^2 = \|\Phi(X)\Phi(X)^T w - \Phi(X)Y\|^2$$

$$\mathbb{D}(1.0,\Phi,e) = \left\|\nabla_{w|w=1.0}R^e(w\cdot\Phi)\right\|^2$$

惩罚项最终表达式

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

IRMv1

IRM实验一

□ IRM提出ColoredMNIST数据集：目标是预测数字，二分类任务

Dataset	X_v	X_s	Training	Testing	Bias Ratio: corr(Y, X_s)	corr(Y, X_v)
ColoredMNIST	Digit	Color			(0.9, 0.8, 0.1)	0.75
ColoredObject	Object	Background			(0.999, 0.7, 0.1)	0.95
CIFARMNIST	CIFAR	MNIST			(0.999, 0.7, 0.1)	0.90

- **IRM特征学习：输入$X(n, 14*14)$，学习MLP：Φ，深度学习输出$z(n, 1)$，分类器$w(1,)$，$X\Phi w = Y^{\wedge}$，$Y(n, 1)$**
- **ICP特征选择**

> **IRM特征学习对数据集要求：**
>
> - 明确 X_v 和 X_s
> - 划分环境使 P(Y | Xs) 发生变化
>
> 现实情况很难满足

IRM实验—环境划分

东南大学 SOUTHEAST UNIVERSITY

```python
def make_environment(images, labels, e):
  def torch_bernoulli(p, size):
    return (torch.rand(size) < p).float()
  def torch_xor(a, b):
    return (a-b).abs() # Assumes both inputs are either 0 or 1
  # 2x subsample for computational convenience
  images = images.reshape((-1, 28, 28))[:, ::2, ::2]
  # Assign a binary label based on the digit; flip label with probability 0.25
  labels = (labels < 5).float()
  labels = torch_xor(labels, torch_bernoulli(0.25, len(labels)))
  # Assign a color based on the label; flip the color with probability e
  colors = torch_xor(labels, torch_bernoulli(e, len(labels)))
  # Apply the color to the image by zeroing out the other color channel
  images = torch.stack([images, images], dim=1)
  images[torch.tensor(range(len(images))), (1-colors).long(), :, :] *= 0
  return {
    'images': (images.float() / 255.).cuda(),
    'labels': labels[:, None].cuda()
  }

envs = [
  make_environment(mnist_train[0][::2], mnist_train[1][::2], 0.2),
  make_environment(mnist_train[0][1::2], mnist_train[1][1::2], 0.1),
  make_environment(mnist_val[0], mnist_val[1], 0.9)
]
```

- 将60000张MNIST训练集中前50000张作为训练集，后10000张作为测试集。
- 对于每张28✕28图片：通过子采样得到14✕14图片；若图片中数字0-4则 $\tilde{y}=0$，否则$\tilde{y}=1$，再以**0.25**的概率翻转 $\tilde{y}$ 得到最终标签 y。
- 对标签 y 以概率**e**进行翻转得到颜色标签 z，$z=1$ 图像为红色，否则为绿色。

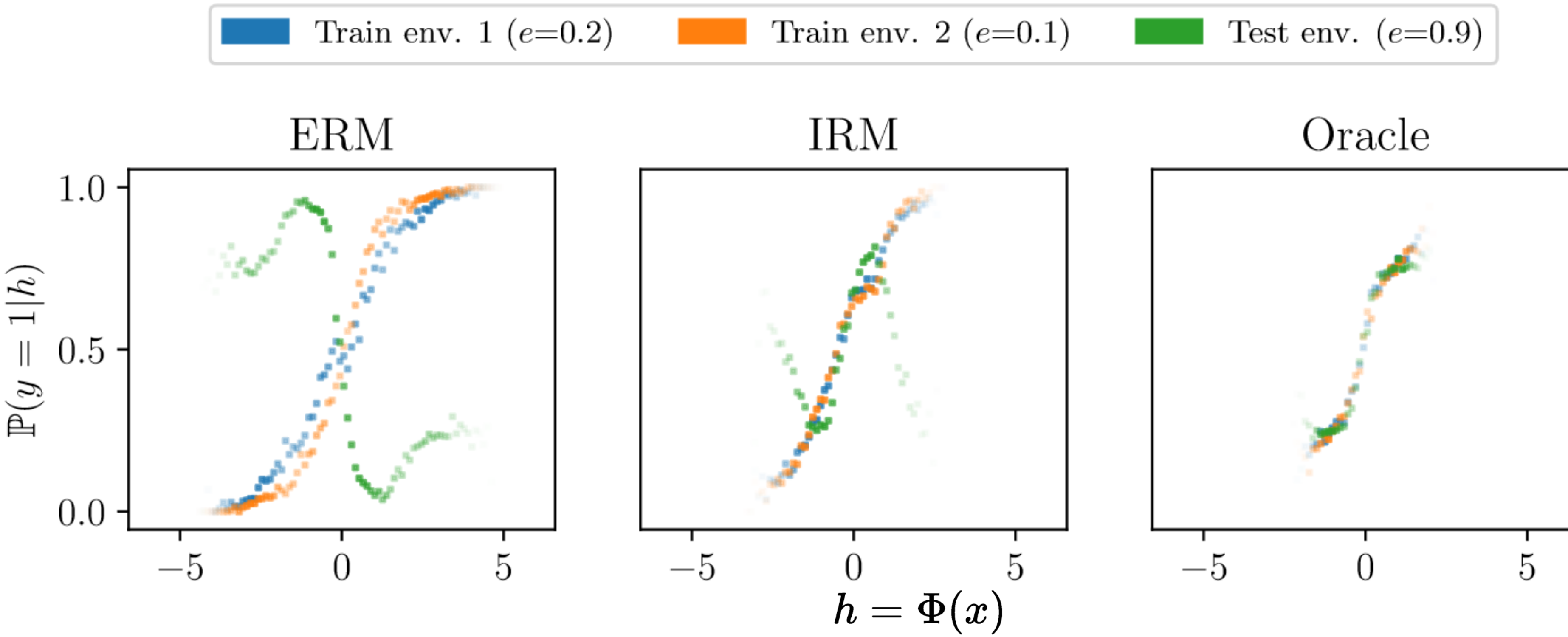
IRM实验—结果

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

□ 实验结论：

- **ERM**主要根据颜色进行分类，训练集精度高，测试集精度低。
- **IRM**训练集表现较差，但对颜色依赖较少，可以更好泛化到测试集。
- 忽略颜色信息的ERM（**oracle**）：训练集和测试集表现略微优于IRM。

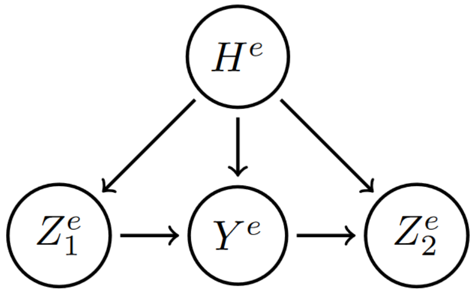
IRM实验—结果



□ 实验结论：

- IRM模型比ERM更易实现不变性。
- IRM模型并未实现完美的不变性，可能是由于**有限样本**问题。

IRM实验二-合成实验



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

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

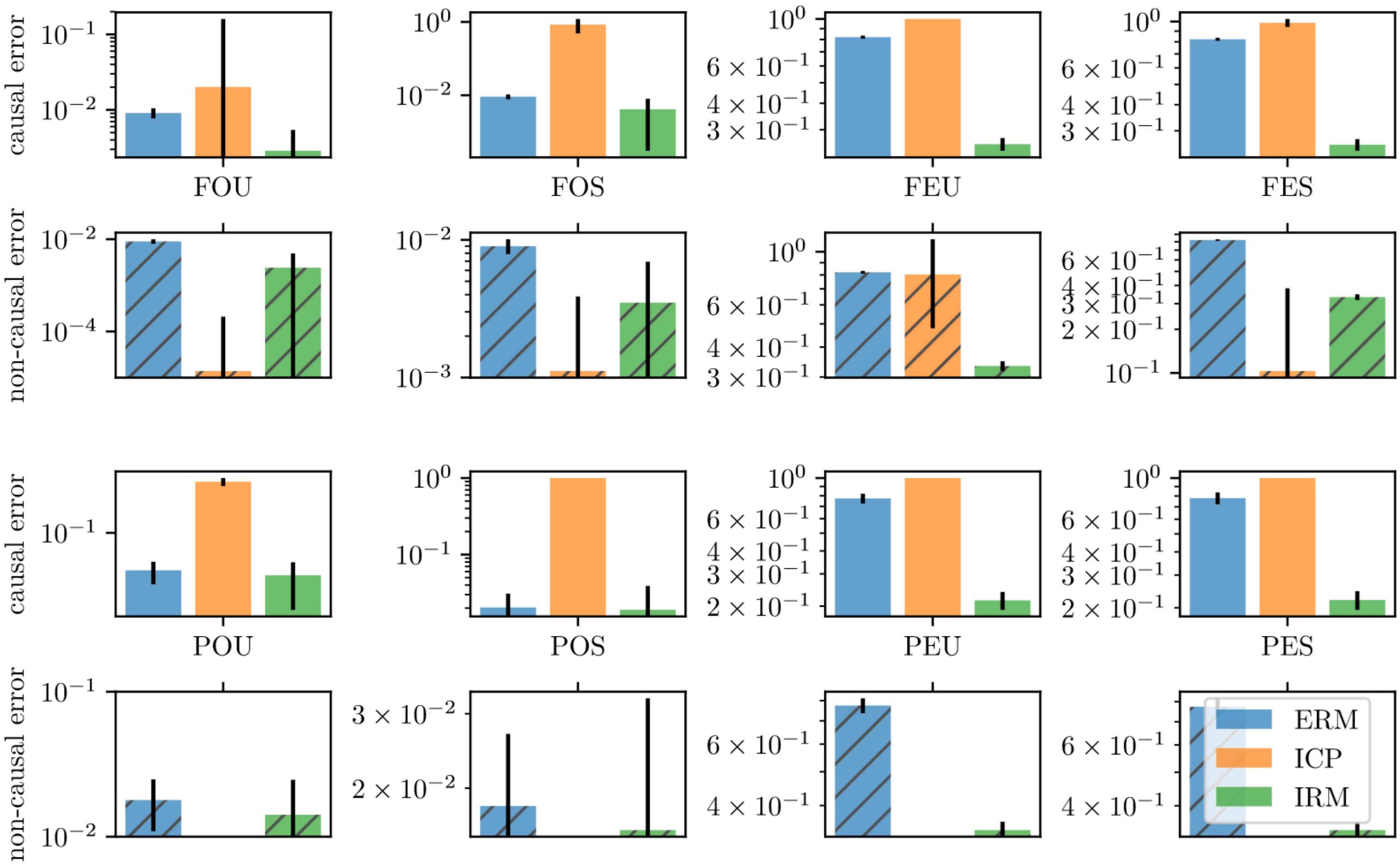
- *Scrambled* (S) observations, where S is an orthogonal matrix, or *unscrambled* (U) observations, where $S = I$.
- *Fully-observed* (F) graphs, where $W_{h \to 1} = W_{h \to y} = W_{h \to 2} = 0$, or *partially-observed* (P) graphs, where $(W_{h \to 1}, W_{h \to y}, W_{h \to 2})$ are Gaussian.
- *Homoskedastic*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

- **IRM：输入$X(n,10)$，学习矩阵：$\Phi(10,10)$，输出$z(n,10)$，分类器$w(10,)$， $X\Phi w = Y^{\wedge}(n,1)$**

IRM将维数设置很低，是为了和ICP作对比

[Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization[J]. arXiv preprint arXiv:1907.02893, 2019]

IRM实验二结果



[Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization[J]. arXiv preprint arXiv:1907.02893, 2019]

IRM与ICP中环境的区别

➢ IRM：允许Y的噪声方差在有限范围内变化

Definition 7. *Consider a SEM $\mathcal{C}$ governing the random vector $(X_1, \ldots, X_d, Y)$, and the learning goal of predicting Y from X. Then,* the set of all environments $\mathcal{E}_{\text{all}}(\mathcal{C})$ *indexes all the interventional distributions $P(X^e, Y^e)$ obtainable by* valid interventions e. *An intervention $e \in \mathcal{E}_{all}(\mathcal{C})$ is valid as long as (i) the causal graph remains acyclic, (ii) $\mathbb{E}[Y^e|\text{Pa}(Y)] = \mathbb{E}[Y|\text{Pa}(Y)]$, and (iii) $\mathbb{V}[Y^e|\text{Pa}(Y)]$ remains within a finite range.*

➢ ICP：Y的SCM不能变

for all $e \in \mathcal{E}$, X^e has an arbitrary distribution and

$$Y^e = g(X^e_{S^*}, \varepsilon^e), \qquad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*},$$

$$\boxed{Y^e \mid X^e_{S^*} \text{ and } Y^f \mid X^f_{S^*} \text{ are identical for all environments } e, f \in \mathcal{E}}$$

[Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization[J]. arXiv preprint arXiv:1907.02893, 2019]