MODELING THE DATA-GENERATING PROCESS IS NECESSARY FOR OUT-OF-DISTRIBUTION GENERALIZATION

Jivat Neet Kaur, Emre Kıcıman, Amit Sharma
Microsoft Research {t-kaurjivat,emrek,amshar}@microsoft.com

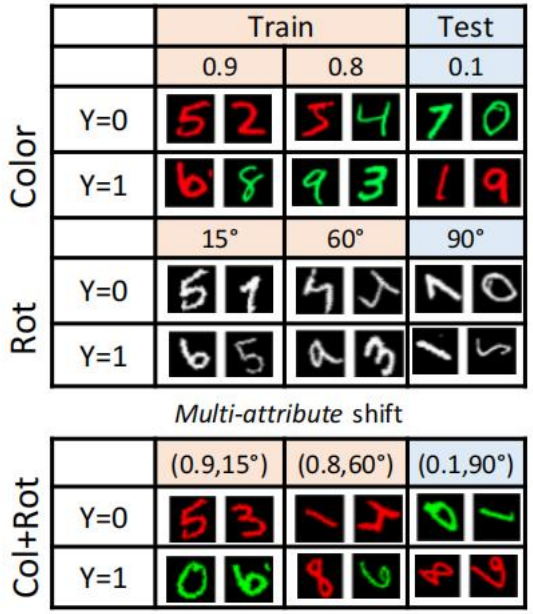
ICLR 2023

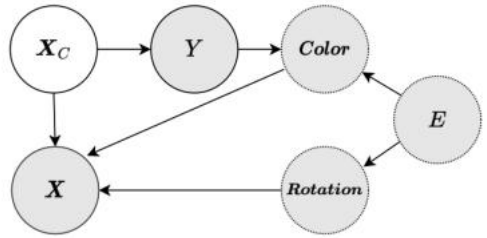
Reporter: 王启迅

2023/3/16

Introduction

Focusing on **multiple** shifts appearing simultaneously in a dataset.



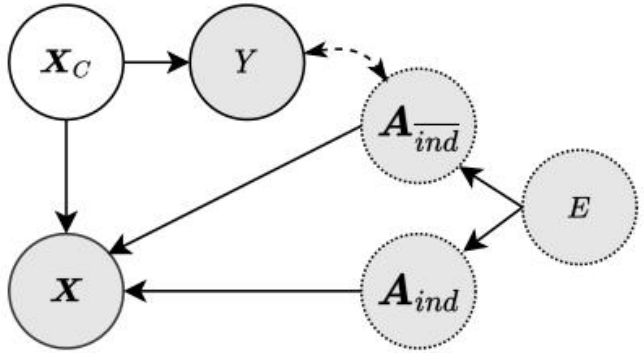


(b)

Algo.	Color	Rotation	Col+Rot
ERM	30.9 ± 1.6	61.9 ± 0.5	25.2 ± 1.3
IRM	50.0 ± 0.1	61.2 ± 0.3	39.6 ± 6.7
MMD	29.7 ± 1.8	62.2 ± 0.5	24.1 ± 0.6
C-MMD	29.4 ± 0.2	62.3 ± 0.4	32.2 ± 7.0
CACM	$\mathbf{70.4 \pm 0.5}$	$\mathbf{62.4 \pm 0.4}$	$\mathbf{54.1 \pm 1.3}$

Main idea to solve multiple shifts

1. Construct a *canonical* causal graph to specify the common data-generating processes. Different shifts lead to different realization of the DAGs (need prior knowledge).
2. Once the DAG is constructed, we can infer the conditional independence constraints on the causal feature X_c using d-separation, then we can add the conresponding regularization of the learned feature Φ(x). If there are multiple shifts, multiple regularizations are needed.



Problem formulation

Two concepts:

Definition 2.1. ***Optimal Risk Invariant Predictor for*** $\mathcal{P}$ ***(from (Makar et al., 2022))*** *Define the risk of predictor* g *on distribution* $P \in \mathcal{P}$ *as* $R_P(g) = \mathbb{E}_{\boldsymbol{x},y\sim P}\ell(g(\boldsymbol{x}), y)$ *where* ℓ *is cross-entropy or another classification loss. Then, the set of risk-invariant predictors obtain the same risk across all distributions* $P \in \mathcal{P}$*, and set of the optimal risk-invariant predictors is defined as the risk-invariant predictors that obtain minimum risk on all distributions.*

$$g_{rinv} \in \arg\min_{g\in G_{rinv}} R_P(g)\ \forall P \in \mathcal{P} \text{ where } G_{rinv} = \{g : R_P(g) = R_{P'}(g) \forall P, P' \in \mathcal{P}\} \quad (1)$$

Definition 2.2. ***Generalization under Multi-attribute shifts.*** *Given a target label* Y*, input features* $\boldsymbol{X}$*, attributes* $\boldsymbol{A}$*, and latent causal features* $\boldsymbol{X}_c$*, consider a set of distributions* $\mathcal{P}$ *such that* $P(Y|\boldsymbol{X}_c)$ *remains invariant while* $P(\boldsymbol{A}|Y)$ *changes across individual distributions. Using a training dataset* $(\boldsymbol{x}_i, \boldsymbol{a}_i, y_i)_{i=1}^n$ *sampled from a subset of distributions* $\mathcal{P}_{\mathcal{E}tr} \subset \mathcal{P}$*, the generalization goal is to learn an optimal risk-invariant predictor over* $\mathcal{P}$*.*

X_c is the unobserved causal feature

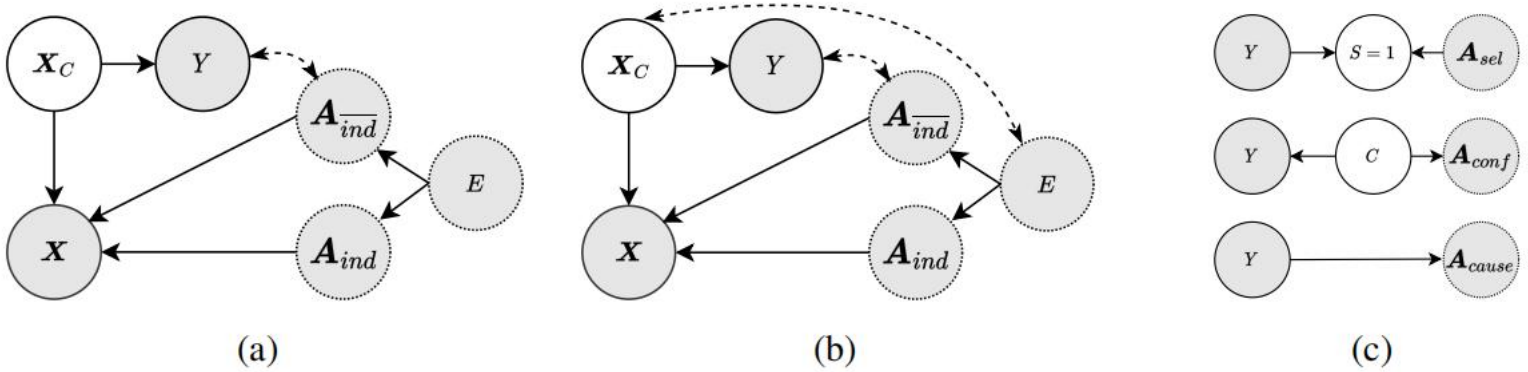
Key challenge:

Learn unobserved X_c using observed $(\boldsymbol{X}, Y, \boldsymbol{A})$

Idependence constraints are necessary

Theorem 2.1. *Consider a causal DAG $\mathcal{G}$ over $\langle \boldsymbol{X}_c, \boldsymbol{X}, \boldsymbol{A}, Y \rangle$ and a corresponding generated dataset $(\boldsymbol{x}_i, \boldsymbol{a}_i, y_i)_{i=1}^n$, where $\boldsymbol{X}_c$ is unobserved. Assume that graph $\mathcal{G}$ has the following property: $\boldsymbol{X}_c$ is defined as the set of all parents of Y ($\boldsymbol{X}_c \to Y$); and $\boldsymbol{X}_c, \boldsymbol{A}$ together cause $\boldsymbol{X}$ ($\boldsymbol{X}_c \to \boldsymbol{X}$, and $\boldsymbol{A} \to \boldsymbol{X}$). The graph may have any other edges (see, e.g., DAG in Figure 1(b)). Let $\mathcal{P}_\mathcal{G}$ be the set of distributions consistent with graph $\mathcal{G}$, obtained by changing $P(\boldsymbol{A}|Y)$ but not $P(Y|\boldsymbol{X}_c)$. Then the conditional independence constraints satisfied by $\boldsymbol{X}_c$ are necessary for a (cross-entropy) risk-invariant predictor over $\mathcal{P}_\mathcal{G}$. That is, if a predictor for Y does not satisfy any of these constraints, then there exists a data distribution $P' \in \mathcal{P}_\mathcal{G}$ such that predictor's risk will be higher than its risk in other distributions.*

The proposed canonical causal graph



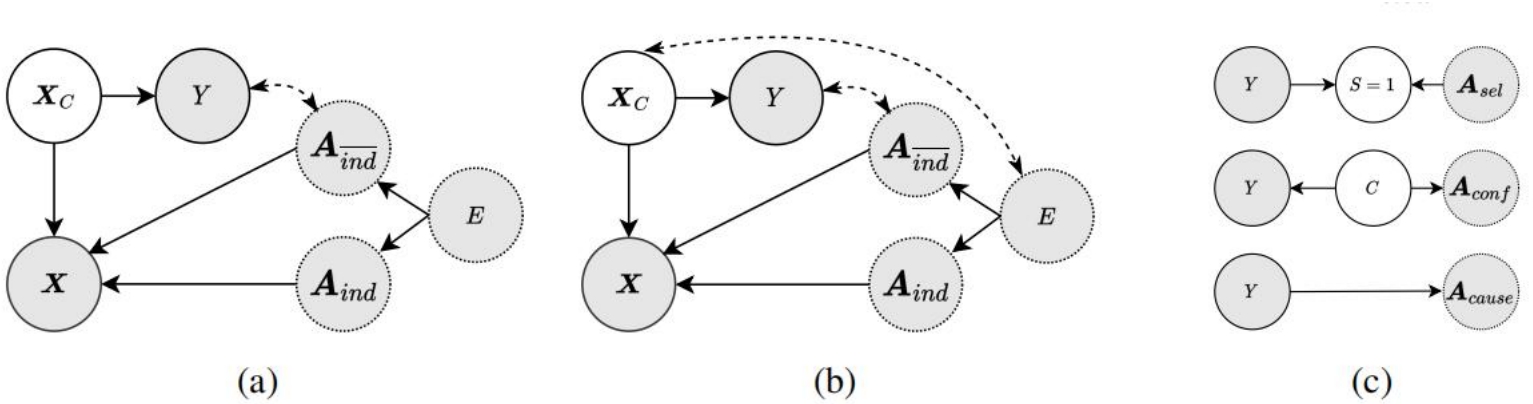
Set of Spurious attributes: $\boldsymbol{A}_{\overline{ind}} \cup \boldsymbol{A}_{\text{ind}} \cup \{E\} = \boldsymbol{A}$ (not all attributes need to be observed)

Assume no label shift ($P(Y)$ is invariant across domains)

(b) Different domains may have different distribution of causal features

(c) Different $Y - \boldsymbol{A}_{\overline{ind}}$ relationship: *Selection during the data-generating process/confounding between* Y *and* $\boldsymbol{A}_{\overline{ind}}$ */ direct-causal*

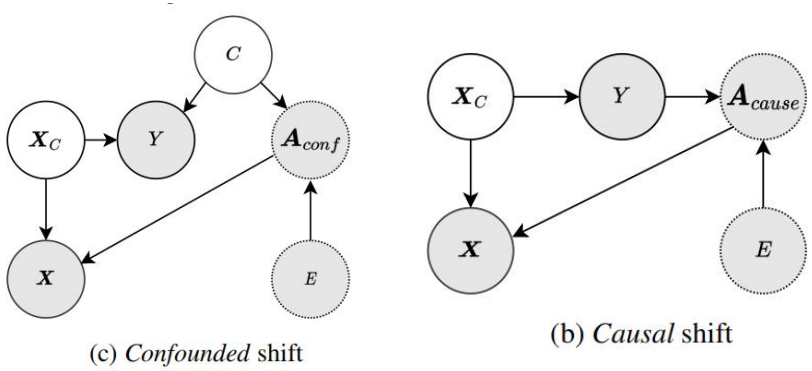
Role of the cononical causal graph



Can help to analysis validity of popular DG methods on certain datasets:

e.g., constraint: $(\phi(\boldsymbol{x}) \perp\!\!\!\perp A \mid Y)$ (here we regard $(\phi(\boldsymbol{x})$ as X_c) is correct for *causal/selected,* but not for *confounding.*

PS: none of these constraints are valid due to a correlation path between X_c and E



Inferring attributes-label relationship type

While the distinction between $\boldsymbol{A}_{ind}$ and $\boldsymbol{A}_{\overline{ind}}$ can be established using a statistical test of independence on a given dataset.

Distinction between $\boldsymbol{A}_{cause}$, $\boldsymbol{A}_{sel}$ and $\boldsymbol{A}_{conf}$ needs to provided based on the knowledge of datasets

learned from data (since $\boldsymbol{A}_{ind} \perp\!\!\!\perp Y$). Under some special conditions with the graph in Figure 2a—assuming all attributes are observed and all attributes in $\boldsymbol{A}_{\overline{ind}}$ are of the same type—we can also identify the type of $\boldsymbol{A}_{\overline{ind}}$ shift: $Y \perp\!\!\!\perp E|\boldsymbol{A}_{\overline{ind}}$ implies *Selected*; if not, then $\boldsymbol{X} \perp\!\!\!\perp E|\boldsymbol{A}_{ind}, \boldsymbol{A}_{\overline{ind}}, Y$ implies *Causal*, otherwise it is *Confounded*. In the general case of Figure 2b, however, it is not possible to differentiate between $\boldsymbol{A}_{cause}$, $\boldsymbol{A}_{conf}$ and $\boldsymbol{A}_{sel}$ using observed data and needs manual

remark: somehow subjective? (see appendix A)

Table 5: Commonly used DG datasets include auxiliary information.

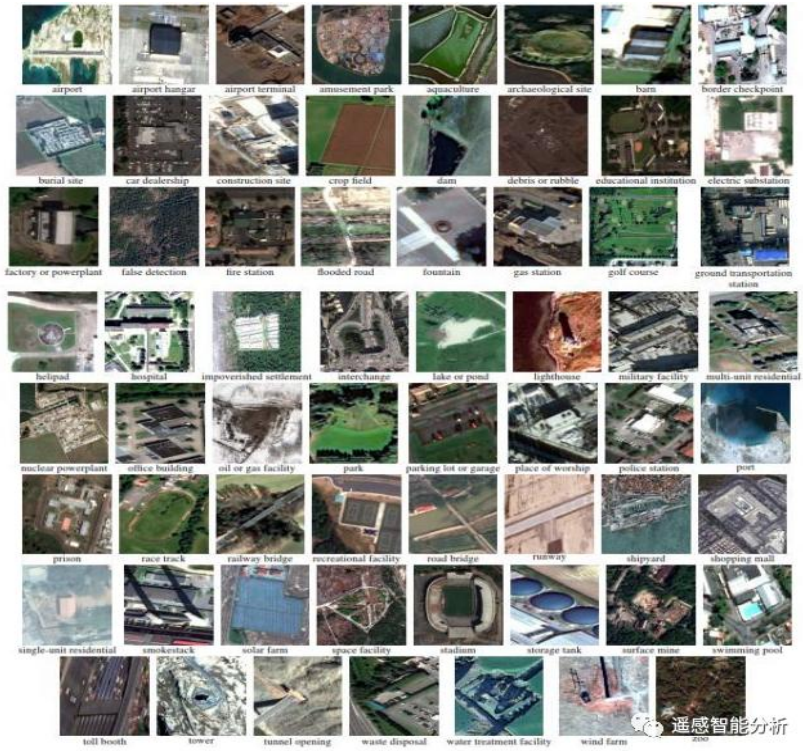
Dataset	**Attribute(s)**	$Y - \boldsymbol{A}$ **relationship**
FMoW-WILDS (Koh et al., 2021)	time	$\boldsymbol{A}_{ind}$
	region	$\boldsymbol{A}_{conf}$
Camelyon17-WILDS (Koh et al., 2021)	hospital	$\boldsymbol{A}_{ind}$
Waterbirds (Sagawa et al., 2020)	background (land/water)	$\boldsymbol{A}_{cause}$
MultiNLI (Sagawa et al., 2020)	negation word	$\boldsymbol{A}_{cause}$
CivilComments-WILDS (Koh et al., 2021)	demographic	$\boldsymbol{A}_{conf}$

For FMoW dataset, *time* can be considered an *Independent* attribute ($\boldsymbol{A}_{ind}$) since it reflects the time at which images are captured which is not correlated with Y; whereas *region* is a *Confounded* attribute since certain regions associated with certain Y labels are over-represented due to ease of data collection. Note that region cannot lead to *Causal* shift since the decision to take images in a region was not determined by the final label nor *Selected* for the same reason that the decision was not taken based on values of Y. Similarly, for the Camelyon17 dataset, it is known that differences in slide staining or image acquisition leads to variation in tissue slides across *hospitals*, thus implying that *hospital* is an *Independent* attribute ($\boldsymbol{A}_{ind}$) (Koh et al., 2021; Komura & Ishikawa, 2018; Tellez et al., 2019); As another example from healthcare, a study in MIT Technology Review[3] discusses biased data where a person's *position* ($\boldsymbol{A}_{conf}$) was spuriously correlated with disease prediction as patients lying down were more likely to be ill. As another example, (Sagawa et al., 2020) adapt MultiNLI dataset for OoD generalization due to the presence of spurious correlation between *negation words* (attribute) and the contradiction label between "premise" and "hypothesis" inputs. Here, negation words are a result of the contradiction label (*Causal* shift), however this relationship between negation words and label may not always hold. Finally, for the CivilComments dataset, we expect the *demographic* features to be *Confounded* attributes as there could be biases which result in spurious correlation between comment toxicity and demographic information.

To provide examples showing the availability of attributes and their type of relationship with the label, Table 5 lists some popular datasets used for DG and the associated auxiliary information present as metadata. In addition to above discussed datasets, weinclude the popularly used Waterbirds dataset (Sagawa et al., 2020) where the type of *background* (land/water) is assigned to bird images based on bird label; hence, being a *Causal* attribute (results on Waterbirds dataset are in Table 2).

no examples of *selected*

FMoW



Independence constraints based on label-attribute relationship

Proposition 3.1. *Given a causal DAG realized by specifying the target-attributes relationship in Figure 2a, the correct constraint depends on the relationship of label Y with the attributes $\boldsymbol{A}$. As shown, $\boldsymbol{A}$ can be split into $\boldsymbol{A}_{\overline{ind}}$, $\boldsymbol{A}_{ind}$ and E, where $\boldsymbol{A}_{\overline{ind}}$ can be further split into subsets that have a causal ($\boldsymbol{A}_{cause}$), confounded ($\boldsymbol{A}_{conf}$), selected ($\boldsymbol{A}_{sel}$) relationship with Y ($\boldsymbol{A}_{\overline{ind}} = \boldsymbol{A}_{cause} \cup \boldsymbol{A}_{conf} \cup \boldsymbol{A}_{sel}$). Then, the (conditional) independence constraints $\boldsymbol{X}_c$ should satisfy are,*

1. *Independent:* $X_c \perp\!\!\!\perp A_{ind}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{ind}|Y$; $X_c \perp\!\!\!\perp A_{ind}|E$; $X_c \perp\!\!\!\perp A_{ind}|Y, E$
2. *Causal:* $X_c \perp\!\!\!\perp A_{cause}|Y$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{cause}|Y, E$
3. *Confounded:* $X_c \perp\!\!\!\perp A_{conf}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{conf}|E$
4. *Selected:* $X_c \perp\!\!\!\perp A_{sel}|Y$; $X_c \perp\!\!\!\perp A_{sel}|Y, E$

Corollary 3.1. *All the above derived constraints are valid for Graph 2a. However, in the presence of a correlation between E and $\boldsymbol{X}_c$ (Graph 2b), only the constraints conditioned on E hold true.*

PS: if E – X_c relationship is not sure, E-conditioned constraints should be used

A fixed constraint cannot work for all shifts

Theorem 3.1. *Under the canonical causal graph in Figure 2(a,b), there exists no (conditional) independence constraint over* $\langle \boldsymbol{X}_c, \boldsymbol{A}, Y \rangle$ *that is valid for all realized DAGs as the type of multi-attribute shifts vary. Hence, for any predictor algorithm for* Y *that uses a single (conditional) independence constraint over its representation* $\phi(\boldsymbol{X})$, $\boldsymbol{A}$ *and* Y, *there exists a realized DAG* $\mathcal{G}$ *and a corresponding training dataset such that the learned predictor cannot be a risk-invariant predictor for distributions in* $\mathcal{P}_{\mathcal{G}}$, *where* $\mathcal{P}_{\mathcal{G}}$ *is the set of distributions obtained by changing* $P(\boldsymbol{A}|Y)$.

Corollary 3.2. *Even when* $|\boldsymbol{A}| = 1$, *an algorithm using a single independence constraint over* $\langle \phi(\boldsymbol{X}), A, Y \rangle$ *cannot yield a risk-invariant predictor for all kinds of single-attribute shift datasets.*

Causality Adaptive Constraint Minimization (CACM)

Phase I: Derive correct independence constraints.

For datasets satisfying the canonical graph, specify the relationship type for each attribute and uses the constraints from Proposition 3.1.

For other datasets, CACM requires a causal graph describing the dataset's DGP and uses the following steps to derive the independence constraints. Let V be the set of observed variables in the graph except Y, and C be the list of constraints.

1. For each observed variable $V \in \mathcal{V}$, check whether $(\boldsymbol{X}_c, V)$ are d-separated. Add $\boldsymbol{X}_c \perp\!\!\!\perp V$ to $\mathcal{C}$.
2. If not, check if $(\boldsymbol{X}_c, V)$ are d-separated conditioned on any subset Z of the remaining observed variables in $\mathcal{Z} = \{Y\} \cup \mathcal{V} \setminus \{V\}$. For each subset Z with d-separation, add $\boldsymbol{X}_c \perp\!\!\!\perp V | Z$ to $\mathcal{C}$.

1. Independent: $X_c \perp\!\!\!\perp A_{\text{ind}}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{\text{ind}} | Y$; $X_c \perp\!\!\!\perp A_{\text{ind}} | E$; $X_c \perp A_{\text{ind}} \mid Y, E$
2. Causal: $X_c \perp A_{\text{cause}} | Y$; $X_c \perp E$; $X_c \perp A_{\text{cause}} | Y, E$

Causality Adaptive Constraint Minimization (CACM)

Phase II: Apply regularization penalty using derived constraints.

Add regularizer (constraints) to ERM loss.

e.g. 1, for $A \in \boldsymbol{A}_{ind}$, we need to force $\phi(\boldsymbol{x}) \perp A$

$$\text{RegPenalty}_{\boldsymbol{A}_{ind}} = \sum_{i=1}^{|\boldsymbol{A}_{\text{ind}}|} \sum_{j>i} \text{MMD}(P(\phi(\boldsymbol{x}) \mid a_{i,ind}), P(\phi(\boldsymbol{x}) \mid a_{j,ind}))$$

PS: when E are available, E is preferred since A may not be unobserved and A may have many values.

e.g. 2, for $A \in \boldsymbol{A}_{\text{cause}}$ (Causal), here we additionally condition on E as there may be a correlation E-X_c

$$\text{RegPenalty}_{\boldsymbol{A}_{\text{cause}}} = \sum_{|E|} \sum_{y \in Y} \sum_{i=1}^{|\boldsymbol{A}_{\text{cause}}|} \sum_{j>i} \text{MMD}(P(\phi(\boldsymbol{x}) \mid a_{i,\text{ cause}}, y), P(\phi(\boldsymbol{x}) \mid a_{j,\text{ cause}}, y))$$

If E is observed, we always condition on E because of Corollary

Algorithm 1 *CACM*

Input: Dataset $(\boldsymbol{x}_i, \boldsymbol{a}_i, y_i)_{i=1}^n$, causal DAG $\mathcal{G}$
Output: Function $g(\boldsymbol{x}) = g_1(\phi(\boldsymbol{x})) : \boldsymbol{X} \to Y$
$\mathcal{A} \leftarrow$ set of observed variables in $\mathcal{G}$ except Y, E (special domain attribute)
$C \leftarrow \{\}$ ▷ mapping of A to $\boldsymbol{A_s}$
Phase I: Derive correct independence constraints
for $A \in \mathcal{A}$ **do**
  if $(\boldsymbol{X_c}, A)$ are d-separated **then**
    $\boldsymbol{X_c} \perp\!\!\!\perp A$ is a valid independence constraint
  else if $(\boldsymbol{X_c}, A)$ are d-separated conditioned on any subset $\boldsymbol{A_s}$ of the remaining observed variables in $\mathcal{A} \setminus \{A\} \cup \{Y\}$ **then**
    $\boldsymbol{X_c} \perp\!\!\!\perp A | \boldsymbol{A_s}$ is a valid independence constraint
    $C[A] = \boldsymbol{A_s}$
  end if
end for
Phase II: Apply regularization penalty using constraints derived
for $A \in \mathcal{A}$ **do**
  if $\boldsymbol{X_c} \perp\!\!\!\perp A$ **then**
    $RegPenalty_A = \sum_{|E|} \sum_{i=1}^{|A|} \sum_{j>i} \text{MMD}(P(\phi(\boldsymbol{x})|A_i), P(\phi(\boldsymbol{x})|A_j))$
  else if A is in C **then**
    $\boldsymbol{A_s} = C[A]$
    $RegPenalty_A = \sum_{|E|} \sum_{a \in \boldsymbol{A_s}} \sum_{i=1}^{|A|} \sum_{j>i} \text{MMD}(P(\phi(\boldsymbol{x})|A_i, a), P(\phi(\boldsymbol{x})|A_j, a))$
  end if
end for
$RegPenalty = \sum_{A \in \mathcal{A}} \lambda_A RegPenalty_A$
$g_1, \phi = \arg\min_{g_1, \phi};$ $\quad \ell(g_1(\phi(\boldsymbol{x})), y) + RegPenalty$

Experiments

Datasets

MNIST. Colored (Arjovsky et al., 2019) and Rotated MNIST (Ghifary et al., 2015) present *Causal* ($\boldsymbol{A}_{cause} = color$) and *Independent* ($\boldsymbol{A}_{ind} = rotation$) distribution shifts, respectively. We combine these to obtain a multi-attribute dataset with $\boldsymbol{A}_{cause}$ and $\boldsymbol{A}_{ind}$ ($col + rot$). For comparison, we also evaluate on single-attribute $\boldsymbol{A}_{cause}$ (Colored) and $\boldsymbol{A}_{ind}$ (Rotated) MNIST datasets.

small NORB (LeCun et al., 2004). This dataset was used by Wiles et al. (2022) to create a challenging DG task with single-attribute shifts, having multi-valued classes and attributes over realistic 3D objects. We create a multi-attribute shift dataset ($light$+azi), consisting of a causal connection, $\boldsymbol{A}_{cause}$ =$lighting$, between lighting and object category Y; and $\boldsymbol{A}_{ind}$ =$azimuth$ that varies independently across domains. We also evaluate on single-attribute $\boldsymbol{A}_{cause}$ ($lighting$) and $\boldsymbol{A}_{ind}$ ($azimuth$) datasets.

Waterbirds. We use the original dataset (Sagawa et al., 2020) where bird type (water or land) (Y) is spuriously correlated with background ($\boldsymbol{A}_{cause}$). To create a multi-attribute setup, we add different weather effects ($\boldsymbol{A}_{ind}$) to train and test data with probability p = 0.5 and 1.0 respectively.

Performance on multi-attribute shift

Table 2: **Colored + Rotated MNIST:** Accuracy on unseen domain for singe-attribute $(color, rotation)$ and multi-attribute $(col + rot)$ distribution shifts; **small NORB:** Accuracy on unseen domain for single-attribute $(lighting, azimuth)$ and multi-attribute $(light + azi)$ distribution shifts. **Waterbirds:** Worst-group accuracy on unseen domain for single- and multi-attribute shifts.

Algo	**Colored+Rotated MNIST** Accuracy			**small NORB** Accuracy			**Waterbirds** Worst-group accuracy	
	color	*rotation*	*col+rot*	*lighting*	*azimuth*	*light+azi*	original	multi-attr
ERM	30.9 ±1.6	61.9 ±0.5	25.2 ±1.3	65.5 ±0.7	78.6 ±0.7	64.0 ±1.2	66.0 ± 3.7	37.0 ± 1.1
IB-ERM	27.8 ±0.7	62.1 ±0.8	41.2 ±4.1	66.0 ±0.9	75.9 ±1.2	61.2 ±0.1	66.9 ± 4.6	40.8 ± 5.6
IRM	50.0 ±0.1	61.2 ±0.3	39.6 ±6.7	66.7 ±1.5	75.7 ±0.4	61.7 ±0.5	61.2 ± 5.2	37.7 ± 1.7
IB-IRM	49.9 ±0.1	61.4 ±0.9	49.3 ±0.3	64.7 ±0.8	77.6 ±0.3	62.2 ±1.2	62.3 ± 7.7	46.9 ± 6.5
VREx	30.3 ±1.6	62.1 ±0.4	23.3 ±0.4	64.7 ±1.0	77.6 ±0.5	62.5 ±1.6	68.8 ± 2.5	38.1 ± 2.3
MMD	29.7 ±1.8	62.2 ±0.5	24.1 ±0.6	66.6 ±1.6	76.7 ±1.1	62.5 ±0.3	68.1 ± 4.4	45.2 ± 2.4
CORAL	28.5 ±0.8	**62.5 ±0.7**	23.5 ±1.1	64.7 ±0.5	77.2 ±0.7	62.9 ±0.3	73.6 ± 4.8	54.1 ± 3.0
DANN	20.7 ±0.8	61.9 ±0.7	32.0 ±7.8	64.6 ±1.4	78.6 ±0.7	60.8 ±0.7	78.5 ± 1.8	55.5 ± 4.6
C-MMD	29.4 ±0.2	62.3 ±0.4	32.2 ±7.0	65.8 ±0.8	76.9 ±1.0	61.0 ±0.9	77.0 ± 1.2	52.3 ± 1.9
CDANN	30.8 ±8.0	61.8 ±0.2	32.2 ±7.0	64.9 ±0.5	77.3 ±0.3	60.8 ±0.9	69.9 ± 3.3	49.7 ± 3.9
DRO	33.9 ±0.4	60.6 ±0.9	25.3 ±0.5	65.5 ±0.7	77.1 ±1.0	62.3 ±0.6	70.4 ± 1.3	53.1 ± 2.2
Mixup	25.1 ±1.2	61.4 ±0.6	21.1 ±1.6	66.2 ±1.3	80.4 ±0.5	57.1 ±1.5	74.2 ± 3.9	64.7 ± 2.4
MLDG	31.0 ±0.3	61.6 ±0.8	24.4 ±0.7	66.0 ±0.7	77.9 ±0.5	64.2 ±0.6	70.8 ± 1.5	34.5 ± 1.7
SagNet	28.2 ±0.8	60.7 ±0.7	23.7 ±0.2	65.9 ±1.5	76.1 ±0.4	62.2 ±0.5	69.1 ± 1.0	40.6 ± 7.1
RSC	29.1 ±1.9	62.3 ±0.4	22.8 ±0.3	62.4 ±0.4	75.6 ±0.6	61.8 ±1.3	64.6 ± 6.5	40.9 ± 3.6
CACM	**70.4 ±0.5**	62.4 ±0.4	**54.1 ±1.3**	**85.4 ±0.5**	**80.5 ±0.6**	**69.6 ±1.6**	**84.5 ± 0.6**	**70.5 ± 1.1**

False constraints hurt generalization

(comparing to ERM?)

2. Causal: $X_c \perp A_{\text{cause}} | Y; X_c \perp E; X_c \perp A_{\text{cause}} | Y, E$

Table 3: small NORB *Causal* shift. Comparing $\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause} | Y, E$ with incorrect constraints.

Constraint	Accuracy
$\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause}$	72.7 ± 1.1
$\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause} \| E$	76.2 ± 0.9
$\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause} \| Y$	79.7 ± 0.9
$\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause} \| Y, E$	$\mathbf{85.4 \pm 0.5}$

Comparison between correct regularization

Table 4: Comparing $\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause} | Y, E$ and $\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause} | Y$ for *Causal* shift in MNIST and small NORB. The constraint implied by E-$\boldsymbol{X}_c$ correlation (Fig. 2b, Prop. 3.1) affects accuracy.

Constraint	MNIST	small NORB
$\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause} \| Y$	69.7 ± 0.2	79.7 ± 0.9
$\boldsymbol{X}_c \perp\!\!\!\perp \boldsymbol{A}_{cause} \| Y, E$	70.4 ± 0.5	85.4 ± 0.5

small NORB has a E-X_c correlation

Some thinking

weakness:

1. just a necessary condition

2. hard to determine if E-X_c correlation exists/type of Y-A_ind correlation