

Domain Counterfactuals for Explainability, Fairness, and Domain Generalization

David I. Inouye

Collaborators: Zeyu Zhou, Ruqi Bai, Tianci Liu, Sean Kulinski,
Yao Ji, Jing Gao, and Murat Kocaoglu



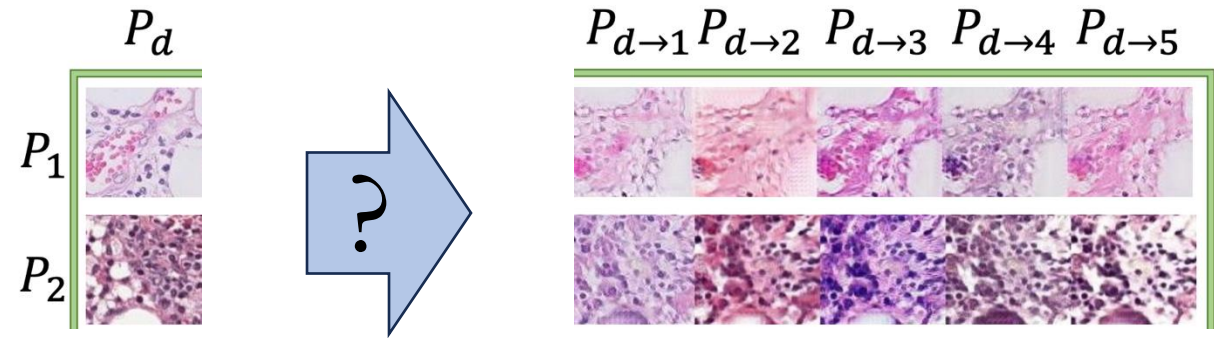
Elmore Family School of Electrical
and Computer Engineering



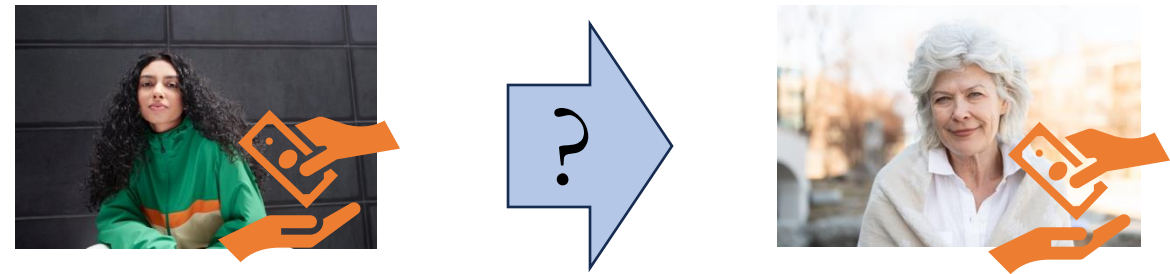
Z.Z., R.B., and D.I. acknowledge support from NSF (IIS-2212097), ARL (W911NF-2020221), and ONR (N00014-23-C-1016). M.K. acknowledges support from NSF CAREER 2239375, IIS 2348717, Amazon Research Award and Adobe Research. T.L. and J.G. acknowledge support from NSF-IIS2226108. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any funding source.

Domain Counterfactuals (DCF): What would a sample look like if it had been generated in a different domain?

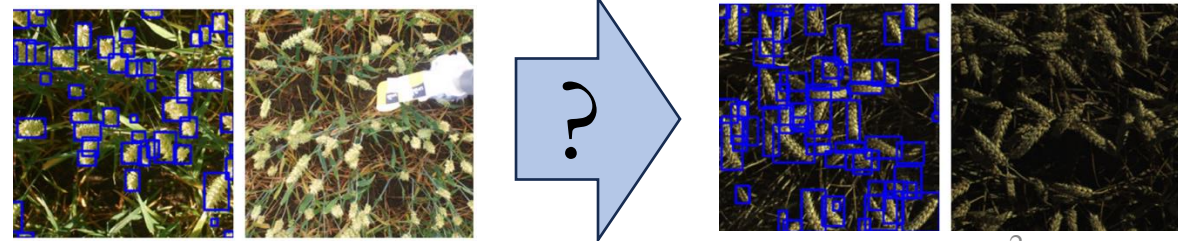
- What would a cell image look like if it had been collected at a different hospital?



- What would this person's loan application look like if they were elderly rather than young?



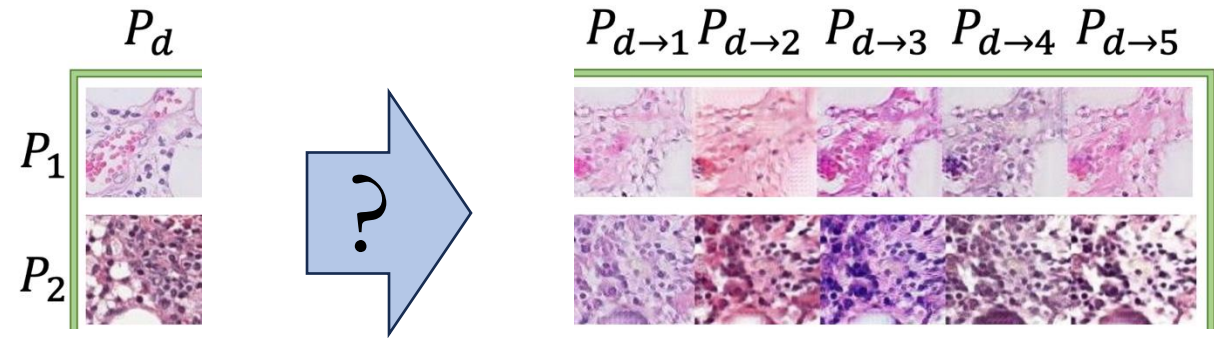
- What would this wheat image from Germany look like if it had been taken in France?



Domain counterfactuals could improve multiple areas of trustworthy ML

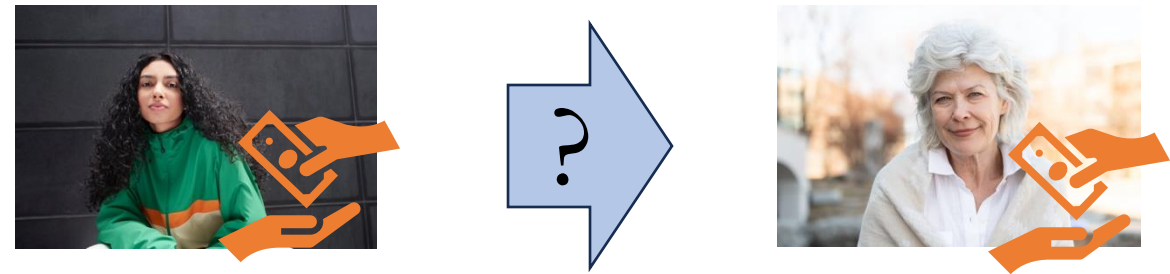
- Explaining distribution shifts

★ [Kulinski & Inouye, 2023]



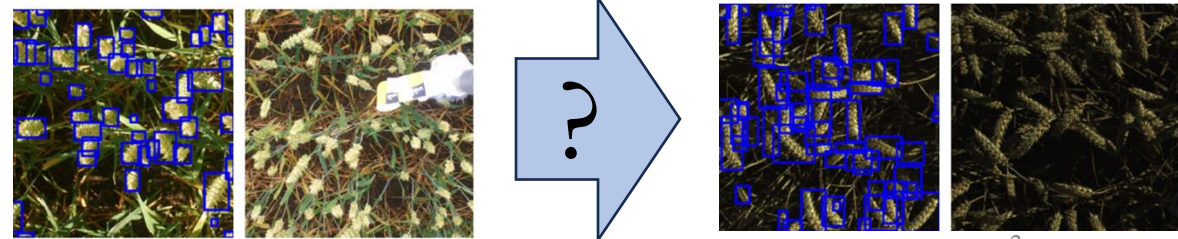
- Counterfactual fairness

★ [Zhou et al., 2024]



- Domain generalization
/ Out-of-distribution robustness

★ [Bai et al., 2024, under submission]



Domain Counterfactual (DCF) Applications and Estimation

DCF Applications

- Explaining distribution shifts
- Counterfactual fairness
- Domain generalization
(i.e., out-of-distribution robustness)

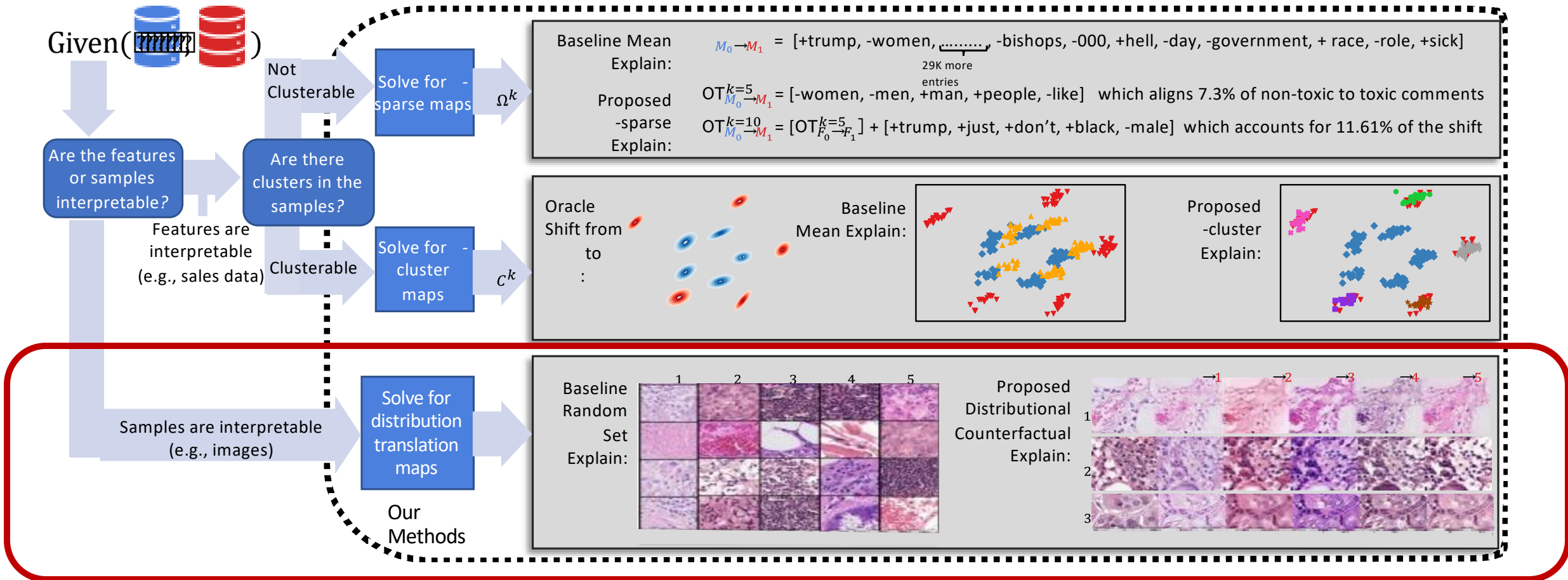
DCF Estimation

- Introduction to DCF estimation
- Theoretic contributions to DCF estimation
- VAE-based practical algorithm for DCF estimation
- Results and discussion

Explaining distribution shifts can help an ML operator mitigate shifts

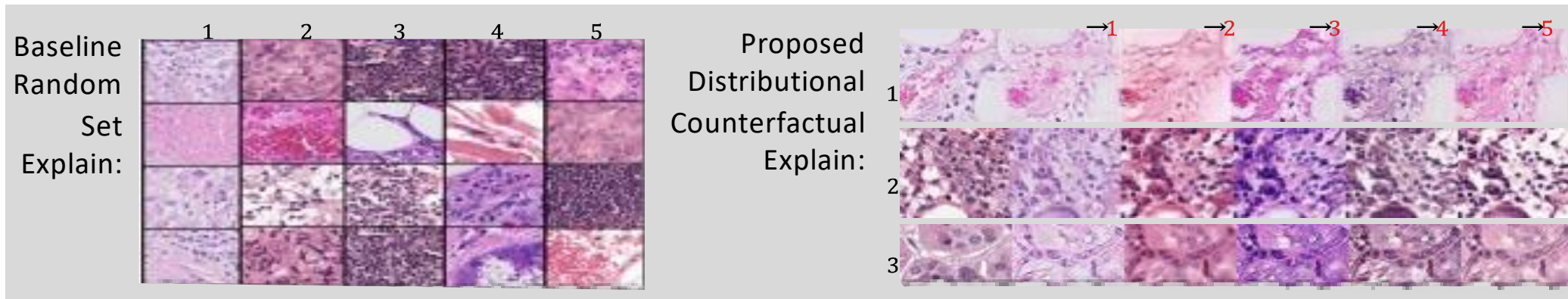
- Distribution shifts (when $P_{train} \neq P_{test}$) can cause serious decreases in model performance during deployment
- **Problem:** Most prior works focus on only *detecting* a shift, and do not help with “How should an ML operator respond?”
 - Should I retrain the model, ignore the shift, gather better data, etc.?
- **Our goal:** Aid the operator by *explaining* how P_{train} shifted to P_{test}

We propose shift explanations based on interpretable optimal transport and image-to-image translation



Yet, our “counterfactuals” were naïve counterfactuals without theoretically grounded understanding

- We merely used a StarGAN approach to translate between images



- This relied only on the inductive biases of the StarGAN architecture and was not grounded in causal theory
- The rest of this presentation will theoretically ground the idea of “counterfactual”

Domain Counterfactual (DCF) Applications and Estimation

DCF Applications

- Explaining distribution shifts
- **Counterfactual fairness**
- Domain generalization
(i.e., out-of-distribution robustness)

DCF Estimation

- Introduction to DCF estimation
- Theoretic contributions to DCF estimation
- VAE-based practical algorithm for DCF estimation
- Results and discussion

Counterfactual fairness requires the same predictions across different (counterfactual) worlds

- **Counterfactual fairness** ensures a model's decision is the same even if we intervene on a protected attribute (e.g., race).
- Example: Law school admission for fictional characters
 - Aladdin (poor) is predicted to have a 50% chance to pass the bar in the future.
 - What would have been the model's prediction if Aladdin was rich?
 - If the same, then the predictor is fair.
- **To answer, we need to formalize causality**



David I. Inouye, Purdue University



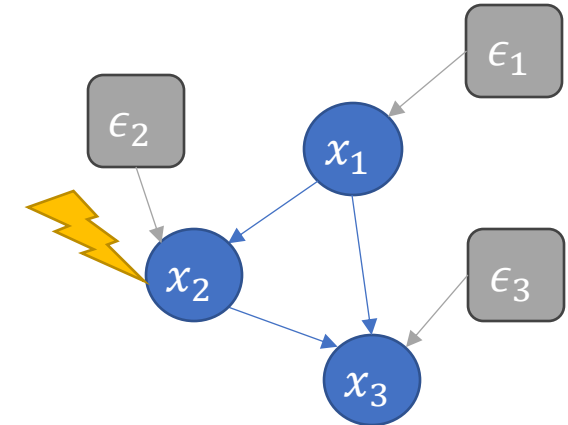
Background: Causality = probability + **interventions**

- Probability alone cannot answer questions about causality.
- Example: The use of umbrellas and rain are highly correlated.
 - Statistical dependency (MI)
 - Prediction
- But do umbrellas cause rain or does rain cause umbrellas?
 - Probability theory cannot help us.
- Interventions to the rescue!



Background: Structural causal models (SCM) enable causal reasoning about interventions

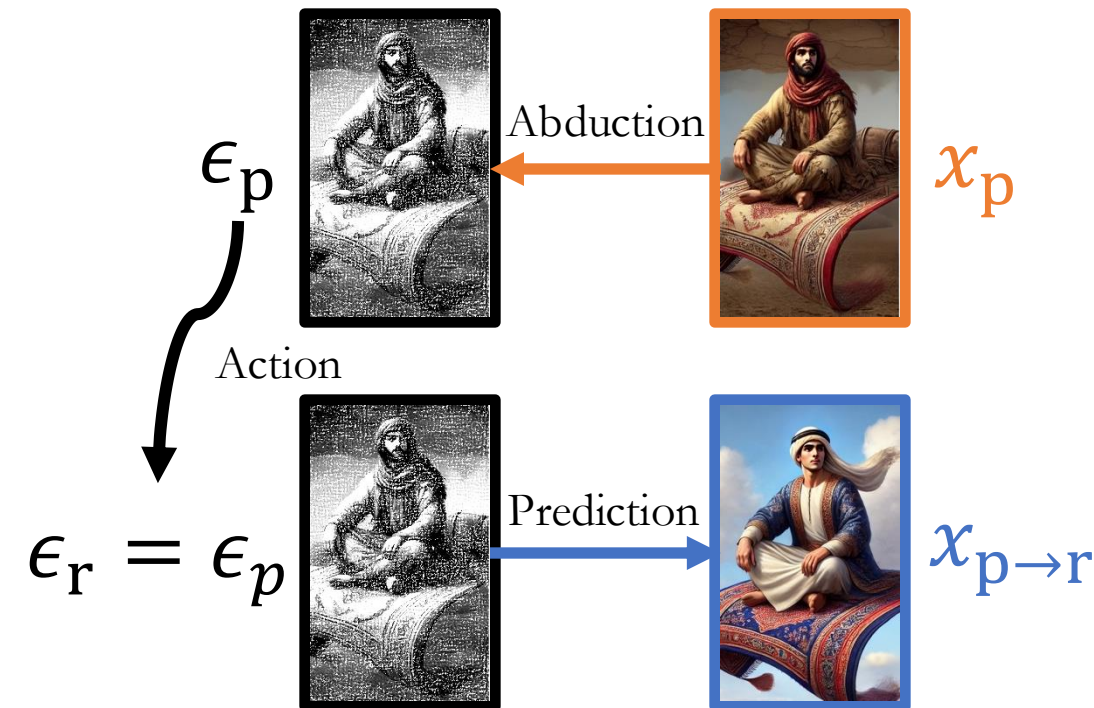
- Each causal variable x_i is assumed to be a **deterministic function** $f^{(i)}$ of its **causal parents** $Pa(x_i)$ and some **exogenous noise** $\epsilon_i \sim \mathcal{N}(0,1)$
- SCM example:
 - $x_1 = f^{(1)}(\epsilon_1) = 2\epsilon_1$
 - $x_2 = f^{(2)}(\epsilon_2, x_1) = x_1 + \epsilon_2$
 - $x_3 = f^{(3)}(\epsilon_3, x_1, x_2) = x_1 x_2 \epsilon_3$
- Intervened SCM:
 - $\tilde{x}_1 = f^{(1)}(\epsilon_1) = 2\epsilon_1$
 - $\tilde{x}_2 = \tilde{f}^{(2)}(\epsilon_2, \tilde{x}_1) = \tilde{x}_1^2 + \epsilon_2$
 - $\tilde{x}_3 = f^{(3)}(\epsilon_3, \tilde{x}_1, \tilde{x}_2) = \tilde{x}_1 \tilde{x}_2 \epsilon_3$



Background:

Counterfactuals bridge two causal worlds

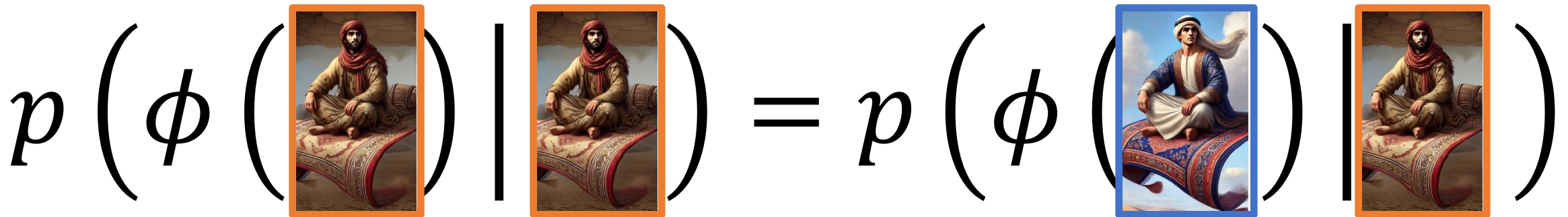
- Counterfactuals consider the distribution of variables in one world *given evidence from an alternate world*
- Counterfactuals can be formed in 3 steps
(Example evidence: $x_1 = 2, x_2 = 1, x_3 = 3$)
 - Abduction – Infer the exogenous noise ϵ_i based on the evidence in original world.
 - $\epsilon_1 = \frac{x_1}{2} = 1, \epsilon_2 = x_2 - x_1 = -1, \epsilon_3 = \frac{x_3}{x_1 x_2} = \frac{3}{2}$
 - Action – Change from original to intervened world.
 - Predict – Generate new values based on inferred exogenous noise.
 - $\tilde{x}_1 = 2\epsilon_1 = 2, \tilde{x}_2 = \tilde{x}_1^2 + \epsilon_2 = 3, \tilde{x}_3 = \tilde{x}_1 \tilde{x}_2 \epsilon_3 = 9$



Total effect (TE) measures the expected difference between factual and counterfactual predictions

- A stochastic predictor $\hat{Y} = \phi(X, A)$ is counterfactually fair if and only if:

$$p(\hat{Y} | X = x, A = a) = p(\hat{Y}_{1-a} | X = x, A = a), \quad \forall(x, a)$$

$$p\left(\phi\left(\text{img1}\right) \mid \text{img2}\right) = p\left(\phi\left(\text{img3}\right) \mid \text{img4}\right)$$


- Total effect (TE) for binary classification quantifies the violation of counterfactual fairness:

$$TE := \mathbb{E}[|\hat{Y} - \hat{Y}_{1-A}|] = \mathbb{E}_{X,A}[|\phi(X, A) - \phi(X_{1-A}, 1 - A)|]$$

The **optimal** counterfactually fair classifier mixes the **factual** and **counterfactual** predictions

- The counterfactually fair prediction problem is:

$$\begin{aligned} \min_{\phi} \mathbb{E}[\ell(\phi(X, A), Y)] \\ \text{s. t. } TE(\phi) = 0 \end{aligned}$$

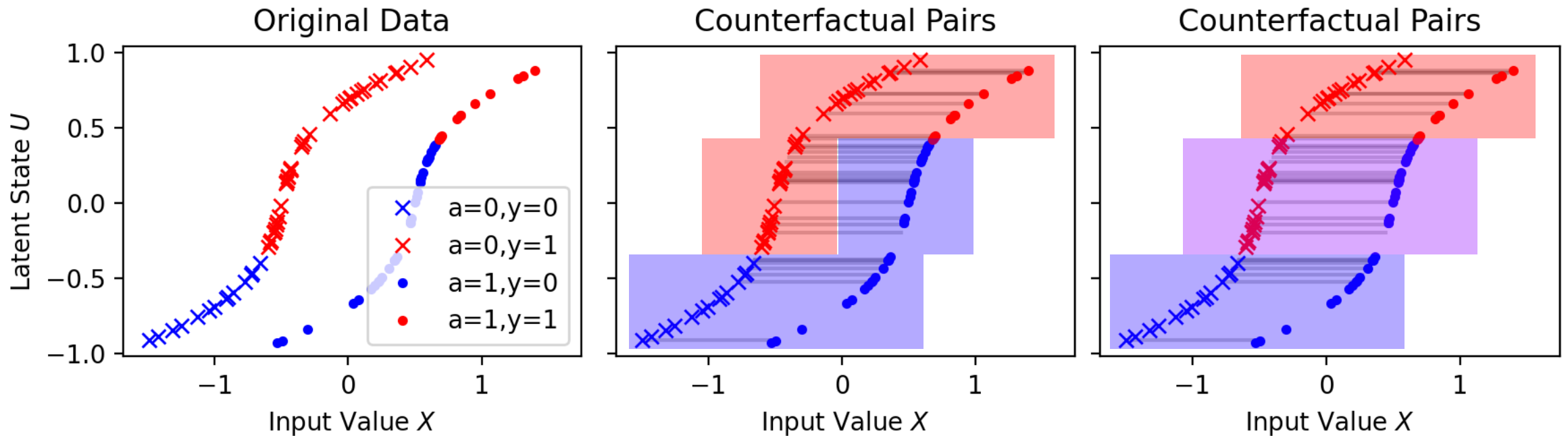
Theorem 3.3 & 3.4 (informal): The optimal counterfactually fair predictor mixes the factual and counterfactual predictions:

$$\phi_{CF}^*(x, a) := p(a)\phi^*(x, a) + p(1-a)\phi^*(x_{1-a}, 1-a),$$

where $\phi^*(x, a) := \operatorname{argmin}_{\phi} \mathbb{E}[\ell(\phi(X, A), Y)]$ is the (unfair) optimal predictor, and the excess risk for classification is:

$$\mathcal{R}_{CF}^* - \mathcal{R}^* = I(A, Y|U).$$

The optimal classifier forces each counterfactual pair to have the same prediction



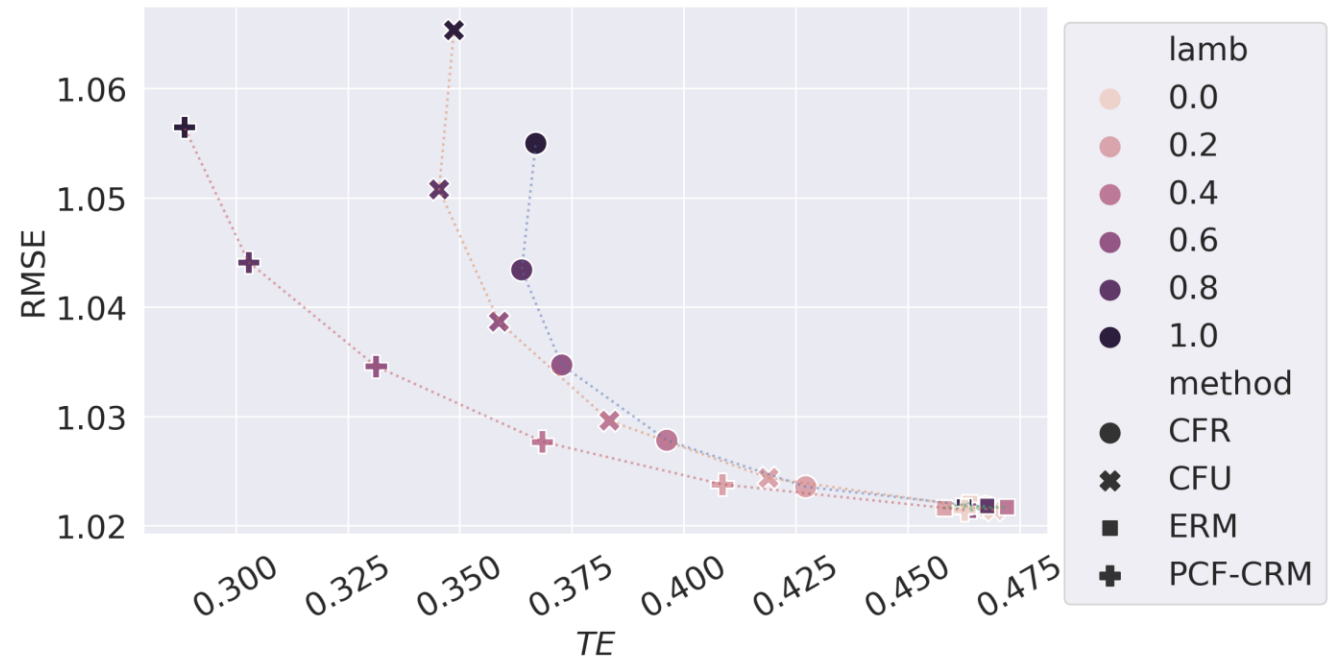
The (unfair) optimal classifier would achieve 100% accuracy on this dataset.

For fairness, the factual and counterfactual predictions must be the equal (i.e., both ends of lines must have same prediction).

Optimal fair must average the factual and counterfactual predictions when they differ.

Our Plug-in Counterfactual (PCF) method enables a better tradeoff between accuracy and fairness

- Plug-in Counterfactual (PCF)
 - Estimate counterfactuals
 - Estimate (unfair) classifier
 - Plug-in estimates to optimal fair predictor formula



Our regression results on a simulated law school fairness dataset demonstrates that PCF can provide a better tradeoff compared to other methods and has the lowest TE (lower is better).

Domain Counterfactual (DCF) Applications and Estimation

DCF Applications

- Explaining distribution shifts
- Counterfactual fairness
- **Domain generalization**
(i.e., out-of-distribution robustness)

DCF Estimation

- Introduction to DCF estimation
- Theoretic contributions to DCF estimation
- VAE-based practical algorithm for DCF estimation
- Results and discussion

Background: Domain generalization (DG) aims to predict accurately even under distribution shift


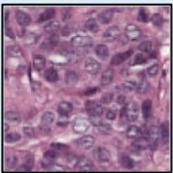
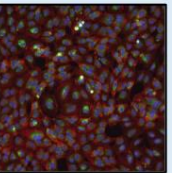
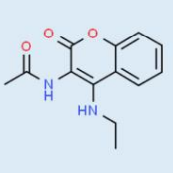




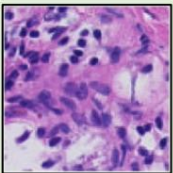
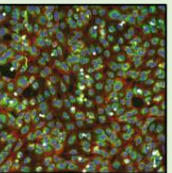
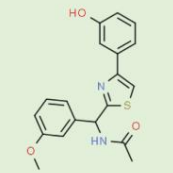



	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000
Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	<pre>import numpy as np ... norm=np.____</pre>
Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	<pre>import subprocess as sp p=sp.Popen() stdout=p.____</pre>
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

Figure from Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., ... & Liang, P. (2021, July). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning* (pp. 5637-5664). PMLR.

★ Bai, R., Bagchi, S., & Inouye, D. I. (2023). Benchmarking Algorithms for Federated Domain Generalization. *arXiv preprint arXiv:2307.04942*.

ERM often wins. Perhaps DG is too difficult. What is the value of different kinds of data?

- Add **labeled** test domain samples
 - Clearly a good idea but not always practical
 - Also, it's not really DG anymore
- Add **unlabeled** test domain samples
 - This becomes multi-source domain adaptation
 - Requires adapting to each new test domain
- Our proposal: Add (approximate) **counterfactual pairs** *within training domains*
 - No test domain data required
 - Insight 1: Matching counterfactual pairs can provably generalize to certain shifts.
 - Insight 2: Only a small number of pairs needed. (few-shot setting)

We focus on spurious correlation DG scenarios for linear SCMs

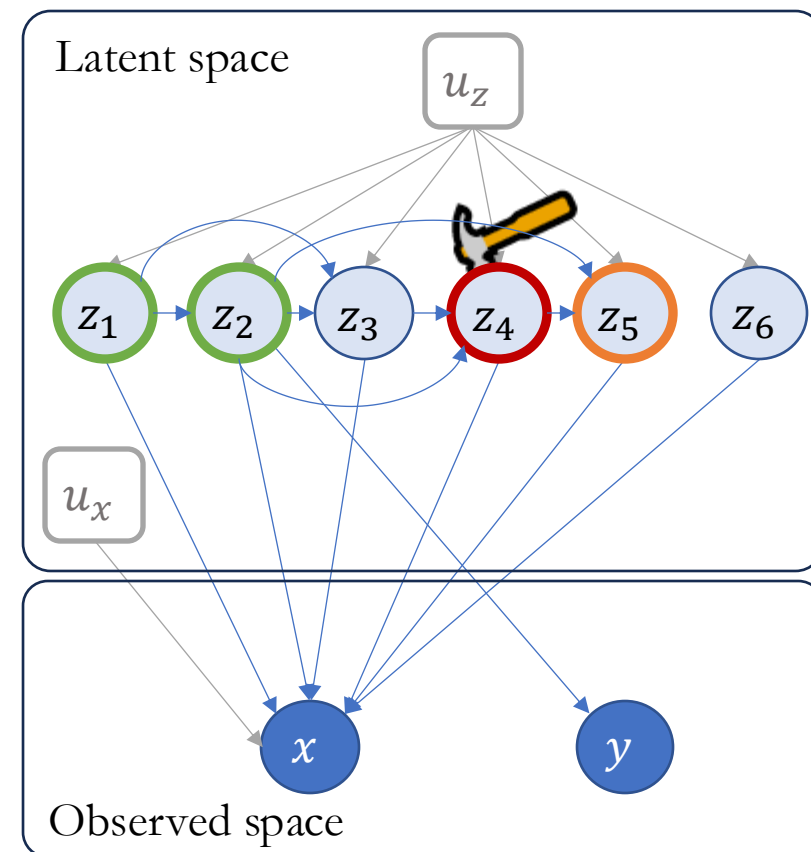
- **Assumption 1:** The intervened variables between domains are not ancestors of the target variable, i.e., intervened are *spuriously correlated*.

$$I(\mathcal{F}) \cap \text{Anc}(y) = \emptyset$$

- **Assumption 2:** Test domains f_+ intervene on same variables or descendants as seen in the training domains \mathcal{F} .

$$I(\mathcal{F} \cup f_+) \subseteq I(\mathcal{F}) \cup \text{Desc}(I(\mathcal{F}))$$

- **Assumption 3:** The SCM is linear.
 - $[u_z, u_x] \sim \text{ExogenousNoiseDistribution}$
 - $z_d = A_d u_z$ (A_d is different between domains)
 - $x_d = B z_d + u_x$ (B is shared between domains)
 - $y \sim \text{Bernoulli}(\sigma(c^T z_d))$ (c is shared between domains)



Counterfactual Matching (CFM) simply adds a counterfactual constraint to ERM

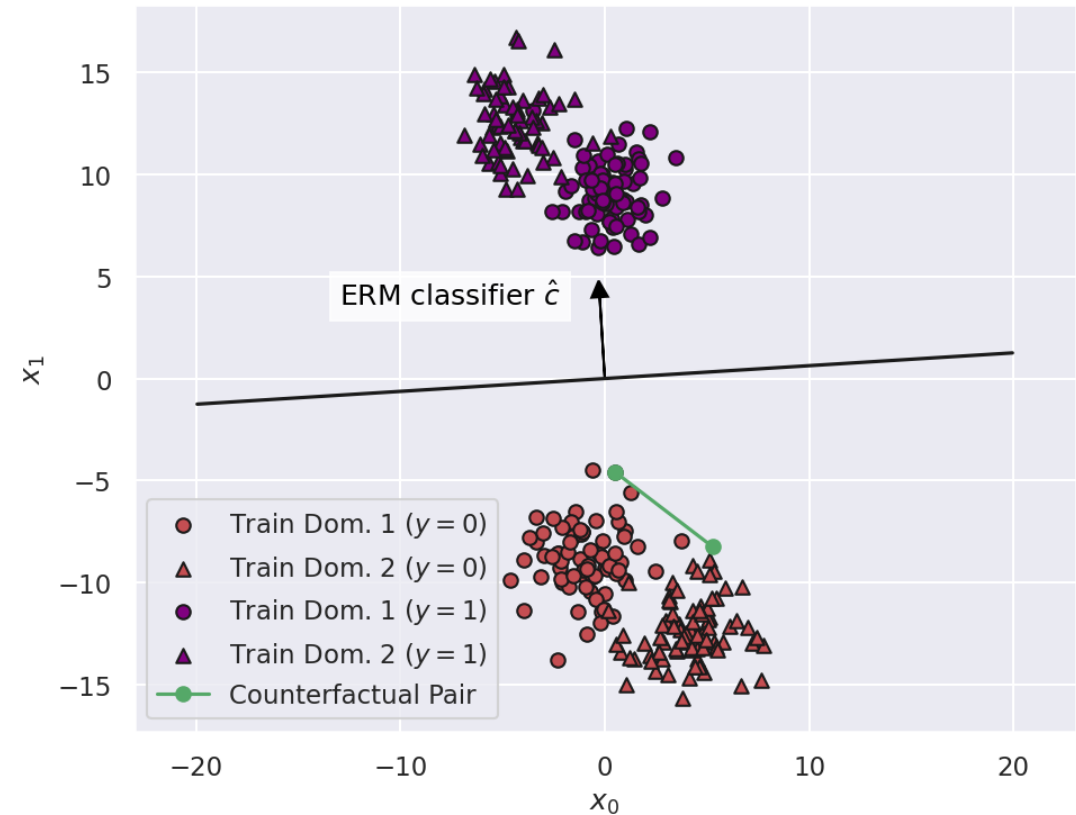
- Given training domain data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ and counterfactual pairs $\{(x_d^{(j)}, x_{d \rightarrow d'}^{(j)})\}_{j=1}^k$, the counterfactual matching problem (CFM) is defined as:

$$\begin{aligned} \min_{\phi} \quad & \frac{1}{n} \sum_{i=1}^n \ell(\phi(x^{(i)}), y^{(i)}) \\ \text{s.t.} \quad & \phi(x_d^{(j)}) - \phi(x_{d \rightarrow d'}^{(j)}) = 0, \forall j \end{aligned}$$

- This is simply ERM + a constraint that predictions for counterfactual pairs match
- Can the learned classifier generalize to new domains?**

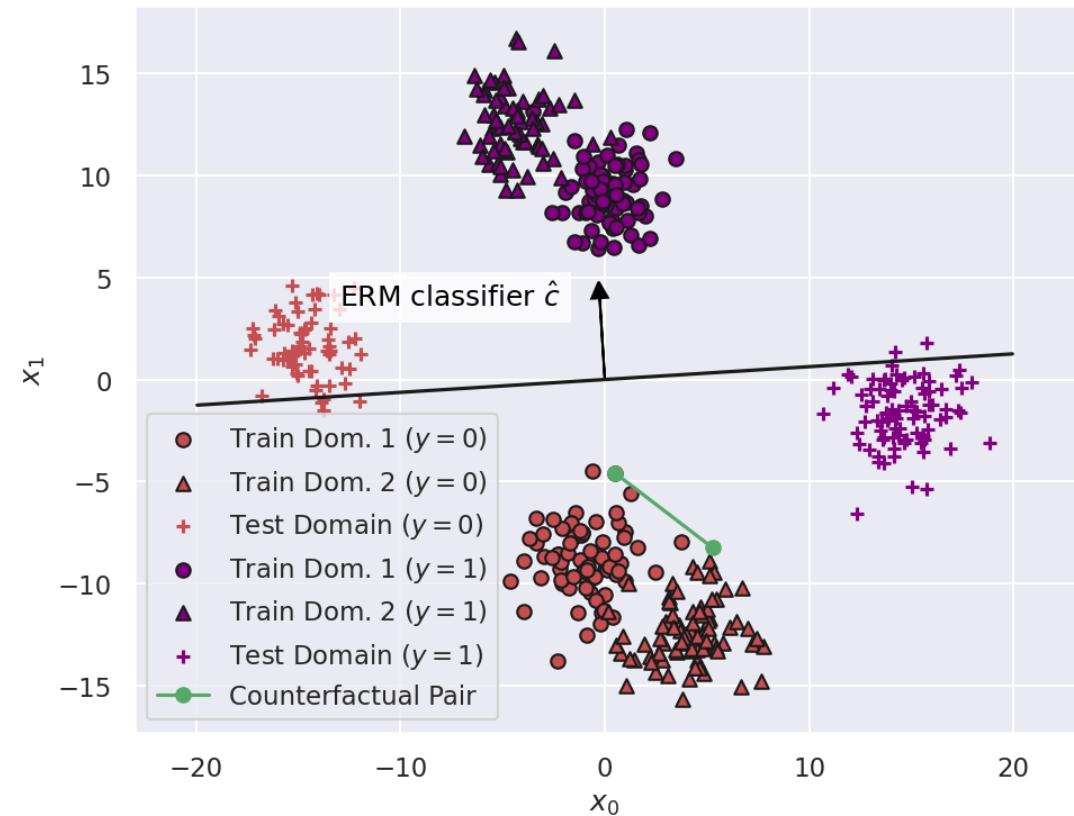
ERM classifier on training domains may depend on spurious features

- The ERM classifier does very well on the training domains



The ERM classifier may not be robust to spurious feature changes

- However, a test domain can clearly show that this classifier is not robust to spurious feature shifts

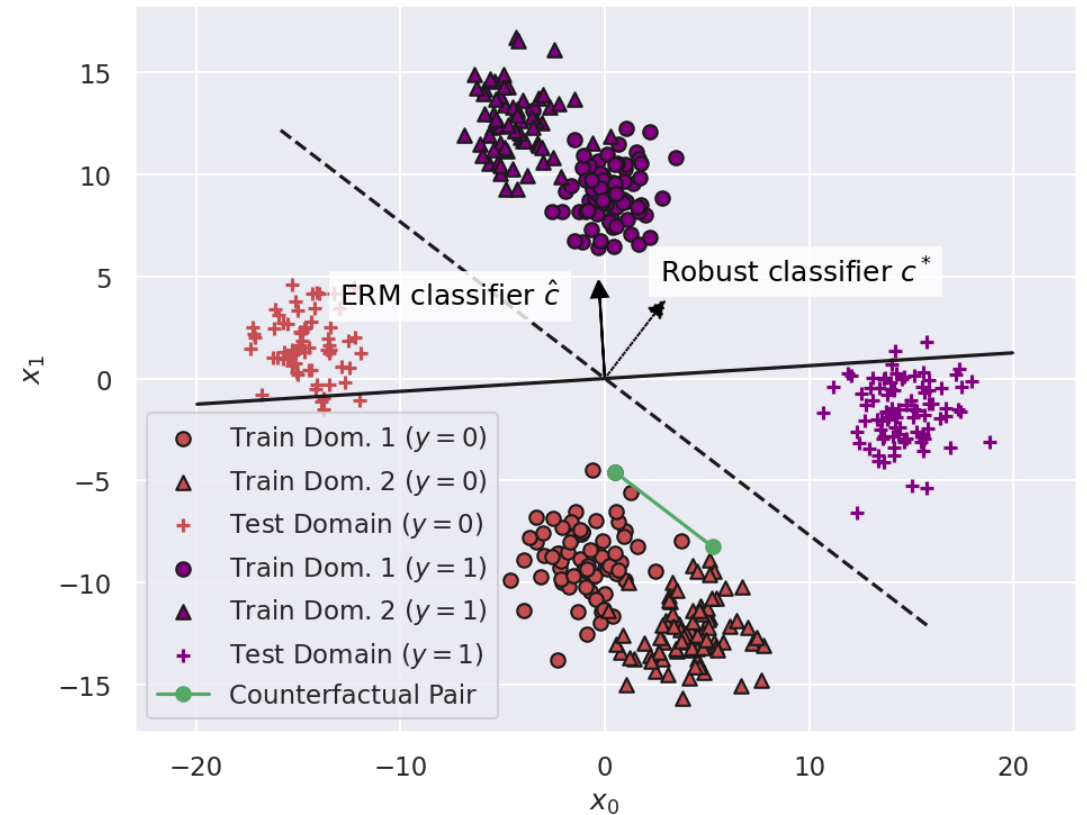


Intuition: Each counterfactual pair can eliminate one spurious dimension

- The CFM constraint forces the classifier to be orthogonal to the counterfactual difference

$$\begin{aligned} \phi\left(x_d^{(j)}\right) - \phi\left(x_{d \rightarrow d'}^{(j)}\right) &= 0 \\ \Leftrightarrow c^{*T} \left(x_d^{(j)} - x_{d \rightarrow d'}^{(j)} \right) &= 0 \end{aligned}$$

- Counterfactuals provide a data-driven constraints that correspond spurious feature directions



Modified CFM finds DG robust solutions even with noisy or approximate counterfactuals

- Intuition
 - First find best rank r subspace of noisy counterfactual differences
 - Make classifier orthogonal to this subspace
- For imperfect counterfactuals $\tilde{x}_{d \rightarrow d'}^{(1)} \approx x_{d \rightarrow d'}^{(1)}$, the modified CFM is:

$$\min_c \frac{1}{n} \sum_{i=1}^n \ell(c^T \mathbf{x}^{(i)}, \mathbf{y}^{(i)})$$
$$s.t. \quad c^T \tilde{\mathbf{U}}_r = 0$$

- Where $\tilde{\mathbf{U}}_r$ are the r largest left singular vectors of $\tilde{\Delta}$

$$\tilde{\Delta} = \begin{bmatrix} x_d^{(1)} - \tilde{x}_{d \rightarrow d'}^{(1)} \\ x_d^{(1)} - \tilde{x}_{d \rightarrow d'}^{(1)} \\ \vdots \\ x_d^{(k)} - \tilde{x}_{d \rightarrow d'}^{(k)} \end{bmatrix}$$

The test domain risk is bounded by the training risk and a subspace comparison term

Lemma: Given assumptions 1-3 and letting $\tilde{\mathcal{S}} := I(\mathcal{F}) \cup \text{Desc}(I(\mathcal{F}))$ denote the intervened features, the test domain risk with MSE is bounded as:

$$\begin{aligned} R_{d^+}(c) &:= \mathbb{E}_{p(x_{d^+}, y)} [\|c^T x_{d^+} - y\|^2] \\ &\leq 2\mathbb{E}_{p(d)p(x_d, y)} [\|c^T x_d - y\|^2] + 2\|c\|^2 \cdot \lambda_1^+ \cdot \|(I - \tilde{U}_r \tilde{U}_r^T) U_{\tilde{\mathcal{S}}}\|^2 \end{aligned}$$

where λ_1^+ is the largest eigenvalue of $M^+ := \mathbb{E}_{p(d)p(x_{d^+}, x_d)} [(x_{d^+} - x_d)(x_{d^+} - x_d)^T]$ and $U_{\tilde{\mathcal{S}}}$ is any orthogonal basis for the subspace corresponding to the $\tilde{\mathcal{S}}$ latent features.

- The first term is simply the training risk.
- λ_1^+ corresponds to the hardness of the test domain.
(i.e., how far it is from the training domains)
- The last part quantifies how much of the spurious feature space is ignored by classifier.
(i.e., the projection of the spurious subspace $U_{\tilde{\mathcal{S}}}$ onto the orthogonal subspace \tilde{U}_r)

We bound the DG test risk via Davis-Kahan subspace perturbation theory

Theorem: Given the same assumptions as before and assuming we observe $k \geq |\tilde{\mathcal{S}}|$ counterfactual pairs, the test domain risk is bounded as:

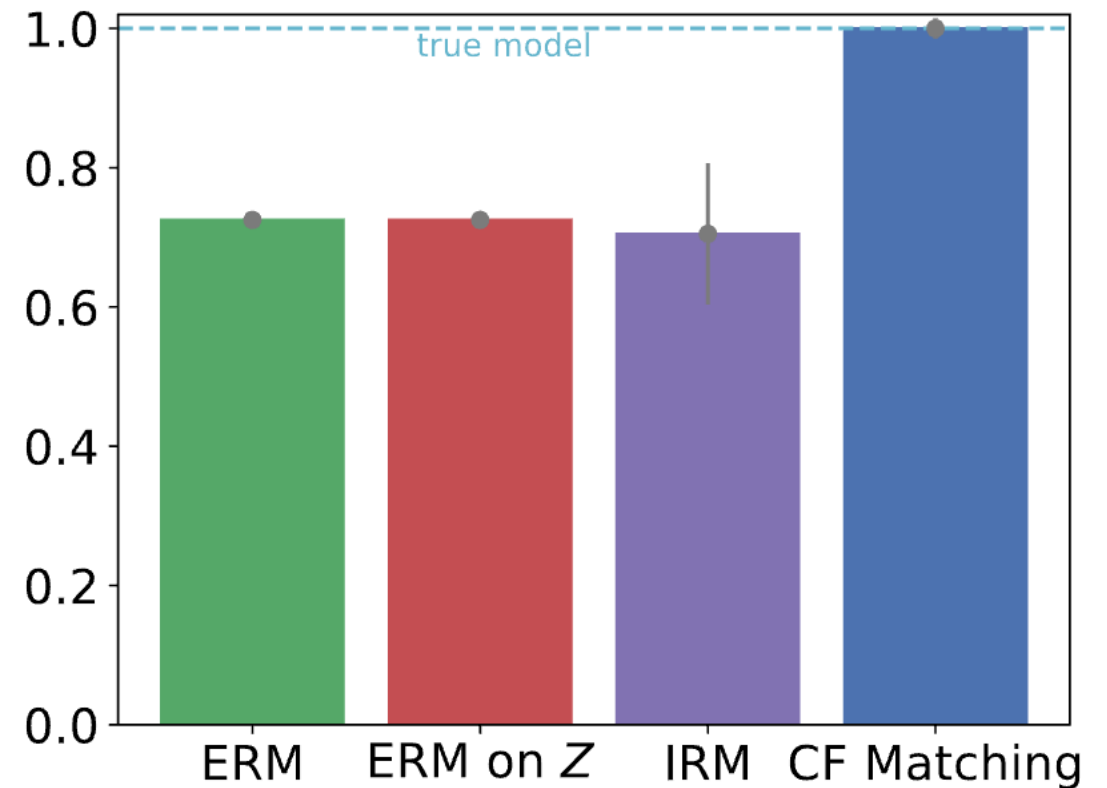
$$\begin{aligned} R_{d^+}(c) &:= \mathbb{E}_{p(x_{d^+}, y)} [\|c^T x_{d^+} - y\|^2] \\ &\leq 2\mathbb{E}_{p(d)p(x_d, y)} [\|c^T x_d - y\|^2] + 2\|c\|^2 \frac{\lambda_1^+ \cdot \|\tilde{\Delta}\tilde{\Delta}^T - \Delta\Delta^T\|^2}{\min_{r+1 \leq j \leq m, 1 \leq j' \leq |\tilde{\mathcal{S}}|} |\tilde{\lambda}_j - \lambda_{j'}|^2} \end{aligned}$$

where Δ corresponds to the matrix of perfect/oracle counterfactual pair differences and $\tilde{\lambda}_j$ and $\lambda_{j'}$ correspond to the eigenvalues of $\tilde{\Delta}\tilde{\Delta}^T$ and $\Delta\Delta^T$ respectively.

- Example: If $\tilde{\Delta} = \Delta$, then this term is 0 for only $k \geq |\tilde{\mathcal{S}}|$ pairs (few-shot setting).
- Example: If $\tilde{\Delta} = \Delta + \epsilon$, then the DG error is based on the variance of ϵ .

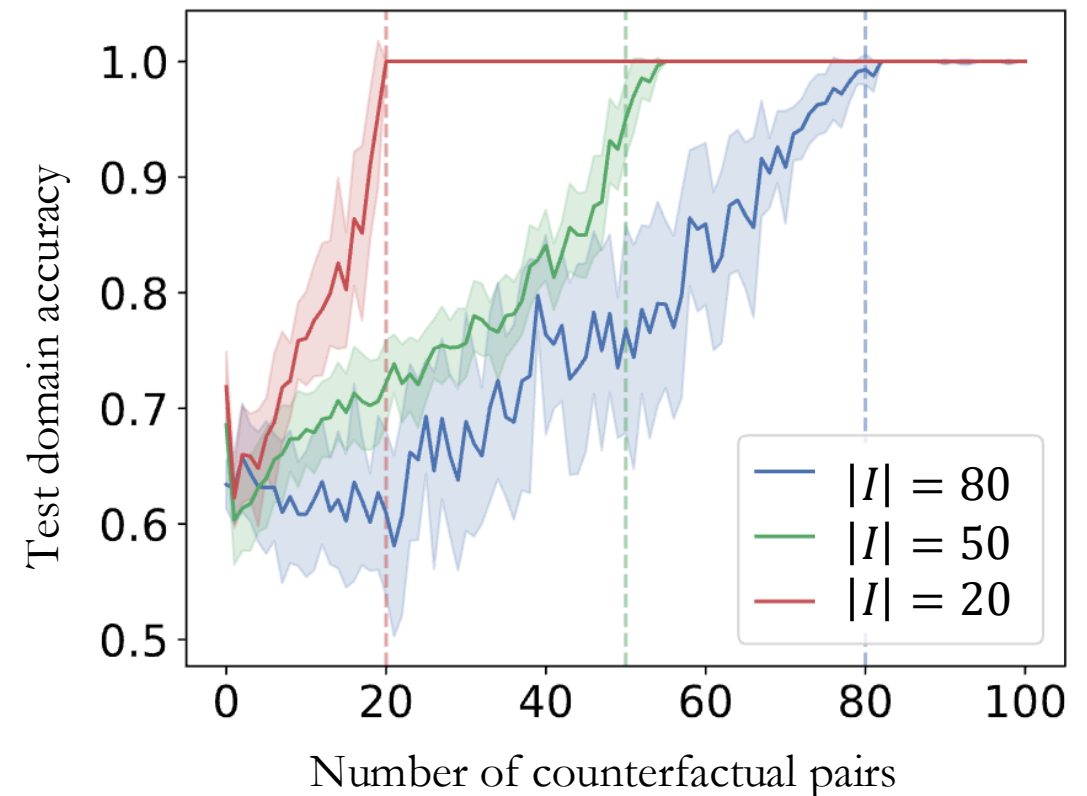
Results on synthetic data confirm theory that CFM is optimal

- Baselines
 - ERM
 - ERM with oracle latent Z
 - Invariant Risk Minimization (IRM)
- Our CFM approach can match oracle model performance in this simple simulated setup



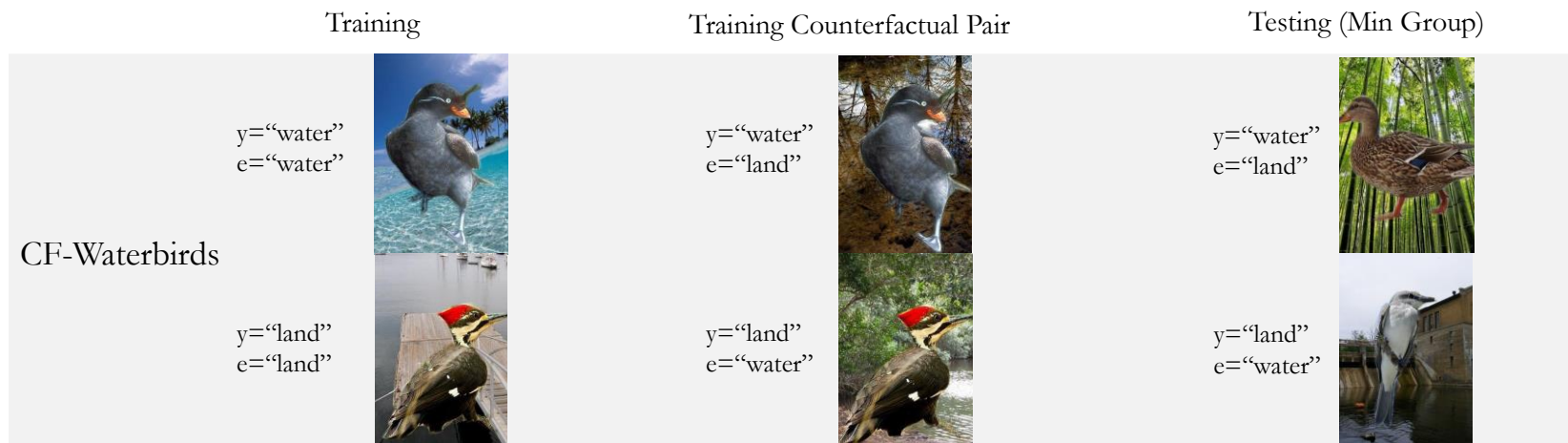
CFM only needs a small number of counterfactual pairs

- For **linear** causal model, we conjecture that only $|I|$ pairs are needed, where $|I|$ is the intervention set size.
- Intuition: k pairs uniquely define a linear transformation.
- Our initial results suggest that this is indeed true.



Beyond linearity, we show significant improvement on realistic tasks with only 240 counterfactual pairs

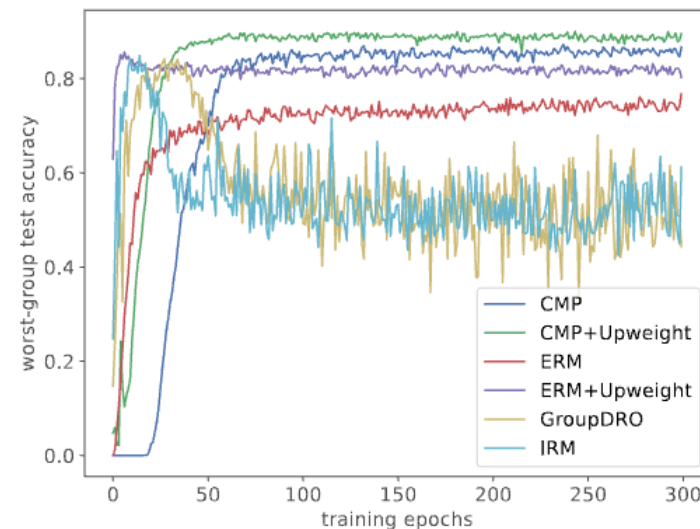
Counterfactual
Waterbirds dataset is
a variant of the well-
known Waterbirds
DG dataset



	adj acc	acc avg	acc wg
ERM	0.978	0.917	0.767
IRM	0.943	0.920	0.849
GroupDRO	0.934	0.907	0.842
ERM+upweighting	0.980	0.936	0.856
CFM (Ours)	0.978	<i>0.953</i>	<i>0.872</i>
CFM+upweighting	0.980	0.958	0.900

We outperform ERM by **10%** and
others by **3-4%**.

Our training method
is significantly more
stable than other
DG methods



Counterfactual DG open questions and concluding thought

- Can the theory be extended to non-linear or invertible causal models?
- How can we elicit approximate counterfactuals in different applications?
- *Hypothesis*: These domain counterfactuals provide a **data-driven** way to *implicitly* specify **task constraints**.
(Analogous to class labels that are a data-driven way to implicitly specify the task goal.)

Domain Counterfactual (DCF) Applications and Estimation

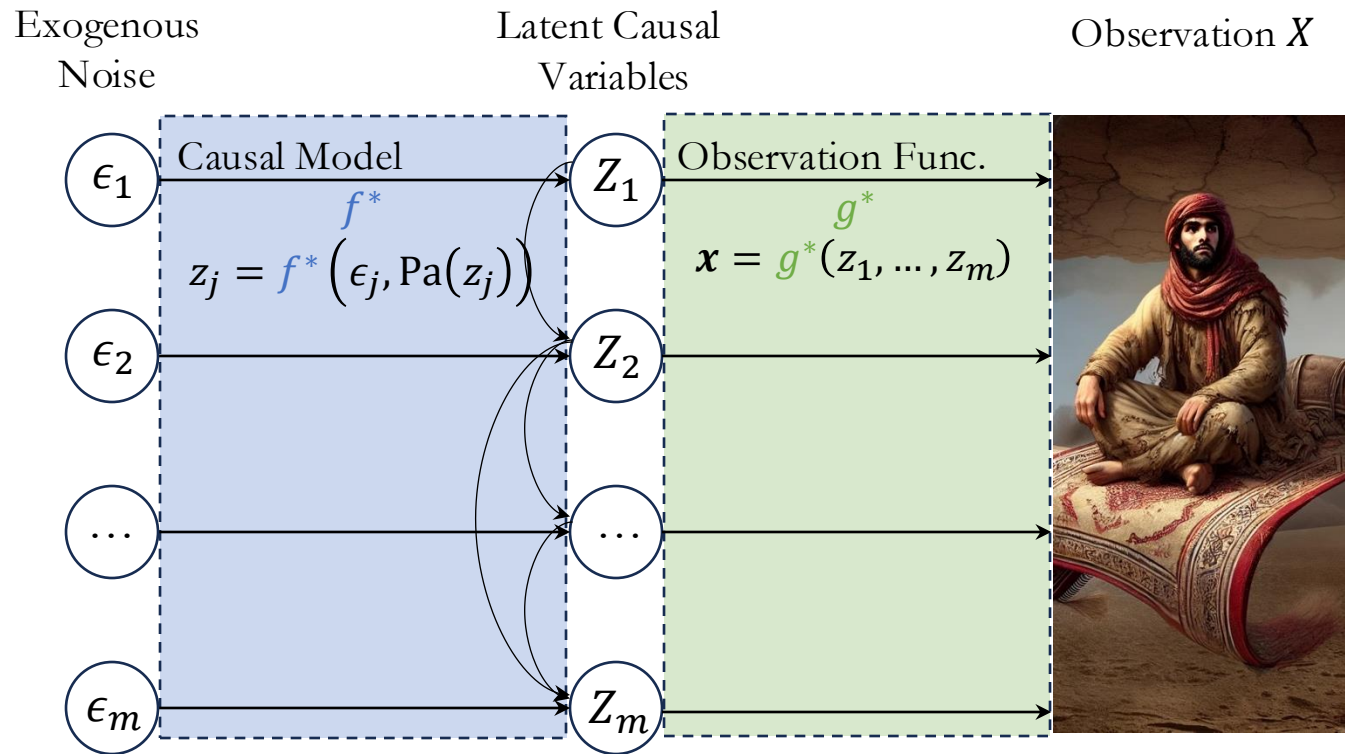
DCF Applications

- Explaining distribution shifts
- Counterfactual fairness
- Domain generalization
(i.e., out-of-distribution robustness)

DCF Estimation

- Introduction to DCF estimation
- Theoretic contributions to DCF estimation
- VAE-based practical algorithm for DCF estimation
- Results and discussion

Background: We consider estimation in the challenging case when the **causal variables are latent**



Latent causal models assume there exists:

1. A **causal model** f^* that maps exogenous noise ϵ to latent causal variables $\mathbf{z} = [z_1, z_2, \dots, z_m]$
2. An **observation function** g^* that maps latent variables to observed variables \mathbf{x} .

Given the ground truth causal models, a domain counterfactual infers the exogenous noise and then constructs counterfactual

Domain counterfactuals can be constructed via two steps

1. Infer exogenous noise from observation using causal model from domain 1.

For invertible models, this is:

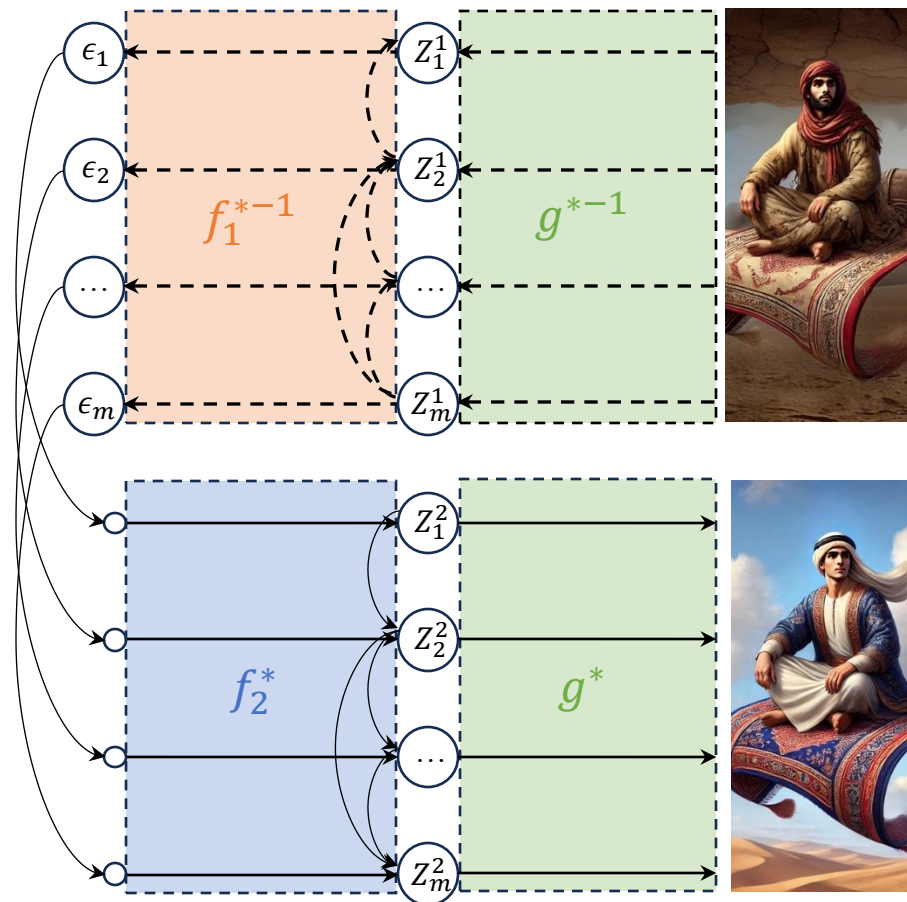
$$\epsilon = f_1^{*-1}(z^1) = f_1^{*-1}(g^{*-1}(x^1))$$

2. Using the recovered exogenous noise ϵ , apply the causal model and shared observation function, i.e.,

$$x_{1 \rightarrow 2} = g^*(z^2) = g^*(f_2^*(\epsilon))$$

For invertible models, these two steps are:

$$x_{1 \rightarrow 2} = g^* \left(f_2^* \left(f_1^{*-1} \left(g^{*-1}(x^1) \right) \right) \right)$$



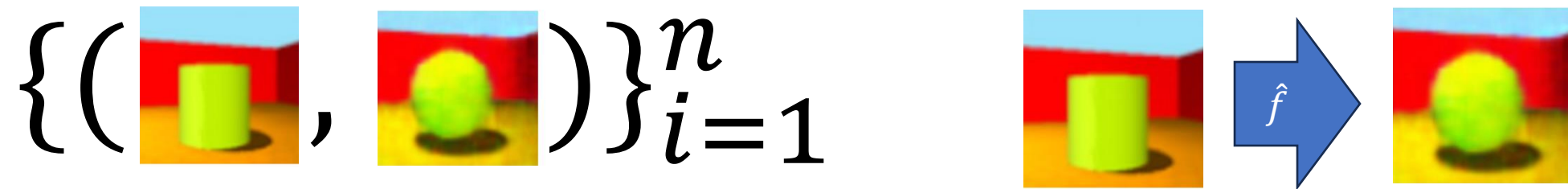
The “factual”
Aladdin was
poor.

What would
Aladdin be like
if he was **rich**
instead?
(counterfactual)

g^* is assumed to be shared
between domains

How do you estimate domain counterfactuals given only samples from each domain?

- If given counterfactual pairs, we could simply use supervised learning



- However, we assume only access to (unpaired) samples from each domain

$$\left\{ \begin{array}{c} \text{Cylinder} \end{array} \right\}_{i=1}^{n_A} \quad \left\{ \begin{array}{c} \text{Sphere} \end{array} \right\}_{i=1}^{n_B}$$

How do you estimate domain counterfactuals given only samples from each domain?

Constructive approach

- Goal: Prove that you can recover g^* and f^* (causal discovery)
 - And then construct counterfactual (causal inference)
- Method: Determine necessary and sufficient conditions for recovery
- Problem: Assumptions are often too strong for realistic problems.

Hope-for-the-best approach

- Goal: Train the best model on the data you have and hope for the best.
- Method: Train a conditional generative model (e.g., VAE) and hope the encoder and decoder approximately recover g^* and f^*
- Problem: This ignores the core challenges and doesn't work in practice.

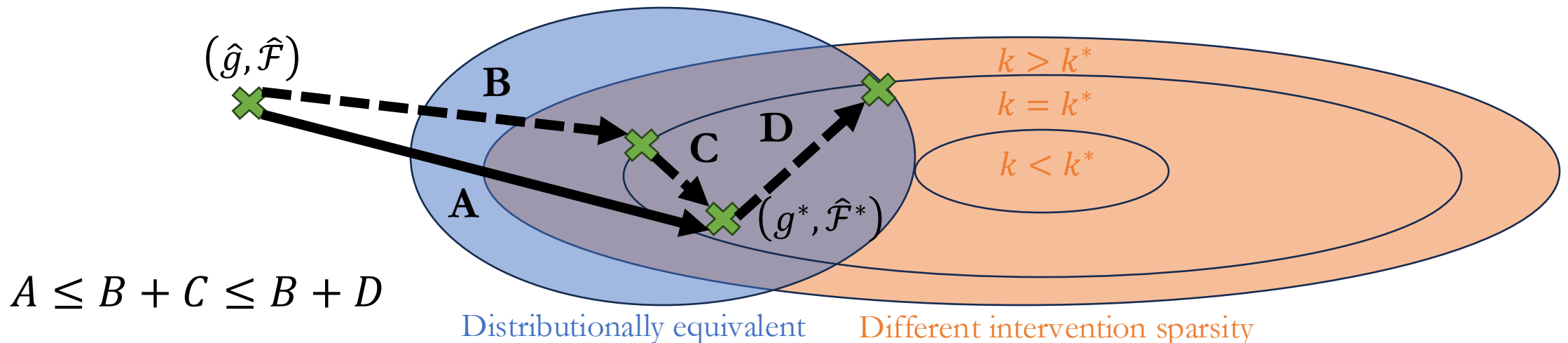
Our work tries to bridge the gap between these two approaches

- Our theory supports the following main claims
 1. Recovering the ground truth causal models f_d^* and g^* is **not necessary** for estimating domain counterfactuals (Theorem 1).
 2. We can bound the estimation error based on (Theorem 2):
 - a) **Distribution fit** – Does the generative model fit the domain data?
 - b) **Intervention sparsity** – If we assume the causal models for each domain only differ w.r.t. to a few variables (i.e., sparse), we can bound the extra error based on this assumption.
 3. We can assume **w.l.o.g.** that all **intervened causal variables** are the last variables.



DCF estimation error is bounded by distribution fit and intervention sparsity

- Informally, the DCF error can be decomposed into two terms:
 - Distribution fit (B) – How far are the generated distributions from the observed distributions?
 - Intervention sparsity (D) – What is the worst case ILD model for a target intervention sparsity k ?

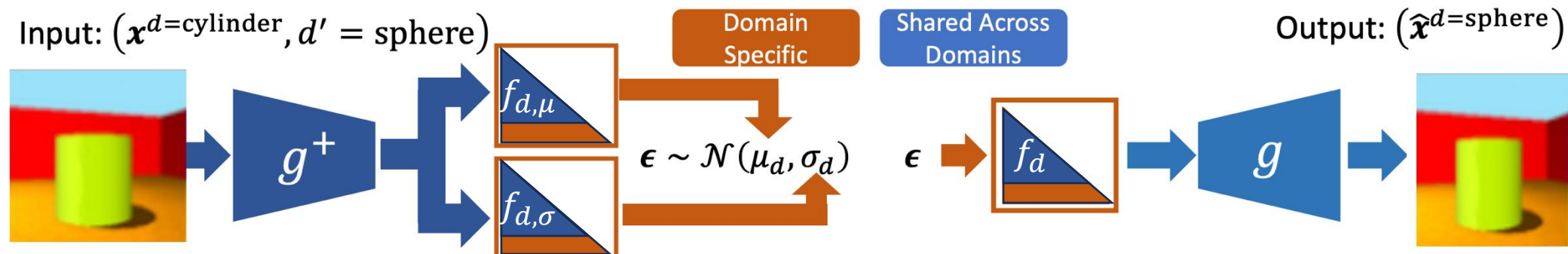


DCF estimation simplifies to optimizing VAE with MLE and sparse intervention constraint

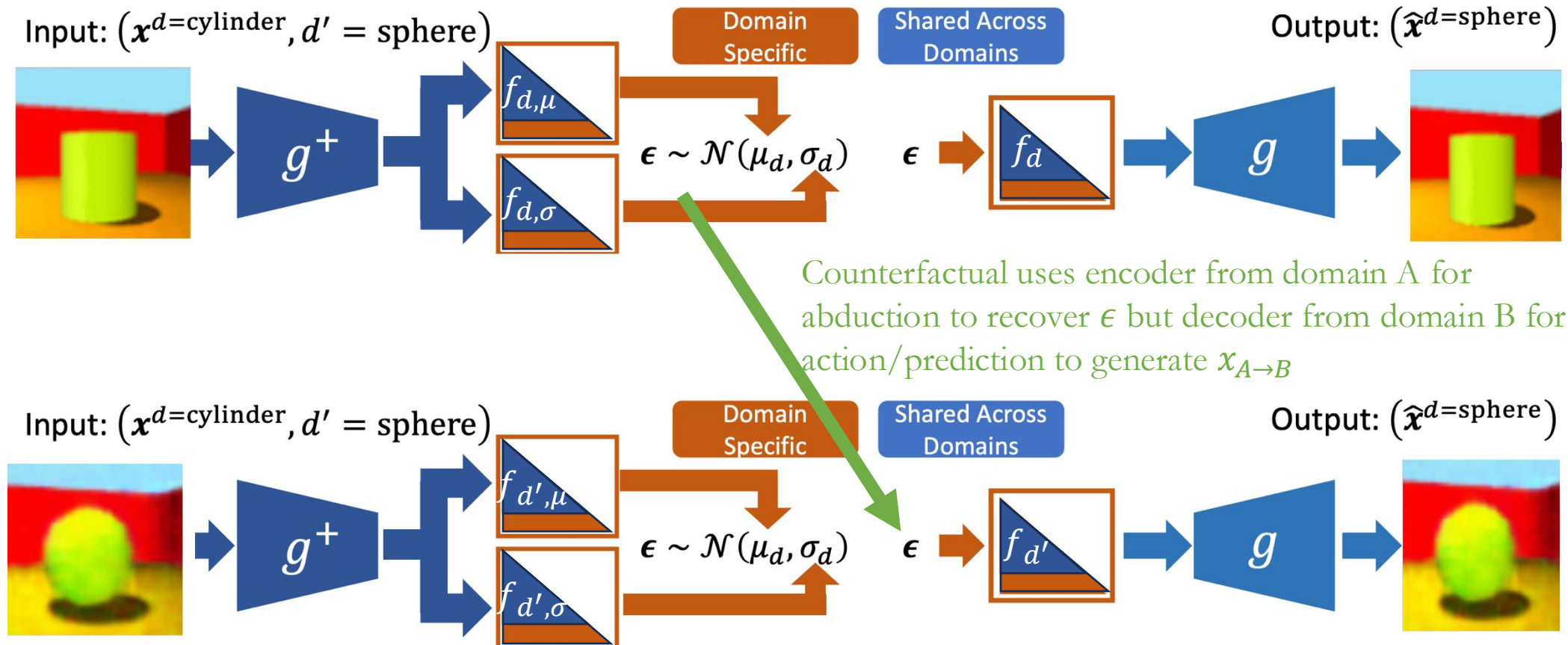
- Our ILD objective given max intervention sparsity k

$$\min_{g, \mathcal{F}} \mathbb{E}_{p(\mathbf{x}, d)} [-\log q_{g, \mathcal{F}}(\mathbf{x}, d)] \quad s. t. [f_d]_{\leq m-k} = [f_{d'}]_{\leq m-k}, \forall d \neq d'.$$

- Normalizing flows or VAEs can be used here
- Shared parameters for g and the first $m - k$ variables of f

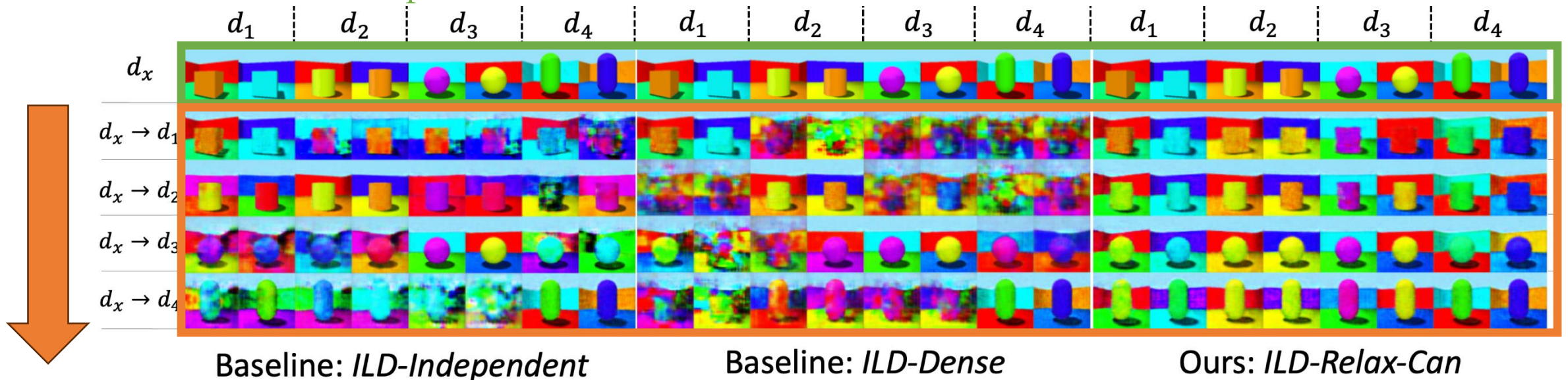


To generate counterfactuals, we use encoder from domain A and decoder from domain B



Qualitatively, our approach preserves changes the domain while preserving other semantic information

2 factual examples from each domain



Counterfactuals - Should only change shape while keeping other variables constant.

Our sparse ILD improves upon all counterfactual metrics compared to baselines

Table 2: Quantitative result for **Composition** (Comp.), **Reversibility** (Rev.), **Preservation** (Pre.), and **Effectiveness** (Eff.), where higher is better. CRMNIST, 3D Shapes, Causal3DIdent are averaged 20, 5, 10 runs respectively. Best models are bold (within 1 standard deviation) and due to space constraint, expanded tables with additional datasets and standard deviation are in Appendix [D.5](#).

	CRMNIST				3D Shapes				Causal3DIdent			
	Comp.	Rev.	Eff.	Pre.	Comp.	Rev.	Eff.	Pre.	Comp.	Rev.	Eff.	Pre.
<i>ILD-Independent</i>	87.24	59.88	94.65	60.39	99.79	32.56	94.97	32.49	88.15	51.43	91.05	51.94
<i>ILD-Dense</i>	88.18	62.29	92.72	59.60	99.76	32.60	80.92	32.64	83.59	49.17	92.17	48.83
<i>ILD-Can</i>	92.10	85.74	94.48	72.95	99.85	79.84	96.72	64.99	86.00	79.73	84.15	79.73

Conclusion and discussion



DCFs can be used for
trustworthy ML
applications

Explaining distribution
shifts
Counterfactual fairness
Domain generalization

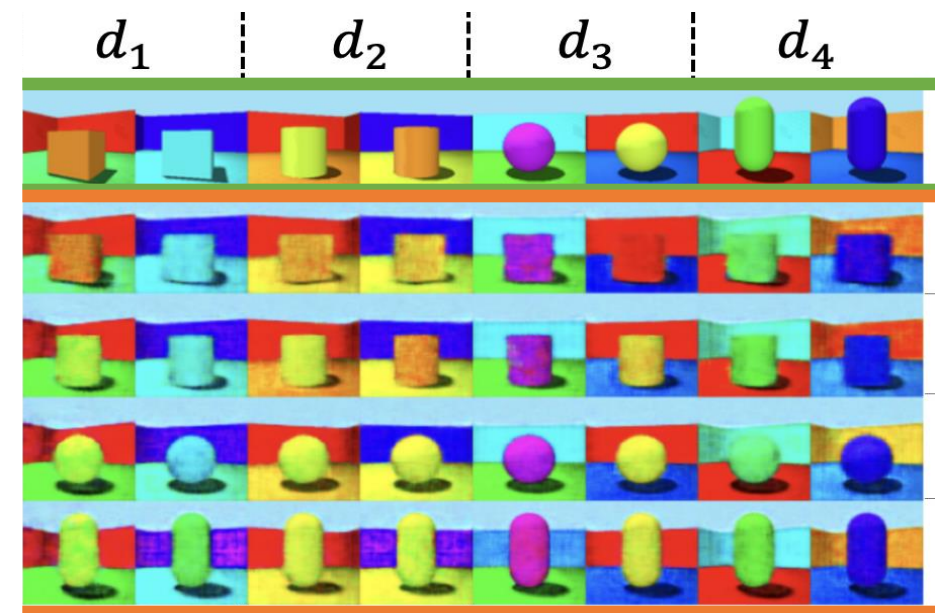
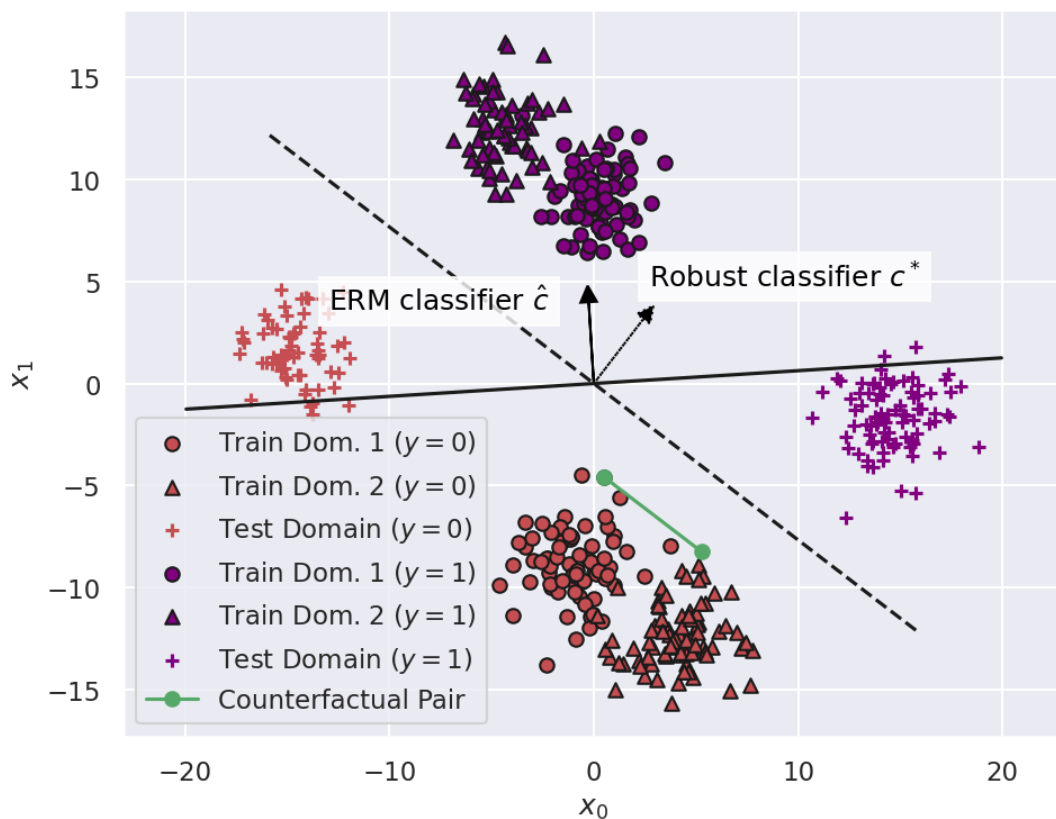


Estimating DCFs is challenging but
potentially feasible in certain
circumstances

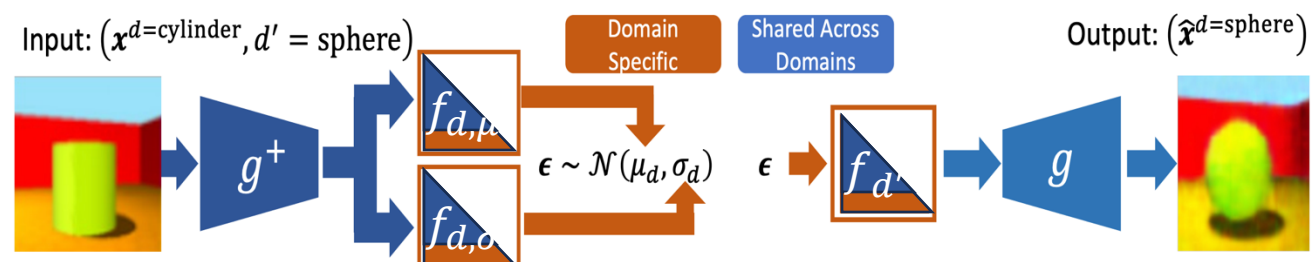


Much more work to be done on the
estimation and use of DCFs!

Thanks for listening!



Ours: *ILD-Relax-Can*



Proof sketch for noisy counterfactual matching theorem

- Decompose into training and counterfactual error

$$\begin{aligned}
 & \mathbb{E}[\|c^T x_{d^+} - y\|^2] \\
 &= \mathbb{E}[\|c^T (x_{d^+} + x_d - x_d) - y\|^2] \\
 &= 2\mathbb{E}[\|c^T x_d - y\|^2] + 2\mathbb{E}[\|c^T (x_{d^+} - x_d)\|^2]
 \end{aligned}$$

- Inflate by $\tilde{U}_r \tilde{U}_r^T$ because of constraint

$$\begin{aligned}
 & \mathbb{E}[\|c^T (x_{d^+} - x_d)\|^2] = c^T M^+ c \\
 &= c^T (I - \tilde{U}_r \tilde{U}_r^T) M^+ (I - \tilde{U}_r \tilde{U}_r^T) c
 \end{aligned}$$

- Notice that true M^+ can be represented by $U_{\tilde{\mathcal{S}}}$

$$\begin{aligned}
 & c^T (I - \tilde{U}_r \tilde{U}_r^T) M^+ (I - \tilde{U}_r \tilde{U}_r^T) c \\
 &= c^T (I - \tilde{U}_r \tilde{U}_r^T) U_{\tilde{\mathcal{S}}} Q \Lambda_{|\tilde{\mathcal{S}}|} Q^T U_{\tilde{\mathcal{S}}}^T (I - \tilde{U}_r \tilde{U}_r^T) c
 \end{aligned}$$

- Use eigen values to extract upper bound

$$\begin{aligned}
 & c^T (I - \tilde{U}_r \tilde{U}_r^T) U_{\tilde{\mathcal{S}}} Q \Lambda_{|\tilde{\mathcal{S}}|} Q U_{\tilde{\mathcal{S}}}^T (I - \tilde{U}_r \tilde{U}_r^T) c \\
 & \leq \|c\|^2 \lambda_1^+ \|(I - \tilde{U}_r \tilde{U}_r^T) U_{\tilde{\mathcal{S}}}\|^2
 \end{aligned}$$

- Use Davis-Kahan bound noticing that oracle version will yield $U_{\tilde{\mathcal{S}}}$

$$\begin{aligned}
 & \|(I - \tilde{U}_r \tilde{U}_r^T) U_{\tilde{\mathcal{S}}}\|^2 \\
 & \leq \frac{\|\tilde{\Delta} \tilde{\Delta}^T - \Delta \Delta^T\|^2}{\min_{r+1 \leq j \leq m, 1 \leq j' \leq |\tilde{\mathcal{S}}|} |\tilde{\lambda}_j - \lambda_{j'}|^2}
 \end{aligned}$$

An ILD model joins a shared observation function g and latent SCMs \mathcal{F}

Defintion 1: An *invertible latent domain causal model* (ILD) (g, \mathcal{F}) joins a shared **invertible** observation function $g: \mathcal{Z} \rightarrow \mathcal{X}$ with a set of N_d domain-specific latent SCMs $\mathcal{F} = \{f_d: \mathbb{R}^m \rightarrow \mathcal{Z}\}_{d=1}^{N_d}$, where f_d are invertible and autoregressive and $\epsilon \sim \mathcal{N}(0, I_m)$.

- Informally, autoregressive means that Jacobian is lower triangular.
- The intervention set $\mathcal{I}(f_d, f_{d'})$ is defined **implicitly**:

$$j \in \mathcal{I}(f_d, f_{d'}) \Leftrightarrow [f_d^{-1}]_j \neq [f_{d'}^{-1}]_j$$

- The intervention set of ILD is defined as $\mathcal{I}(\mathcal{F}) := \bigcup_{d \neq d'} \mathcal{I}(f_d, f_{d'})$
- Two ILDs are **distributionally equivalent** $(g, \mathcal{F}) \simeq_D (g', \mathcal{F}')$ iff

$$p_{\mathcal{N}}(f_d^{-1} \circ g^{-1}(x)) \Big|_{J_{f_d^{-1} \circ g^{-1}}(x)} = p_{\mathcal{N}}(f_{d'}'^{-1} \circ g'^{-1}(x)) \Big|_{J_{f_{d'}'^{-1} \circ g'^{-1}}(x)}, \quad \forall d$$

[Or more formally $(g \circ f_d)_{\#} \mathcal{N}(0, I_m) = (g' \circ f_{d'}')_{\#} \mathcal{N}(0, I_m), \forall d$]

ILD domain counterfactuals are a simple function composition

- ILD domain counterfactuals are a simple composition:

$$x_{d \rightarrow d'} = g \circ f_{d'} \circ f_d^{-1} \circ g^{-1}(x)$$

- Two ILDs are **domain counterfactually equivalent**
 $(g, \mathcal{F}) \simeq_c (g', \mathcal{F}')$ iff all their domain counterfactuals are equal, i.e.,

$$g \circ f_{d'} \circ f_d^{-1} \circ g^{-1} = g' \circ f'_{d'} \circ f'^{-1}_d \circ g'^{-1}, \quad \forall d \neq d'$$

Q1: Is estimating DCFs easier than identifying latent causal representations?

Hypothesis 1: Estimating DCFs is **easier** than estimating latent causal representations.

Counterfactual equivalence characterization proves that identifying causal representations is unnecessary for DCFs

Theorem 1: $(g, \mathcal{F}) \simeq_c (g', \mathcal{F}')$ if and only if there exists invertible functions h_1, h_2 such that:

$$g' = g \circ h_1^{-1} \quad \text{and} \quad f'_d = h_1 \circ f_d \circ h_2, \quad \forall d.$$

- “If” direction is trivial, but “only if” is challenging to prove
- This theorem can be used to
 1. Construct counterfactually equivalent models
 2. Validate if two models are counterfactually equivalent
- Indeed, recovering the latent variables is **unnecessary** because h_1 could be arbitrary

Q2: Can we estimate DCFs by assuming intervention sparsity?

Hypothesis 2: Estimating domain counterfactuals could be feasible with **weaker assumptions**, particularly intervention sparsity.

Counterfactual pseudo-metric measures distance between ILD models w.r.t. DCFs

Definition 2: Given a joint distribution $p(x, d)$, a counterfactual pseudo-metric between two ILDs can be defined as:

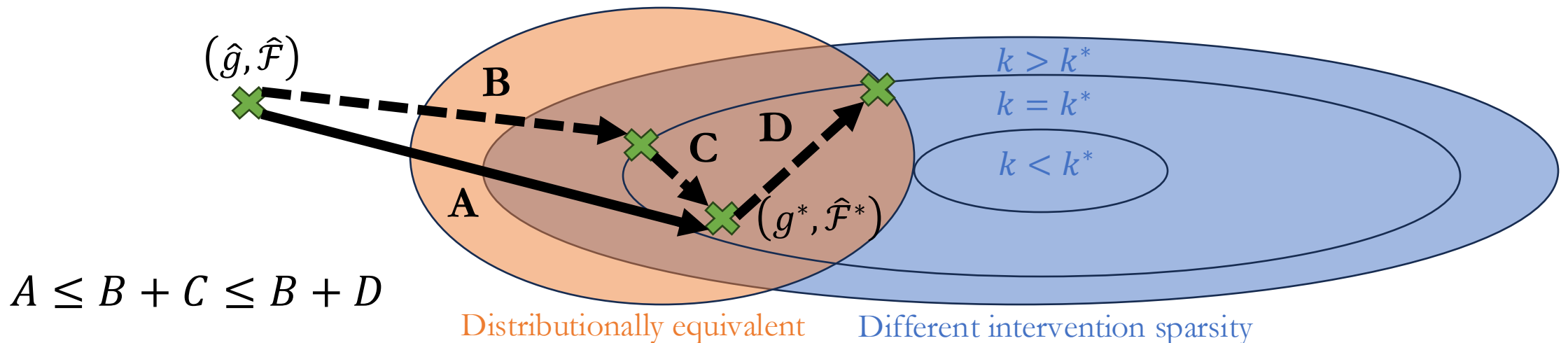
$$d_c((g, \mathcal{F}), (g', \mathcal{F}')) := \sqrt{\mathbb{E}_{p(x, d)p(d')} \left[\|g \circ f_{d'} \circ f_d^{-1} \circ g^{-1}(x) - g' \circ f_{d'} \circ f_d'^{-1} \circ g'^{-1}(x)\|_2^2 \right]}$$

- DCF estimation error is the distance to the ground truth ILD (g^*, \mathcal{F}^*) where $p(x, d)$ is the ground truth distribution:

$$\varepsilon(\hat{g}, \hat{\mathcal{F}}) := d_c((\hat{g}, \hat{\mathcal{F}}), (g^*, \mathcal{F}^*))$$

DCF estimation error is bounded by distribution fit and intervention sparsity

- Informally, the DCF error can be decomposed into two terms:
 - Distribution fit (B) – How far are the generated distributions from the observed distributions?
 - Intervention sparsity (D) – What is the worst case ILD model for a target intervention sparsity k ?



DCF estimation error is bounded by distribution fit and intervention sparsity

Theorem 2 (Counterfactual Error Bound Decomposition). Given a max intervention sparsity $k \geq 0$ and letting

$$\mathcal{M}(k) \triangleq \{(g, \mathcal{F}): (g, \mathcal{F}) \simeq_D (g^*, \mathcal{F}^*), |\mathcal{I}(\mathcal{F})| \leq \max\{k, |\mathcal{I}(\mathcal{F}^*)|\}\},$$

The counterfactual error can be upper bounded as follows:

$$\varepsilon((\hat{g}, \hat{\mathcal{F}})) \leq \underbrace{\min_{(g', \mathcal{F}') \in \mathcal{M}(k)} d_C((\hat{g}, \hat{\mathcal{F}}), (g', \mathcal{F}'))}_{\text{(B) Error due to lack of distribution equivalence}} + \underbrace{\max_{(\tilde{g}, \tilde{\mathcal{F}}) \in \mathcal{M}(k)} d_C((\tilde{g}, \tilde{\mathcal{F}}), (g^*, \mathcal{F}^*))}_{\text{(D) Worst-case error given distribution equivalence}}$$

(B) Error due to lack of distribution equivalence (D) Worst-case error given distribution equivalence

Furthermore, if we assume that the ILD mixing functions are Lipchitz continuous, we can bound the worst-case error (B) as follows:

$$(D) \leq \underbrace{\left[\max_{(\tilde{g}, \tilde{\mathcal{F}}) \in \mathcal{M}(k)} \tilde{k} L_{\tilde{g}}^2 \max_{i \in [m]} \mathbb{E} \left[[\tilde{f}_d(\epsilon) - \tilde{f}_{d'}(\epsilon)]_i^2 \right] \right]}_{\text{Error depends on } k \text{ since } \tilde{k} \leq \max\{k, k^*\}} + \underbrace{\left[k^* L_{g^*}^2 \max_{i \in [m]} \mathbb{E} \left[[f_d^*(\epsilon) - f_{d'}^*(\epsilon)]_i^2 \right] \right]}_{\text{Error only depends on ground truth model}}^{\frac{1}{2}},$$

where $\tilde{k} \equiv |\mathcal{I}(\tilde{\mathcal{F}})|$ and $k^* \equiv |\mathcal{I}(\mathcal{F}^*)|$, and the expectation is over $p(d, d', \epsilon) \triangleq p(d)p(d')p(\epsilon)$, where $p(d) = p(d')$ and $p(\epsilon)$ is a standard normal.

Q3: How do we estimate DCFs practically?

Hypothesis 3: There exist **practical methods** that can estimate domain counterfactuals.

Imposing the sparsity constraint can be challenging

- Given sparsity k , there are $\binom{m}{k}$ sparsity patterns corresponding to different variable subsets
- Naïvely, one could optimize $\binom{m}{k}$ ILD models independently
- However, optimizing over **one** “canonical” sparsity pattern **is sufficient** without loss of generality

For any ILD, an equivalent canonical ILD exists where all interventions are the last variables

Definition 5 (Canonical Domain Counterfactual Model). An ILD (g, \mathcal{F}) is a canonical domain counterfactual model (canonical ILD), denoted by $(g, \mathcal{F}) \in \mathcal{C}$, if and only if the last variables are intervened, i.e.,

$$(g, \mathcal{F}) \in \mathcal{C} \Leftrightarrow \mathcal{I}(\mathcal{F}) = \{m - j : 0 \leq j \leq |\mathcal{I}(\mathcal{F})|\}.$$

Theorem 3 (Existence of Equivalent Canonical ILD). Given an ILD (g, \mathcal{F}) , there exists a canonical ILD that is both counterfactually and distributionally equivalent to (g, \mathcal{F}) while maintaining the size of the intervention set, i.e.,

$$\forall (g, \mathcal{F}), \exists (g', \mathcal{F}') \in \mathcal{C} \text{ s.t. } (g', \mathcal{F}') \simeq_{C,D} (g, \mathcal{F}) \text{ and } |\mathcal{I}(\mathcal{F})| = |\mathcal{I}(\mathcal{F}')|.$$

Thus, we can optimize over canonical ILDs without loss of generality.