Invariance Principle meets Information Bottleneck for OOD generalization

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet

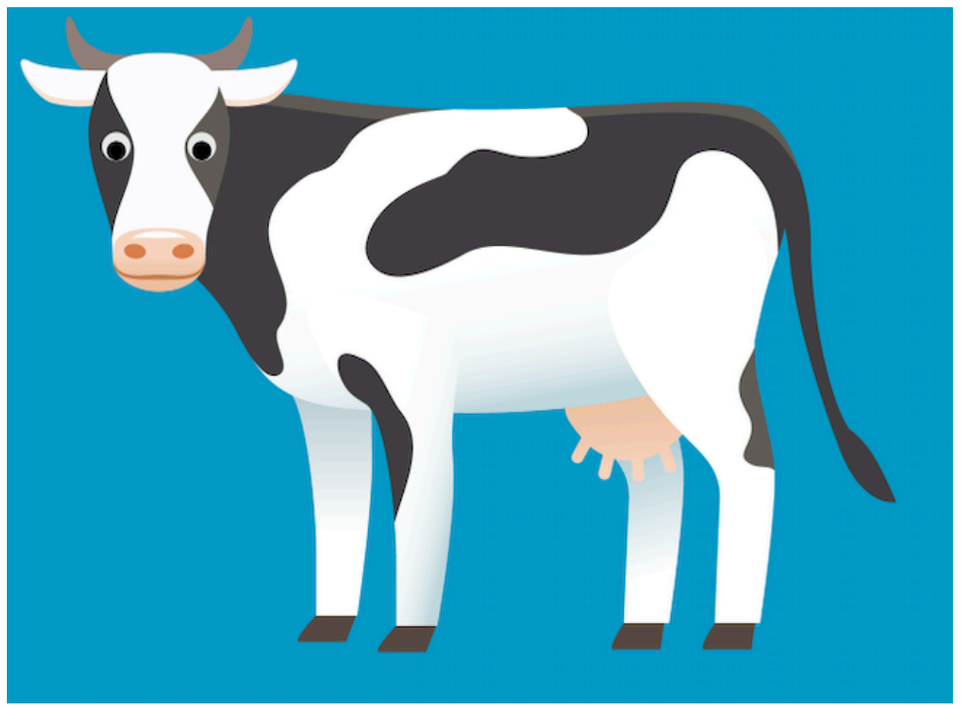
Yoshua Bengio, Ioannis Mitliagkas, Irina Rish

Mila

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Typical cow: green background



Atypical cow: blue background

Deep neural networks use background color to identify cow!

Mila

Explaining the failures
Correlation vs. Causation

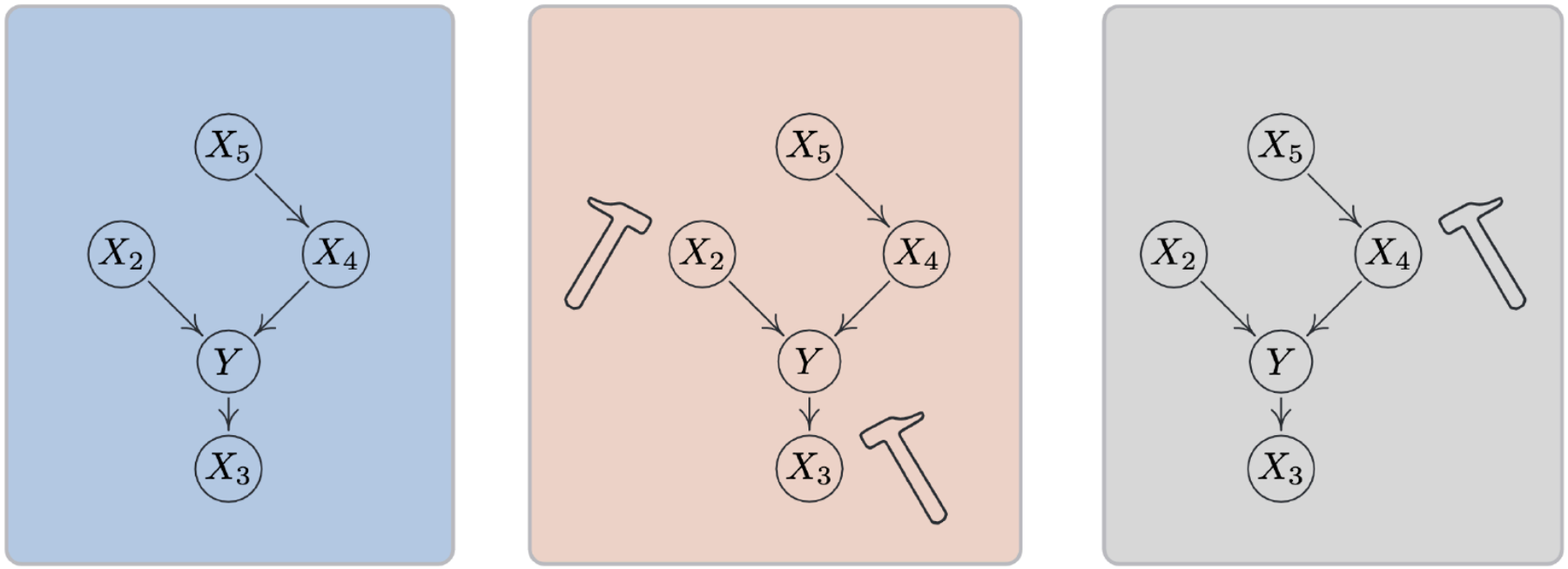
Human → Uses causes (foreground) to label cow

ERM-based model → Uses correlation (background) to label cow

Goal: Ideal model should use **causes** to make predictions

Mila

Invariance Principle & Causation



Illustrating valid interventions (Peters et al.)

Intervention: Models distribution shifts (e.g., images from different locations)

Invariance principle and causation:

Y is caused by $X_S \iff \mathbb{P}(Y|X_S)$ is invariant across all interventions (except on Y)

Mila

Problem Formulation

- Data is gathered from multiple environments (interventions)
- $\mathscr{E}_{tr}$ — training environments
- $\mathscr{E}_{all}$ — test (all) environments

- Construct a predictor $f : \mathscr{X} \to \mathscr{Y}$

- $R^e(f) = \mathbb{E}[\ell(f(X^e), Y^e)]$ is the risk achieved by the predictor in environment e

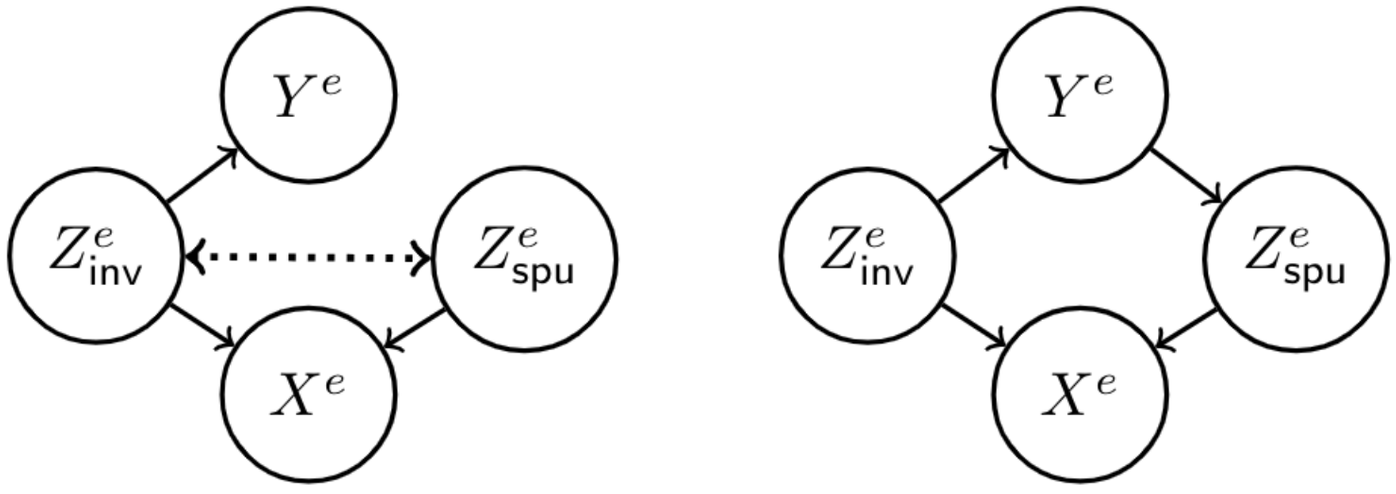
OOD generalization objective

$$\min_{f} \max_{e \in \mathscr{E}_{all}} R^e(f)$$

Mila

Optimal Causal Predictor is OOD optimal

Data generation process



Fully informative invariant feature (FIIF)

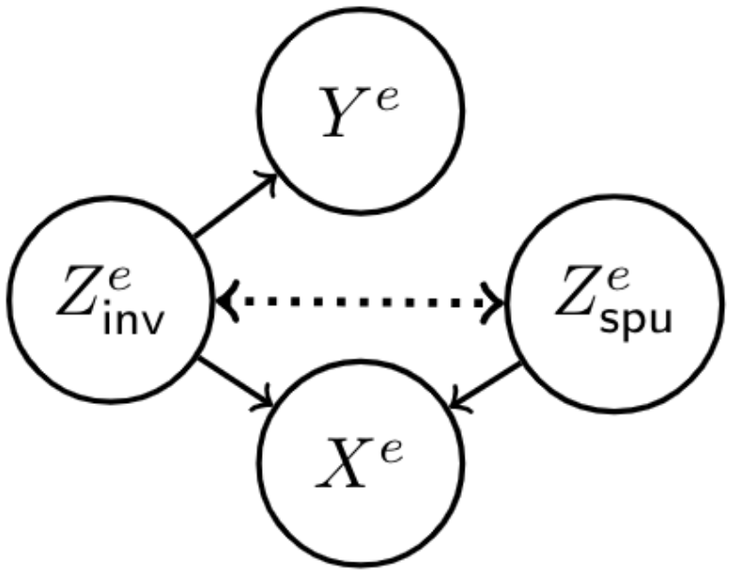
Partially informative invariant feature (PIIF)

Theorem [Arjovsky et al., Ahuja et al.]
Optimal predictor that only relies on causes of the label is OOD optimal.

Challenge. How to learn the OOD optimal predictor from a few training distributions?

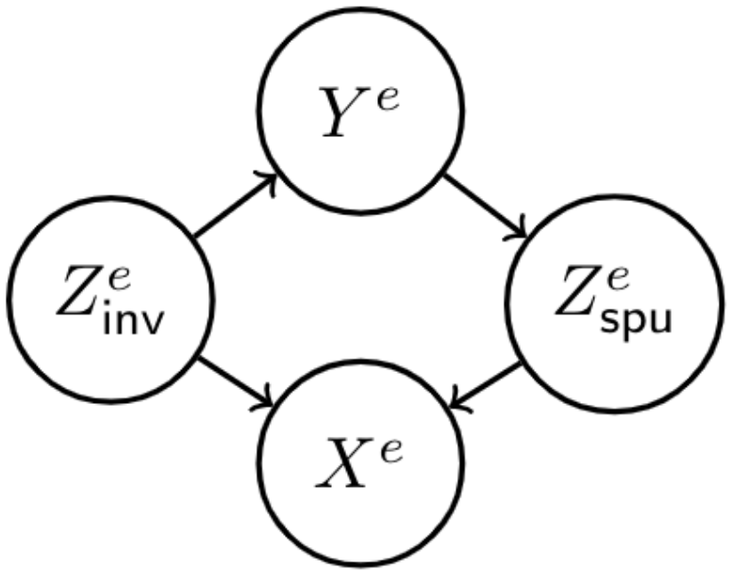
Mila

Failure of ERM and IRM



Fully informative invariant feature (FIIF)

IRM and ERM both fail!

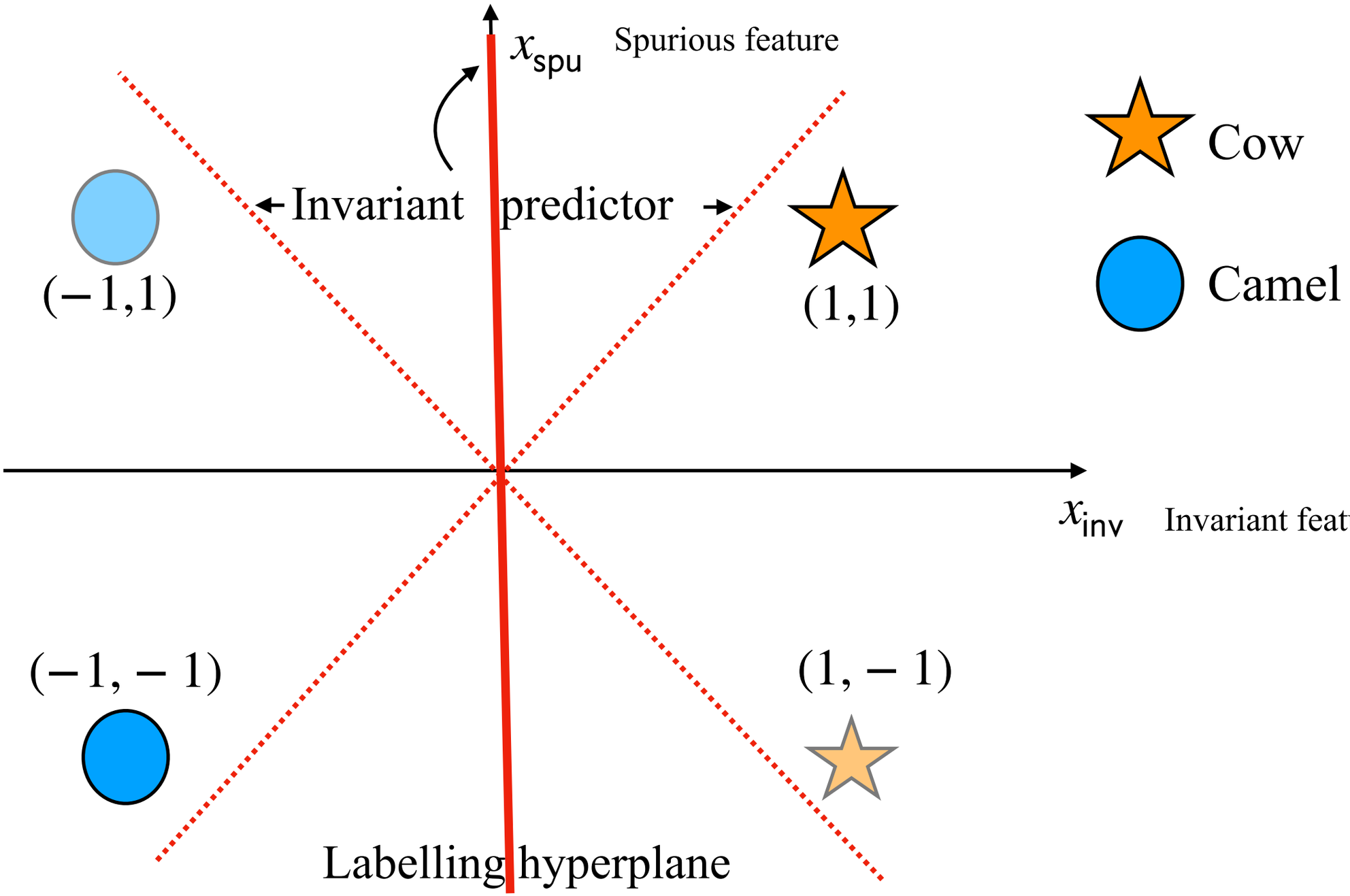


Partially informative invariant feature (PIIF)

ERM fails and IRM can succeed!

Mila

Linear Classification (FIIF)



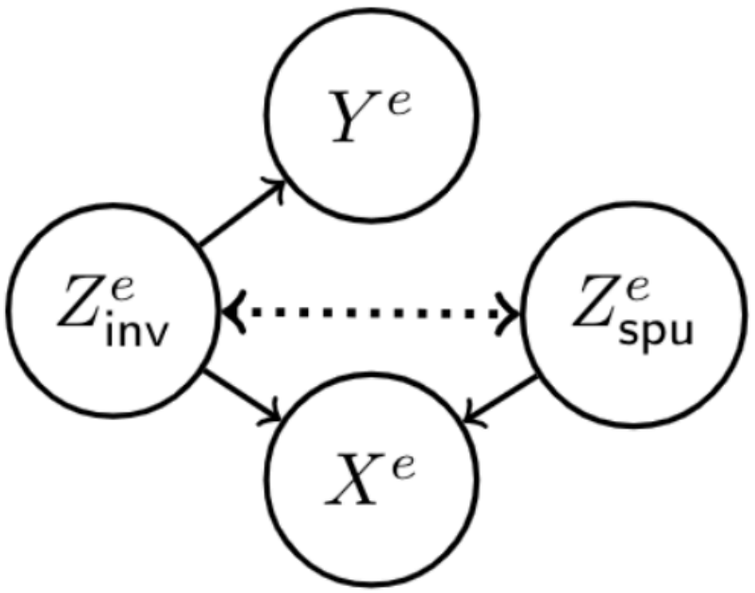
Example based on [Nagarajan et al.]

OOD Generalization for Linear Classification

- **Linear classification SEM for environment** $e \in \mathscr{E}_{all}$

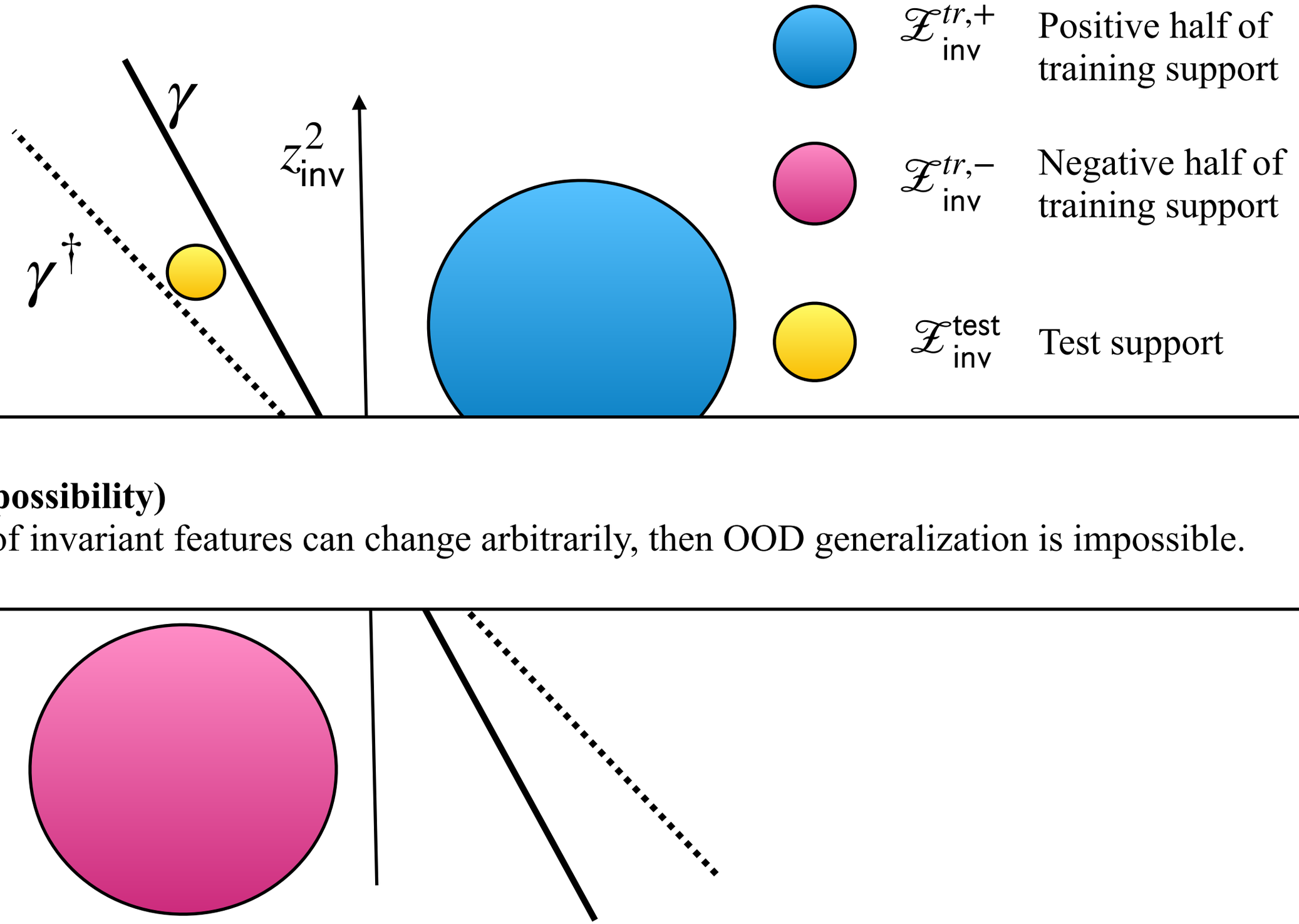
$$X^e \leftarrow S(Z^e_{\text{inv}}, Z^e_{\text{spu}})$$

$$Y^e \leftarrow \mathtt{I}(\gamma^\top Z^e_{\text{inv}}) \oplus N^e,\ N^e \sim \text{Bernoulli}(q),\ N^e \perp (Z^e_{\text{inv}}, Z^e_{\text{spu}})$$



Fully informative invariant feature (FIIF)

Interventions for Classification Tasks



Theorem (Impossibility)
If the support of invariant features can change arbitrarily, then OOD generalization is impossible.

Mila

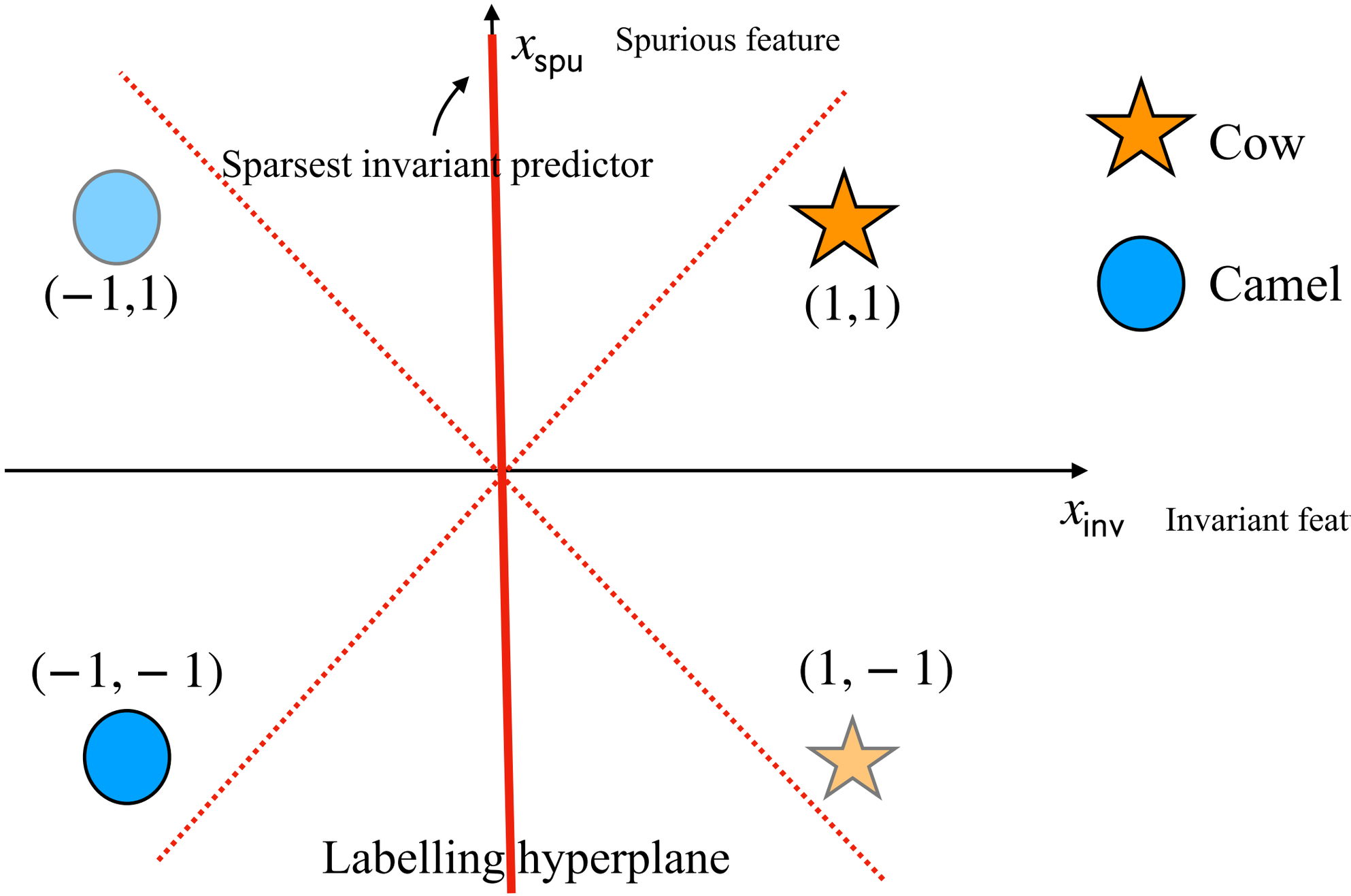
Role of Support Overlap in OOD generalization

Theorem (Failure of ERM & IRM): If the support of invariant features do not change but the spurious features can change, then **ERM and IRM fail** to achieve OOD optimality.

Theorem (Success of ERM & IRM): If the support for both invariant features and spurious features do not change, then **ERM and IRM** achieve OOD optimality.

Mila

Invariance + Sparsity Constraints



Mila

Invariance + Information Bottleneck

- **Latent features not accessible thus cannot impose sparsity directly!**
- **Mutual information** $I(\phi(X); X)$

Information-bottleneck based IRM (IB-IRM)

$$\min_{w,\Phi} \sum_{e \in \mathscr{E}_{tr}} h^e(w \circ \phi(X^e))$$

$$\sum_e R^e(w \circ \phi) \leq r^{\text{th}}$$

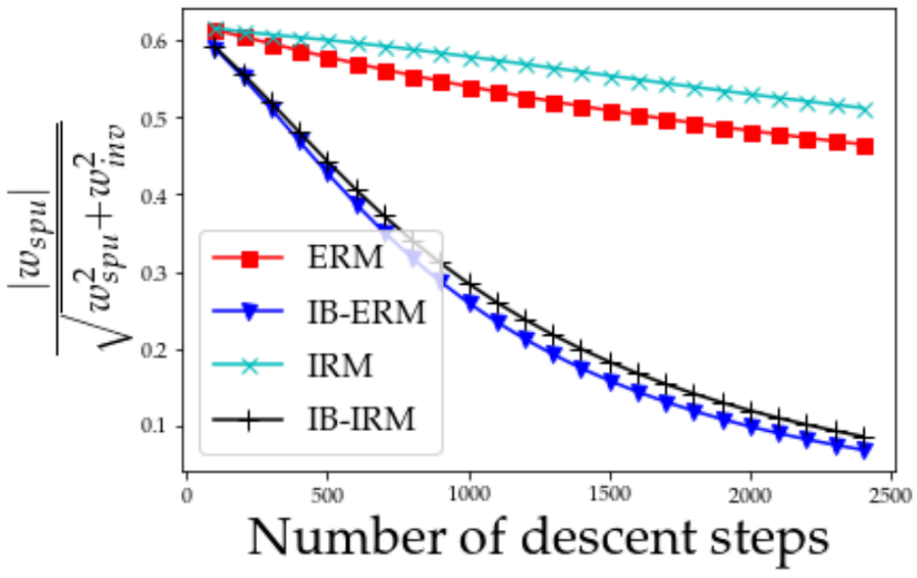
$w \circ \phi$ is an invariant predictor

Theorem (Success of IB-IRM): If the support of invariant features do not change but the spurious features can change, then **IB-IRM** achieves OOD optimality.

IB-IRM Objective

$$R^{IB-IRM}(\phi) = \sum_{e \in \mathscr{E}_{tr}} R^e(\Phi) \quad + \lambda \|\nabla_{w=1.0} R^e(w \cdot \phi)\|^2 \quad + \nu \mathsf{Var}(\phi)$$

ERM loss **IRMv1 penalty** **IB penalty**



Theorem: Continuous time gradient descent on IB-ERM converges exponentially faster than ERM.

Mila

Experiments

Dataset	ERM	IB-ERM	IRM	IB-IRM
Example 1 (Regression PIIF)	13.36	12.96	11.15	11.68
Example 1S (Regression PIIF)	13.33	12.92	11.02	11.74
Example 2 (Classification FIIF)	0.42	0.00	0.45	0.00
Example 2S (Classification FIIF)	0.45	0.00	0.45	0.06

Error (MSE or classification error) comparisons on linear unit tests

Experiments

Dataset	ERM	IB-ERM	IRM	IB-IRM
CS-CMNIST (FIIF)	60.27	71.8	61.49	**71.79**
AC-CMNIST (PIIF)	16.82	50.84	66.98	**67.67**
Terra incognita (FIIF)	49.80	**56.40**	54.60	54.10
COCO (FIIF)	22.70	**31.66**	18.47	25.10

Accuracy comparisons on colored MNIST variants

Mila

Thank you!

Mila