The Risks of Invariant Risk Minimization

Elan Rosenfeld

Pradeep Ravikumar

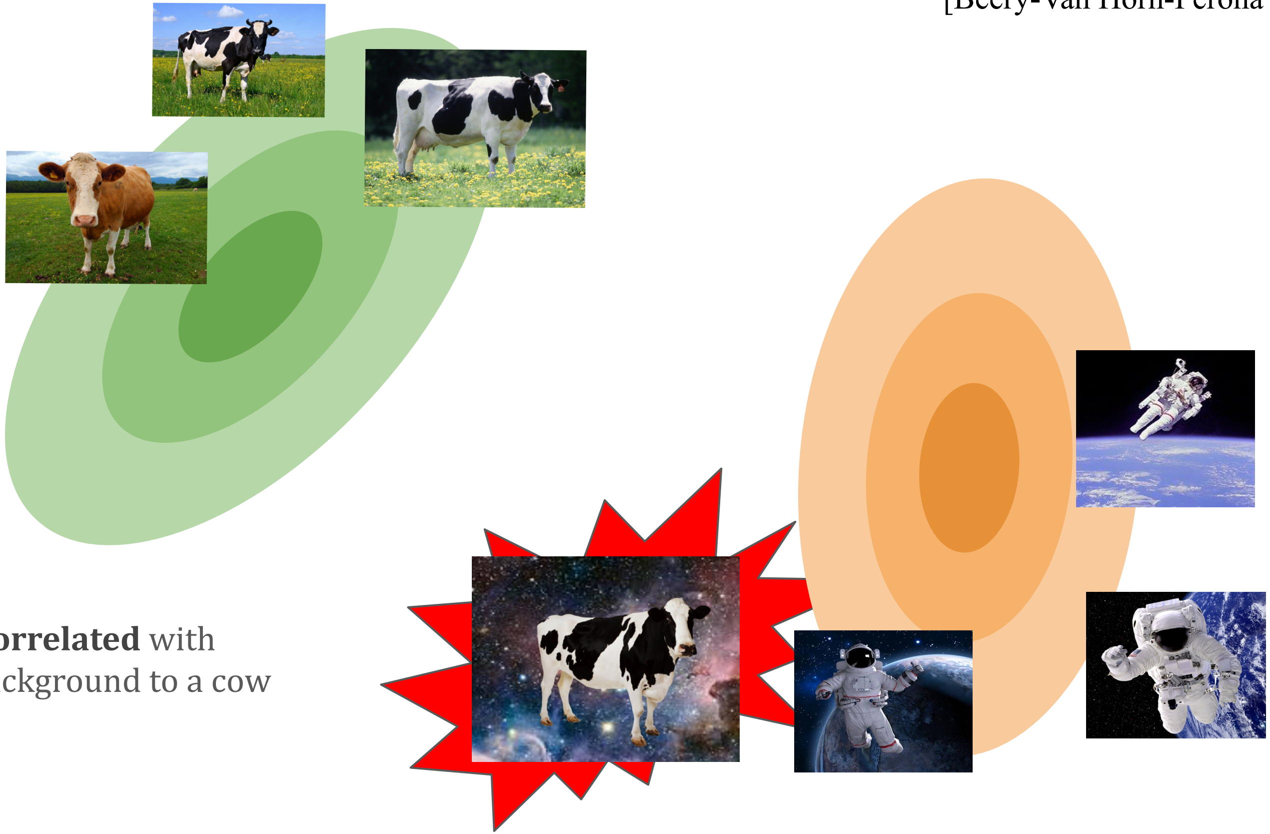
Andrej Risteski

Carnegie Mellon University

Goal: Out-of-Distribution Generalization

(Loosely based on)
[Beery-Van Horn-Perona'18]

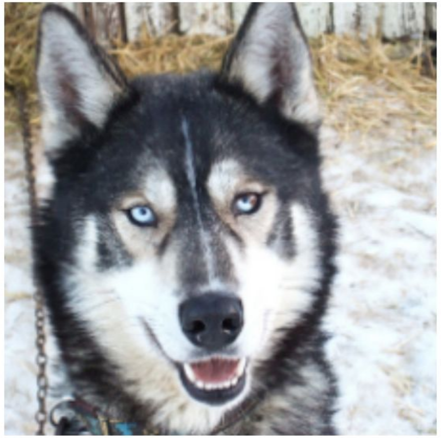
Consider the following classification problem:



"Space background" is **strongly correlated** with "astronaut". But adding a space background to a cow **doesn't make it an astronaut.**

Goal: Out-of-Distribution Generalization

[Ribeiro-Singh-Guestrin'16]



(a) Husky classified as wolf

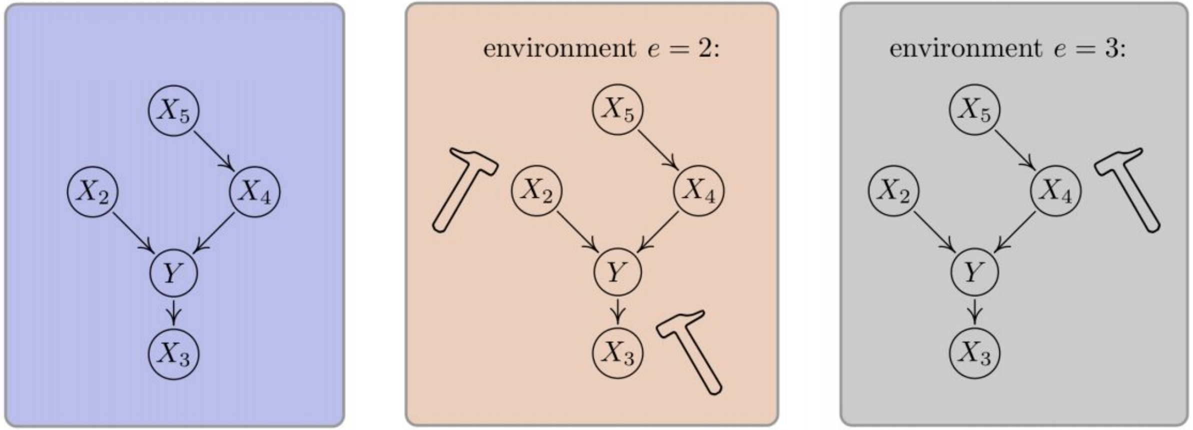
(b) Explanation

Key Question: How can a classifier generalize to distributions where these correlations **do not hold** (or are even **reversed**)?

Invariant Causal Prediction (ICP)

[Peters–Bühlmann–Meinshausen’15]

Assume some Structural Causal Model (SCM): $X_i = f_i(\text{Parents}(X_i); \epsilon_i)$



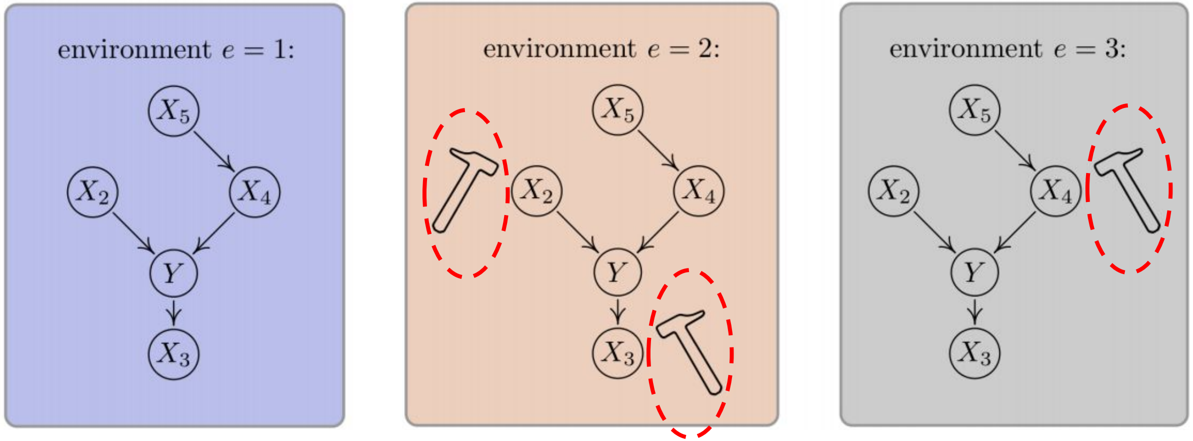
Further assume that the data can be partitioned into **environments**.

Subsets generated by the **same SCM**, but each has a distinct **intervention** on some of the covariates.

Invariant Causal Prediction (ICP)

[Peters–Bühlmann–Meinshausen’15]

Assume some Structural Causal Model (SCM): $X_i = f_i(\mathrm{Parents}(X_i); \epsilon_i)$



Further assume that the data can be partitioned into **environments**.

Subsets generated by the **same SCM**, but each has a distinct **intervention** on some of the covariates.

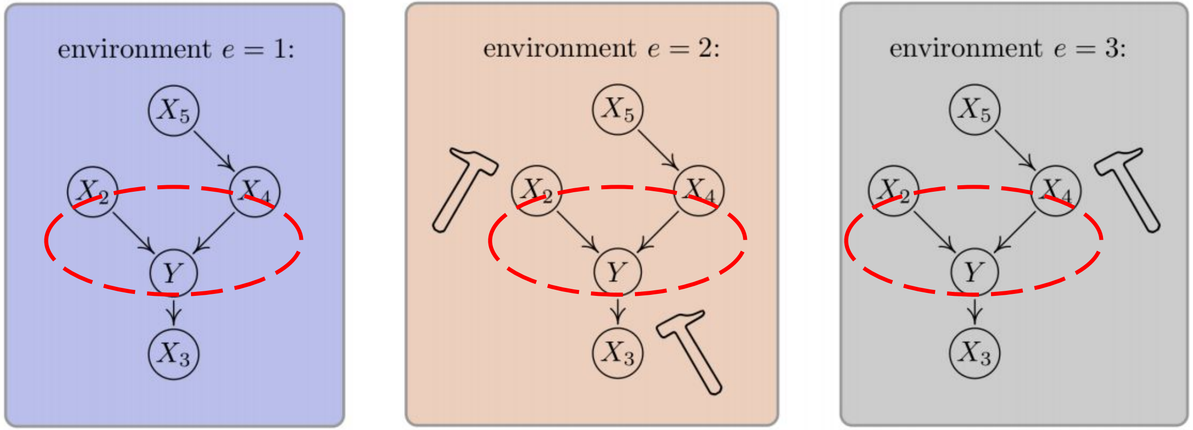
Arbitrary modification of the distribution of some covariates, maintaining all other functional mechanisms.

Invariant Causal Prediction (ICP)

[Peters–Bühlmann–Meinshausen'15]

Assume some Structural Causal Model (SCM): $X_i = f_i(\text{Parents}(X_i); \epsilon_i)$

P(*Y* | *Parents*(*Y*)) is **invariant**.



If we predict using only the direct parents of the target, predictor is **minimax**.

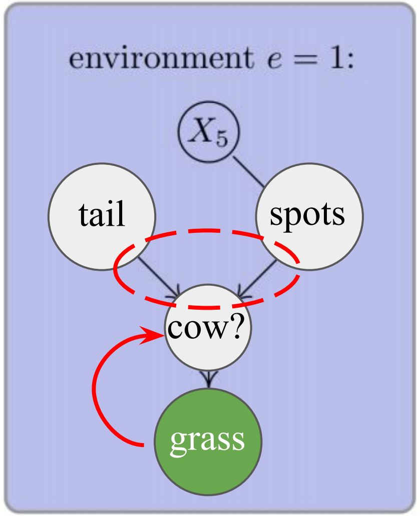
For **any predictor** which uses a non-parent, $\exists$ an intervention which causes it to fail.

Invariant Causal Prediction (ICP)

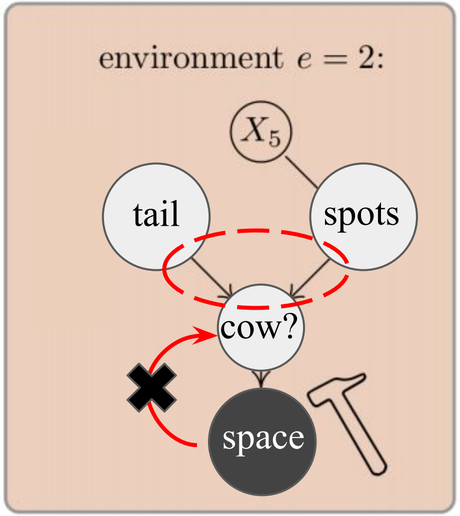
[Peters–Bühlmann–Meinshausen’15]

Returning to our example of cows vs. astronauts...

Train



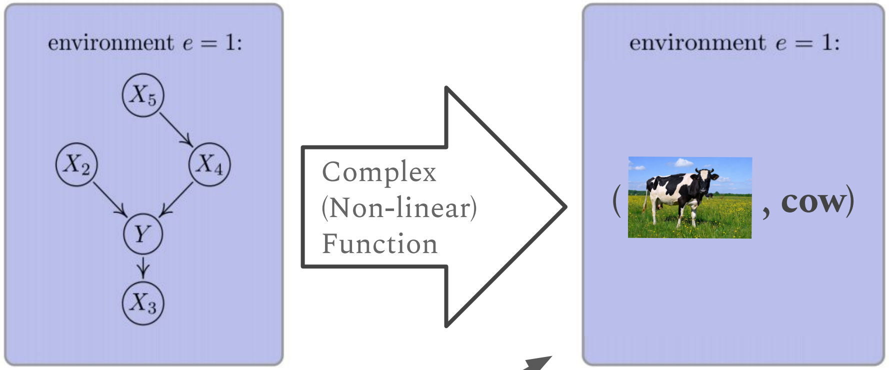
Test



Deep Invariant Feature Learning

Works well for fully observed features.

What about when covariates are **unobserved**?



Plenty of algorithms for provably uncovering structure here.

Not so here.

New Question: How can we identify the invariant features when the covariates are **latent**?

Deep Invariant Feature Learning

This talk **does not attempt** to answer this question.

Rather, this talk shows that

existing objectives intended to solve this problem
do not behave as expected.

New Question: How can we identify the invariant features when the covariates are **latent**?

Deep Invariant Feature Learning

This talk **does not attempt** to answer this question.

Rather this talk shows that

We prove that solving the proposed objectives can **rarely, if ever** ensure outperforming ERM at test time.

New Question: How can we identify the invariant features when the covariates are **latent**?

Outline

1. IRM and Variations
2. Latent Variable Model
3. Formal Results
   a. Linear Setting
   b. Non-Linear Setting

Outline

1. IRM and Variations
2. Latent Variable Model
3. Formal Results
   a. Linear Setting
   b. Non-Linear Setting

Invariant Risk Minimization (IRM)

[Arjovsky–Bottou–Gulrajani–Lopez-Paz'19]

⚠ None of this is intended to be formal!

Consider a feature embedder $\boldsymbol{\Phi}$ and regression vector $\boldsymbol{\beta}$.

Linear Regression: $E[Y \mid \Phi(x)] = \beta^T \Phi(x)$

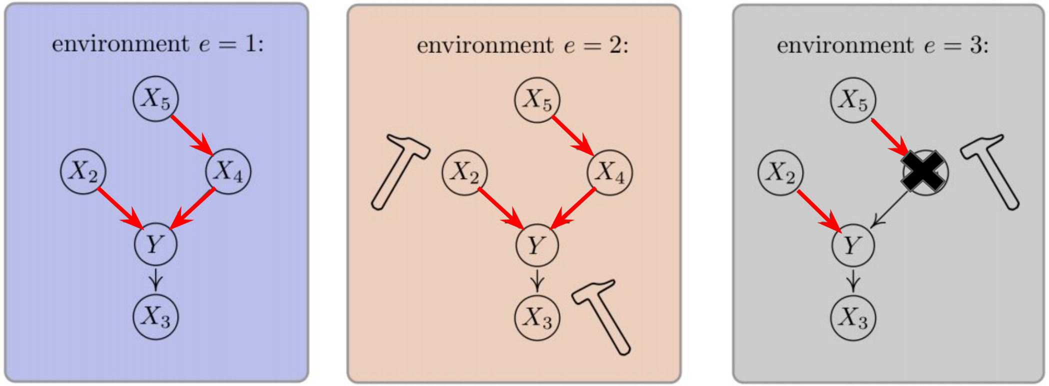
Intervention is *arbitrary*.

$\Phi(X) = [X_2, X_5]$

$E[Y \mid \Phi(X)] = \beta_1^{*T}[X_2, X_5]$

$E[Y \mid \Phi(X)] = \beta_2^{*T}[X_2, X_5]$

$E[Y \mid \Phi(X)] = \beta_3^{*T}[X_2, X_5]$



$$\beta_1^* = \beta_2^* \neq \beta_3^*$$

Invariant Risk Minimization (IRM)

[Arjovsky–Bottou–Gulrajani–Lopez-Paz'19]

⚠ None of this is intended to be formal!

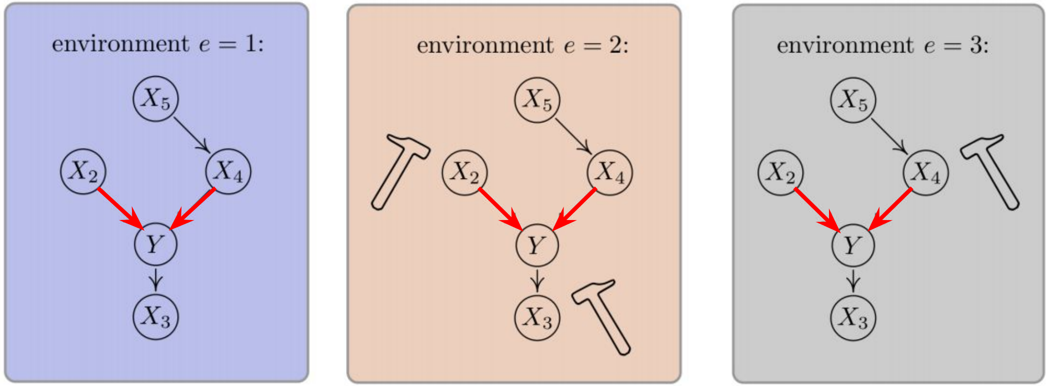
Consider a feature embedder $\mathbf{\Phi}$ and regression vector $\boldsymbol{\beta}$.

Linear Regression: $E[Y \mid \Phi(x)] = \beta^T \Phi(x)$

$\cancel{\Phi(X) = [X_2, X_5]}$

$\Phi(X) = [X_2, X_4]$

Optimal vector is
equal in all environments!



$$\beta_1^* = \beta_2^* = \beta_3^*$$

Invariant Risk Minimization (IRM)

[Arjovsky–Bottou–Gulrajani–Lopez-Paz'19]

⚠ None of this is intended to be formal!

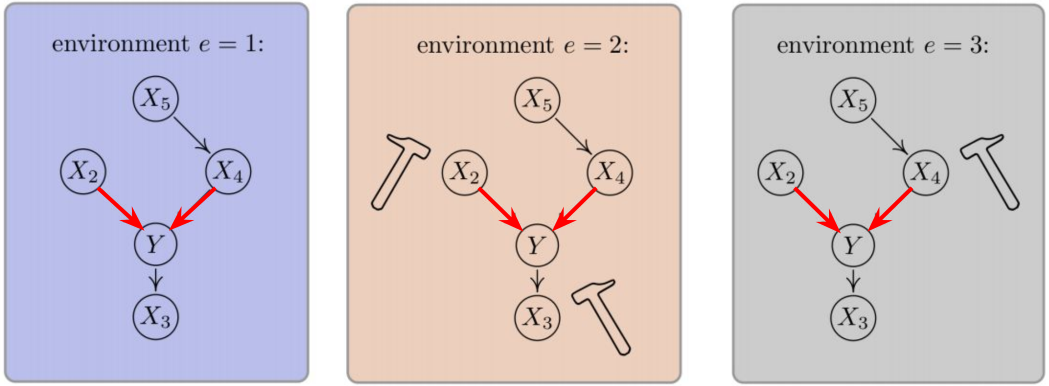
Consider a feature embedder $\boldsymbol{\Phi}$ and regression vector $\boldsymbol{\beta}$.

Linear Regression: $E[Y \mid \Phi(x)] = \beta^T \Phi(x)$

$\cancel{\Phi(X) = [X_2, X_5]}$

$\Phi(X) = [X_2, X_4]$

Optimal vector is equal in all environments!



Expect E[Y | $\boldsymbol{\Phi}(X)$] is **invariant** if and only if $\boldsymbol{\Phi}$ recovers the invariant features.

Invariant Risk Minimization (IRM)

[Arjovsky–Bottou–Gulrajani–Lopez-Paz'19]

Idea: The optimal $\boldsymbol{\beta}$ should be **the same** for all environments.

Standard ERM risk term

IRM Objective to enforce this:

$$\min_{\Phi,\,\beta} \sum_{e\in\mathcal{E}} R^e(\beta\circ\Phi)$$

$$s.t.\ \beta \in \arg\min_{\hat\beta} R^e\left(\hat\beta\circ\Phi\right),\ \forall e\in\mathcal{E}$$

Invariance requirement

Side note: This is *not* just "regularized ERM".

(Problems with) Invariant Risk Minimization

$$\min_{\Phi,\,\beta} \sum_{e\in\mathcal{E}} R^e(\beta \circ \Phi)$$

$$s.t.\ \beta \in \arg\min_{\hat{\beta}} R^e\left(\hat{\beta} \circ \Phi\right),\ \forall e \in \mathcal{E}$$

- ❖ One formal result regarding solution invariance
  - ➢ More of a motivation than a justification
  - ➢ Only for **fully linear** regression
  - ➢ Could require as many environments as **ambient dimension** (think images)

- ❖ What about other forms of invariance? **Possible misspecification**
  - ➢ Lots of suggested variations
  - ➢ *Still* no rigorous analysis...

[Gulrajani & Lopez-Paz'21]
(extensive experiments suggest that none of these beat ERM)

Second Moment Invariance	Feature Distribution Invariance
REx [KCJZ+'20] RVP [XCLL'20] Gradient Norm [BS'20]	Causal Matching [MTS'20] MMD/KL Penalty [GZLK'21]

Outline

1. IRM and Variations
2. Latent Variable Model
3. Formal Results
    a. Linear Setting
    b. Non-Linear Setting

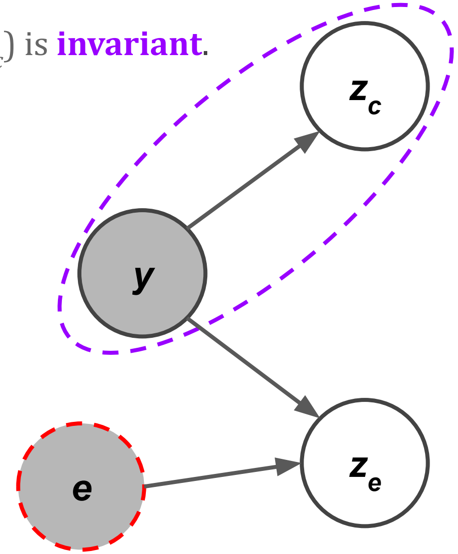
A Formal Model of Latent Invariant Features

For each environment *e*,

$$y = \begin{cases} 1 & w.p.\ \eta \\ -1 & otherwise \end{cases}$$

$$z = \begin{bmatrix} z_c \\ z_e \end{bmatrix} \begin{cases} z_c \sim \mathcal{N}(\mu_c(y),\, \Sigma_c) \\ z_e \sim \mathcal{N}(\mu_e(y),\, \Sigma_e) \end{cases}$$

p(y, z_c) is **invariant**.



We'll call these features "invariant"

And these "environmental"

A Formal Model of Latent Invariant Features

For each environment *e*,

$$y = \begin{cases} 1 & w.p.\ \eta \\ -1 & otherwise \end{cases}$$

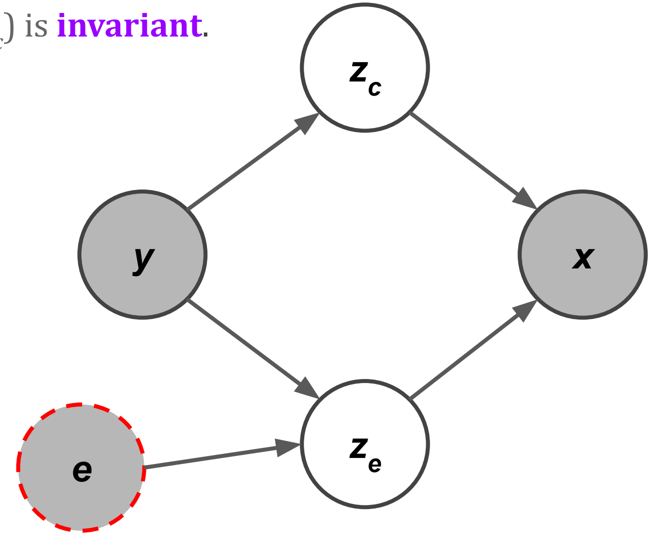
$$z = \begin{bmatrix} z_c \\ z_e \end{bmatrix} \begin{cases} z_c \sim \mathcal{N}(\mu_c(y),\, \Sigma_c) \\ z_e \sim \mathcal{N}(\mu_e(y),\, \Sigma_e) \end{cases}$$

$$x = f(z)$$

Our model allows for *any* invertible *f*.

f can be **non-linear** and **arbitrarily complex**.

p(y, z_c) is **invariant**.



Generalization of the "Gaussian model"

[Schmidt–Santurkar–Tsipras–Talwar–Mądry'18]

[Sagawa–Raghunathan–Koh–Liang'20]

A Formal Model of Latent Invariant Features

For each environment *e*,

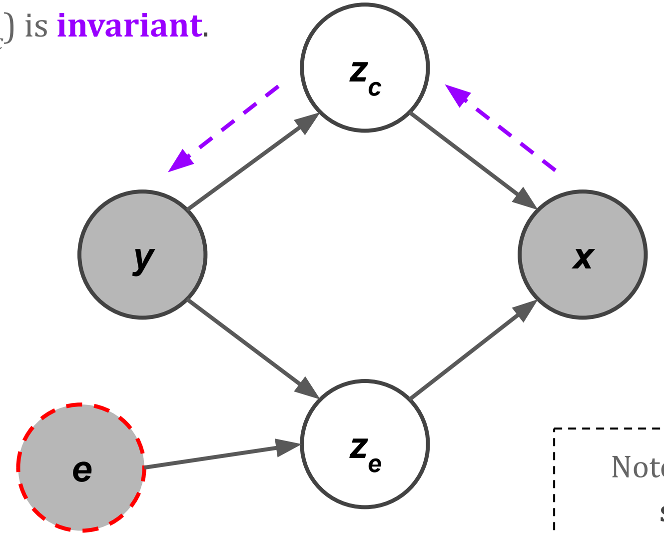
$$y = \begin{cases} 1 & w.p.\ \eta \\ -1 & otherwise \end{cases}$$

$$z = \begin{bmatrix} z_c \\ z_e \end{bmatrix} \begin{cases} z_c \sim \mathcal{N}(\mu_c(y),\, \Sigma_c) \\ z_e \sim \mathcal{N}(\mu_e(y),\, \Sigma_e) \end{cases}$$

$$x = f(z)$$

$$d_e := \dim(z_e) \qquad \left[d_e \gg E \right]$$

p(y, z_c) is **invariant**.



Note that **this is purely a statistical model**.

We assume a finite number of environments ***E***, but infinite observations from each environment.

A Formal Model of Latent Invariant Features

For each environment *e*,

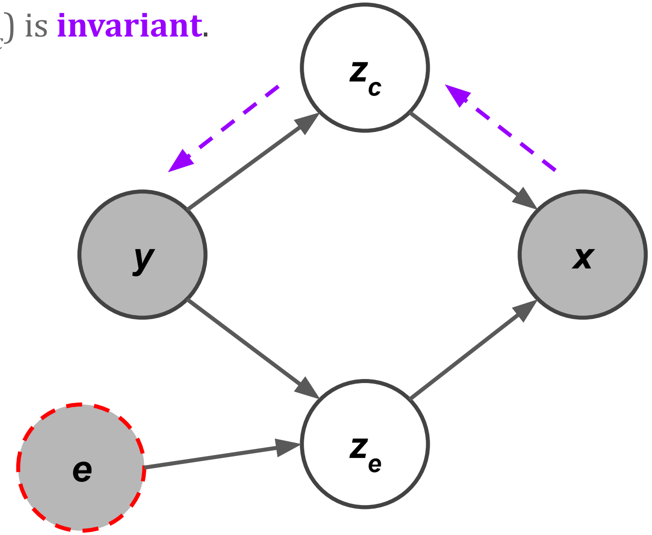
$$y = \begin{cases} 1 & w.p.\ \eta \\ -1 & otherwise \end{cases}$$

$$z = \begin{bmatrix} z_c \\ z_e \end{bmatrix} \begin{cases} z_c \sim \mathcal{N}(\mu_c(y), \Sigma_c) \\ z_e \sim \mathcal{N}(\mu_e(y), \Sigma_e) \end{cases}$$

$$x = f(z)$$

$d_e := \dim(z_e)$ $\quad \left[d_e \gg E \right]$

p(y, z_c) is **invariant**.



$$\boxed{\text{Ultimate goal: } \Phi^*(x) = z_c}$$

$$\beta^* = \arg\min_\beta R(\beta \circ \Phi^*)$$

We call $\mathbf{\Phi}^*, \boldsymbol{\beta}^*$ the "Optimal Invariant Predictor" (OIP)

Outline

1. IRM and Variations
2. Latent Variable Model
3. Formal Results
   a. Linear Setting
   b. Non-Linear Setting

Linear Setting

Regressing on *learned* features $\hat{z} = \Phi(x) = \Phi(f(z))$

Restrict *f*, $\boldsymbol{\Phi}$ to be **linear**

Thus, our features can be written $\hat{z} = \Phi F z = A z_c + B z_e$

If we want **feature invariance**, our goal should be **B = 0**.

To capture all invariant information, A should be full rank.

Linear Setting

Recall: $\hat{z} = \Phi F z = A z_c + B z_e$

Theorem: *Suppose we observe* ***E*** *environments. Then under minor non-degeneracy conditions:*

Linear Setting

Recall: $\hat{z} = \Phi F z = A z_c + B z_e$

Theorem: *Suppose we observe* $\boldsymbol{E}$ *environments. Then under minor non-degeneracy conditions:*

- If $\boldsymbol{E} \leq \boldsymbol{d_e}$, *there is a* ***feasible*** *linear* $\boldsymbol{\Phi}$

Linear Setting

Recall: $\hat{z} = \Phi F z = A z_c + B z_e$

Theorem: *Suppose we observe **E** environments. Then under minor non-degeneracy conditions:*

- If $\boldsymbol{E} \leq \boldsymbol{d}_e$, *there is a **feasible** linear* $\boldsymbol{\Phi}$ *inducing **B** with **non-zero rank***

Not **minimax!**

Linear Setting

Recall: $\hat{z} = \Phi F z = A z_c + B z_e$

Theorem: *Suppose we observe* ***E*** *environments. Then under minor non-degeneracy conditions:*

- If $\boldsymbol{E \leq d_e}$, *there is a* ***feasible*** *linear* $\boldsymbol{\Phi}$ *inducing* ***B*** *with* **non-zero rank** *and* ***lower training risk*** *than the Optimal Invariant Predictor.*

Preferable solution.

Linear Setting

Recall: $\hat{z} = \Phi F z = A z_c + B z_e$

Theorem: *Suppose we observe* ***E*** *environments. Then under minor non-degeneracy conditions:*

- If $\boldsymbol{E \leq d_e}$, *there is a* ***feasible*** *linear* $\boldsymbol{\Phi}$ *inducing* ***B*** *with* ***non-zero rank*** *and* ***lower training risk*** *than the Optimal Invariant Predictor.*
- *If* $\boldsymbol{E > d_e}$, *any feasible linear* $\boldsymbol{\Phi}$ *must have* $\boldsymbol{B = 0}$.

Corollary: The Optimal Invariant Predictor is the **global minimum** of the IRM objective **if and only if** $E > d_e$.

Linear Setting

Proof sketch:

Any embedder $\Phi(x) = [z_c,\ Bz_e]$ where $\boldsymbol{B \neq 0}$ will have **lower risk**.

Remains to show that such a $\boldsymbol{\Phi}$ can be feasible...

Linear f, $\boldsymbol{\Phi} \Rightarrow \hat{z}$ Gaussian.

Optimal vector $\boldsymbol{\beta}^*$ is available in closed form: $\Sigma^{-1}(\mu(1) - \mu(-1))$

This will be a function of B

Linear Setting

Proof sketch:

Any embedder $\Phi(x) = [z_c,\ Bz_e]$ where $\boldsymbol{B \neq 0}$ will have **lower risk**.

Remains to show that such a $\boldsymbol{\Phi}$ can be feasible...

We construct B as a function of $\mu_e,\ \Sigma_e$ such that:

- B has rank $d_e - E + 1$ (*depends on non-invariant features*)
- β^* is **the same** for all environments (*feasible*)

For $E > d_e$: we show β^* **invariant** $\implies$ $\boldsymbol{B = 0}$.

Linear Setting

Takeaway: Invariant prediction is *difficult, but possible* in the linear setting.

$E > d_e$ seems unachievable in practice; linear dependence probably unavoidable without stronger assumptions.

Arjovsky et al. provide a similar upper bound in the linear setting.

- Ours is **sharper**, more intuitive.
- We also give a matching **lower bound**.

Both proofs show that each environment restricts a "degree of freedom" of $\mathbf{\Phi}$.

Natural Question: Does this intuition extend to **non-linear** observations?

Outline

1. IRM and Variations
2. Latent Variable Model
3. Formal Results
   a. Linear Setting
   b. Non-Linear Setting

Non-Linear Setting

Still regress on $\hat{z} = \Phi(x) = \Phi(f(z))$, but f, $\boldsymbol{\Phi}$ are **arbitrarily complex.**

We study the Lagrangian used in practice:

$$\text{IRMv1} := \min_{\Phi,\beta} \; \sum_{e\in\mathcal{E}} R^e(\beta \circ \Phi) \; + \; \lambda \sum_{e\in\mathcal{E}} \|\nabla R^e(\beta \circ \Phi)\|_2^2$$

For convex risk, equivalent to IRM for $\lambda \to \infty$.

Non-Linear Setting

Theorem: *For* any *invertible f there is a predictor* $\boldsymbol{\Phi}, \boldsymbol{\beta}$ *with the following properties:*

1. *The penalty term is* ***exponentially small*** *in* d_e.

2. *The predictor matches the Optimal Invariant Predictor on all but an* ***exponentially small*** *fraction of the training data.*

3. *On any test distribution* ***slightly different*** *from the training distributions, the predictor behaves* exactly like the ERM solution *on all but an exponentially small fraction.*

Non-Linear Setting

Theorem: *For any invertible f there is a predictor* $\boldsymbol{\Phi}, \boldsymbol{\beta}$ *with the following properties:*

1. *The penalty term is* ***exponentially small*** *in* d_e.
2. *The predictor matches the Optimal Invariant Predictor on all but an* ***exponentially small*** *fraction of the training data.*
   - *(Polynomial # of samples* $\Rightarrow$ ***totally indistinguishable!)***
3. *On any test distribution* ***slightly different*** *from the training distributions, the predictor behaves* exactly like the ERM solution *on all but an exponentially small fraction.*

For any *f*, there exists a predictor which is **practically indistinguishable** from the Optimal Invariant Predictor at train time.

Non-Linear Setting

Theorem: *For any invertible f there is a predictor **Φ**, β with the following properties:*

1. *The penalty term is **exponentially small** in* d_e.
2. *The predictor matches the Optimal Invariant Predictor on all but an **exponentially small** fraction of the training data.*
   - *(Polynomial # of samples ⇒ **totally indistinguishable!**)*
3. *On any test distribution **slightly different** from the training distributions, the predictor behaves* exactly like the ERM solution *on all but an exponentially small fraction.*

At test time, this predictor will perform **almost exactly** like the predictor learned with ERM.

Non-Linear Setting

Takeaway: Even if we *could* solve these objectives, there is **no reason** to believe that we've recovered more useful features than ERM.

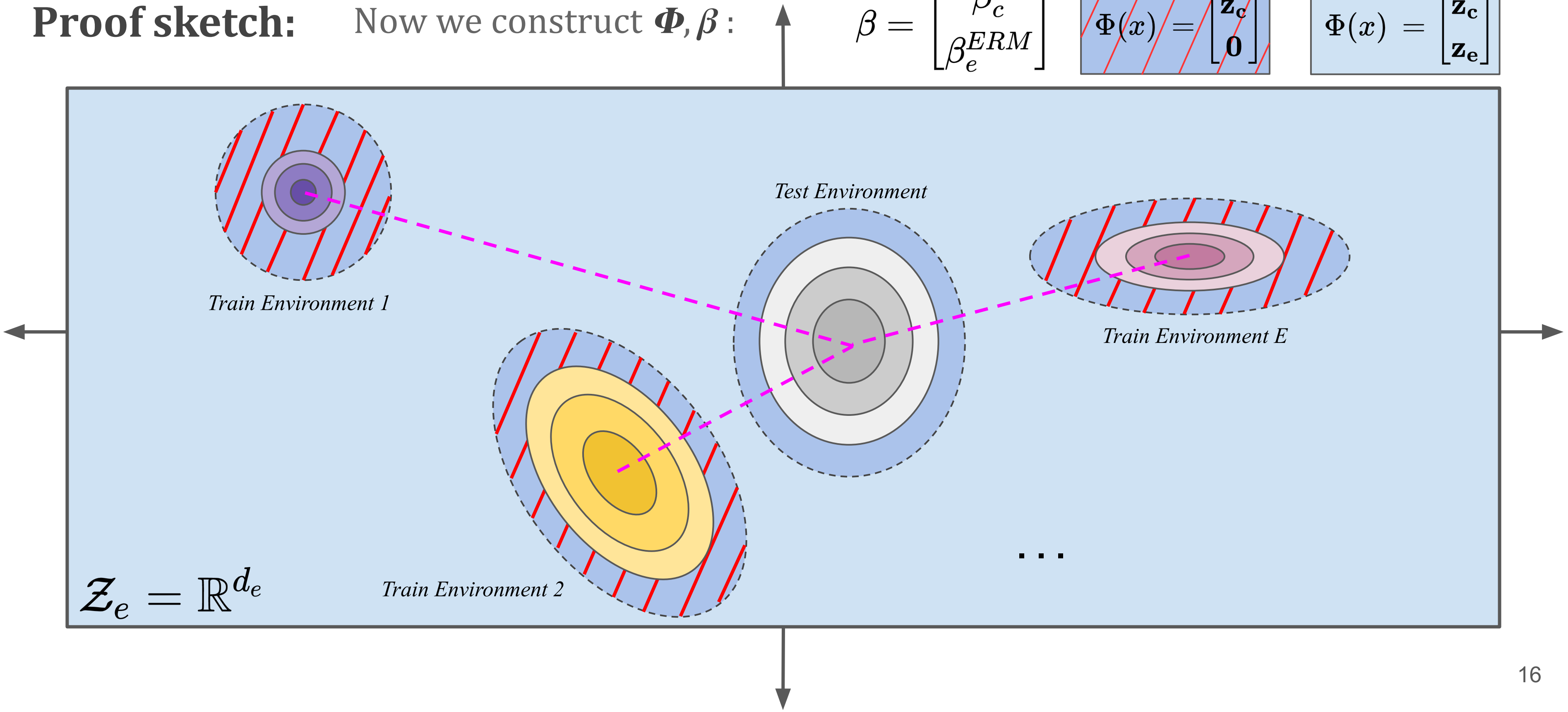
1. *The penalty term is **exponentially small** in d_e.*
2. *The predictor matches the Optimal Invariant Predictor on all but an **exponentially small** fraction of the training data.*
   - *(Polynomial # of samples $\Rightarrow$ **totally indistinguishable!**)*
3. *On any test distribution **slightly different** from the training distributions, the predictor behaves* exactly like the ERM solution.

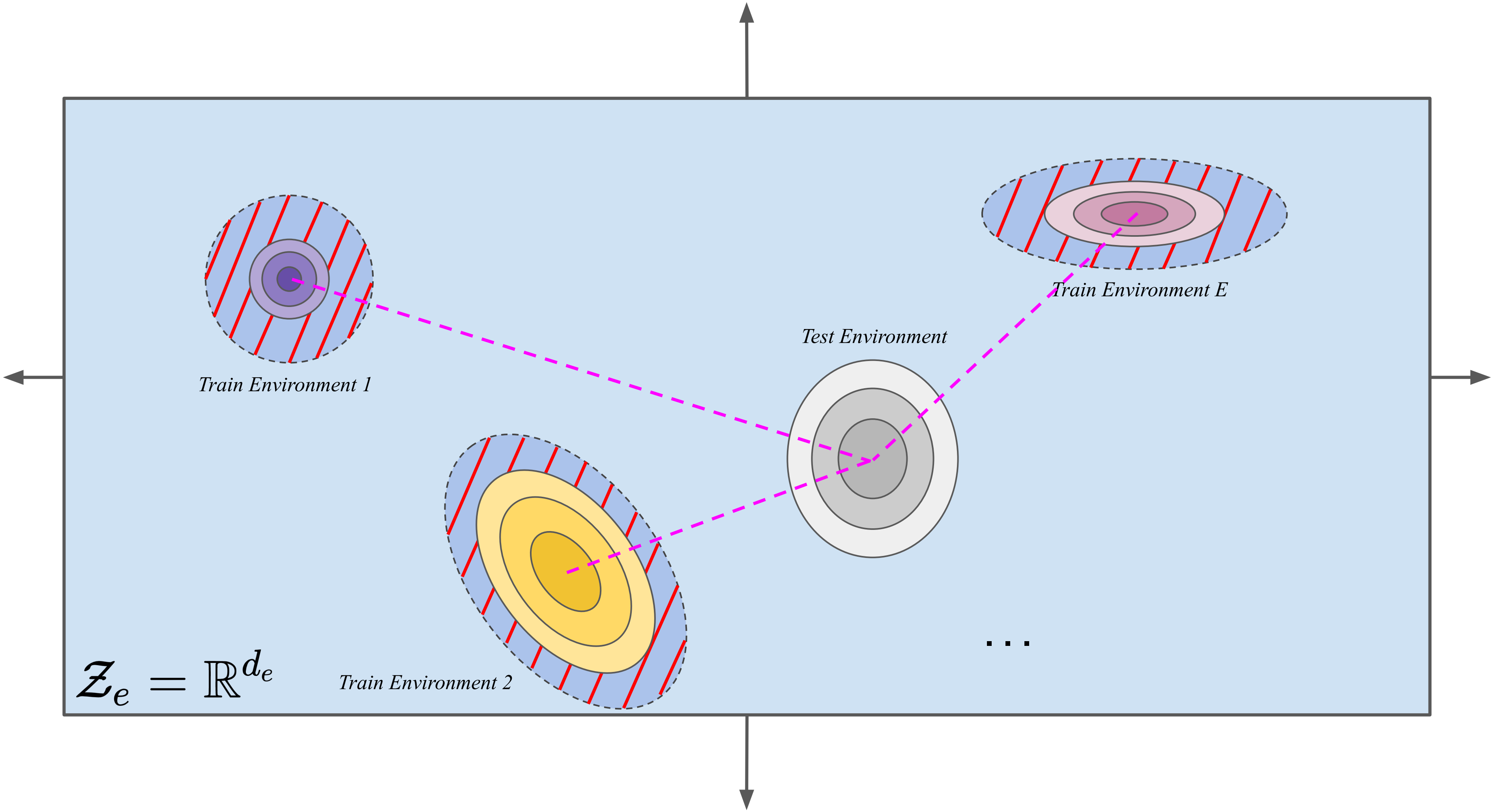
For these objectives to work, we'd need to observe enough environments to "cover" the space of features. But then *ERM will **generalize just as well!***

Non-Linear Setting

Proof sketch: Now we construct $\boldsymbol{\Phi}, \boldsymbol{\beta}$:

$$\beta = \begin{bmatrix} \beta_c^* \\ \beta_e^{ERM} \end{bmatrix} \qquad \Phi(x) = \begin{bmatrix} \mathbf{z_c} \\ \mathbf{0} \end{bmatrix} \qquad \Phi(x) = \begin{bmatrix} \mathbf{z_c} \\ \mathbf{z_e} \end{bmatrix}$$





Train Environment E
Test Environment
Train Environment 1
Train Environment 2
...
$\mathcal{Z}_e = \mathbb{R}^{d_e}$

Implications and Future Work

Causal reasoning for invariant prediction remains a promising approach for generalizing to unseen domains.

But when the features are latent, more care is needed to ensure our objectives actually work!
(Especially for complex, non-linear data)

Open questions:

- ❖ Can we avoid the linear dependence on environmental dimension?
  - ➢ With the right assumptions, $O(\sqrt{d_e})$ or even $O(\log d_e)$ seem feasible.

- ❖ In the non-linear case, can we do **anything at all** to improve over ERM?
  - ➢ Perhaps by limiting the complexity of $\mathbf{\Phi}$.

**Also, we need to stop tuning on the test set.*