Is Spurious Correlation Removal Always Learnable?

Yibo Zhou, Bo Li, Hai-Miao Hu, Hanzi Wang,
Xiaokang Zhang, Ruifan Zhang

- Invariant structure may be statistically **identifiable**
- But efficient recovery can still be **computationally hard**
- Environment **diversity** controls identifiability and sample complexity

ICML Poster Presentation

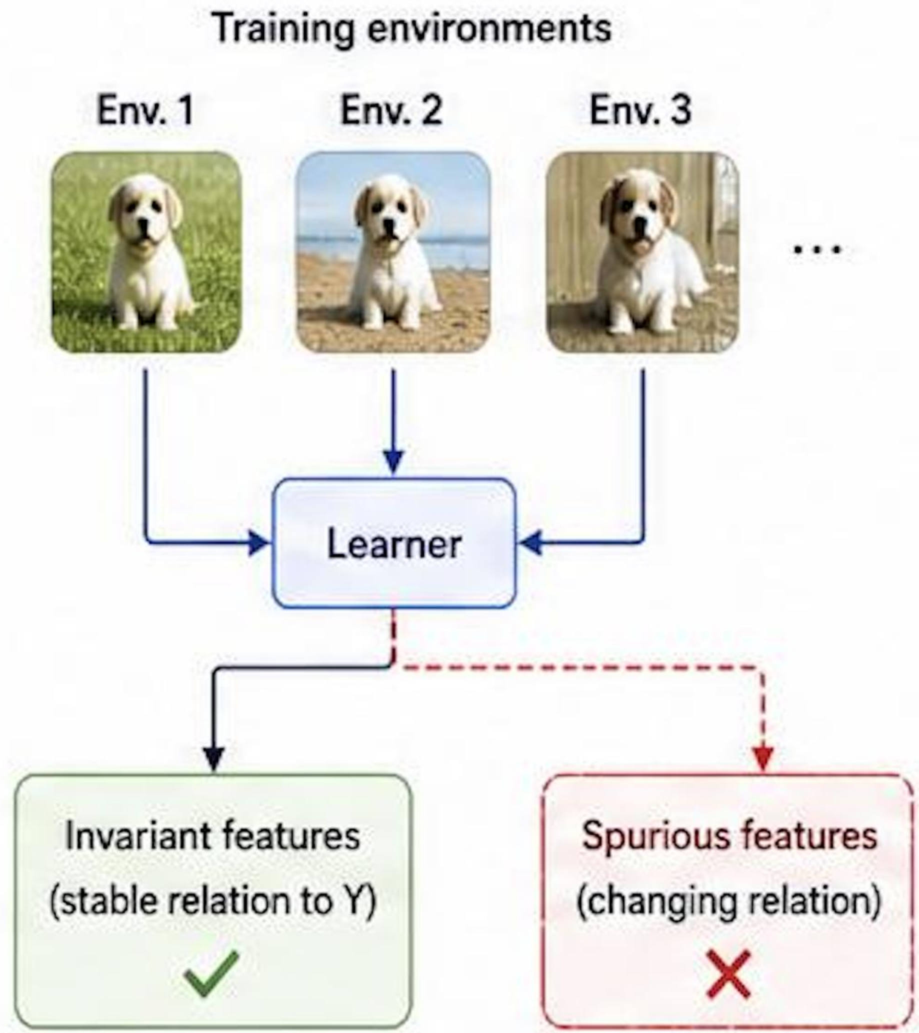
2 Motivation and Problem

Spurious correlations under distribution shift

- Models often rely on unstable features
- Example: background, color, co-occurrence bias
- Invariant learning uses multiple environments
- Goal: recover features whose relation to the label is stable

Question

If the invariant features are identifiable, are they efficiently learnable?

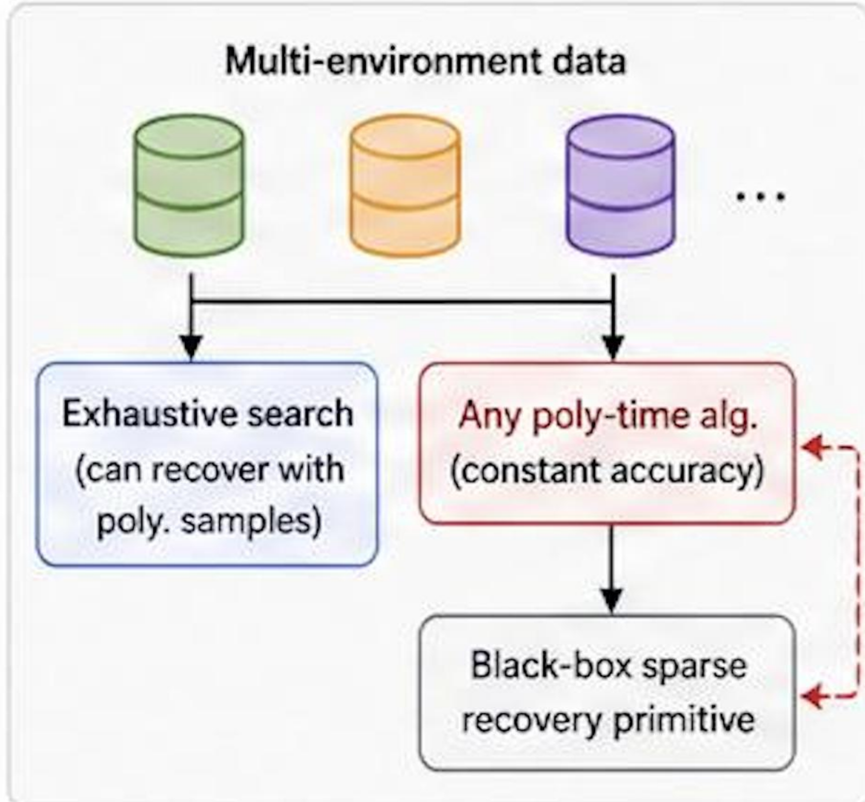


3 Main Theoretical Result: A Computational-Statistical Gap

Conditional hardness result

We construct multi-environment instances where:

- The invariant subspace is identifiable
- Exhaustive search recovers it with polynomial samples
- Any polynomial-time constant-accuracy recovery algorithm would contradict a black-box sparse recovery primitive



Key implication

Statistical learnability does not imply efficient learnability.

6 Takeaways

What should we learn from this?

Invariant learning failures can have different causes:

- Insufficient samples
- Insufficient environment diversity
- Computational hardness

Diversity should be diagnosed before blaming optimization.

Practical guidance

1. Estimate diversity in learned representations
2. If diversity is low, collect more diverse environments
3. If diversity is high but performance is poor, use stronger algorithms or more compute

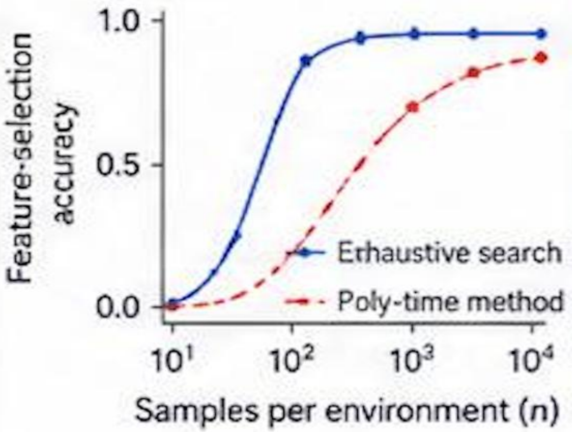
Conclusion

Spurious correlation removal is identifiable only under **diversity**, and learnable only when **statistics and computation** both **cooperate**.

5 Experiments

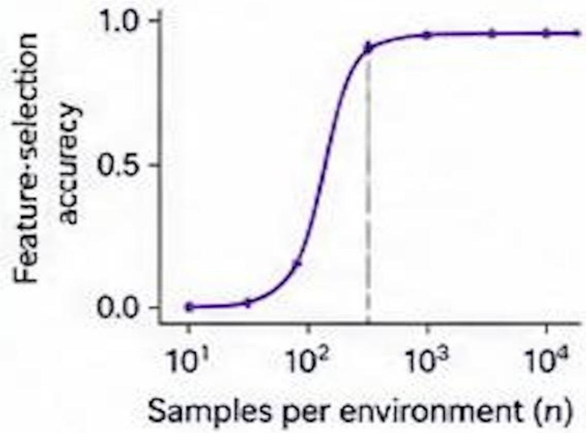
Compute-sample gap

Exhaustive search needs fewer samples than polynomial-time methods.



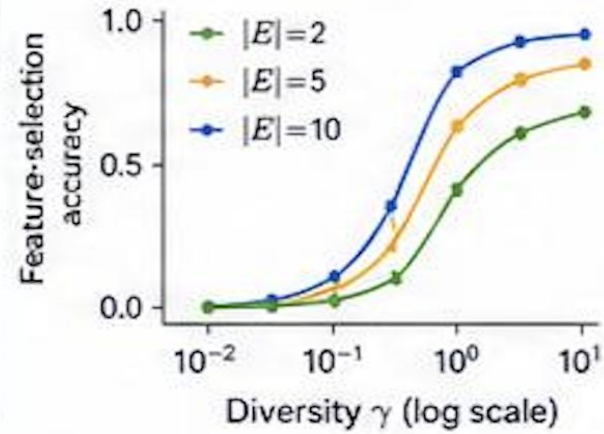
Phase transition

Recovery improves sharply around predicted sample threshold.



Diversity matters

Performance improves as environment diversity increases.



Benchmarks

4 Environment Diversity Controls Identifiability

Diversity parameter

γ measures how differently environments perturb spurious correlations.

Results

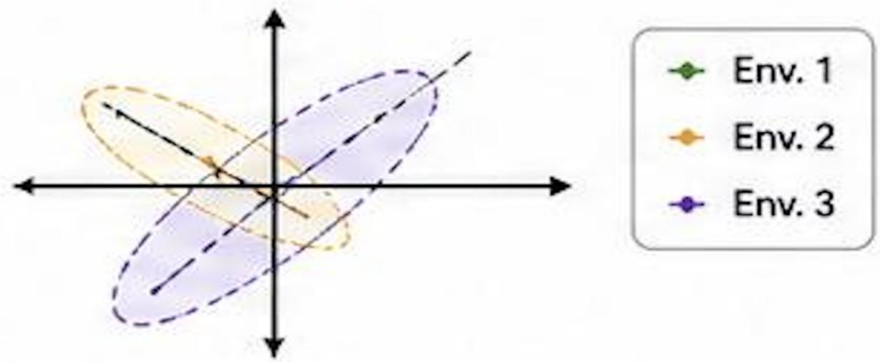
- If diversity is zero, invariant subspace may be unidentifiable
- With sufficient diversity:

$$\mathbb{E}\left[dist(\hat{V}, V_{\text{inv}})^2\right] = \Theta\left(\frac{k(d-k)}{n|E|}\right)$$

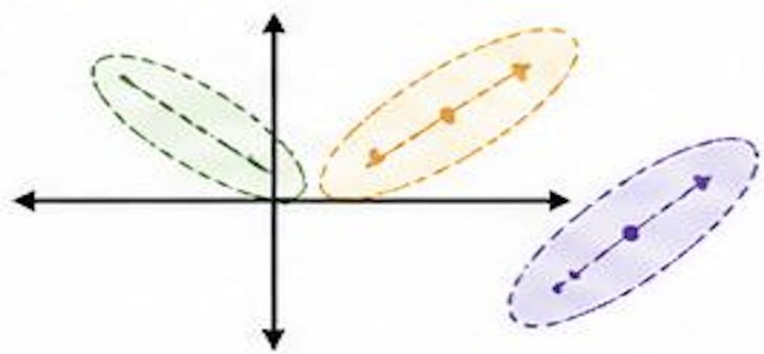
- Under label-induced shifts:

$$n^* \propto \frac{k(d-k)}{|E|\gamma^2}$$

Low diversity ($\gamma \approx 0$)



High diversity (large γ)



Takeaway

A few **diverse environments** can be better than many similar ones.