Iterative Feature Matching: Toward Provable Domain Generalization with Logarithmic Environments

Yining Chen[1], Elan Rosenfeld[2], Mark Sellke[1], Tengyu Ma[1], Andrej Risteski[2]
Stanford University[1], Carnegie Mellon University[2]

Introduction

Domain generalization[1]

Training environments

Test environments

Conditional feature matching algorithms

- Learn Φ such that $P[\Phi(x)|y]$ invariant
- CORAL[2], DANN[3], MMD[4]…
- Popular but lack formal guarantees

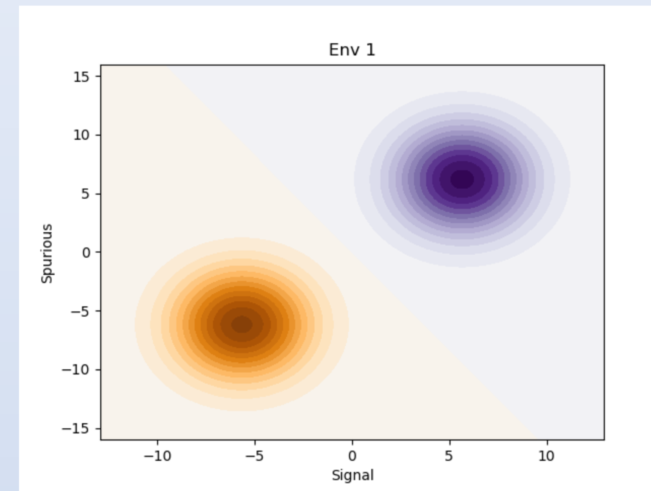
Invariant risk minimization (IRM)[5]

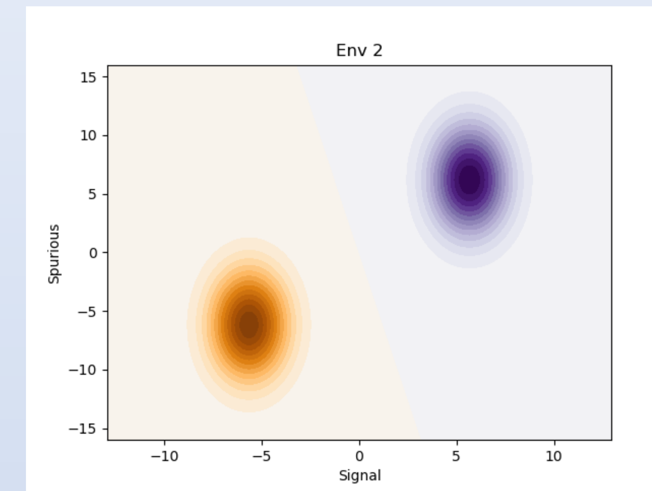
- Learn Φ such that $E[y|\Phi(x)]$ invariant
- Empirically doesn't work better than ERM[6,7]
- Theoretically requires many training environments even on a simple linear data model[8]

Our Contributions

- First positive guarantee for a feature matching algorithm
- Separation of feature matching vs IRM under a variant of data model in [Rosenfeld et al.]:
  - Spurious dimension d_s
  - Feature matching algorithm generalizes with $O(\log d_s)$ training environments
  - IRM / ERM requires $\Omega(d_s)$
- Empirical validation on synthetic datasets

Data Model





Label $Y \sim unif\{\pm 1\}$

Invariant Latents $Z_1|Y \sim N(Y \cdot \mu_1, \Sigma_1)$

Spurious Latents $Z_2|Y \sim N(Y \cdot \mu_2^e, \Sigma_2^e) \in \mathbb{R}^{d_s}$

Test Env $Z_2|Y \sim N(Y \cdot -\mu_2^e, \Sigma_2^e) \in \mathbb{R}^{d_s}$

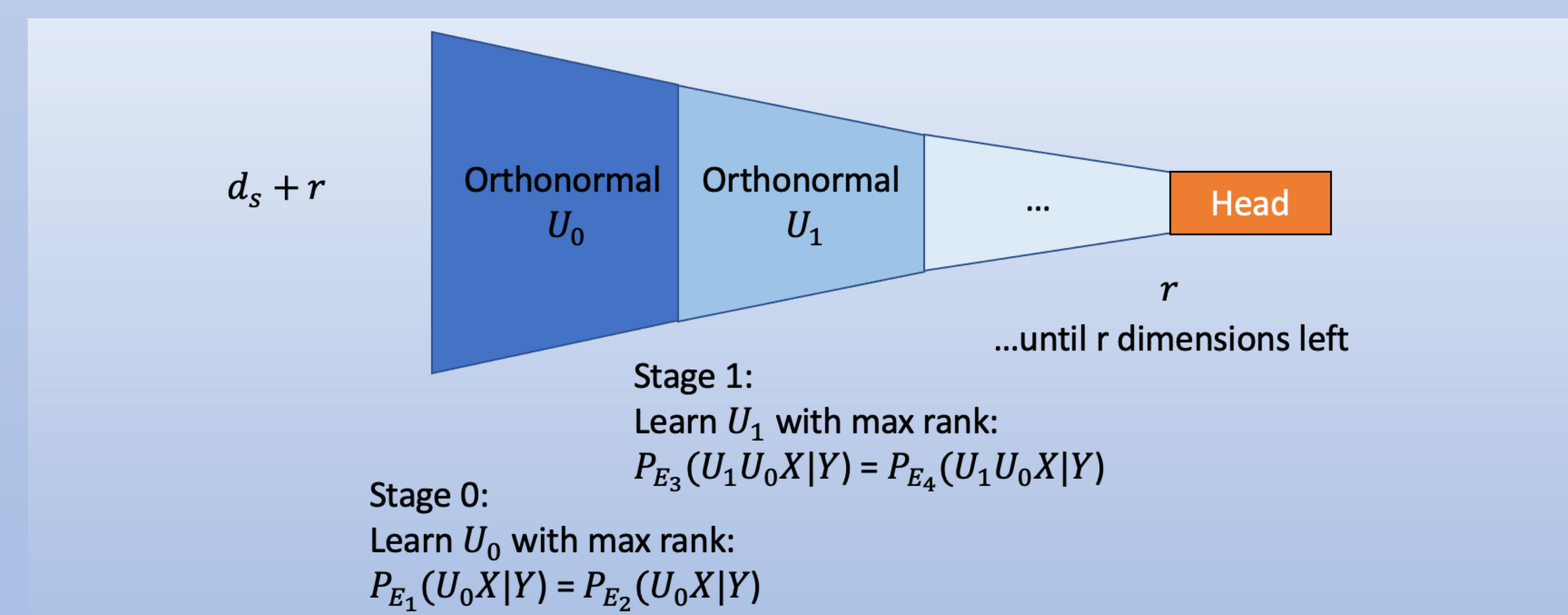
Observable Features $X = S\,[Z_1, Z_2] \in \mathbb{R}^d$

[Rosenfeld et al.] assumes isotropic spurious covariance $\Sigma_2^e = {\sigma_2^e}^2 I$

- IRM / ERM has $\Omega(d_s)$ environment complexity

We assume $\Sigma_2^e = arbitrary\ covariance\ \overline{\Sigma_2^e} + Gaussian\ noise\ G_e {G_e}^T$

Iterative Feature Matching (IFM) algorithm



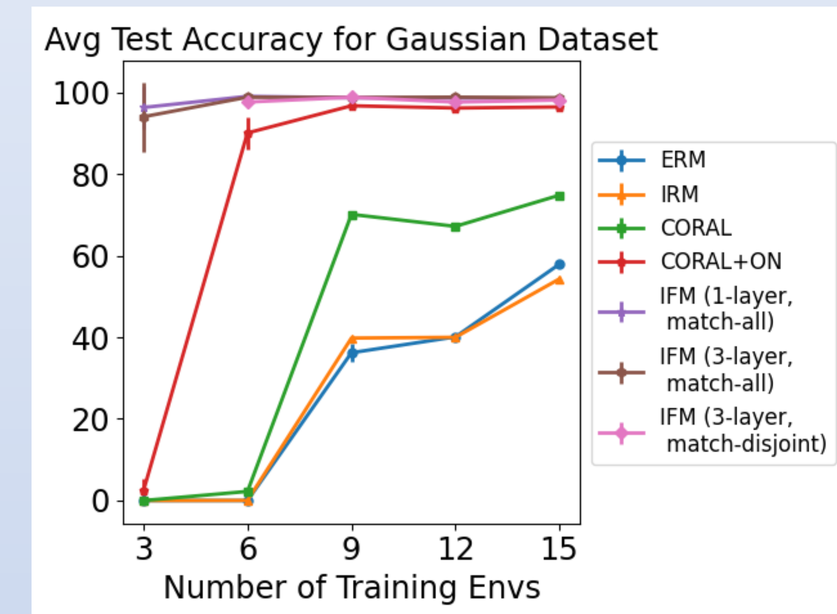
Assume infinite Gaussian data, match the second moment $\Sigma^e = E[X^e X^e | Y]$

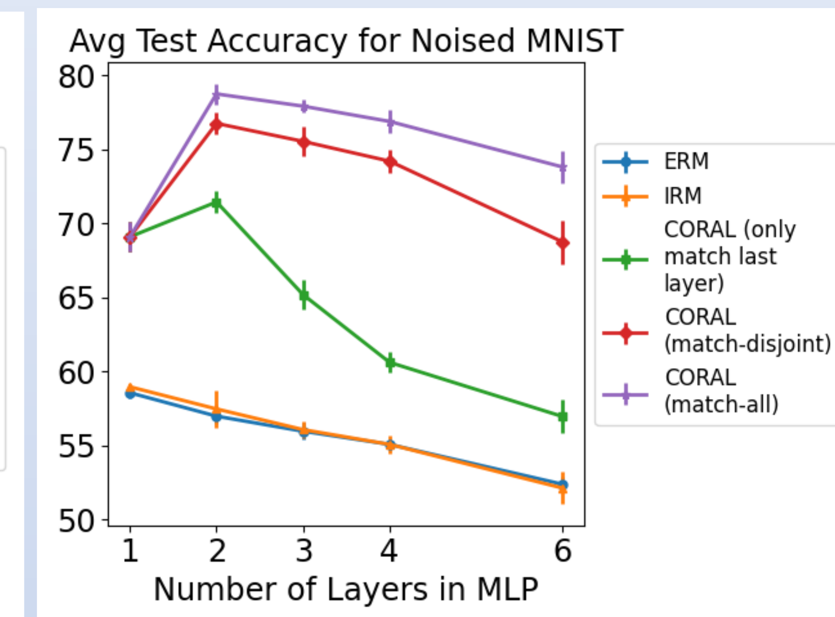
$$\max_U rank(U)\ such\ that \sum_{e,e'} \|U(\Sigma^e - \Sigma^{e'})U^T\|_2^2 + \|UU^T - I\|_2^2 = 0$$

Main Theorem

Using $O(1)$ environments each round, with high prob IFM terminates in $O(\log d_s)$ rounds and outputs the optimal invariant classifier.

Experiments





Proof Sketch

- Intuition: Independent random subsets prevents collusion
- Suppose exists rank-r U where $U(\Sigma^e - \Sigma^{e'})U^T = 0$

  $\Longrightarrow$ Exists $r \times$r orthonormal Q where $Q(\Sigma^e - \Sigma^{e'})Q^T = 0$
- Discretize over space of Q, suppose for fixed Q:

  $\Longrightarrow Q(G^e {G^e}^T - G^{e'} {G^{e'}}^T + \Delta_{e,e'})Q^T = 0$
- Random Gaussian vector $G^e q_i = v_i$, $G^{e'} q_i = u_i$:

  $\Longrightarrow \sum_{i \neq j} (v_i^T v_j - u_i^T u_j + cij)^2 = 0$
- LHS concentrates around its positive mean, so unlikely to be 0.

References

[1] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization.
[2] Sun, Baochen, and Kate Saenko. "Deep coral: Correlation alignment for deep domain adaptation." *European conference on computer vision*. Springer, Cham, 2016.
[3] Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *The journal of machine learning research* 17.1 (2016): 2096-2030.
[4] Li, Haoliang, et al. "Domain generalization with adversarial feature learning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[5] Arjovsky, Martin, et al. "Invariant risk minimization." *arXiv preprint arXiv:1907.02893* (2019).
[6] Aubin, Benjamin, et al. "Linear unit-tests for invariance discovery." *arXiv preprint arXiv:2102.10867* (2021).
[7] Gulrajani, Ishaan, and David Lopez-Paz. "In search of lost domain generalization." *arXiv preprint arXiv:2007.01434* (2020).
[8] Rosenfeld, Elan, Pradeep Ravikumar, and Andrej Risteski. "The risks of invariant risk minimization." *arXiv preprint arXiv:2010.05761* (2020).