Iterative Feature Matching: Toward Provable Domain Generalization with Logarithmic Environments

Yining Chen1, Elan Rosenfeld2, Mark Sellke1, Tengyu Ma3, Andrej Risteski2
Stanford University1, Carnegie Mellon University2

Introduction

Domain generalization

- Training environments
- Test environments

Conditional feature matching algorithms

- Learn $\Phi$ such that $P[\Phi(x)|y]$ invariant
- CORAL2, DANN3, MMD4...
- Popular but lack formal guarantees

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

- Invariant risk minimization (IRM)2
  - Learn $\Phi$ such that $E[y|\Phi(x)]$ invariant
  - Empirically doesn’t work better than ERM5,6
  - Theoretically requires many training environments even on a simple linear data model8

Invariant Latents

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

Spurious Latents

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

Test Envs

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

Observables

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

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

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

We assume $\Sigma_2^2$ is arbitrary covariance $\Sigma_2^2 + \text{Gaussian noise } G_0e_G^T$

Iterative Feature Matching (IFM) algorithm

\[d_s + r\]

Stage 0:

$$\text{Learn } U_0 \text{ with rank: } P_{\Sigma_2}(U_0^t X Y) = P_{\Sigma_2}(U_0^t U_0 X Y)$$

Stage 1:

$$\text{Learn } U_1 \text{ with rank: } P_{\Sigma_2}(U_1^t X Y) = P_{\Sigma_2}(U_1^t U_1 X Y)$$

$$\text{Orthogonal } U_k$$

$$\text{Head}$$

$$\ldots \text{until } r \text{ dimensions left}$$

Assume infinite Gaussian data, match the second moment $\Sigma_0 = E[X^t X^t Y]$ as $\max \text{rank}(U)$ such that

$$\sum \frac{1}{2} \| U^t (x^t - x^t e_0) u^2 \|^2 + \| U^t (Y - 1) \|^2 = 0$$

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

- Avg Test Accuracy for Gaussian Dataset
- Avg Test Accuracy for Noised MNIST

Proof Sketch

- Intuition: Independent random subsets prevents collusion
- Suppose exists rank-$r$ $U$ where $U (\Sigma_e - \Sigma^e)^{1/2} Y^T = 0$
  $\Rightarrow$ Exists $r \times r$ orthonormal $Q$ where $Q (\Sigma_e - \Sigma^e)^{1/2} Y^T = 0$
- Discretize over space of $Q$, suppose for fixed $Q$:
  $Q (\Sigma_e G^t G^t - G^t G^t e)^{1/2} Y^T = 0$
- Random Gaussian vector $G^t q_i = v_i, G^t q_i = u_i, u_i$:
  $\Rightarrow \Sigma_{ij}(v_i^t v_j - u_i^t u_j + c i j)^2 = 0$
- LHS concentrates around its positive mean, so unlikely to be 0.

References