Fundamental Limits and Tradeoffs in Invariant Representation Learning

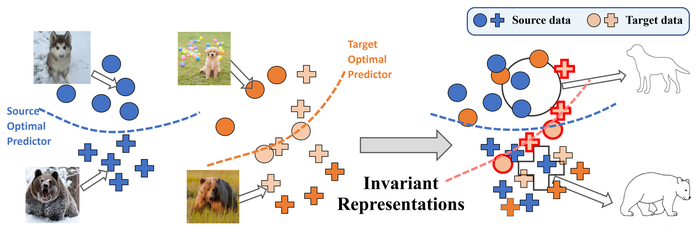
Han Zhao[†], Chen Dan[‡], Bryon Aragam[⋆], Tommi Jaakkola[§], Geoffrey Gordon[‡], Pradeep Ravikumar[‡]

[†]**University of Illinois Urbana Champaign,** [‡]**Carnegie Mellon University,** [⋆]**University of Chicago,** [§]**MIT**

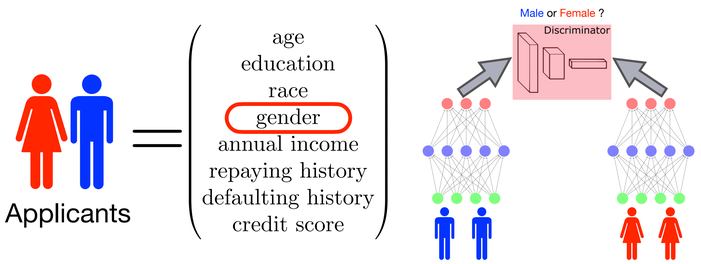
hanzhao@illinois.edu, {cdan,pradeepr,ggordon}@cs.cmu.edu, bryon@chicagobooth.edu, tommi@csail.mit.edu

Example and Applications

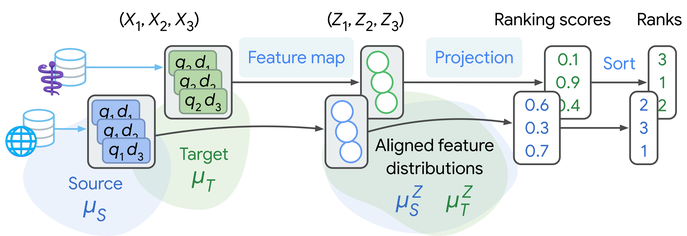
Domain Adaptation/Generalization: mitigating distribution shifts



Fair Representations: mitigating bias in data



Learning to Rank: matching query/document distributions



Research Questions

Question:

Is there any limitation on using invariant representations? If yes, what is the fundamental tradeoff between utility (accuracy) and invariance (distribution matching)?

Our Answer: Yes in general, and we can characterize the tradeoff on an **information plane**, where the optimal tradeoff depends on the coupling between the target Y and the attribute A.

Carnegie Mellon University · University of Illinois Urbana-Champaign · The University of Chicago · Massachusetts Institute of Technology

Invariant Representation Learning

Problem Setup: Given a joint distribution μ over (X, A, Y), we would like to learn a (randomized) mapping $g : \mathcal{X} \times \mathcal{A} \mapsto \mathcal{Z}$ such that the marginal distribution of Z is **invariant to the attribute** A:

$$\Pr_\mu(Z \mid A = a) = \Pr_\mu(Z),\ \forall a \in \mathcal{A}$$

- $\mathcal{X} \subseteq \mathbb{R}^d$: input space of the data
- $\mathcal{A}$: attribute space (classification: $\mathcal{A} = \{0, 1\}$; regression: $\mathcal{A} = \mathbb{R}$)
- $\mathcal{Y}$: label space (classification: $\mathcal{Y} = \{0, 1\}$; regression: $\mathcal{Y} = \mathbb{R}$)

Practical Implementation via Adversarial Training:

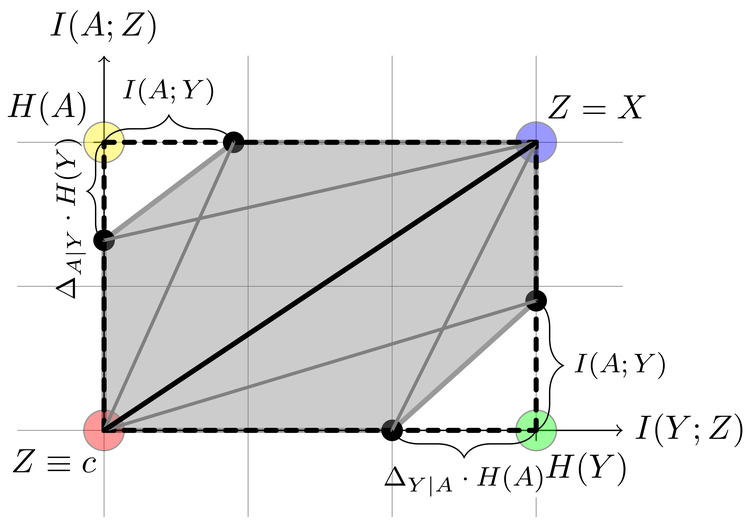
$$\min_{g,h} \max_{h'} \quad \ell_Y(h(g(X, A)), Y) - \lambda \cdot \ell_A(h'(g(X, A)), A)$$

- Classification: $\ell_Y = \ell_A =$ cross-entropy loss
- Regression: $\ell_Y = \ell_A =$ mean-squared error
- λ: tradeoff parameter: $\lambda = 0$ (accuracy); $\lambda \to \infty$ (invariance)

To focus on the fundamental tradeoff, we focus on the **noiseless setting** and we assume both h and h' have infinite capacity, i.e., they are **perfect predictors**.

Classification

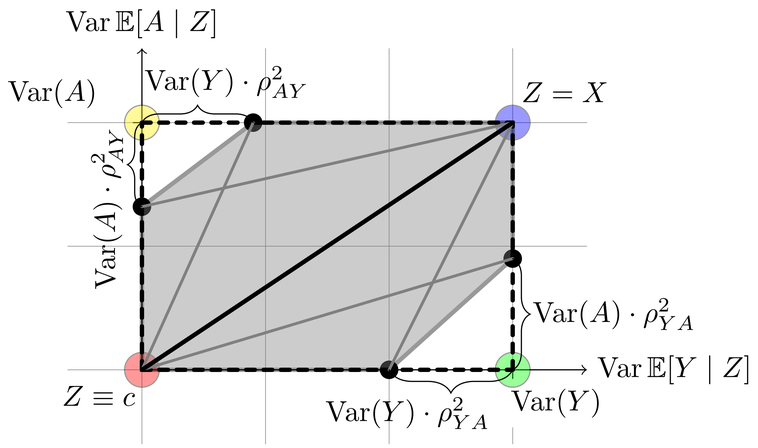
Under the noiseless setting, the tradeoff problem has the following form: $\max_{Z=g(X,A)} I(Y; Z) - \lambda I(A; Z)$. We define the **information plane** for classification problems as the feasible region of the tradeoff problem: $\mathcal{R}_{\text{CE}} := \{(I(Y; Z), I(A; Z)) : Z = g(X, A)\}$.



- $\Delta_{Y|A} := |\Pr_\mu(Y = 1 \mid A = 0) - \Pr_\mu(Y = 1 \mid A = 1)|$.

Regression

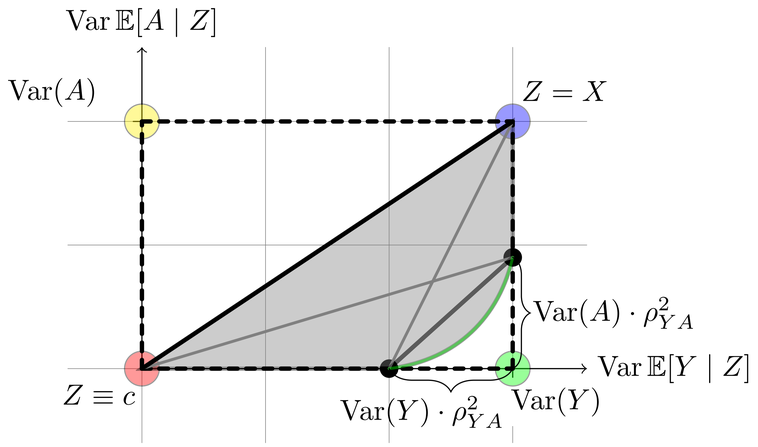
Under the noiseless setting, the tradeoff problem has the following form: $\max_{Z=g(X,A)} \operatorname{Var} \mathbb{E}[Y|Z] - \lambda \operatorname{Var} \mathbb{E}[A|Z]$. Define the **information plane** for regression problems as the feasible region of the tradeoff problem: $\mathcal{R}_{\text{LS}} := \{(\operatorname{Var} \mathbb{E}[Y|Z], \operatorname{Var} \mathbb{E}[A|Z]) : Z = g(X, A)\}$.



- $\rho_{YA} :=$ correlation coefficient between Y and A.

In regression problems, we can also precisely characterize the **Pareto frontier** of the following problem:

$$\max_{Z=g(X,A)} \operatorname{Var} \mathbb{E}[Y \mid Z], \quad \text{s.t. } \operatorname{Var} \mathbb{E}[A \mid Z] \le c$$



Equation for the **Pareto frontier**:

$$\operatorname{Var} \mathbb{E}[Y|Z] \le \operatorname{Var}(Y) \left(2\rho_{YA}\sqrt{(1-\rho^2_{YA})\alpha(1-\alpha)} + 1 - \alpha - \rho^2_{YA} + 2\alpha\rho^2_{YA}\right)$$

- $\alpha := c / \operatorname{Var}(A)$.