Causal Inference and Stable Learning

Peng Cui
Tsinghua University

Zheyan Shen
Tsinghua University

Outline

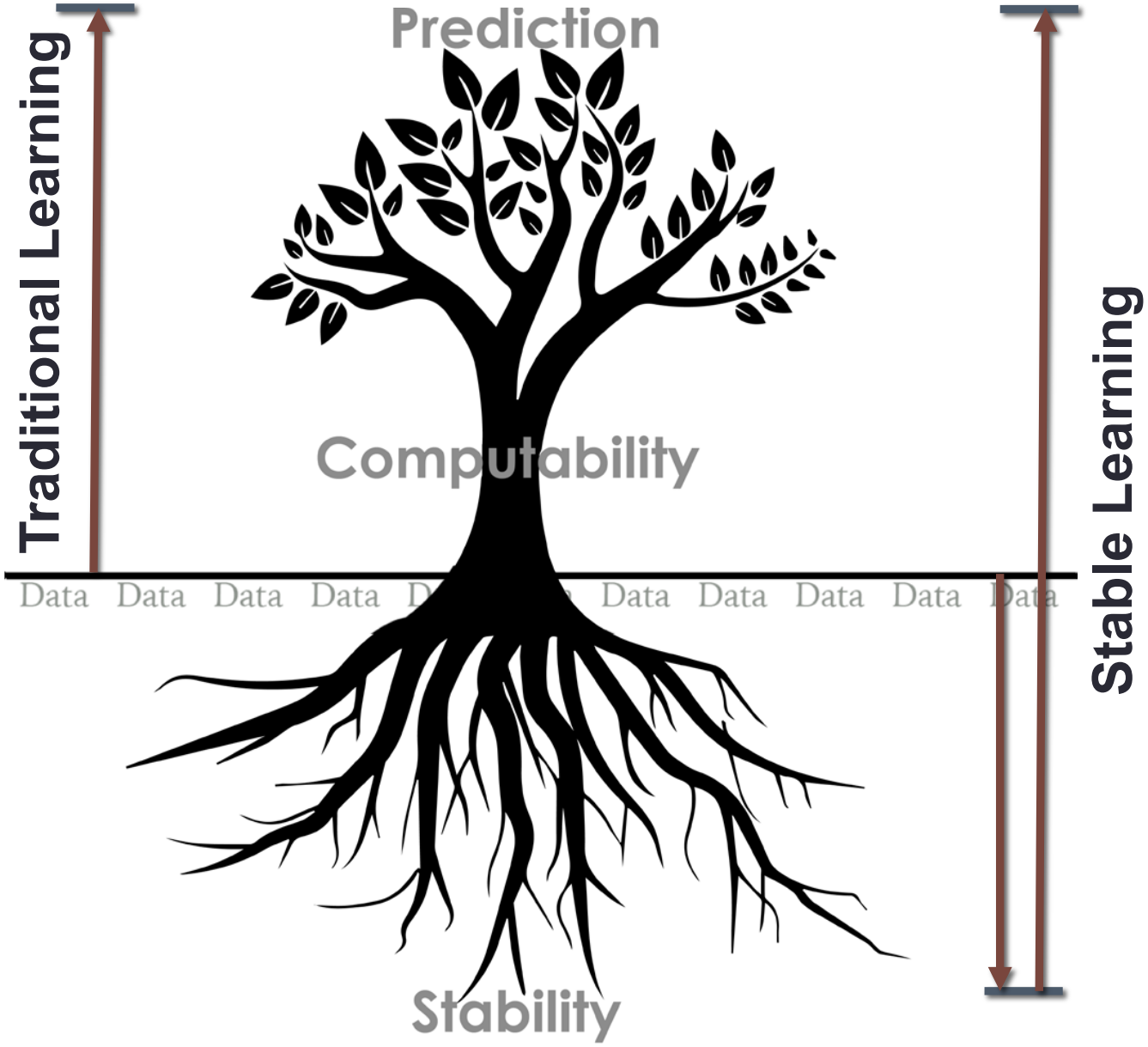
- Stable Learning: Definition and Related Problems
- Stable Learning: From Causally-Oriented Perspective
- Stable Learning: From Statistical Learning Perspective
- Beyond Structural Data: Stable Learning on Graph
- NICO: A Benchmark and Baseline for Stable Learning
- Conclusions

Stability and Prediction

Prediction Performance

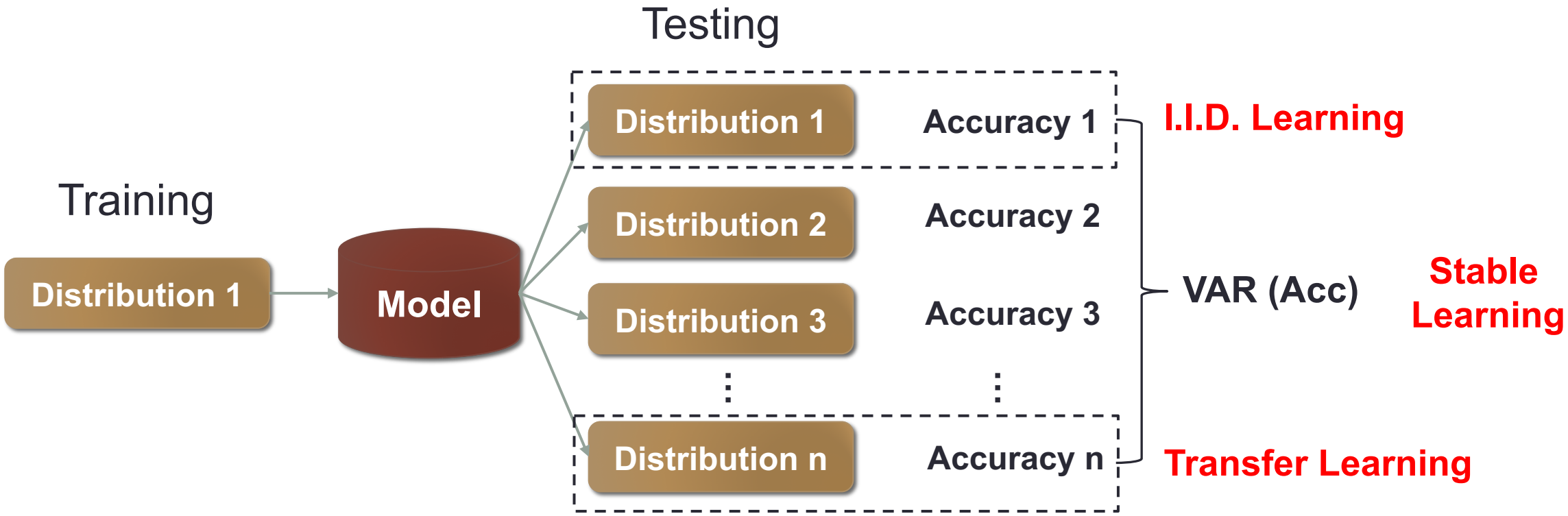
Learning Process

True Model



Bin Yu (2016), Three Principles of Data Science: predictability, computability, stability

Stable Learning: Definition



Stable Learning: Achieve uniformly good performance on any distribution

Stability and Robustness

- Robustness
  - More on prediction performance over data perturbations
  - ***Prediction*** performance-driven
- Stability
  - More on the true model
  - Lay more emphasis on ***Bias***
  - Sufficient for robustness

Stable learning is a (intrinsic?) way to realize robust prediction

Distributionally Robust Optimization

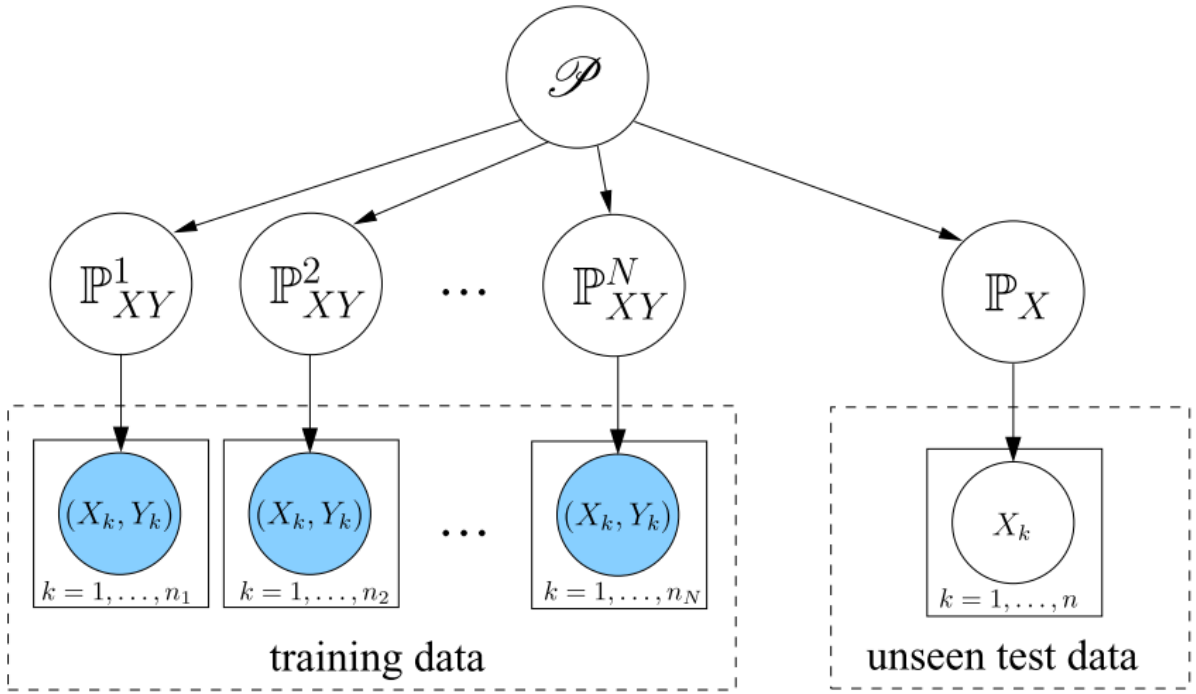
- Problem Definition:

$$\underset{\theta \in \Theta}{\text{minimize}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)]$$

where $\mathcal{P}$ is a class of distributions around the data-generating distribution P_0

- Idea: if class $\mathcal{P}$ contains all distributions under shift-interventions or do-interventions, then causal parameter θ_{causal} is the distributionally robust parameter.

Domain Generalization / Invariant Learning



- Given data from different observed environments $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim F^e, \quad e \in \mathcal{E}$$

- The task is to predict Y given X such that the prediction works well (is “robust”) for “all possible” (including unseen) environments

Domain Generalization

- **Assumption**: the conditional probability P(Y|X) is stable or invariant across different environments.
- **Idea**: taking knowledge acquired from a number of related domains and applying it to previously unseen domains
- **Theorem**: Under reasonable technical assumptions. Then with probability at least $1-\delta$

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \mathbb{E}^*_{\mathscr{P}} \mathbb{E}_{\mathbb{P}} \ell(f(\tilde{X}_{ij}), Y_i) - \mathbb{E}_{\hat{\mathbb{P}}} \ell(f(\tilde{X}_{ij}), Y_i) \right|^2$$

$$\leq c_1 \cdot \underbrace{\mathbb{V}_{\mathcal{H}}(\mathbb{P}^1, \mathbb{P}^2, \ldots, \mathbb{P}^N)}_{\textit{distributional variance}} + \underbrace{c_2 \frac{N \cdot (\log \delta^{-1} + 2 \log N)}{n} + c_3 \frac{\log \delta^{-1}}{N} + \frac{c_4}{N}}_{\textit{vanish as } N,n \to \infty}$$

Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature. ICML 2013.

Invariant Prediction

- **Invariant Assumption**: There exists a subset $S \in X$ is causal for the prediction of Y, and the conditional distribution P(Y|S) is stable across all environments.

$$\text{for all } e \in \mathcal{E},\ X^e \text{ has an arbitrary distribution and}$$

$$Y^e = g(X^e_{S^*}, \varepsilon^e), \qquad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}$$

- **Idea: Linking to causality**
  - Structural Causal Model (Pearl 2009):

$$Y^e \leftarrow \sum_{k \in \mathrm{pa}(Y)} \underbrace{\beta_{Y,k}}_{\forall e} X^e_k + \underbrace{\varepsilon^e_Y}_{\sim F_\varepsilon \forall e \in \mathcal{G}}$$

  - The parent variables of Y in SCM satisfies Invariant Assumption
  - The causal variables lead to invariance w.r.t. “all” possible environments

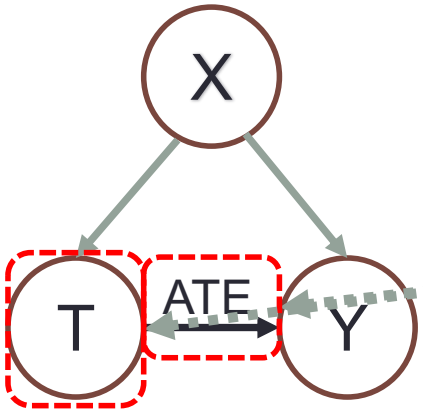
Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2016*

Outline

- Stable Learning: Definition and Related Problems
- Stable Learning: From Causally-Oriented Perspective
- Stable Learning: From Statistical Learning Perspective
- Beyond Structural Data: Stable Learning on Graph
- NICO: A Benchmark and Baseline for Stable Learning
- Conclusions

Sample Reweighting: Bridge from Causality to ML

Causal Problem



Typical Causal Framework

Causal Effect

Confounder Balancing
with Sample Reweighting

Learning Problem

$$y = x^{\top}\overline{\beta}_{1:p} + \overline{\beta}_0$$

Typical Regression Framework

After confounder balancing, partial effect can be regarded as causal effect. Predicting with causal variables is stable across different environments.

From Direct Balancing to Global Balancing

Directly Confounder Balancing

Given a feature T

↓

Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Over-parametrization and infeasible in high-dimensional setting!

Global Balancing

Given **ANY** feature T

↓

Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Removing confounding bias with an unique set of global weights.

Theoretical Guarantee

PROPOSITION 3.3. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^T \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2, \quad (4)$$

↓

0

PROOF. Since $\|\cdot\| \geq 0$, Eq. (8) can be simplified to $\forall j, \forall k \neq j$

$$\lim_{n\to\infty} \left(\frac{\sum_i \mathbf{X}_{i,k}=1, \mathbf{X}_{i,j}=1 W_i}{\sum_i \mathbf{X}_{i,j}=1 W_i} - \frac{\sum_i \mathbf{X}_{i,k}=1, \mathbf{X}_{i,j}=0 W_i}{\sum_i \mathbf{X}_{i,j}=0 W_i} \right) = 0$$

with probability 1. For W^*, from Lemma 3.1, $0 < P(\mathbf{X}_i = x) < 1$, $\forall x, \forall i, t = 1$ or 0,

$$\begin{aligned} \lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_i \mathbf{X}_{i,j}=t \, W_i^* &= \lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_x \sum_{x_j=t} \sum_i \mathbf{X}_i=x \, W_i^* \\ &= \lim_{n\to\infty} \textstyle\sum_x \sum_{x_j=t} \tfrac{1}{n} \sum_i \mathbf{X}_i=x \, \tfrac{1}{P(\mathbf{X}_i=x)} \\ &= \lim_{n\to\infty} \textstyle\sum_x \sum_{x_j=t} P(\mathbf{X}_i = x) \cdot \tfrac{1}{P(\mathbf{X}_i=x)} = 2^{p-1} \end{aligned}$$

with probability 1 (Law of Large Number). Since features are binary,

$$\lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_i \mathbf{X}_{i,k}=1, \mathbf{X}_{i,j}=1 \, W_i^* = 2^{p-2}$$

$$\lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_i \mathbf{X}_{i,j}=0 \, W_i^* = 2^{p-1}, \quad \lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_i \mathbf{X}_{i,k}=1, \mathbf{X}_{i,j}=0 \, W_i^* = 2^{p-2}$$

and therefore, we have following equation with probability 1:

$$\lim_{n\to\infty} \left(\frac{\mathbf{X}_{\cdot,k}^T (W^* \odot \mathbf{X}_{\cdot,j})}{W^{*T} \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,k}^T (W^* \odot (1-\mathbf{X}_{\cdot,j}))}{W^{*T} (1-\mathbf{X}_{\cdot,j})} \right) = \frac{2^{p-2}}{2^{p-1}} - \frac{2^{p-2}}{2^{p-1}} = 0.$$

□

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

Causal Regularizer for Global Balancing

Set feature *j* as treatment variable

$$\sum_{j=1}^{p} \left\| \frac{X_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|_2^2,$$

All features excluding treatment *j*

Sample Weights

Indicator of treatment status

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM**, 2018*.

Causally Regularized Logistic Regression (CRLR)

$$\min \quad \sum_{i=1}^{n} W_i \cdot \log(1+\exp((1-2Y_i)\cdot(x_i\beta))),$$

$$s.t. \quad \sum_{j=1}^{p} \left\| \frac{X_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j}^T \cdot (W \odot (1-I_j))}{W^T \cdot (1-I_j)} \right\|_2^2 \leq \lambda_1,$$

$$W \geq 0, \quad \|W\|_2^2 \leq \lambda_2, \quad \|\beta\|_2^2 \leq \lambda_3, \quad \|\beta\|_1 \leq \lambda_4,$$

$$(\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_5,$$

Sample reweighted logistic loss

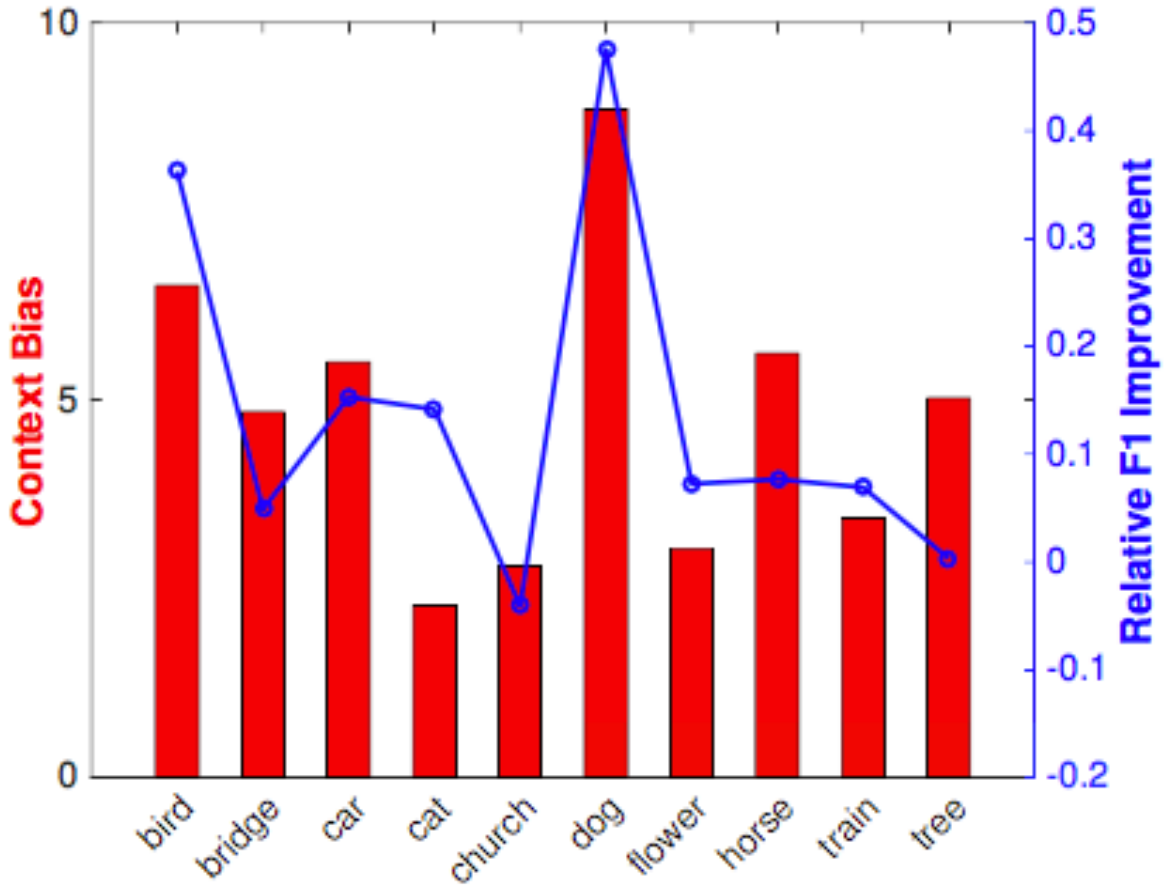
Causal Contribution

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM**, 2018*.

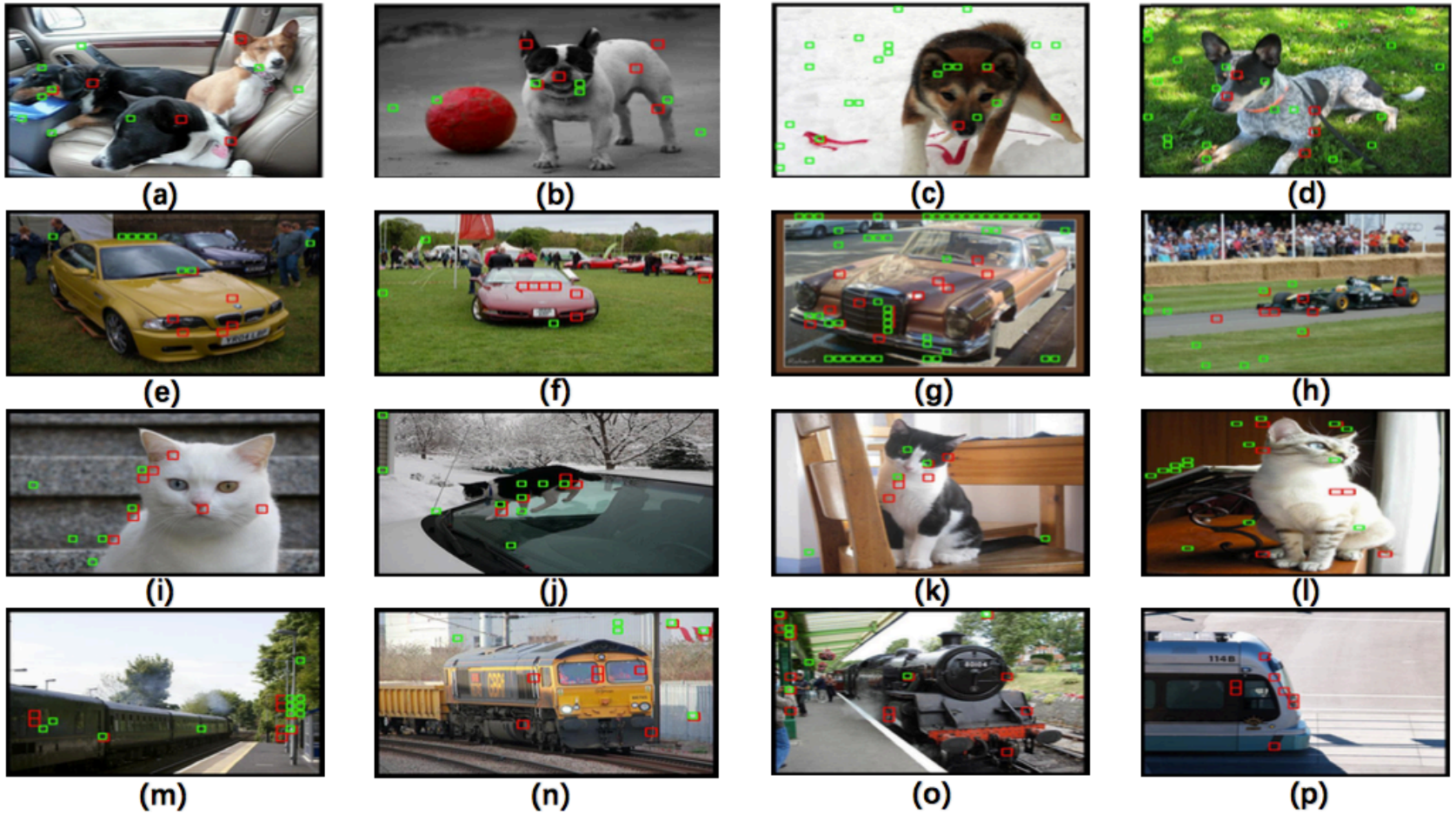
Experiment – Non-i.i.d. image classification

- Source: ***YFCC100M***
- Type: high-resolution and multi-tags
- Scale: 10-category, each with nearly 1000 images
- Method: select 5 ***context tags*** which are frequently co-occurred with the ***major tag*** (category label)

Experimental Result - insights



Experimental Result - insights



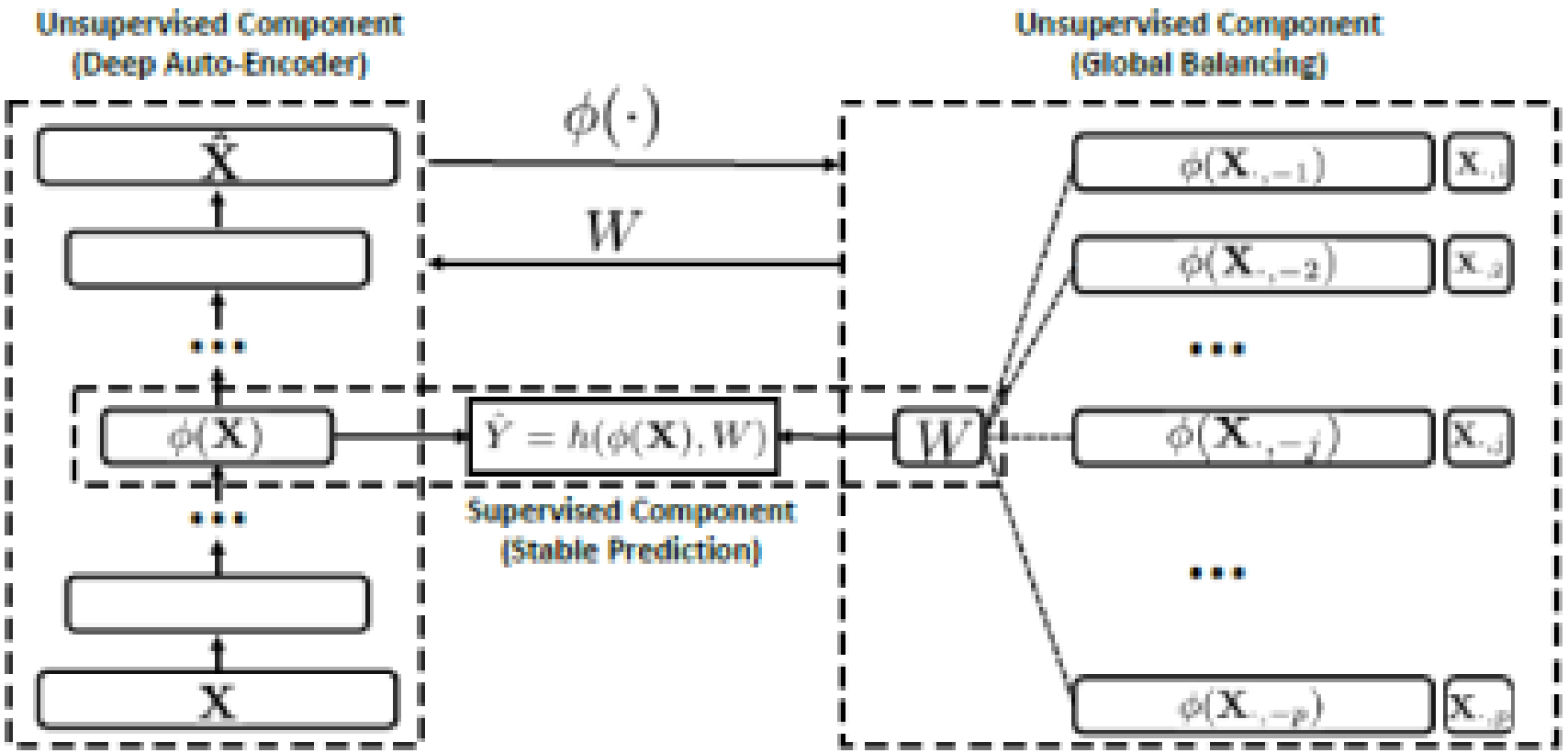
Limitations of Global Balancing

- A hidden assumption for Global Balancing to work

Assumption 2 (Overlap) *For any variable* $\mathbf{X}_{\cdot,j}$ *when setting it as the treatment variable, it has* $\forall j, 0 < P(\mathbf{X}_{\cdot,j} = 1 | \mathbf{X}_{\cdot,-j}) < 1$.

- Practical constraints
  - High dimensional features (potential treatment)
  - Sparsity of real world data
  - Possible interactions between features
  - More complex data type: categorical and continuous

From Shallow to Deep - DGBR



Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

From Shallow to Deep - DGBR

- Deep Global Balancing Regression (DGBR) Algorithm

$$\min \quad \sum_{i=1}^{n} W_i \cdot \log(1+\exp((1-2Y_i)\cdot(\phi(\mathbf{X}_i)\beta))), \tag{7}$$

$$s.t. \quad \sum_{j=1}^{p} \left\| \frac{\phi(\mathbf{X}_{\cdot,-j})^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\phi(\mathbf{X}_{\cdot,-j})^T \cdot (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^T \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2 \leq \lambda_1,$$

$$\|(W \cdot \mathbf{1}) \odot (X - \hat{X})\|_F^2 \leq \lambda_2, \quad W \geq 0, \quad \|W\|_2^2 \leq \lambda_3,$$

$$\|\beta\|_2^2 \leq \lambda_4, \quad \|\beta\|_1 \leq \lambda_5, \quad (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_6$$

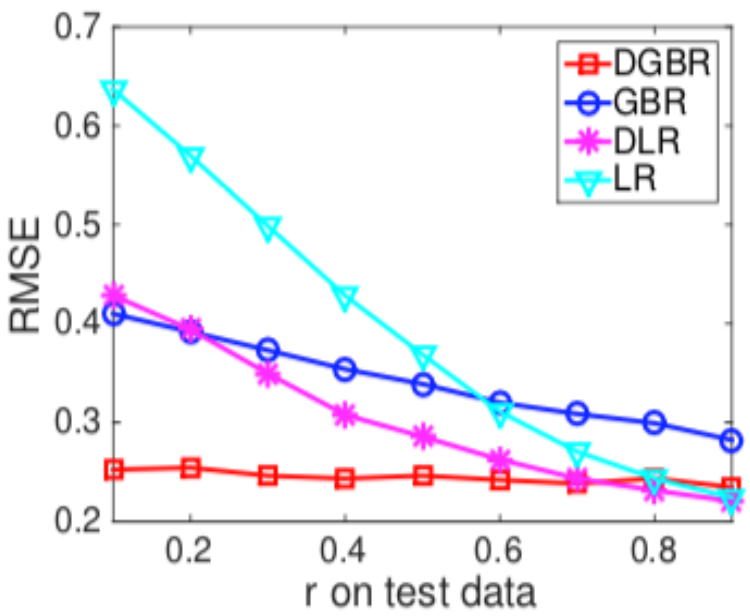
$$\sum_{k=1}^{K} (\|A^{(k)}\|_F^2 + \|\hat{A}^{(k)}\|_F^2) \leq \lambda_7,$$

Deep Auto-Encoder

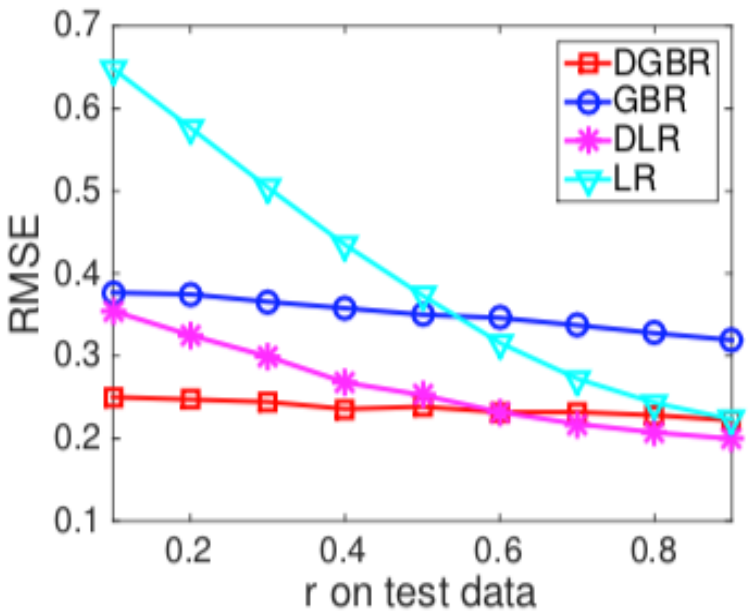
Global Balancing

Stable Prediction

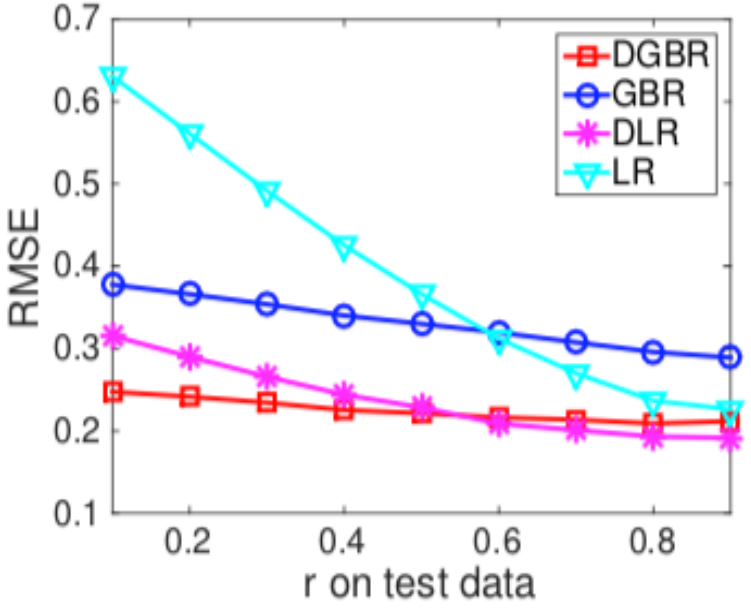
Experiments on Synthetic Data



(b) Trained on $n = 1000$, $p = 20$, $r = 0.75$



(e) Trained on $n = 2000$, $p = 20$, $r = 0.75$



(h) Trained on $n = 4000$, $p = 20$, $r = 0.75$

The RMSE of DGBR is consistently stable and small across environments under all settings.

From Binary to Continuous Variable - DWR

Independence condition for continuous variable

For all $a, b \in \mathbb{N}$, $\mathbb{E}[\mathbf{X}_{,j}^{a}\mathbf{X}_{,k}^{b}] = \mathbb{E}[\mathbf{X}_{,j}^{a}]\mathbb{E}[\mathbf{X}_{,k}^{b}]$

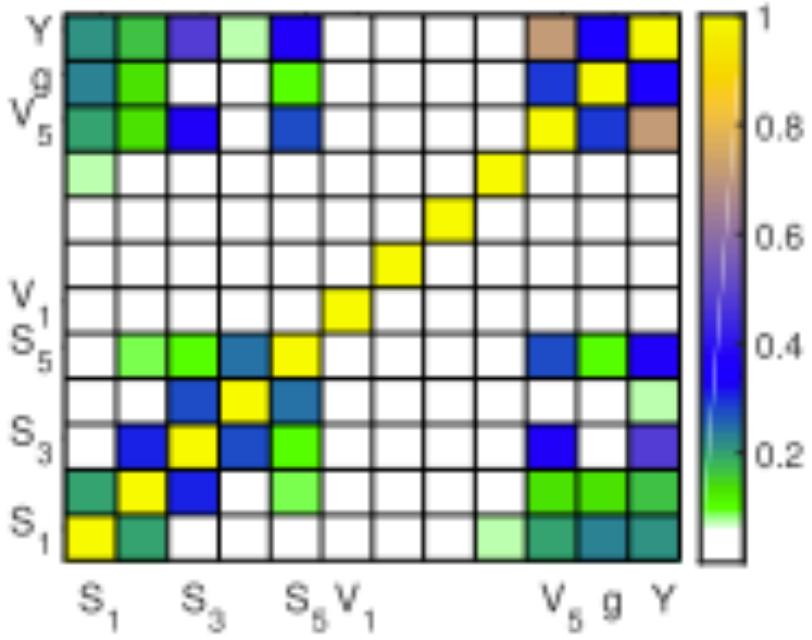
Causal Regularizer for Continuous Variable

$$\min_{W} \sum_{j=1}^{p} \left\| \mathbb{E}[\mathbf{X}_{,j}^{T}\mathbf{\Sigma}_{W}\mathbf{X}_{,-j}] - \mathbb{E}[\mathbf{X}_{,j}^{T}W]\mathbb{E}[\mathbf{X}_{,-j}^{T}W] \right\|_2^2$$

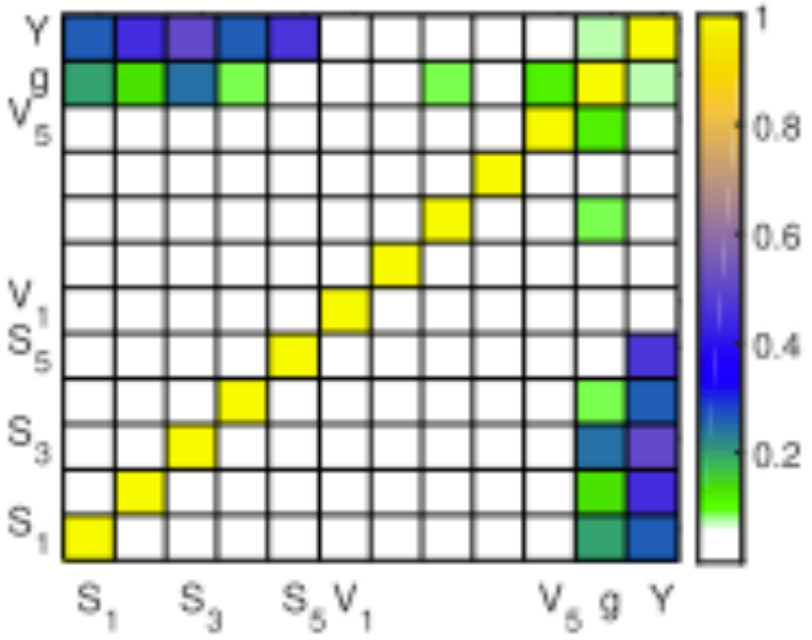
Decorrelated Weighted Regression:

$$\begin{aligned} & \min_{W,\beta} \sum_{i=1}^{n} W_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2 \\ s.t \quad & \sum_{j=1}^{p} \left\| \mathbf{X}_{,j}^{T}\mathbf{\Sigma}_{W}\mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^{T}W/n \cdot \mathbf{X}_{,-j}^{T}W/n \right\|_2^2 < \lambda_2 \\ & |\beta|_1 < \lambda_1, \quad \tfrac{1}{n}\sum_{i=1}^{n} W_i^2 < \lambda_3, \\ & (\frac{1}{n}\sum_{i=1}^{n} W_i - 1)^2 < \lambda_4, \quad W \succeq 0, \end{aligned}$$

De-confounding for continuous variable



(a) On raw data



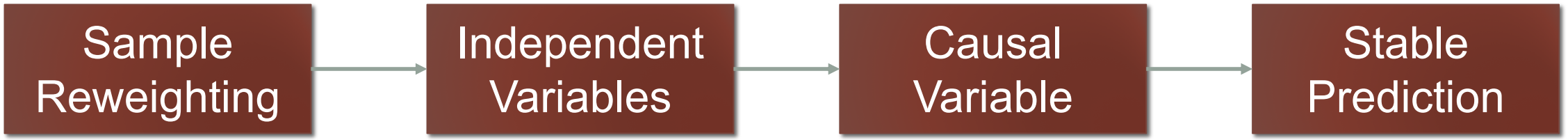
(b) On the weighted data

Outline

- Stable Learning: Definition and Related Problems
- Stable Learning: From Causally-Oriented Perspective
- Stable Learning: From Statistical Learning Perspective
- Beyond Structural Data: Stable Learning on Graph
- NICO: A Benchmark and Baseline for Stable Learning
- Conclusions

From ***Causal*** problem to ***Learning*** problem

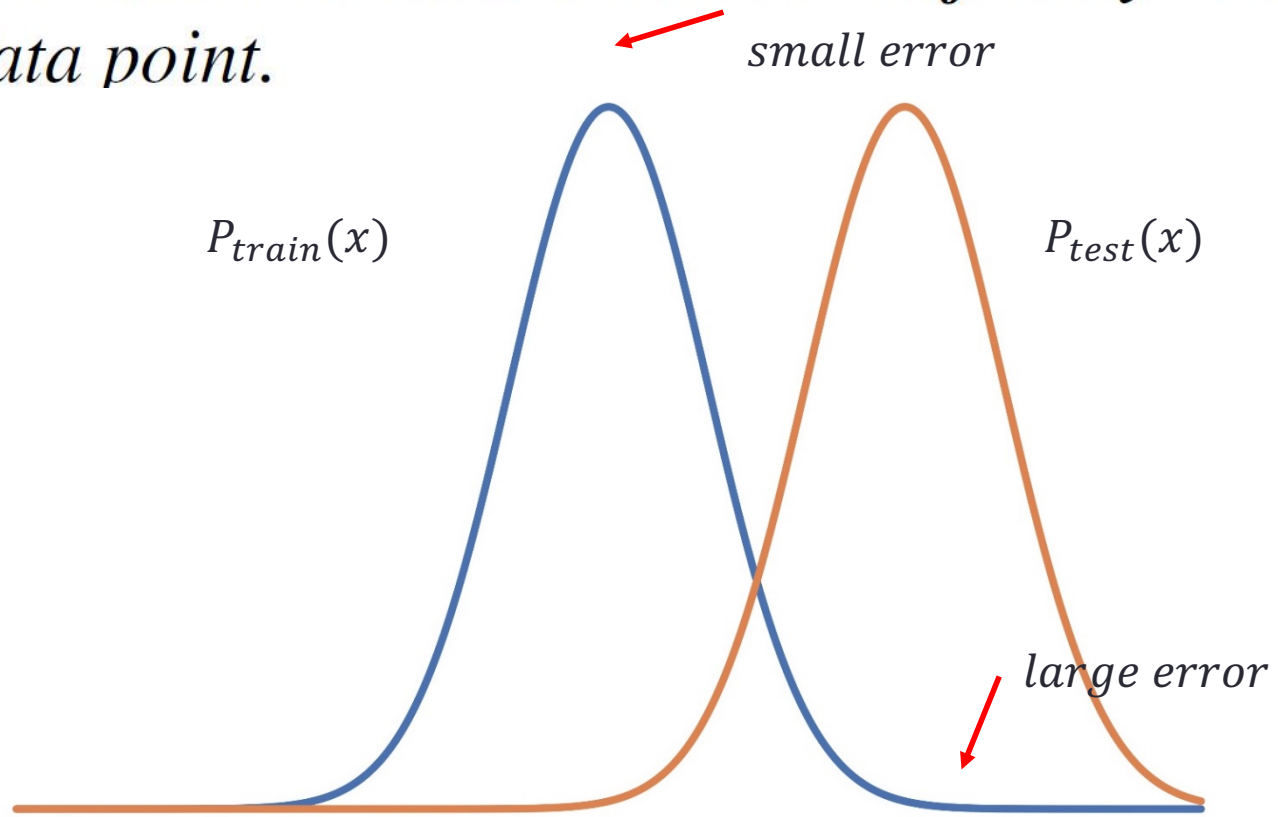
- Previous logic:



- More direct logic:

Thinking from the ***Learning*** end

Problem 1. ***(Stable Learning)*** *: Given the target y and p input variables $x = [x_1, \ldots, x_p] \in \mathbb{R}^p$, the task is to learn a predictive model which can achieve* ***uniformly*** *small error on* ***any*** *data point.*



Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting.

Stable Learning of Linear Models

- Consider the linear regression with misspecification bias

$$y = x^\top \overline{\beta}_{1:p} + \overline{\beta}_0 + b(x) + \epsilon$$

Goes to infinity when perfect collinearity exists!

Bias term with bound $b(x) \leq \delta$

- By accurately estimating $\overline{\beta}$ with the property that $b(x)$ is uniformly small for all x, we can achieve stable learning.
- However, the estimation error caused by misspecification term can be as bad as $\|\hat{\beta} - \overline{\beta}\|_2 \leq 2(\delta/\gamma) + \delta$, where γ^2 is the smallest eigenvalue of centered covariance matrix.

Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting.

Toy Example

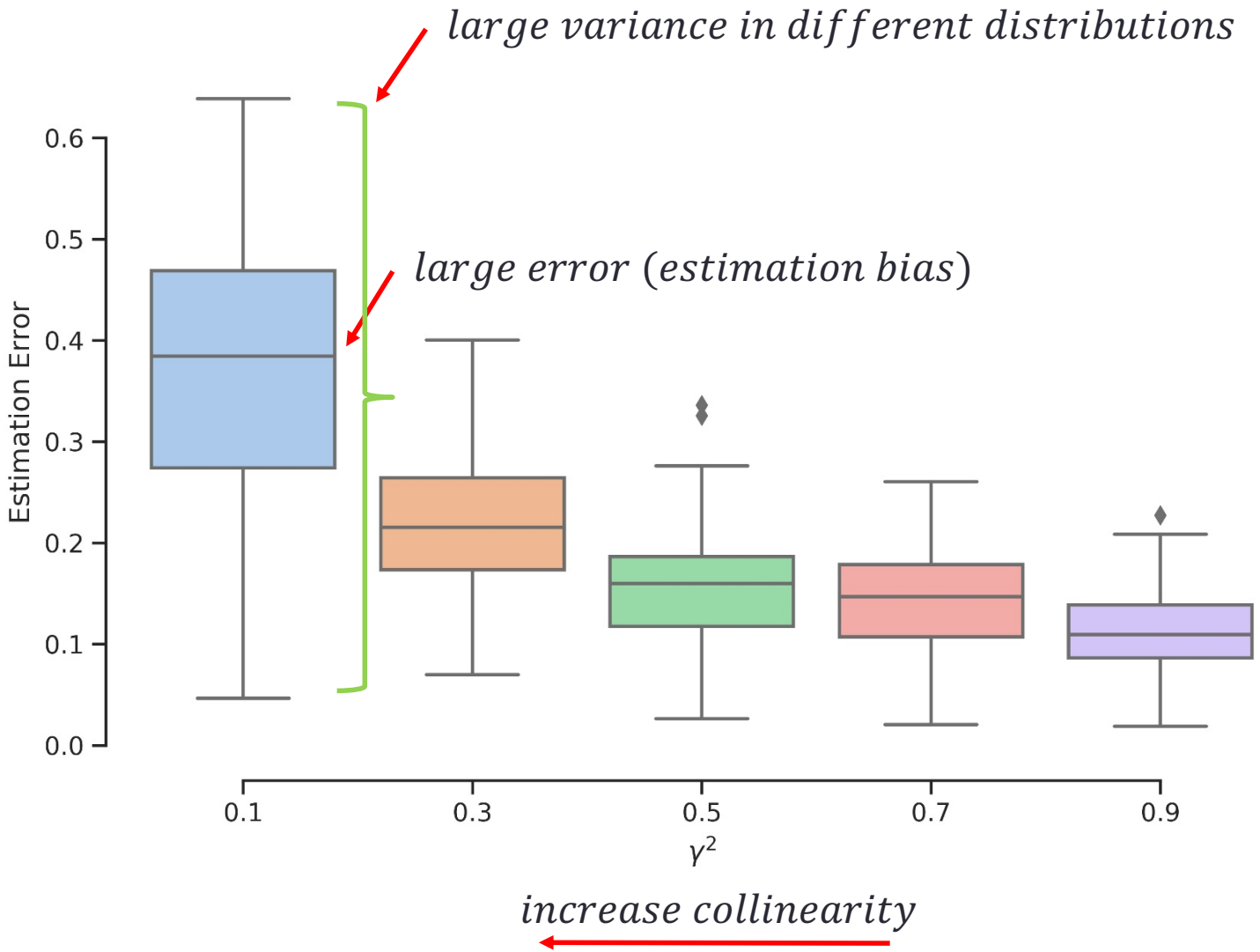
- Assume the design matrix X consists of two variables X_1, X_2, generated from a multivariate normal distribution:

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- By changing ρ, we can simulate different extent of collinearity.
- To induce bias related to collinearity, we generate bias term $b(X)$ with $b(X) = Xv$, where v is the eigenvector of centered covariance matrix corresponding to its smallest eigenvalue γ^2.
- The bias term is sensitive to collinearity.

Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting.

Simulation Results



Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting.

Reducing collinearity by sample reweighting

Idea: Learn a new set of ***sample weights*** $w(x)$ to decorrelate the input variables and increase the smallest eigenvalue

- Weighted Least Square Estimation

$$\hat{\beta} = \arg\min_{\beta} \mathbf{E}_{(x)\sim D} w(x) \left(x^\top \beta_{1:p} + \beta_0 - y\right)^2$$

which is equivalent to

$$\hat{\beta} = \arg\min_{\beta} \mathbf{E}_{(x)\sim \tilde{D}} \left(x^\top \beta_{1:p} + \beta_0 - y\right)^2$$

So, how to find an "oracle" distribution $\tilde{D}$ which holds the desired property?

Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting.

Sample Reweighted Decorrelation Operator (cont.)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \xrightarrow{\text{Decorrelation}} \tilde{\mathbf{X}} = \begin{pmatrix} x_{i1} & \cdots & x_{rl} & \cdots \\ x_{j1} & \cdots & x_{sl} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & \cdots & x_{tl} & \cdots \end{pmatrix}$$

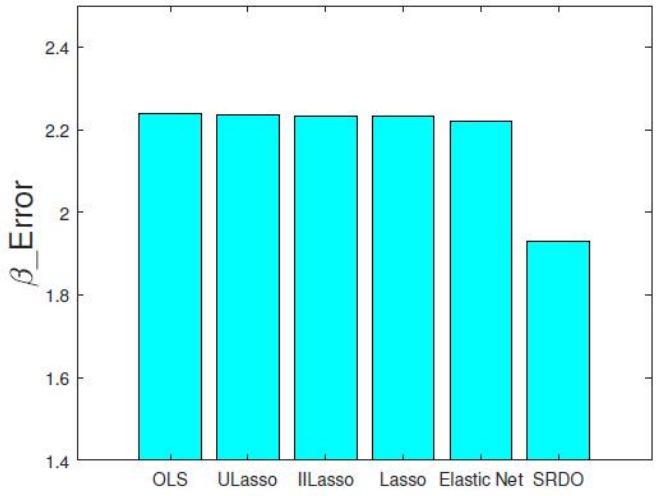
where i, j, k, r, s, t are drawn from $1 \ldots n$ at random

- By treating the different columns independently while performing random resampling, we can obtain a column-decorrelated design matrix with the same marginal as before.
- Then we can use density ratio estimation to get $w(x)$.

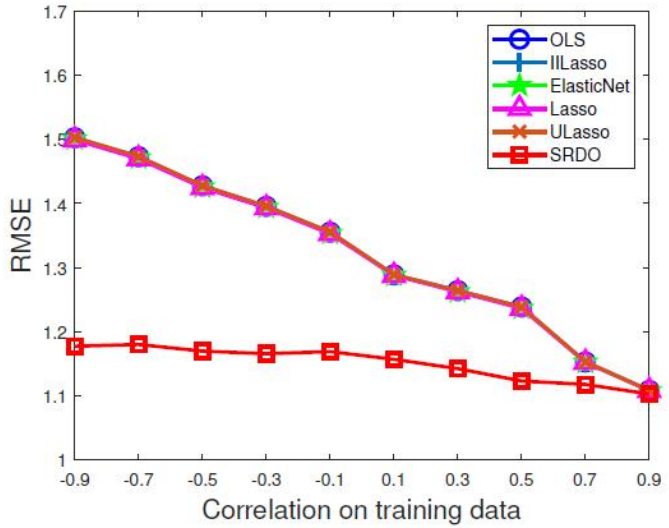
Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting.

Experimental Results

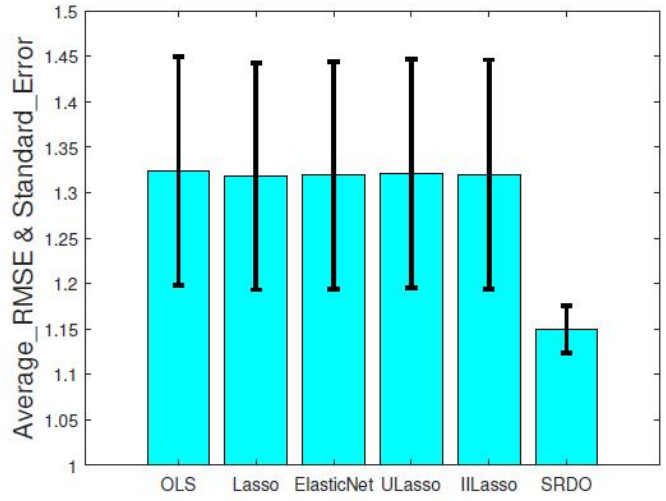
- Simulation Study



(a) Estimation error



(b) Prediction error over different test environments

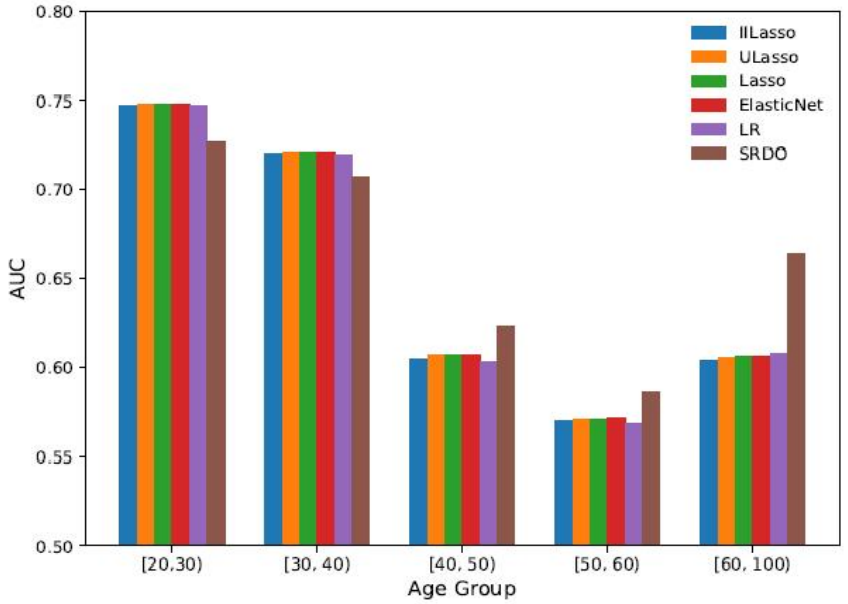


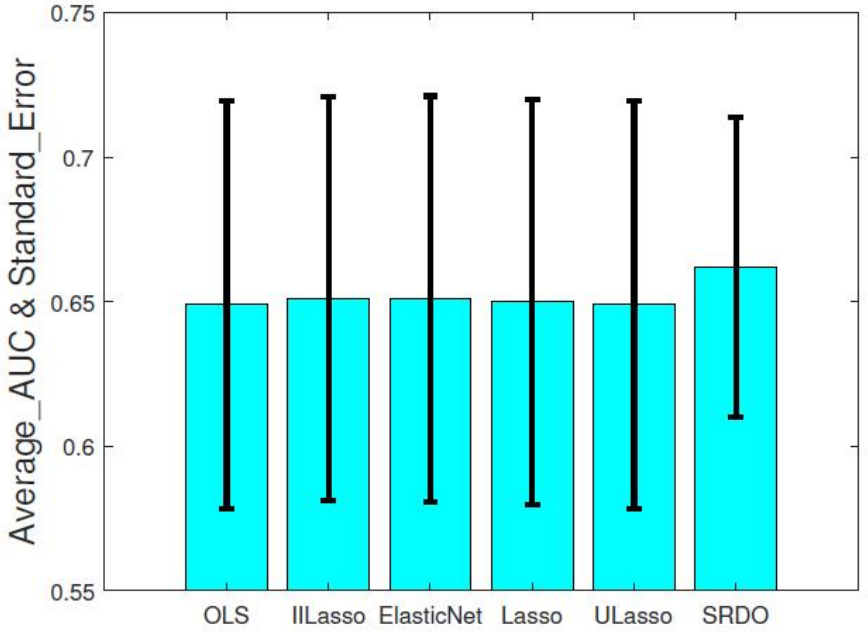
(c) Average prediction error&stability

Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting.

Experimental Results

- Regression
- Classification





(a) AUC over different test environments. (b) Average AUC of all the environments and stability.

Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting.

Stable Learning of Sparse Linear Models

- Suppose $X=\{S,V\}$, and $Y=f(S)+\varepsilon$
- S: set of ***stable (causal) features***, i.e., eyes, ears of dog
- V: set of ***unstable (contextual) features***, i.e., grass, ground
- We assume the outcome is determined by sparse stable signals S regardless of V

Key reason of instability: Spurious correlation between V and Y

Theoretical Analysis

$$\hat{\beta}_{V_{OLS}} = \beta_V + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^T\mathbf{v}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^T g(\mathbf{s}_i)\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^T\mathbf{v}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^T\mathbf{s}_i\right)\left(\beta_S - \hat{\beta}_{S_{OLS}}\right),$$

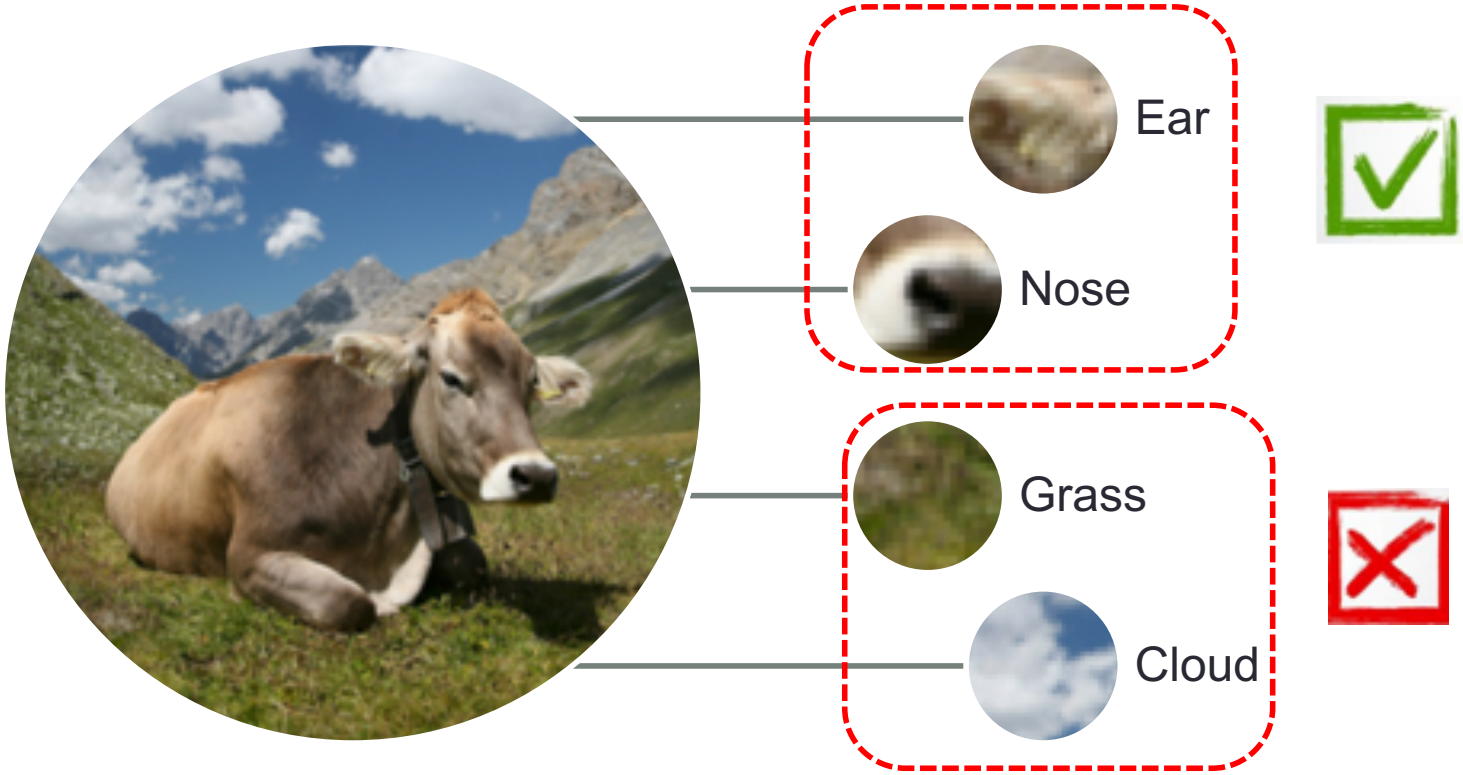
$$\hat{\beta}_{S_{OLS}} = \beta_S + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}_i^T\mathbf{s}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}_i^T g(\mathbf{s}_i)\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}_i^T\mathbf{s}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}_i^T\mathbf{v}_i\right)\left(\beta_V - \hat{\beta}_{V_{OLS}}\right)$$

- The estimation error is induced by
  - Cov(S, V)
  - Cov(V, g(S))
  - Cov(S, g(S))

Spurious correlation between V and S may shift due to different time spans, regions and data collecting strategies, leading to unstable performance.

Our Idea – Heterogeneity & Modularity

ASSUMPTION 3. *The variables* $\mathbf{X} = \{X_1, X_2, \ldots X_p\}$ *could be partitioned into* k *distinct groups* $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_k$. *For* $\forall i, j, i \neq j$ *and* $X_i, X_j \in \mathbf{G}_l, l \in \{1, 2, \ldots, k\}$, *we have* $P^e_{X_i X_j} = P_{X_i X_j}$.



Clustering?

Differentiated Variable Decorrelation

- Feature Partition by Stable Correlation Clustering
  - Define the dissimilarity of two variables:

$$Dis(X_i, X_j) = \sqrt{\frac{1}{M-1}\sum_{l=1}^{M}\left(Corr(X_i^l, X_j^l) - Ave\_Corr(X_i, X_j)\right)^2},$$

- Remove the correlation between variables via sample reweighting:

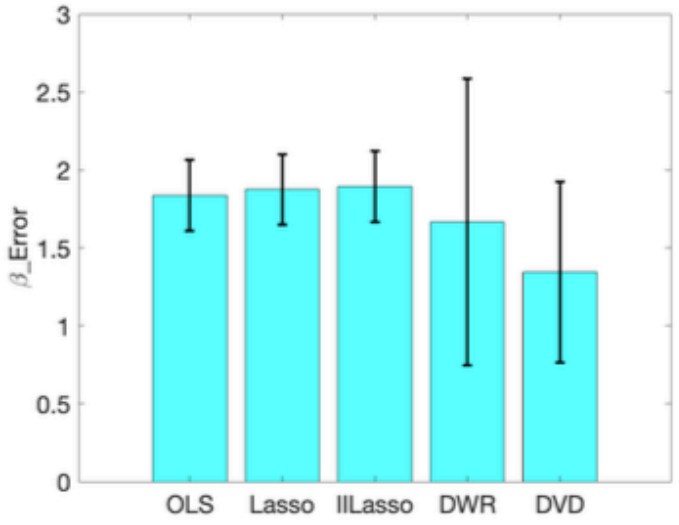
$$\min_{W} \sum_{i \neq j} \mathbb{I}(i,j) \left\| (\mathbf{X}_{,i}^T \Sigma_W \mathbf{X}_{,j}/n - \mathbf{X}_{,i}^T W/n \cdot \mathbf{X}_{,j}^T W/n) \right\|_2^2$$

$$\text{s.t} \ \frac{1}{n}\sum_{i=1}^{n} W_i^2 < \gamma_1, \quad \left(\frac{1}{n}\sum_{i=1}^{n} W_i - 1\right)^2 < \gamma_2, \quad W \geq 0$$

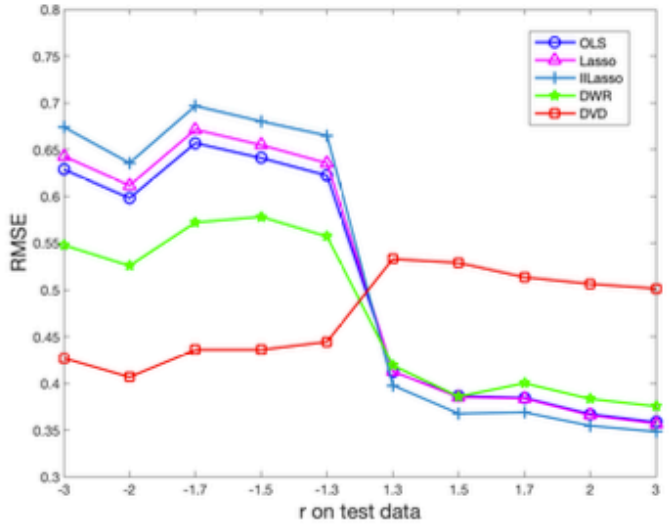
Experiments on Simulation Data

- Data generation
  - $\mathbf{X} \sim N(0, \Sigma)$, $\mathbf{X} = \{\mathbf{S}_{\cdot,1}, \cdots, \mathbf{S}_{\cdot,p_s}, \mathbf{V}_{\cdot,1}, \cdots, \mathbf{V}_{\cdot,p_v}\}$
  - $Y_{poly} = f(\mathbf{S}) + \varepsilon = [\mathbf{S}, \mathbf{V}] \cdot [\beta_s, \beta_v]^T + \mathbf{S}_{\cdot,1}\mathbf{S}_{\cdot,2}\mathbf{S}_{\cdot,3} + \epsilon$
- Simulate Spurious Correlation
  - For each sample, select it with probability $Pr = \prod_{\mathbf{V}_i \in \mathbf{V}_b} |r|^{-5*D_i}$, $D_i = |f(\mathbf{S}) - sign(r) * \mathbf{V}_i|$
  - $r > 1$ indicates positive correlation between outcome and unstable variable
- Simulate changing environments
  - By manipulating the value and sign of bias rate r, we can generate different environments with different correlation.

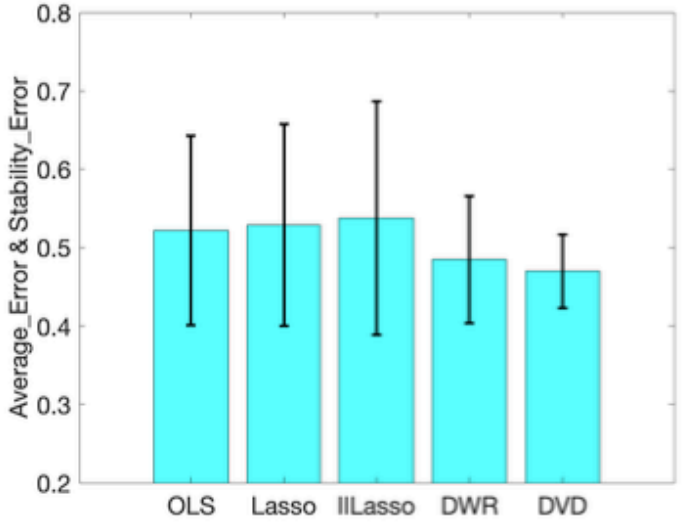
Experimental Results



(a) Estimation error



(b) Prediction error over different test environments

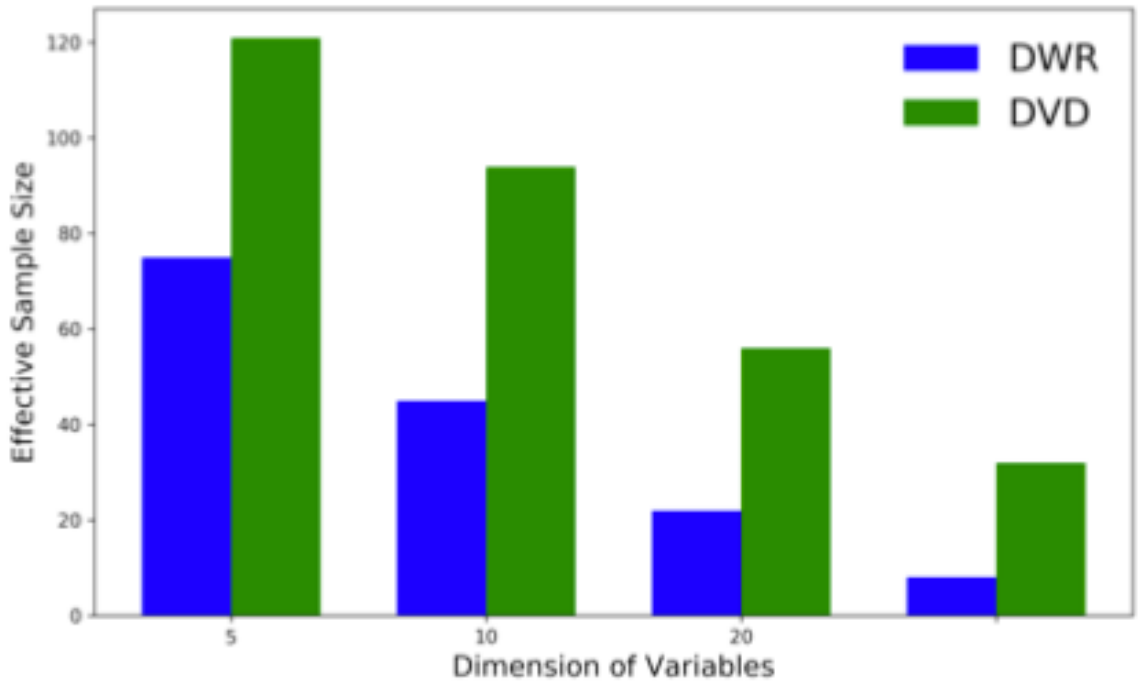


(c) Average prediction error&stability

1. The baseline methods suffer from unstable performance across different test environments, and the performance deteriorate as the discrepancy between training and test data getting larger.
2. DVD can improve estimation error on training data and relieve the variance inflation compared with sample reweighted baseline DWR, leading to its more stable performance across different environments.

Experimental Results

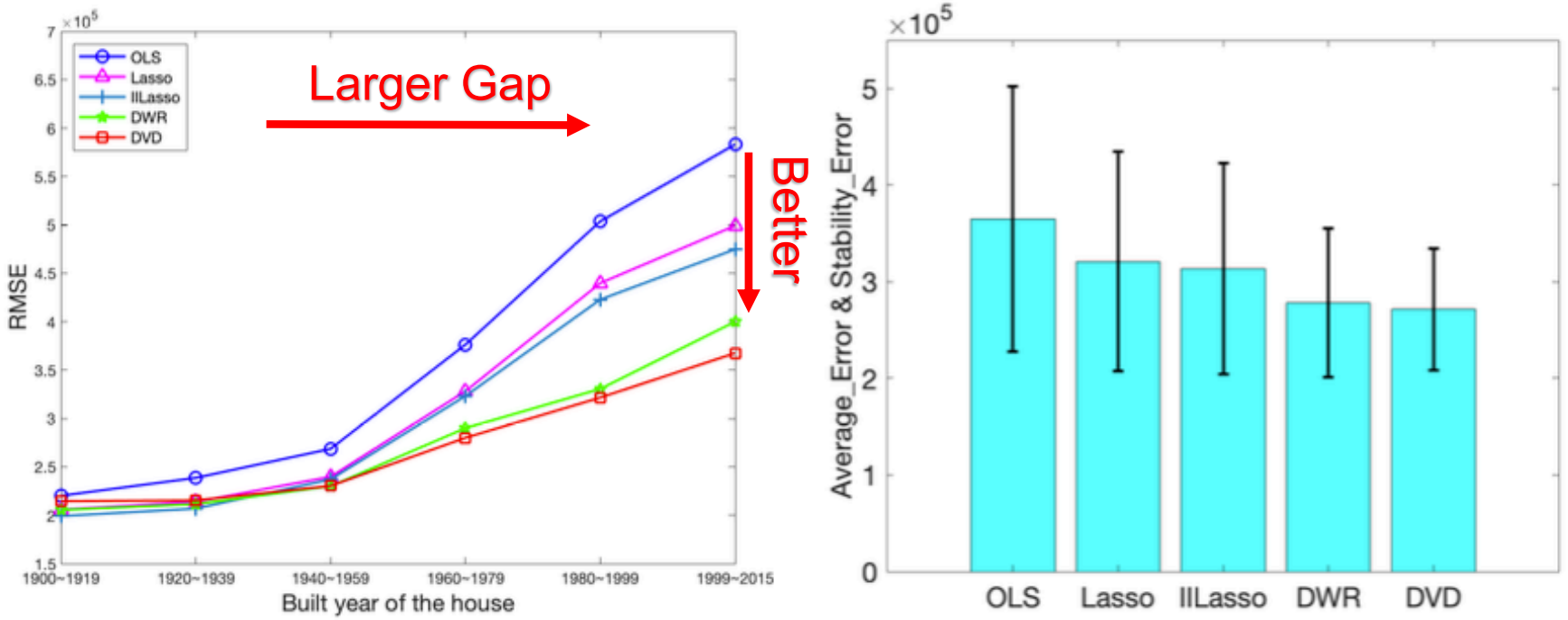
Scenario 1: varying sample size n						
n, p_{v_b}, r	$n = 120, p_{v_b} = p * 0.2, r = 1.9$			$n = 160, p_{v_b} = p * 0.2, r = 1.9$		
Methods	β_Error	Average_Error	Stability_Error	β_Error	Average_Error	Stability_Error
OLS	1.988	0.470	0.087	1.870	0.489	0.105
Lasso	2.021	0.476	0.092	1.905	0.494	0.110
IILasso	2.035	0.475	0.094	1.920	0.498	0.113
DWR	2.012	0.545	0.099	1.991	0.502	0.076
Our	**1.892**	**0.469**	**0.040**	**1.741**	**0.489**	**0.050**
Scenario 2: varying number of unstable variables p_{v_b}						
n, p_{v_b}, r	$n = 200, p_{v_b} = p * 0.2, r = 1.9$			$n = 200, p_{v_b} = p * 0.3, r = 1.9$		
Methods	β_Error	Average_Error	Stability_Error	β_Error	Average_Error	Stability_Error
OLS	1.839	0.522	0.121	2.128	0.563	0.179
Lasso	1.876	0.529	0.129	2.176	0.571	0.186
IILasso	1.894	0.538	0.149	2.196	0.575	0.191
DWR	1.656	0.485	0.081	1.881	0.469	0.092
Our	**1.369**	**0.476**	**0.042**	**1.641**	**0.460**	**0.064**
Scenario 3: varying bias rate r on training data						
n, p_{v_b}, r	$n = 200, p_{v_b} = p * 0.2, r = 1.6$			$n = 200, p_{v_b} = p * 0.2, r = 1.8$		
Methods	β_Error	Average_Error	Stability_Error	β_Error	Average_Error	Stability_Error
OLS	1.296	**0.452**	0.064	1.780	0.510	0.117
Lasso	1.321	0.455	0.067	1.812	0.516	0.123
IILasso	1.339	0.457	0.070	1.829	0.519	0.125
DWR	**1.153**	0.457	0.033	1.262	0.458	0.035
Our	1.236	0.463	**0.021**	**1.236**	**0.450**	**0.023**



Effective Sample Size

Experiments on Real World Data

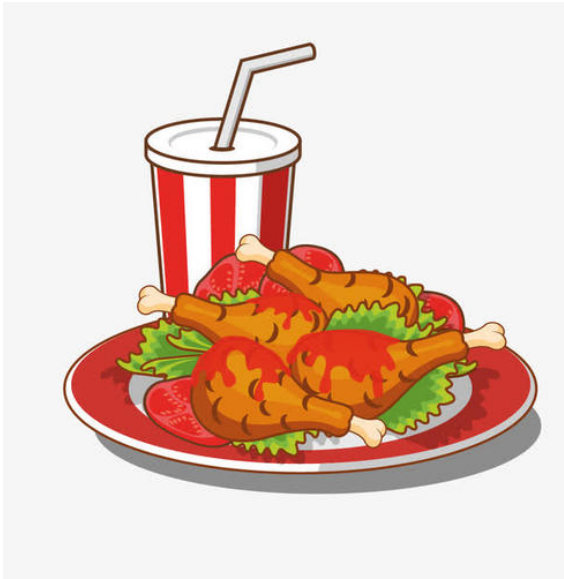
- King county house sales price
  - Features: attributes of house
  - Outcome: the transaction price of house
  - Environments: different periods of house built years

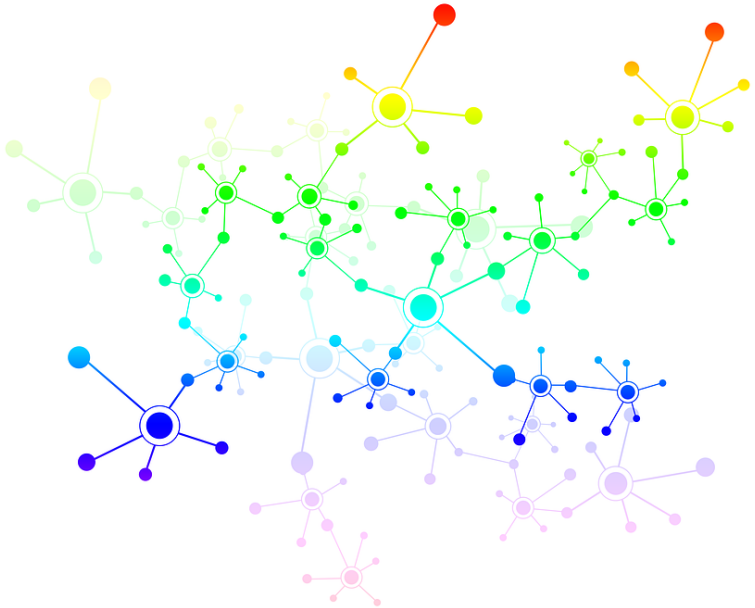


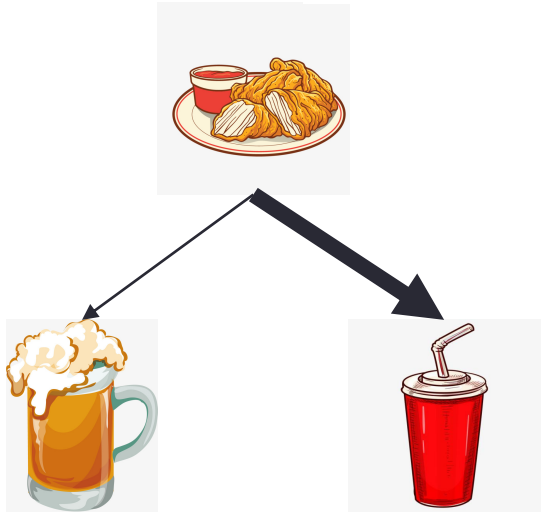
Outline

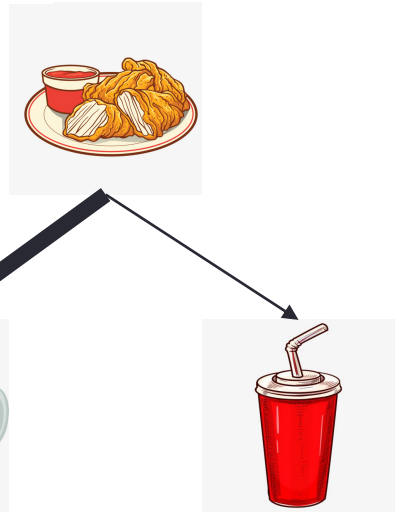
- Stable Learning: Definition and Related Problems
- Stable Learning: From Causally-Oriented Perspective
- Stable Learning: From Statistical Learning Perspective
- Beyond Structural Data: Stable Learning on Graph
- NICO: A Benchmark and Baseline for Stable Learning
- Conclusions

Stability on Graph

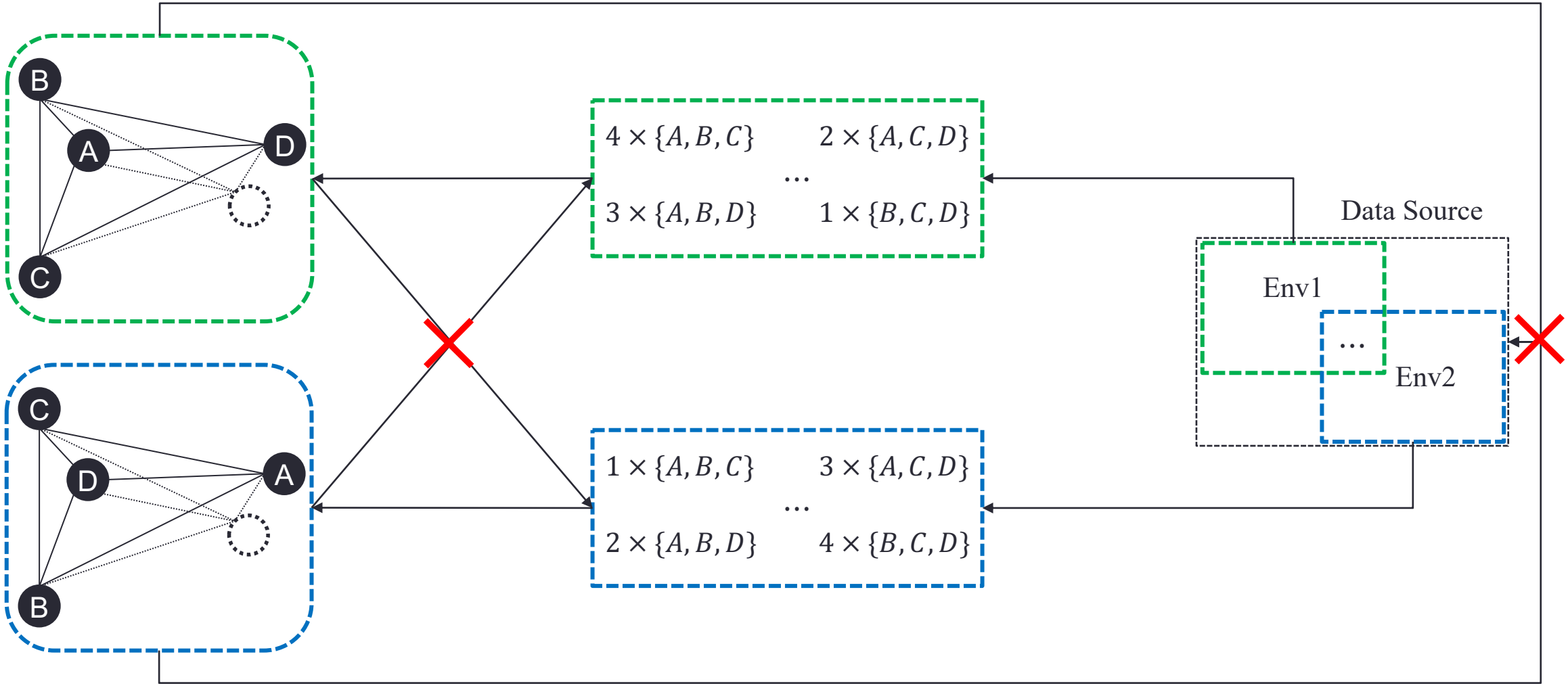






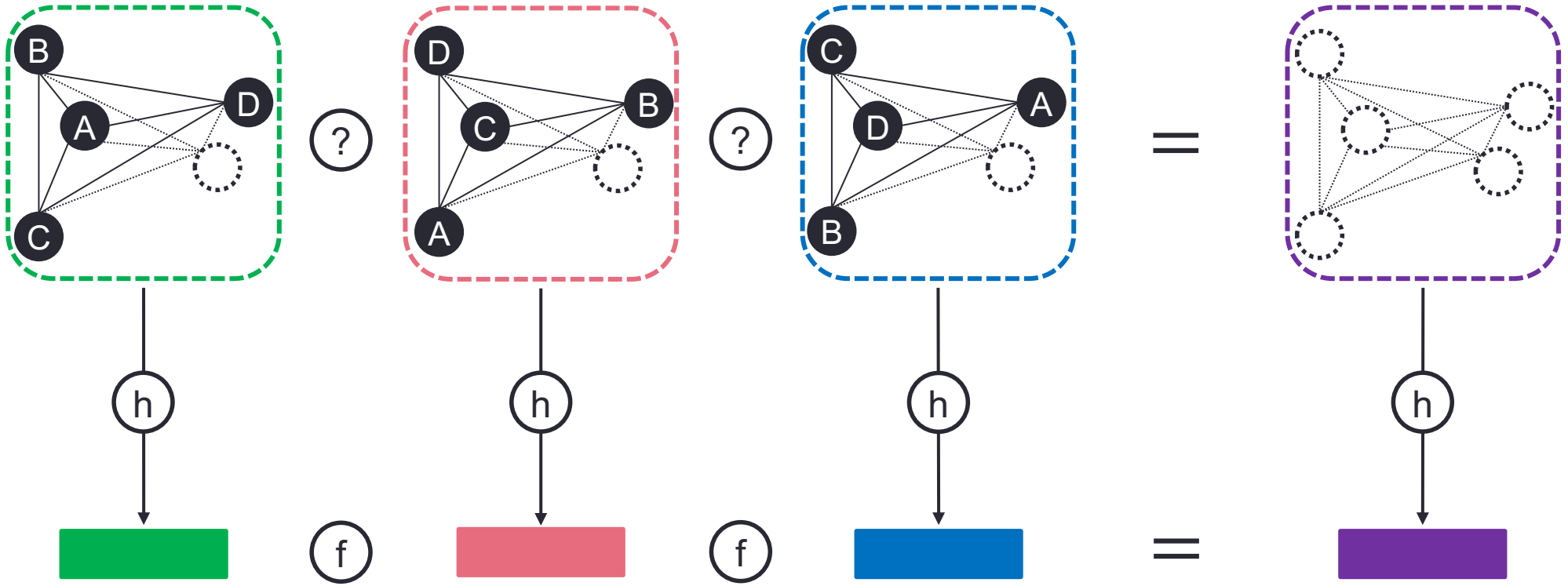


Source of Instability: Selection Bias

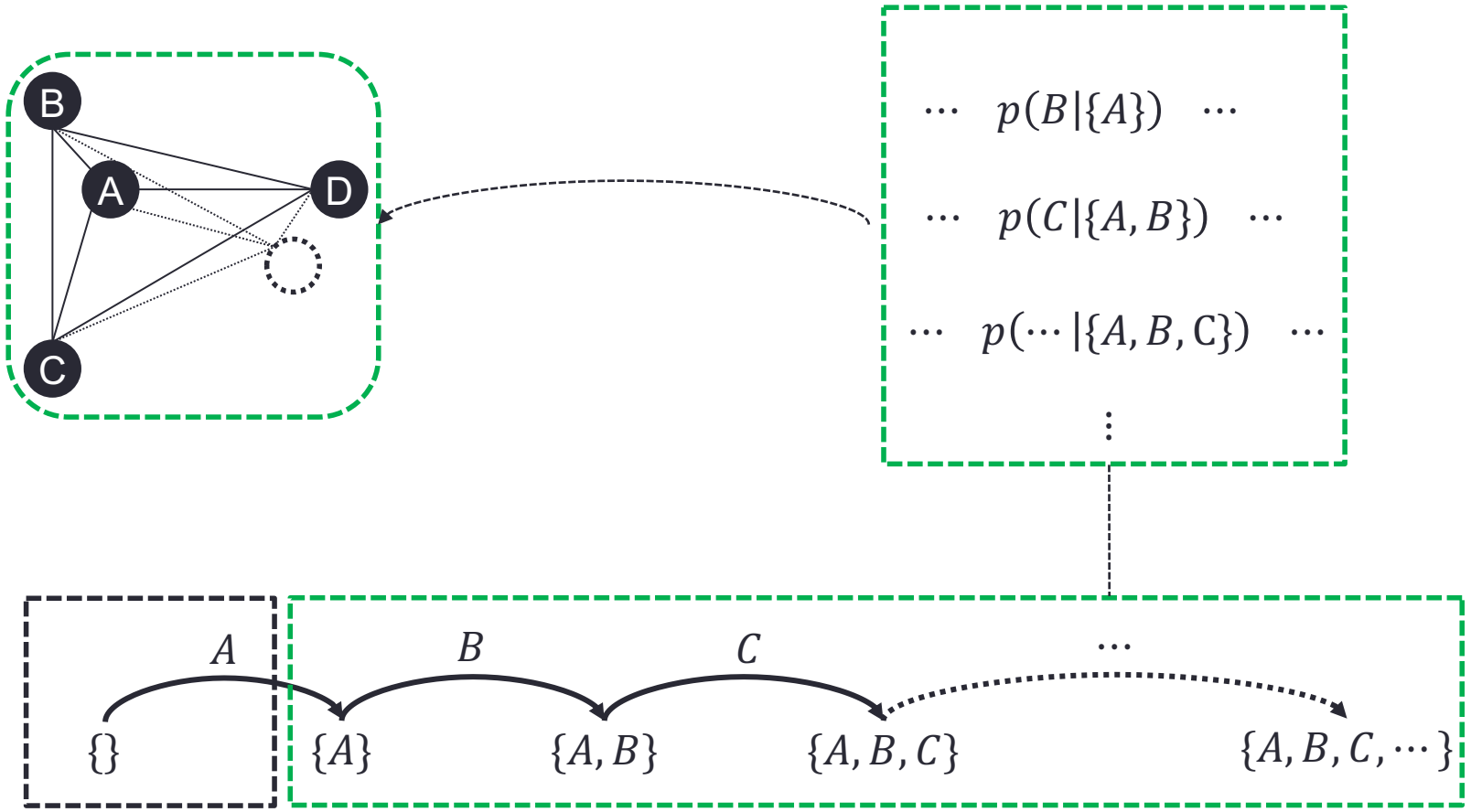


Challenge

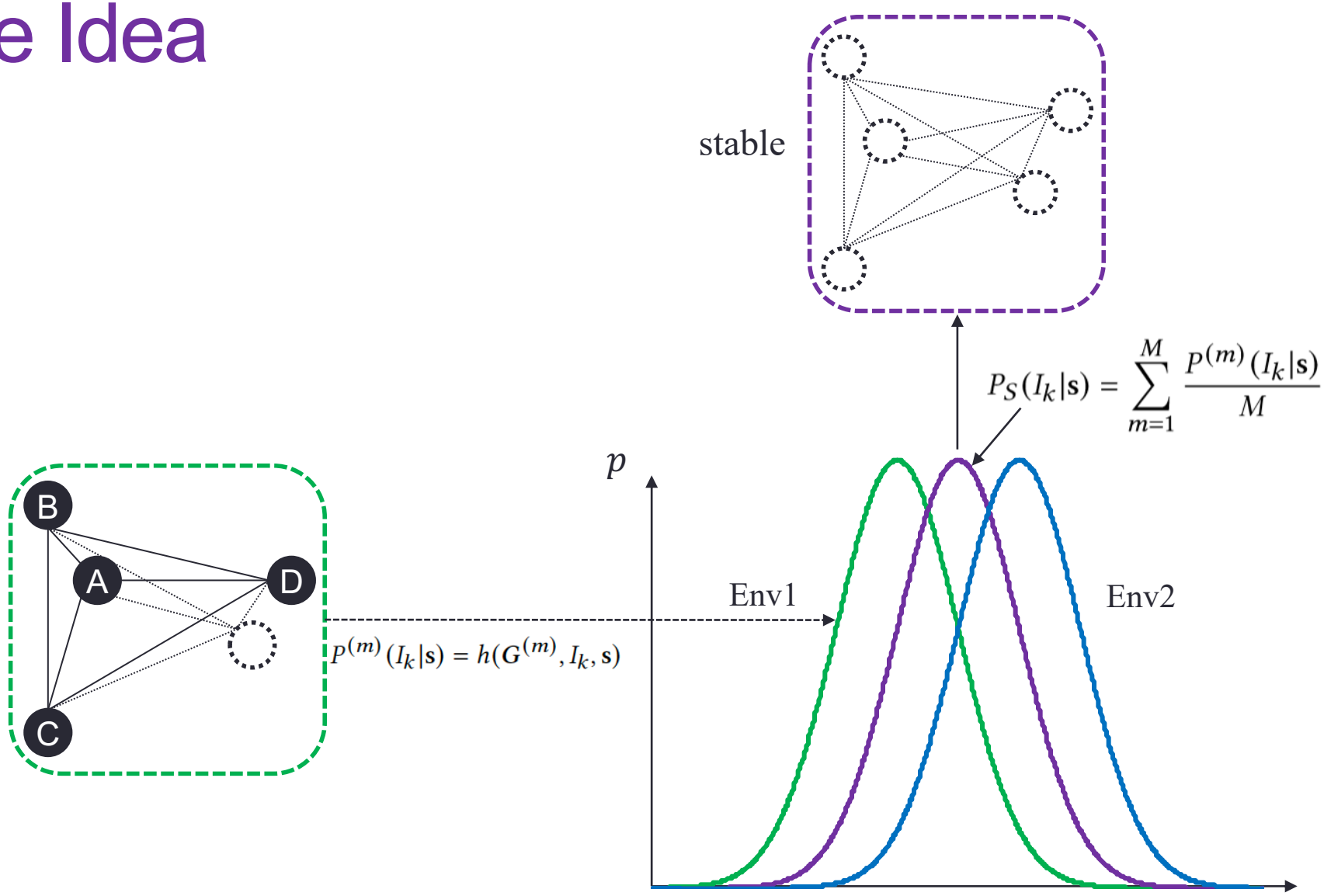
- High Ordered
- Non-Linear
- Large Scale



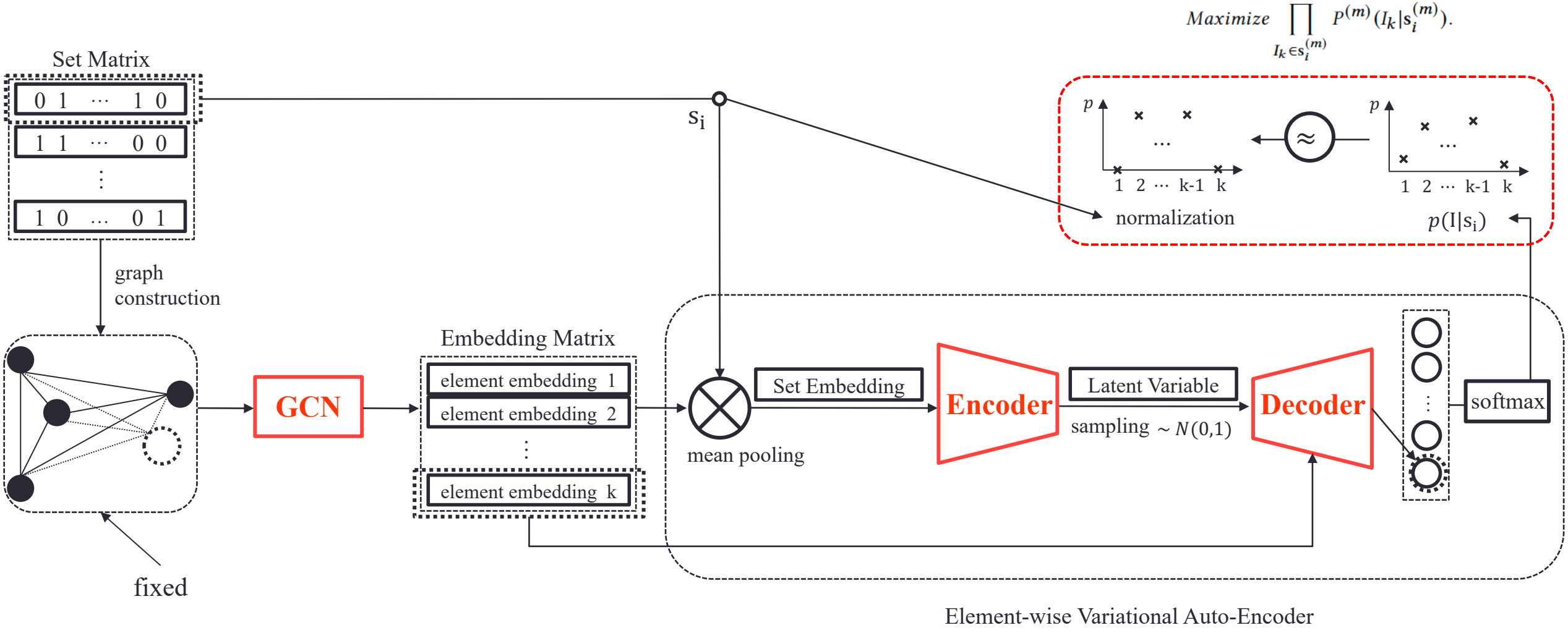
Core Idea


B
A
D
C
$\cdots \quad p(B|\{A\}) \quad \cdots$
$\cdots \quad p(C|\{A,B\}) \quad \cdots$
$\cdots \quad p(\cdots|\{A,B,\mathrm{C}\}) \quad \cdots$
$\vdots$
A
B
C
$\cdots$
$\{\}$
$\{A\}$
$\{A,B\}$
$\{A,B,C\}$
$\{A,B,C,\cdots\}$

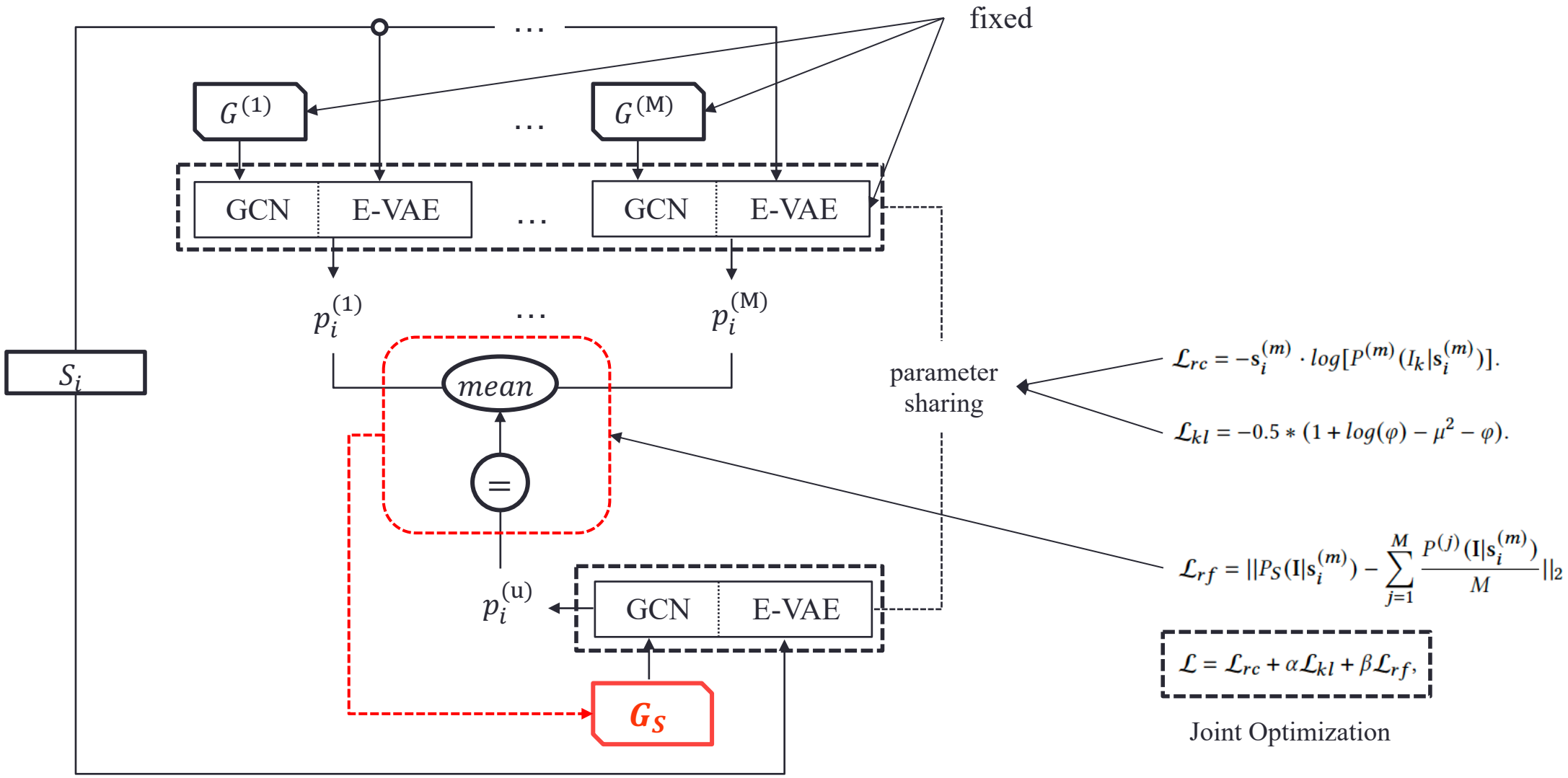
Core Idea


stable
$P_S(I_k|\mathbf{s}) = \sum_{m=1}^{M} \frac{P^{(m)}(I_k|\mathbf{s})}{M}$
B
A
D
C
$P^{(m)}(I_k|\mathbf{s}) = h(\boldsymbol{G}^{(m)}, I_k, \mathbf{s})$
p
Env1
Env2

Graph Based Set Generation in Single Environment



Stable Graph Learning from Multiple Environment



Experiment: Simulation Data

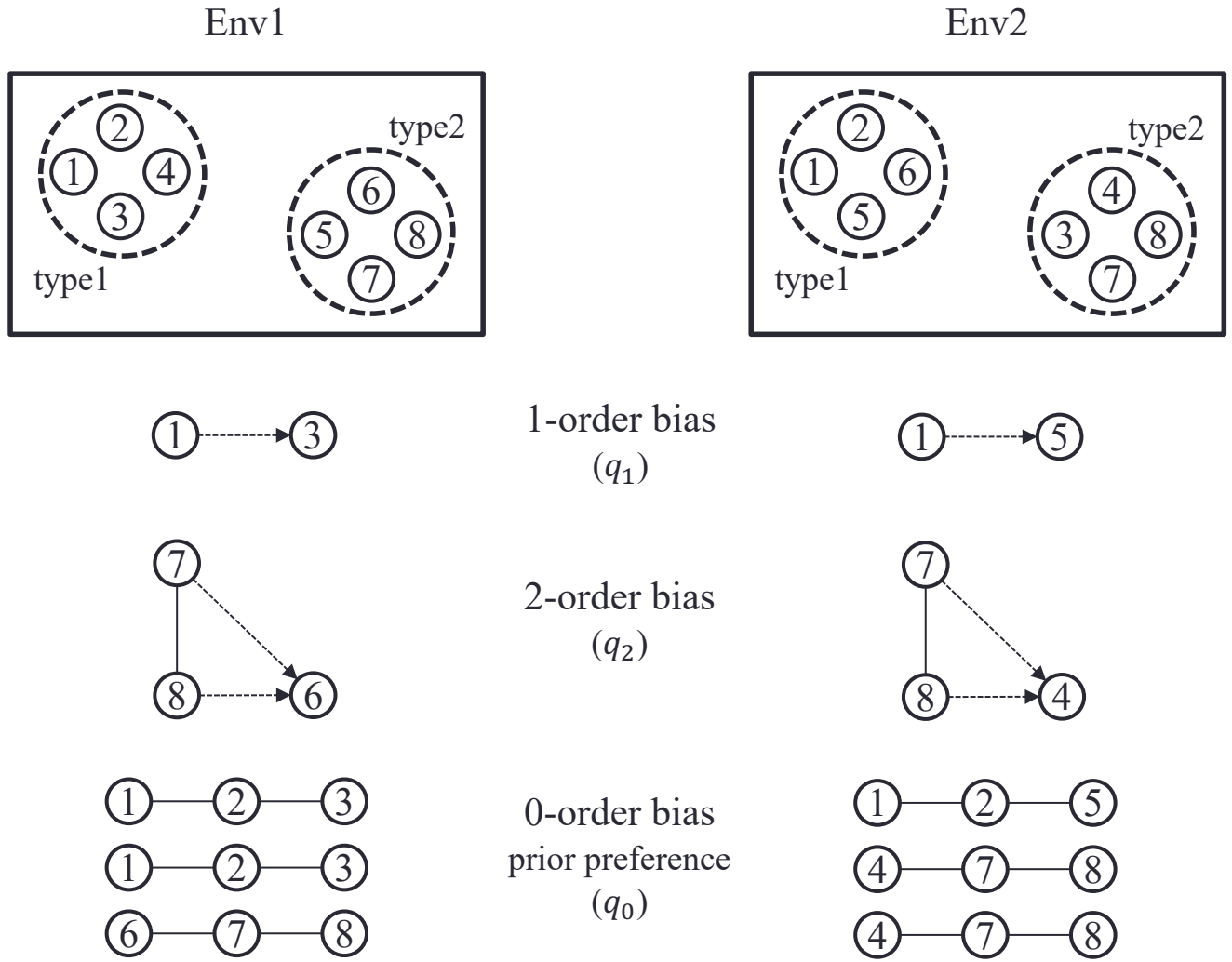
Biased Weighted Random Walk:

Different correlations in different environments

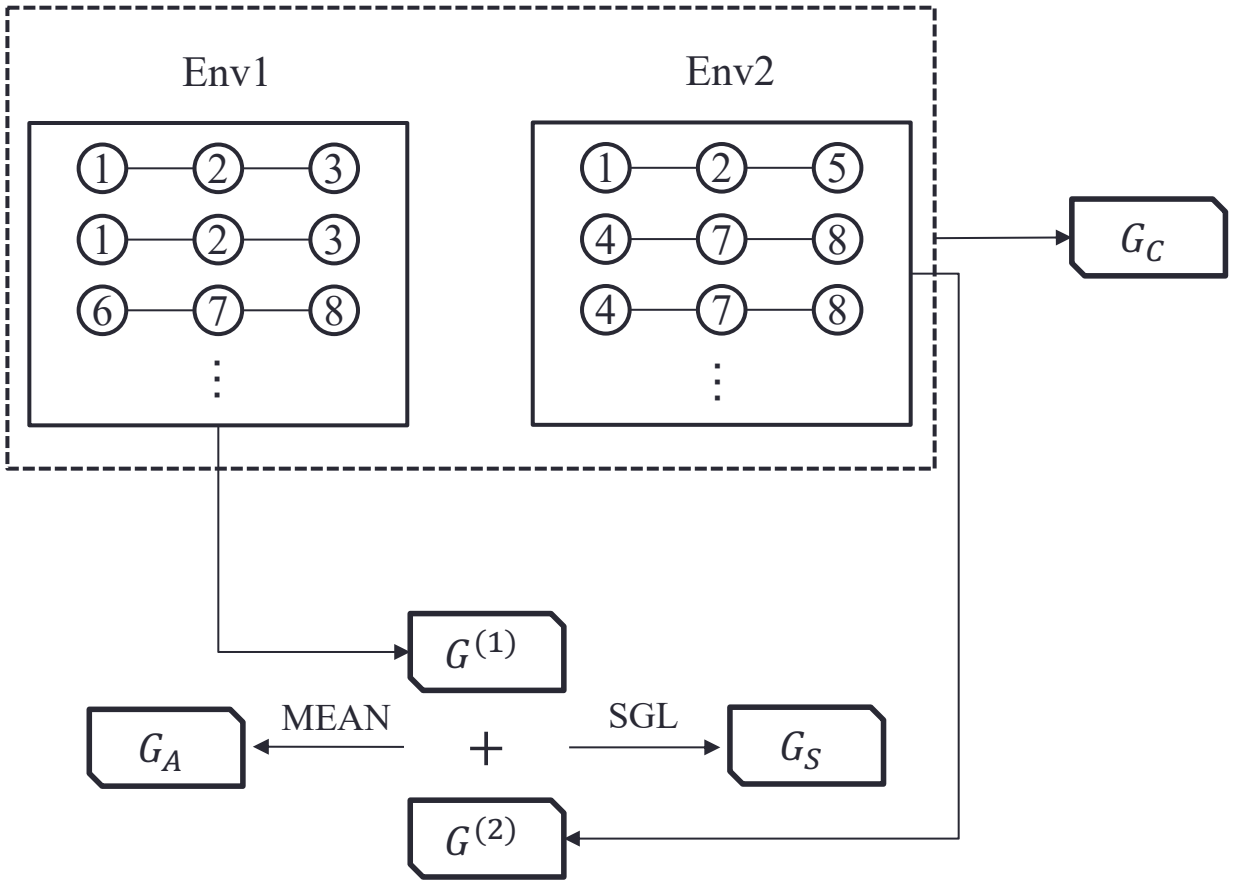
$$Q(t(v), v, u) = \begin{cases} w_{v,u} * q_2 & If\ d(t(v), u) <= 1\ \ and \\ & t(v),\ v,\ u\ are\ the\ same\ type \\ w_{v,u} * q_1 & If\ v,\ u\ are\ the\ same\ type \\ w_{v,u} & Otherwise \end{cases}$$

Moving Probability

$$\pi'_{v,u} = \frac{\boxed{Q(t(v), v, u)} * w_{v,u}}{\sum_{u' \in N(v)} Q(t(v), v, u') * w_{v,u'}}$$



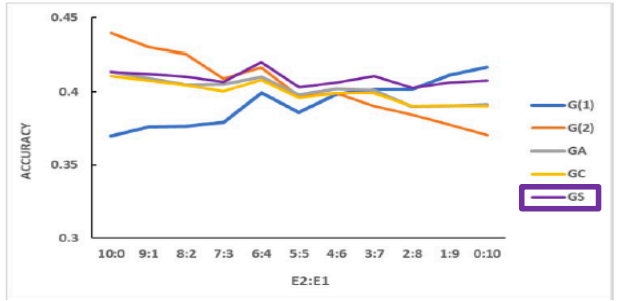
Experiment: Simulation Data



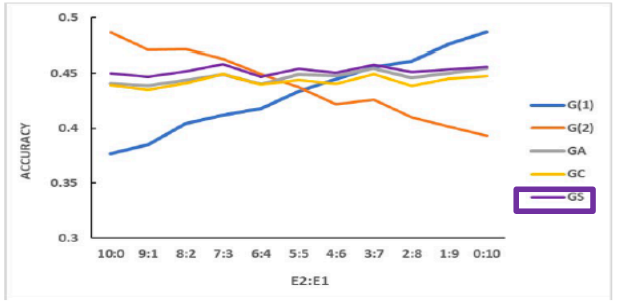
MEAN of ACCURACY				
	$q_0 = 1$			
	$q_1 = 2, q_2 = 4$	$q_1 = 2, q_2 = 6$	$q_1 = 3, q_2 = 6$	$q_1 = 3, q_2 = 9$
$G^{(1)}$	39.24%	43.23%	44.03%	45.18%
$G^{(2)}$	40.36%	43.93%	44.29%	45.34%
G_A	40.14%	44.69%	43.94%	47.62%
G_C	39.98%	44.28%	44.38%	46.96%
G_S	**40.91%**	**45.27%**	**45.02%**	**48.38%**
	$q_0 = 3$			
	$q_1 = 2, q_2 = 4$	$q_1 = 2, q_2 = 6$	$q_1 = 3, q_2 = 6$	$q_1 = 3, q_2 = 9$
$G^{(1)}$	39.58%	40.31%	43.70%	44.97%
$G^{(2)}$	39.43%	40.89%	43.67%	43.78%
G_A	38.40%	39.92%	44.02%	46.64%
G_C	38.68%	40.34%	43.23%	46.68%
G_S	**39.90%**	**41.45%**	**44.99%**	**48.79%**

11 testing datasets: mixing of Env1 and Env2 (10:0 to 0:10)

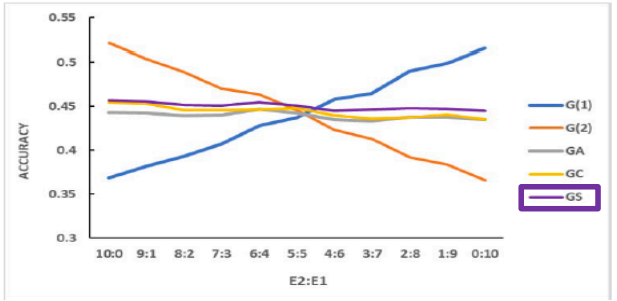
Experiment: Simulation Data



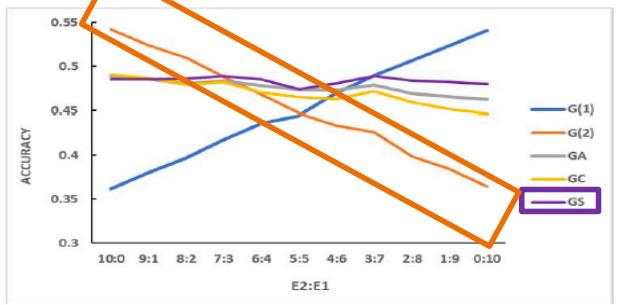
(a) $q_0 = 1, q_1 = 2, q_1 = 4$



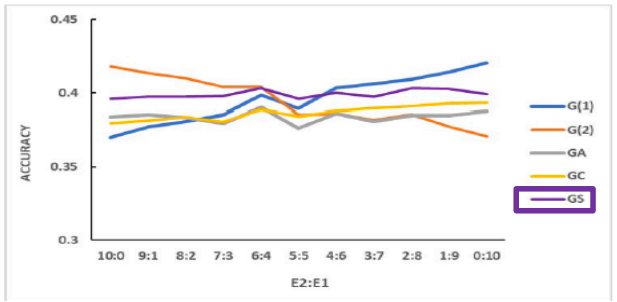
(b) $q_0 = 1, q_1 = 2, q_2 = 6$



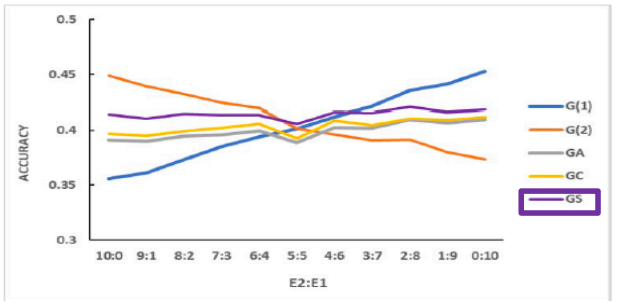
(c) $q_0 = 1, q_1 = 3, q_2 = 6$



(d) $q_0 = 1, q_1 = 3, q_2 = 9$

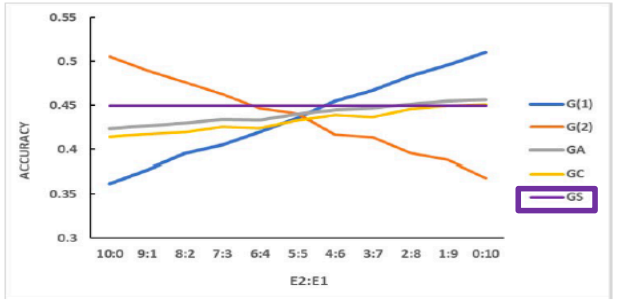


(e) $q_0 = 3, q_1 = 2, q_1 = 4$

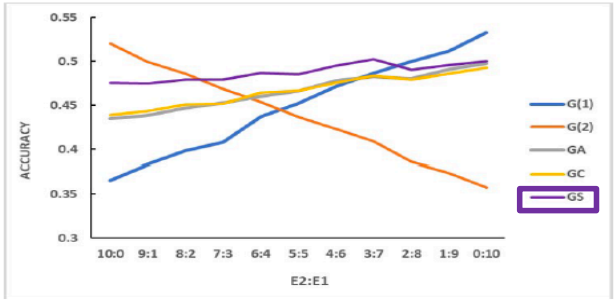


(f) $q_0 = 3, q_1 = 2, q_2 = 6$

Stability Improvement

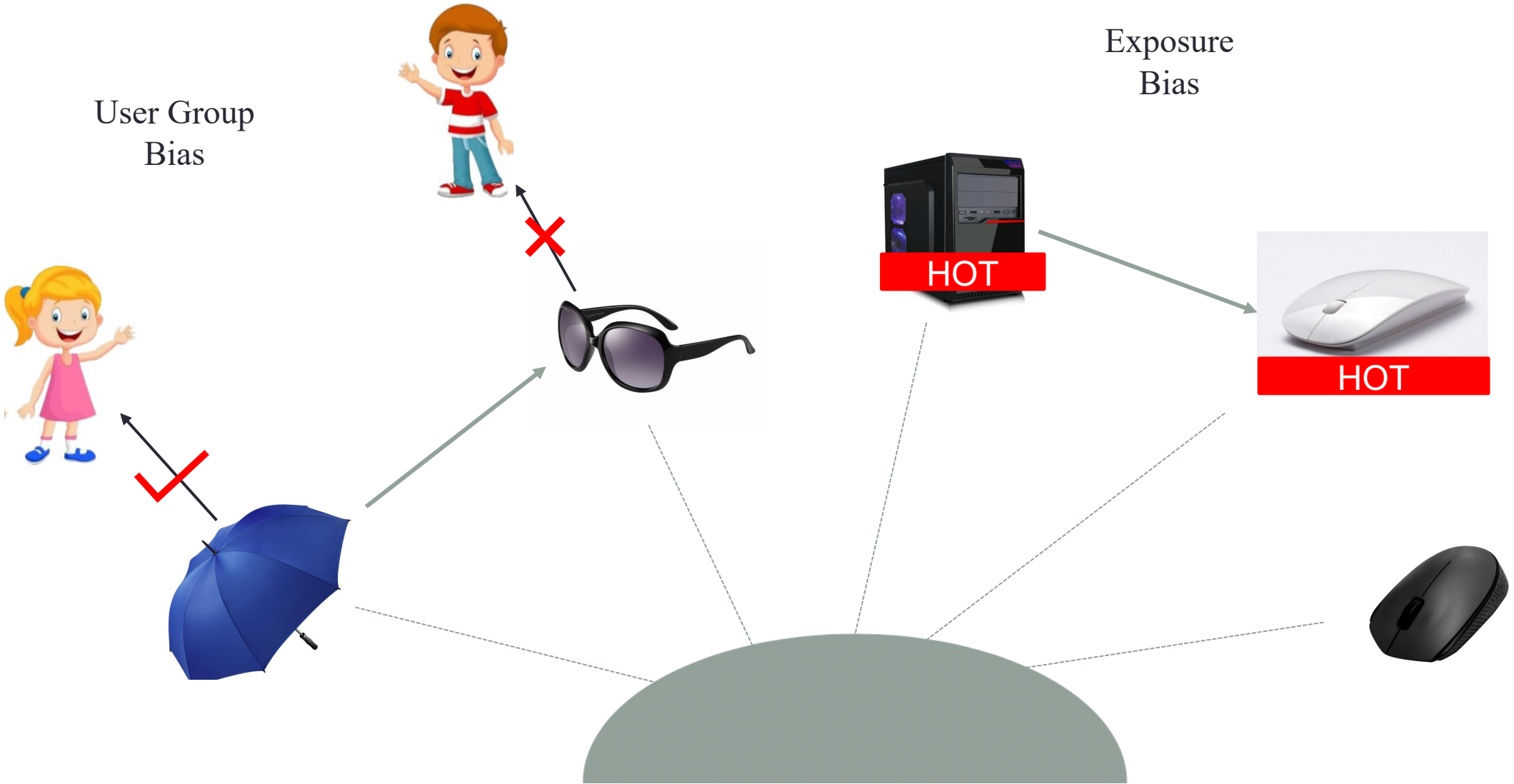


(g) $q_0 = 3, q_1 = 3, q_2 = 6$

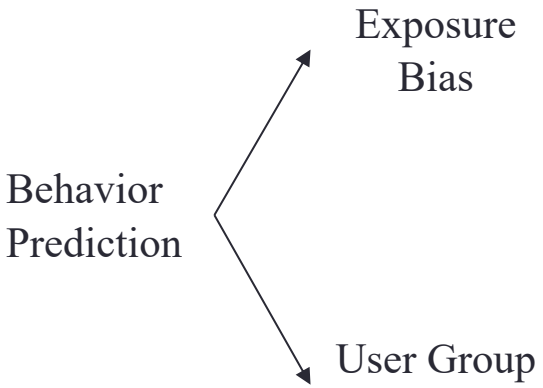


(h) $q_0 = 3, q_1 = 3, q_2 = 9$

Experiment on Commodity Recommendation



Experiment on Commodity Recommendation



	Mean ACC	STD	Env1:Env2=0:10	1:9	2:8	3:7	4:6	5:5	6:4	7:3	8:2	9:1	10:0
Env1 $G^{(1)}$	12.93%	**0.0050**	13.54%	13.42%	13.62%	12.62%	13.49%	12.87%	12.92%	12.81%	12.16%	12.17%	12.62%
Env2 $G^{(2)}$	15.96%	0.0485	23.12%	22.21%	20.64%	18.83%	17.91%	16.11%	14.04%	13.67%	11.62%	9.46%	7.56%
G_A	18.09%	0.0347	23.21%	22.74%	21.76%	19.68%	19.46%	17.71%	16.68%	16.87%	14.98%	13.34%	12.57%
G_C	16.15%	0.0310	20.50%	20.57%	19.23%	17.49%	17.47%	16.08%	14.79%	15.24%	13.26%	11.78%	11.23%
G_S	**18.64%**	**0.0288**	22.90%	22.44%	21.81%	19.71%	19.98%	18.53%	17.31%	17.57%	15.91%	14.68%	14.22%

Table 4: Purchasing behavior prediction with exposure bias using item embeddings learnt from commodity network. The environment 1 consists of shopping logs mainly with unpopular items and env 2 consists of logs mainly with popular items.

	Mean ACC	STD	Env1:Env2=0:10	1:9	2:8	3:7	4:6	5:5	6:4	7:3	8:2	9:1	10:0
Env1 $G^{(1)}$	17.69%	0.0148	15.85%	16.03%	16.44%	16.64%	16.94%	16.67%	18.22%	18.87%	18.87%	19.99%	20.03%
Env2 $G^{(2)}$	17.46%	**0.0063**	16.86%	16.97%	16.56%	16.87%	17.17%	18.16%	16.99%	18.07%	18.10%	18.09%	18.27%
G_A	18.51%	0.0132	16.79%	16.94%	17.24%	17.60%	17.94%	18.50%	18.24%	19.78%	19.43%	20.46%	20.70%
G_C	18.56%	0.0127	16.84%	16.97%	17.34%	17.63%	17.81%	18.68%	18.94%	19.50%	19.53%	20.33%	20.62%
G_S	**20.17%**	**0.0092**	19.09%	19.01%	19.14%	19.51%	19.77%	20.02%	20.29%	20.84%	21.02%	21.62%	21.53%

Table 5: Purchasing behavior prediction in different gender groups using item embeddings learnt from commodity network. The environment 1 consists of shopping logs of females and env 2 consists of logs of males.

- Behaviors in different environments have different prediction difficulties.
- Stable graph can reach the highest mean accuracy more stably.

Outline

- Stable Learning: Definition and Related Problems
- Stable Learning: From Causally-Oriented Perspective
- Stable Learning: From Statistical Learning Perspective
- Beyond Structural Data: Stable Learning on Graph
- NICO: A Benchmark and Baseline for Stable Learning
- Conclusions

Existence of Non-I.I.Dness

- One metric (NI) for Non-I.I.Dness

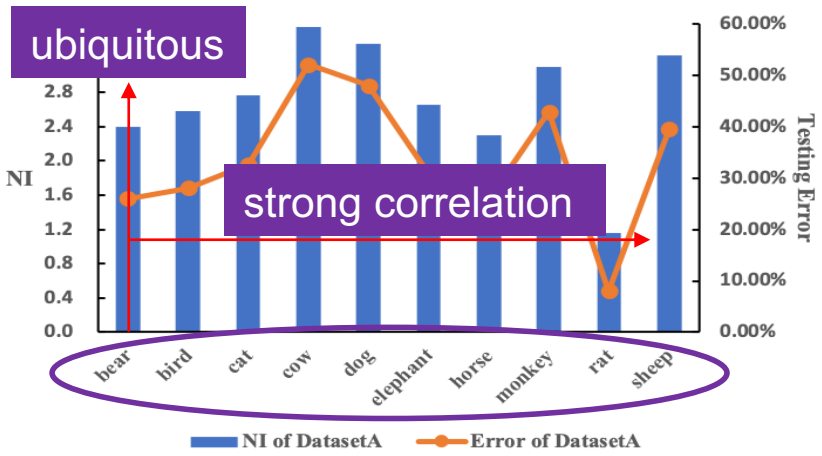
Definition 1 ***Non-I.I.D. Index (NI)*** *Given a feature extractor* $g_\varphi(\cdot)$ *and a class* C, ***the degree of distribution shift*** *between training data* D^C_{train} *and testing data* D^C_{test} *is defined as:*

$$NI(C) = \left\| \frac{\overline{g_\varphi(X^C_{train})} - \overline{g_\varphi(X^C_{test})}}{\sigma(g_\varphi(X^C_{train} \cup X^C_{test}))} \right\|_2,$$

Distribution shift

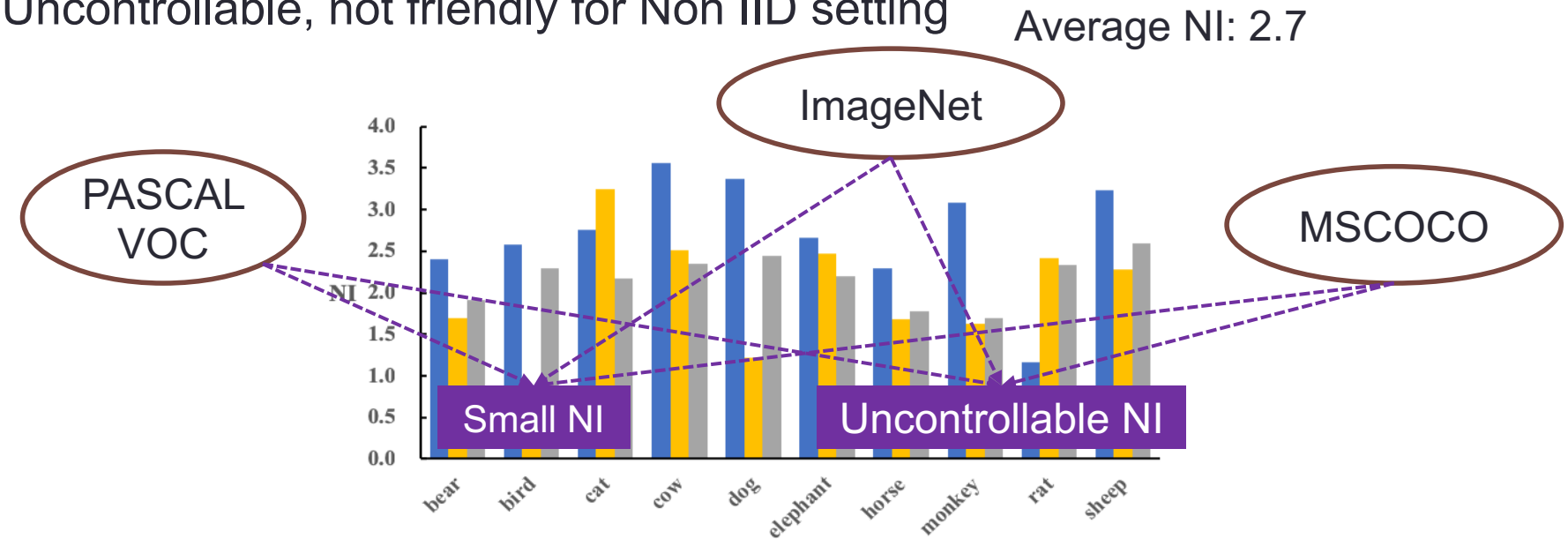
For normalization

- Existence of Non-I.I.Dness on Dataset consisted of 10 subclasses from ImageNet
- For each class
  - Training data
  - Testing data
  - CNN for prediction



Related Datasets

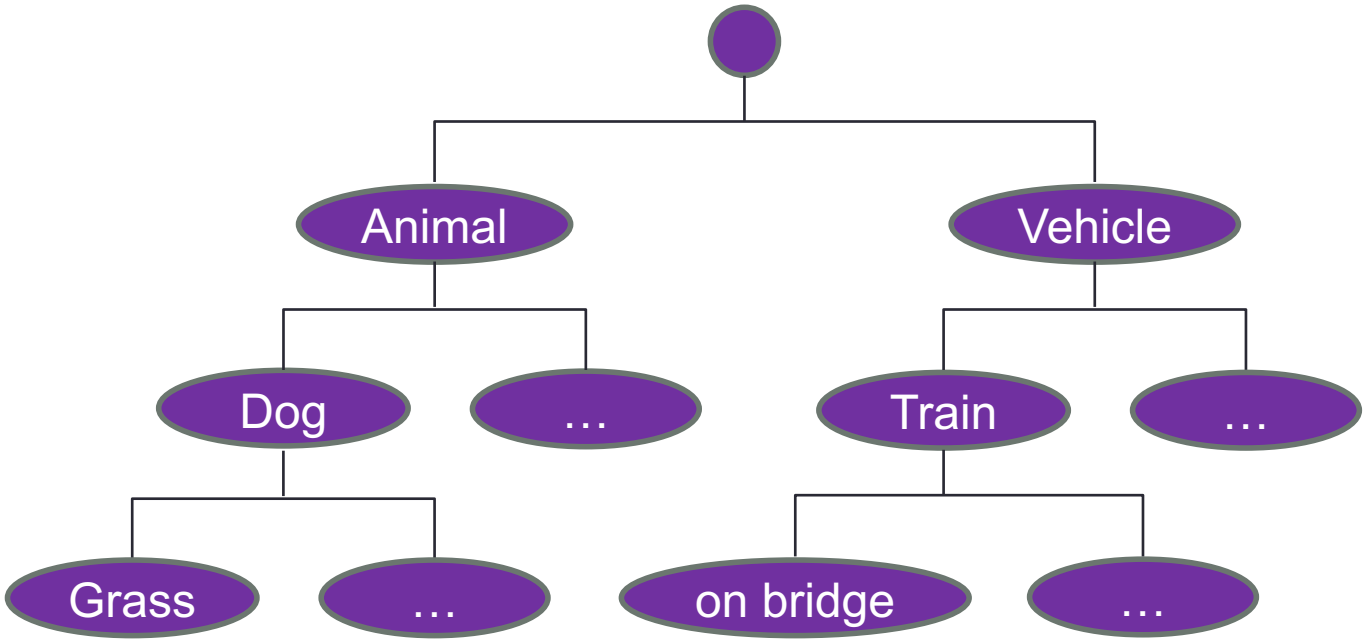
- DatasetA & DatasetB & DatasetC
  - NI is ubiquitous, but small on these datasets
  - NI is Uncontrollable, not friendly for Non IID setting

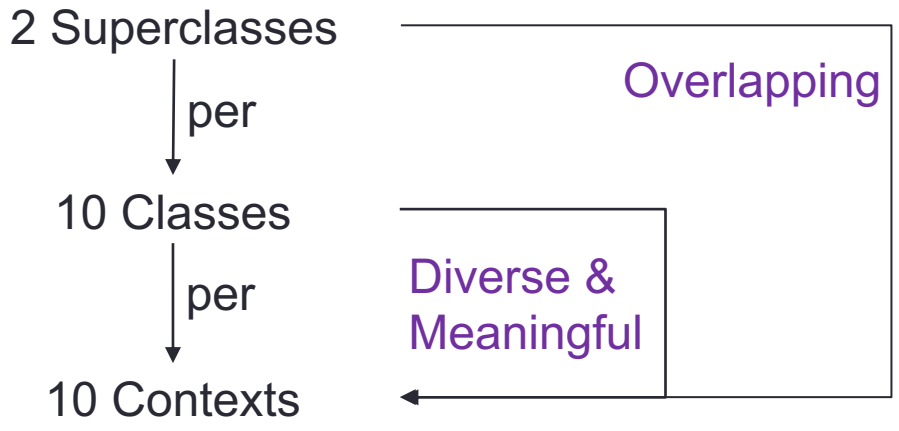


A benchmark for visual stable learning is demanded!

NICO - Non-I.I.D. Image Dataset with Contexts

- **NICO** Datasets:
- Object label: e.g. dog
- Contextual labels (Contexts)
  - the background or scene of a object, e.g. grass/water
- Structure of NICO





NICO - Non-I.I.D. Image Dataset with Contexts

- Data size of each class in NICO
  - Sample size: thousands for each class
  - Each superclass: 10,000 images
  - Sufficient for some basic neural networks (CNN)
- Samples with contexts in NICO

Animal	Data Size	*Vehicle*	Data Size
Bear	1609	Airplane	930
Bird	1590	Bicycle	1639
Cat	1479	Boat	2156
Cow	1192	Bus	1009
Dog	1624	Car	1026
Elephant	1178	Helicopter	1351
Horse	1258	Motorcycle	1542
Monkey	1117	Train	750
Rat	846	Truck	1000
Sheep	918		

Controlling NI on NICO Dataset

- Minimum Bias (comparing with ImageNet)
- Proportional Bias (controllable)
  - Number of samples in each context
- Compositional Bias (controllable)
  - Number of contexts that observed

Minimum Bias

- In this setting, the way of random sampling leads to minimum distribution shift between training and testing distributions in dataset, which simulates a nearly i.i.d. scenario.
  - 8000 samples for training and 2000 samples for testing in each superclass (ConvNet)

	Average NI	Testing Accuracy
Animal	3.85	49.6%
Vehicle	3.20	63.0%

Average NI on ImageNet: 2.7

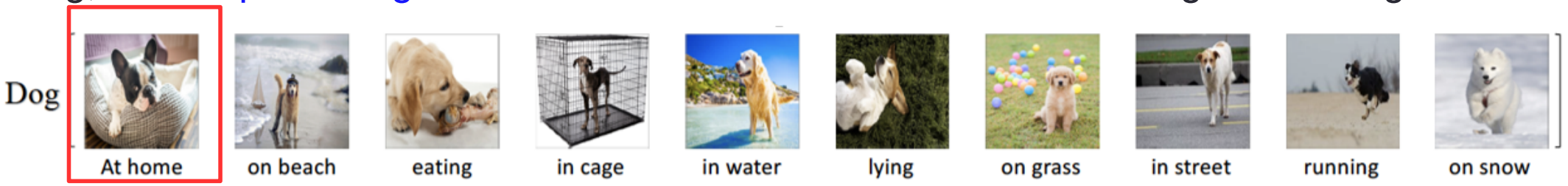
Images in NICO are with rich contextual information

more challenging for image classification

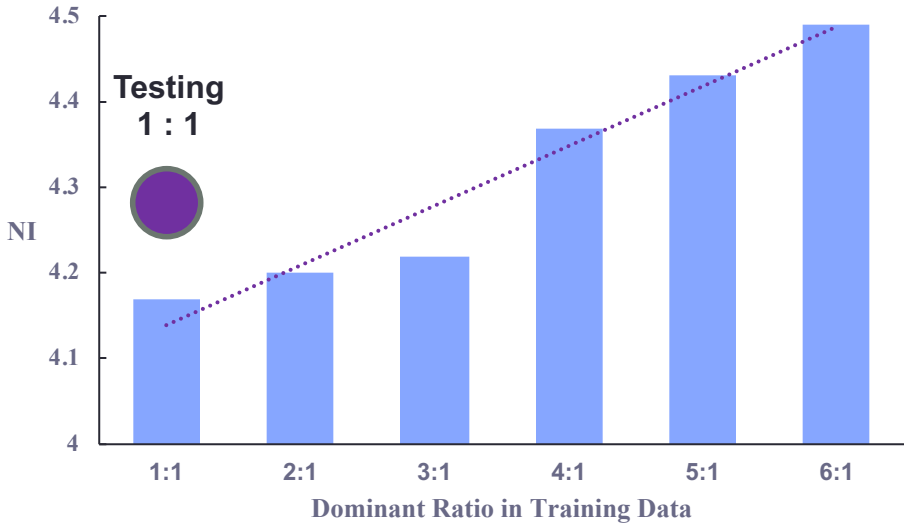
Our NICO data is more Non-iid, more challenging

Proportional Bias

- Given a class, when sampling positive samples, we use all contexts for both training and testing, but the percentage of each context is different between training and testing dataset.



$$Dominant\ Ratio = \frac{N_{dominant}}{N_{minor}}$$

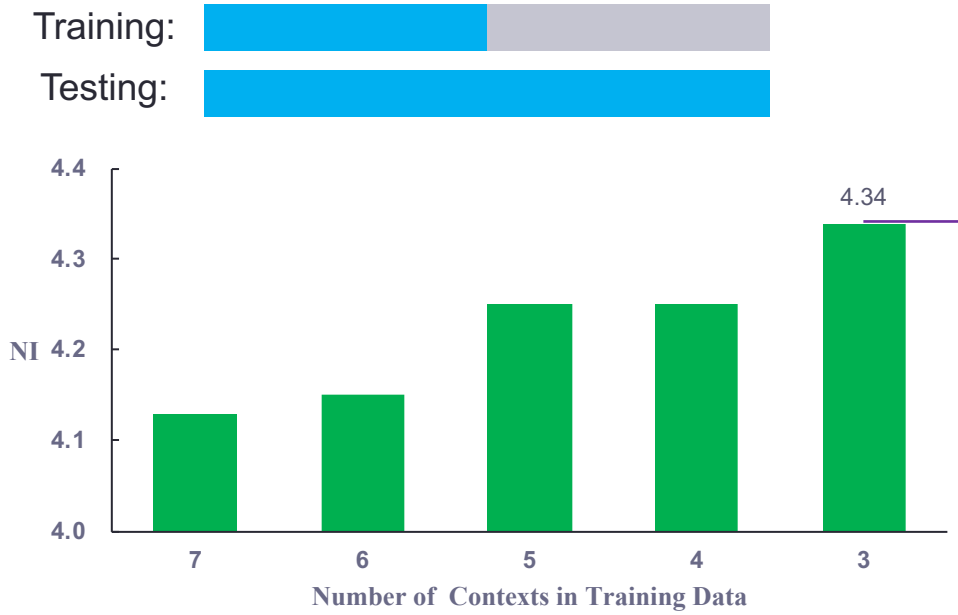


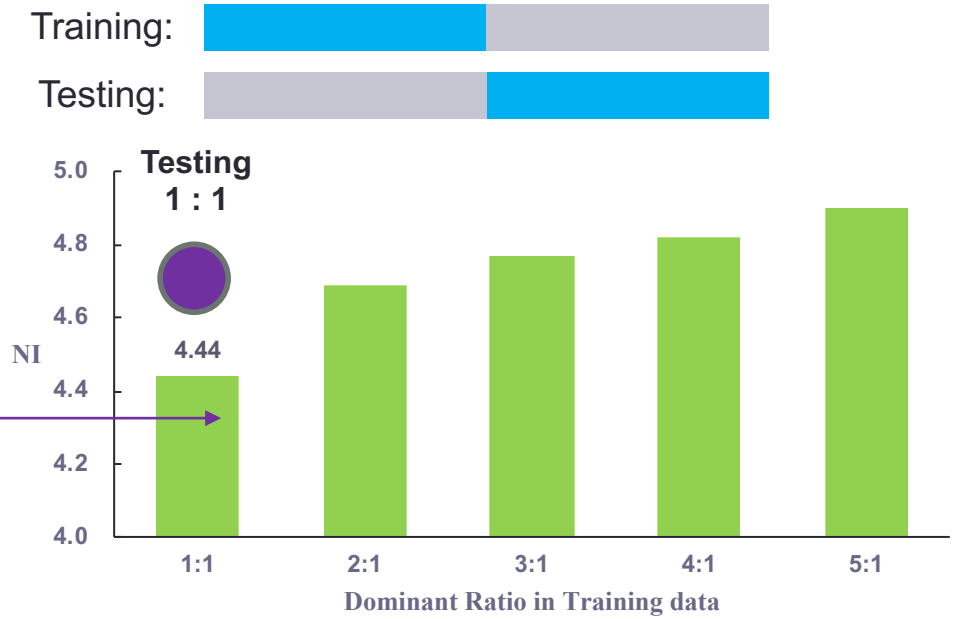
We can control NI by varying dominate ratio

Compositional Bias

$$Dominant\ Ratio = \frac{N_{dominant}}{N_{minor}}$$

- Given a class, the observed contexts are different between training and testing data.



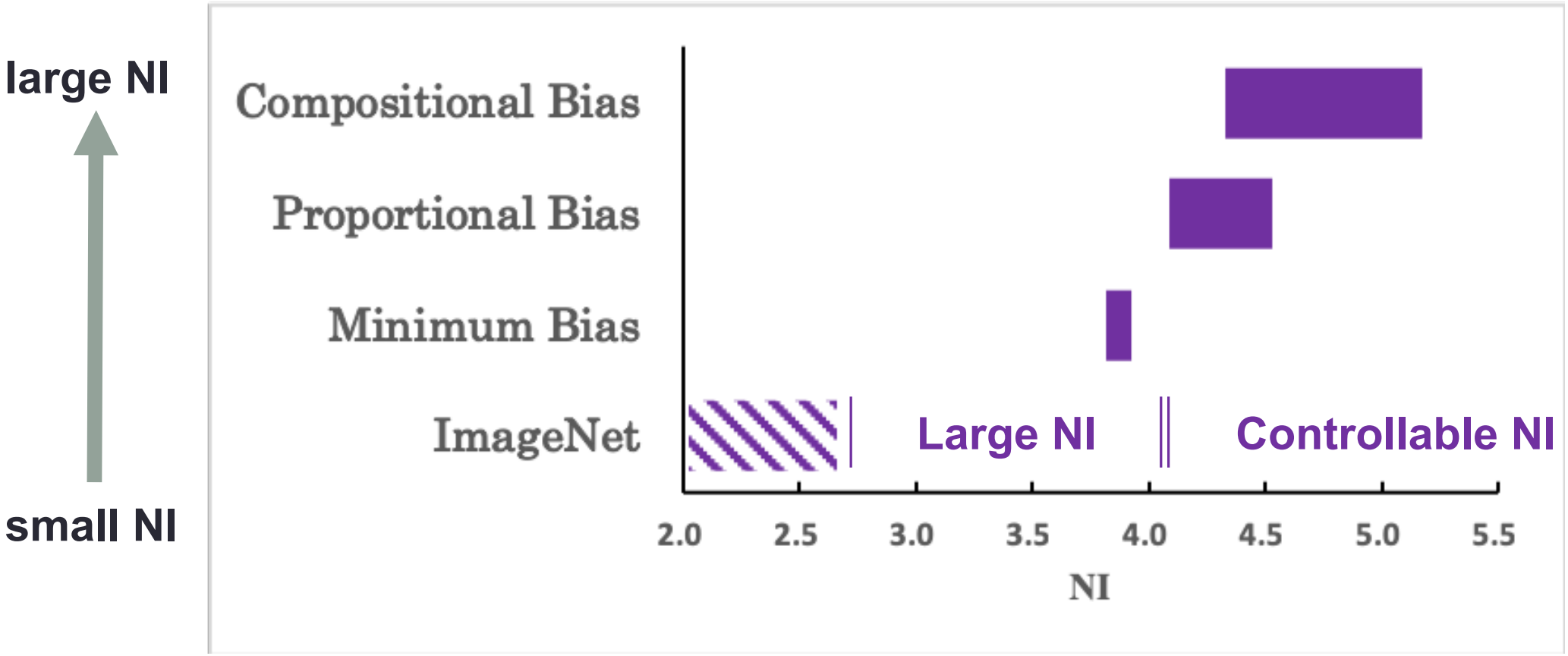


Moderate setting **(Overlap)**

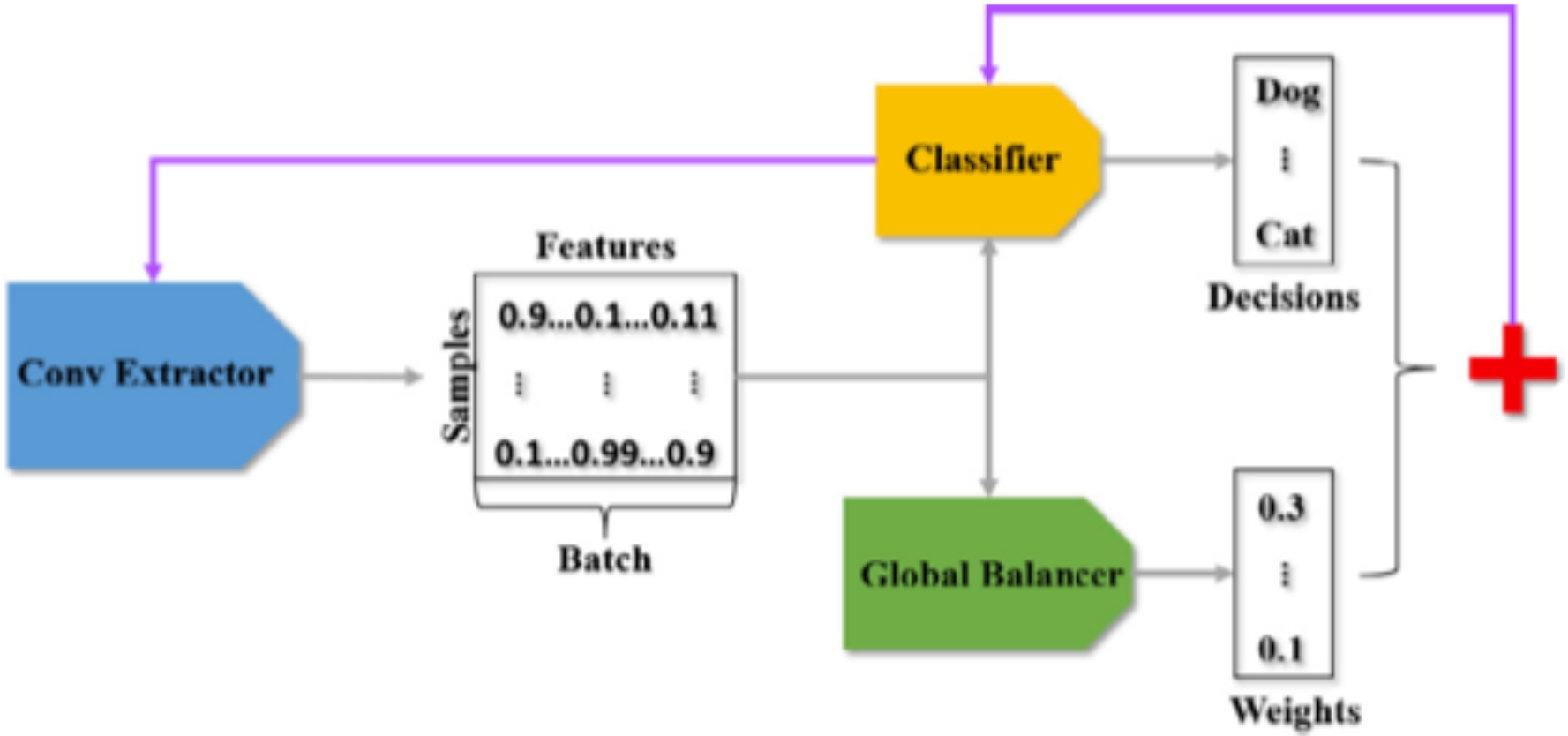
Radical setting (No Overlap & Dominant ratio)

NICO - Non-I.I.D. Image Dataset with Contexts

- Large and controllable NI



ConvNet with Batch Balancing (CNBB)



Experimental Results on NICO

Table 2
Performances of different methods on test accuracy (%) for proportional bias in *Animal* superclass.

Exp2	1 : 5	1 : 1	2 : 1	3 : 1	4 : 1
CNN	37.17	37.80	41.46	42.50	43.23
CNN+BN	38.70	**39.60**	41.64	42.00	43.85
CNBB	**39.06**	**39.60**	**42.12**	**43.33**	**44.15**

Table 3
Performances of different methods on test accuracy (%) for composional bias in *Vehicle* superclass.

Exp3	3	4	5	6	7
CNN	40.61	42.32	43.34	44.03	44.03
CNN+BN	**41.98**	38.85	43.12	44.71	44.31
CNBB	41.41	**43.34**	**44.54**	**45.96**	**45.16**

Table 4
Performances of different methods of test accuracy (%) for combined proportional & compositional bias in *Vehicle* superclass.

Exp4	1 : 1	2 : 1	3 : 1	4 : 1	5 : 1
CNN	37.07	35.20	34.53	34.13	33.73
CNN + BN	33.87	32.93	31.20	30.93	30.67
CNBB	**38.98**	**36.89**	**35.87**	**35.33**	**35.02**

Outline

- Stable Learning: Definition and Related Problems
- Stable Learning: From Causally-Oriented Perspective
- Stable Learning: From Statistical Learning Perspective
- Beyond Structural Data: Stable Learning on Graph
- NICO: A Benchmark and Baseline for Stable Learning
- Conclusions

Conclusions

- **Stable Learning** cares about not only the prediction accuracy but also the prediction stability across different distributions.
- Causality shed light on the intrinsic mechanism of stable learning and provide firm soil for the upcoming research.
- How to marry causality with predictive modeling effectively and further connect it to the representation learning is still an open problem.