Stable Learning and its Causal Implication: Foundations, Frontiers and Applications

Peng Cui
Tsinghua University
Now we are stepping into risk-sensitive areas.

Shifting from *Performance Driven* to *Risk Sensitive*
Problems of today’s ML - **Explainability**

Most machine learning models are black-box models

Unexplainable

Human in the loop

Health  Military  Finance  Industry
Problems of today’s ML - **Stability**

Most ML methods are developed under I.I.D hypothesis

**OOD Generalization Problem**
Problems of today’s ML - *Stability*
Problems of today’s ML - Stability

- Cancer survival rate prediction

City Hospital
Higher income, higher survival rate.

Testing Data
City Hospital

University Hospital
Survival rate is not so correlated with income.
Problems of today’s ML - *Fairness*
Problems of today’s ML - **Verifiability**

Above the surface you see the **Symptoms** of the problem

Dig deeper to find the **Root Cause** of the problem
A plausible reason: Correlation

Correlation is the very basics of machine learning.
Correlation is not explainable

People who drowned after falling out of a fishing boat correlates with Marriage rate in Kentucky

- Kentucky marriages
- Fishing boat deaths

1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

0 deaths 10 deaths 20 deaths

11 per 1,000 10 per 1,000 9 per 1,000 8 per 1,000 7 per 1,000

tylervigon.com
Correlation is ‘unstable’
It's not the fault of **correlation**, but the way we use it it.

- Three sources of correlation:
  - **Causation**
    - Causal mechanism
    - Stable and explainable
  - **Confoundning**
    - Ignoring X
    - Spurious Correlation
  - **Sample Selection Bias**
    - Conditional on S
    - Spurious Correlation
A Practical Definition of Causality

Definition: T causes Y if and only if changing T leads to a change in Y, while keeping everything else constant.

Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

*Interventionist definition [http://plato.stanford.edu/entries/causation-mani/]*
The **benefits** of bringing causality into learning

Causal Framework

Grass—Label: Strong correlation
Weak causation

Dog nose—Label: Strong correlation
Strong causation

More **Explainable** and More **Stable**
Explainability with Causality

Application --- visibility fluent reasoning

• introduce a Causal And-Or Graph (C-AOG) to represent the causal-effect relations between an object’s visibility fluent and its actions

Explainability with Causality

Application --- counterfactual visual explanations

• A causal explanation: why the example image was classified as class $c$ instead of $c'$?
  • If the bird on the left had a similar beak to that on the right, then the system would have output the right class.

Explainability with Causality

Application --- causal recommendation

Causal structure among user features and item features

Example

Explainability and OOD

OOD ← Causality → Explainability

• Explainability would be a side product when pursuing OOD with causality
Outline

- Brief introduction to causal inference
- Stable learning and its development
- Positioning stable learning in OOD generalization
- Benchmark and dataset
Paradigms - Structural Causal Model

A graphical model to describe the causal mechanisms of a system

- Causal Identification with back door criterion
- Causal Estimation with do calculus

How to discover the causal structure?
Paradigms – Structural Causal Model

- Causal Discovery
  - Constraint-based: conditional independence
  - Functional causal model based

A generative model with strong expressive power. But it induces high complexity.
Paradigms - Potential Outcome Framework

• A simpler setting
  • Suppose the confounders of T are known a priori

• The computational complexity is affordable
  • Under stronger assumptions
  • E.g. all confounders need to be observed

More like a **discriminative** way to estimate treatment’s partial effect on outcome.
Causal Effect Estimation

- Treatment Variable: $T = 1$ or $T = 0$
- Treated Group ($T = 1$) and Control Group ($T = 0$)
- Potential Outcome: $Y(T = 1)$ and $Y(T = 0)$
- Average Causal Effect of Treatment (ATE):

$$ATE = E[Y(T = 1) - Y(T = 0)]$$
Counterfactual Problem

- Two key points for causal effect estimation
  - Changing T
  - Keeping everything else constant

- For each person, observe only one: either $Y_{T=1}$ or $Y_{T=0}$
- For different group (T=1 and T=0), something else are not constant

<table>
<thead>
<tr>
<th>Person</th>
<th>T</th>
<th>$Y_{T=1}$</th>
<th>$Y_{T=0}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>0.4</td>
<td>?</td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td>?</td>
<td>0.6</td>
</tr>
<tr>
<td>P3</td>
<td>1</td>
<td>0.3</td>
<td>?</td>
</tr>
<tr>
<td>P4</td>
<td>0</td>
<td>?</td>
<td>0.1</td>
</tr>
<tr>
<td>P5</td>
<td>1</td>
<td>0.5</td>
<td>?</td>
</tr>
<tr>
<td>P6</td>
<td>0</td>
<td>?</td>
<td>0.5</td>
</tr>
<tr>
<td>P7</td>
<td>0</td>
<td>?</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Ideal Solution: Counterfactual World

- Reason about a world that does not exist
- Everything in the counterfactual world is the same as the real world, except the treatment

\[ Y(T = 1) \quad Y(T = 0) \]
Randomized Experiments are the “Gold Standard”

- Drawbacks of randomized experiments:
  - Cost
  - Unethical
  - Unrealistic
Causal Inference with Observational Data

• Counterfactual Problem:
  \[ Y(T = 1) \quad \text{or} \quad Y(T = 0) \]

• Can we estimate ATE by directly comparing the average outcome between treated and control groups?
  • Yes with randomized experiments (X are the same)
  • No with observational data (X might be different)
Confounding Effect

Balancing Confounders’ Distribution
Methods for Causal Inference

- Matching
- Propensity Score
- Directly Confounder Balancing
Matching

\[ T = 0 \]

\[ T = 1 \]
Matching
Matching

- Identify pairs of treated (T=1) and control (T=0) units whose confounders X are similar or even identical to each other

\[ \text{Distance}(X_i, X_j) \leq \epsilon \]

- Paired units guarantee that the everything else (Confounders) approximate constant
- Small \( \epsilon \): less bias, but higher variance
- Fit for low-dimensional settings
- But in high-dimensional settings, there will be few exact matches
Methods for Causal Inference

- Matching
- Propensity Score
- Directly Confounder Balancing
Propensity Score Based Methods

• Propensity score \( e(X) \) is the probability of a unit to get treated

\[
e(X) = P(T = 1|X)
\]

• Then, Donald Rubin shows that the propensity score is sufficient to control or summarize the information of confounders

\[
T \perp X | e(X) \quad \Rightarrow \quad T \perp (Y(1), Y(0)) | e(X)
\]

• Propensity scores cannot be observed, need to be estimated
Propensity Score Matching

- Estimating propensity score: \( \hat{e}(X) = P(T = 1|X) \)
  - **Supervised learning**: predicting a known label \( T \) based on observed covariates \( X \).
  - Conventionally, use logistic regression
- Matching pairs by distance between propensity score:
  \[
  \text{Distance}(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|
  \]
- High dimensional challenge: from matching to PS estimation
- But this is a ‘hard’ solution.

Inverse of Propensity Weighting (IPW)

• Why weighting with inverse of propensity score?
  • Propensity score induces the distribution bias on confounders $X$

$$e(X) = P(T = 1|X)$$

<table>
<thead>
<tr>
<th>Unit</th>
<th>$e(X)$</th>
<th>$1 - e(X)$</th>
<th>#units</th>
<th>#units (T=1)</th>
<th>#units (T=0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.7</td>
<td>0.3</td>
<td>10</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>0.6</td>
<td>0.4</td>
<td>50</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>0.8</td>
<td>40</td>
<td>8</td>
<td>32</td>
</tr>
</tbody>
</table>

Reweighting by inverse of propensity score:

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

Confounders are the same!

Distribution Bias

Inverse of Propensity Weighting (IPW)

• Estimating ATE by IPW [1]:

\[ \text{ATE}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(X_i)} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i)Y_i}{1 - \hat{e}(X_i)} \]

• Interpretation: IPW creates a pseudo-population where the confounders are the same between treated and control groups.
• But requires correct model specification for propensity score
• High variance when \( e \) is close to 0 or 1

Non-parametric solution

• Model specification problem is inevitable
• Can we directly learn sample weights that can balance confounders’ distribution between treated and control groups?
Methods for Causal Inference

- Matching
- Propensity Score
- Directly Confounder Balancing
Directly Confounder Balancing

**Motivation:** The collection of all the moments of variables uniquely determine their distributions.

**Methods:** Learning sample weights by directly balancing confounders’ moments as follows (ATT problem)

\[
\min_{W} \left\| \begin{bmatrix} \bar{X}_t \\ X^T_c \end{bmatrix} \right\|_2^2
\]

The first moments of $X$ on the **Treated** Group

The first moments of $X$ on the **Control** Group

With moments, the sample weights can be learned without any model specification.

Entropy Balancing

\[
\begin{align*}
\min_{W} & \quad W \log(W) \\
\text{s.t.} & \quad \| \hat{X}_t - X^T_c W \|_2^2 = 0 \\
& \quad \sum_{i=1}^n W_i = 1, W \geq 0
\end{align*}
\]

• Directly confounder balancing by sample weights W
• Minimize the entropy of sample weights W

Either know confounders a priori or regard all variables as confounders. All confounders are balanced equally.

The *gap* between causality and learning

- How to evaluate the outcome?
- Wild environments
  - High-dimensional
  - Highly noisy
  - Little prior knowledge (model specification, confounding structures)
- Targeting problems
  - Understanding v.s. Prediction
  - Depth v.s. Scale and Performance

How to bridge the gap between *causality* and *learning*?
Outline

➢ Brief introduction to causal inference
➢ Stable learning and its development
➢ Positioning stable learning in OOD generalization
➢ Benchmark and dataset
Stability and Prediction

Prediction Performance

Learning Process

True Model

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

Training

- Distribution 1

Model

Testing

- Distribution 1
- Distribution 2
- Distribution 3
- Distribution n

Accuracy 1
Accuracy 2
Accuracy 3
Accuracy n

I.I.D. Learning

VAR (Acc)

Stable Learning

Transfer Learning
Sample reweighting can make a variable independent of other variables.
The core idea of stable learning: Sample Reweighting

Typical Causal Framework

Analogy of A/B Testing

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 $Y$)

If all variables are independent after sample reweighting, 

**Correlation = Causality**
Theoretical Guarantee

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

\[
\sum_{j=1}^p \left\| \frac{X_{t-j}^T (W \otimes X_{t-j})}{W^T X_{t-j}} - \frac{X_{t-j}^T (W \otimes (1-X_{t-j}))}{W^T (1-X_{t-j})} \right\|_2^2, \quad (4)
\]

---

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

\[
\lim_{n \to \infty} \left( \frac{\sum \mathbb{I}(X_{t-k} = 0) W_{j,k} - \sum \mathbb{I}(X_{t-k} = 1) W_{j,k}}{\sum \mathbb{I}(X_{t-k} = 0) W_{j,k}} \right) = 0
\]

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

\[
\lim_{n \to \infty} \frac{1}{n} \sum_x \mathbb{I}(X_{t-j} = x) W_t^* = \lim_{n \to \infty} \frac{1}{n} \sum_x \mathbb{I}(X_{t-j} = x) \frac{1}{\hat{P}(X_{t-j} = x)}
\]

\[
= \lim_{n \to \infty} \sum_x \mathbb{I}(X_{t-j} = x) \frac{1}{\hat{P}(X_{t-j} = x)} = \frac{1}{\hat{P}(X_{t-j} = x)} = 2^{p-1}
\]

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

\[
\lim_{n \to \infty} \frac{1}{n} \sum \mathbb{I}(X_{t-k} = 0, X_{t-j} = 1) W_t^* = 2^{p-2}
\]

\[
\lim_{n \to \infty} \frac{1}{n} \sum \mathbb{I}(X_{t-k} = 0, X_{t-j} = 0) W_t^* = 2^{p-1}, \quad \lim_{n \to \infty} \frac{1}{n} \sum \mathbb{I}(X_{t-k} = 1, X_{t-j} = 0) W_t^* = 2^{p-2}
\]

and therefore, we have following equation with probability 1:

\[
\lim_{n \to \infty} \frac{X_{t-j}^T (W^* \otimes X_{t-j})}{W^T X_{t-j}} = \frac{X_{t-j}^T (W^* \otimes (1-X_{t-j}))}{W^T (1-X_{t-j})} = \frac{2^{p-2}}{2^{p-2}} - \frac{2^{p-2}}{2^{p-2}} = 0.
\]

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

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

- 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)

\[
\begin{align*}
\min & \quad \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (x_i^T \beta))) , \\
\text{s.t.} & \quad \sum_{j=1}^{p} \left\| \frac{X_{-j} \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j} \cdot (W \odot (1-I_j))}{W^T \cdot (1-I_j)} \right\|_2^2 \leq \lambda_1 , \\
& \quad W \geq 0 , \quad \|W\|_2^2 \leq \lambda_2 , \quad \|\beta\|_2^2 \leq \lambda_3 , \quad \|\beta\|_1 \leq \lambda_4 , \\
& \quad (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_5 ,
\end{align*}
\]

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

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

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

\[
\begin{align*}
\min & \quad \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (\phi(X_i)\beta))) \\
\text{s.t.} & \quad \sum_{j=1}^{p} \left\| \frac{\phi(X_{.-j})^T \cdot (W \odot X_{.-j})}{W^T \cdot X_{.-j}} - \frac{\phi(X_{.-j})^T \cdot (W \odot (1-X_{.-j}))}{W^T \cdot (1-X_{.-j})} \right\|_2^2 \leq \lambda_1, \\
& \quad \|(W \cdot 1) \odot (X - \hat{X})\|_F^2 \leq \lambda_2, \quad W \succeq 0, \quad \|W\|_2^2 \leq \lambda_3, \\
& \quad \|\beta\|_2^2 \leq \lambda_4, \quad \|\beta\|_1 \leq \lambda_5, \quad (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_6, \\
& \quad \sum_{k=1}^{K} (\|A^{(k)}\|_F^2 + \|\hat{A}^{(k)}\|_F^2) \leq \lambda_7,
\end{align*}
\]

Deep Auto-Encoder \hspace{2cm} Global Balancing \hspace{2cm} Stable Prediction
Experiments on Synthetic Data

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 \( a, b \in \mathbb{N} \), \( \mathbb{E}[X_{i,j}^a X_{i,k}^b] = \mathbb{E}[X_{i,j}^a] \mathbb{E}[X_{i,k}^b] \)

Causal Regularizer for Continuous Variable

\[
\min_W \sum_{j=1}^p \left\| \mathbb{E}[X_{i,j}^T \Sigma W X_{i,-j}^T] - \mathbb{E}[X_{i,j}^T W] \mathbb{E}[X_{i,-j}^T W] \right\|_2^2
\]

Decorrelated Weighted Regression:

\[
\min_{W, \beta} \sum_{i=1}^n W_i \cdot (Y_i - X_{i,\beta})^2
\]

s.t. \( \sum_{j=1}^p \left\| X_{i,j}^T \Sigma W X_{i,-j}^T / n - X_{i,j}^T W / n \cdot X_{i,-j}^T W / n \right\|_2^2 < \lambda_2 \)

\( |\beta|_1 < \lambda_1, \frac{1}{n} \sum_{i=1}^n W_i^2 < \lambda_3, \)

\( \left( \frac{1}{n} \sum_{i=1}^n W_i - 1 \right)^2 < \lambda_4, \quad W \succeq 0, \)
Stable Learning with *Linear* model

De-confounding for continuous variable

(a) On raw data  
(b) On the weighted data
From *Causal* problem to *Learning* problem

- Previous logic:
  - Sample Reweighting → Independent Variables → Causal Variable → Stable Prediction

- More direct logic:
  - Sample Reweighting → Independent Variables → Stable Prediction
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.

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 $\gamma^2$ is the smallest eigenvalue of centered covariance matrix.

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 $\rho$, we can simulate different extent of collinearity.
- To induce bias related to collinearity, we generate bias term $b(X)$ with $b(X) = X\nu$, where $\nu$ is the eigenvector of centered covariance matrix corresponding to its smallest eigenvalue $\gamma^2$.
- The bias term is sensitive to collinearity.

Simulation Results

large variance in different distributions

large error (estimation bias)

increase collinearity

Stable Learning of Sparse Linear Models

• Suppose $X = \{S, V\}$, and $Y = f(S) + \varepsilon$

• $S$: set of \textbf{stable (causal) features}, i.e., eyes, ears of dog

• $V$: set of \textbf{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: \textbf{Spurious correlation} between $V$ and $Y
Theoretical Analysis

\[ \hat{\beta}_{V_{OLS}} = \beta_V + \left( \frac{1}{n} \sum_{i=1}^{n} v_i^T v_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} v_i^T g(S_i) \right) \]

\[ \hat{\beta}_{S_{OLS}} = \beta_S + \left( \frac{1}{n} \sum_{i=1}^{n} s_i^T s_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} s_i^T g(S_i) \right) \]

\[ + \left( \frac{1}{n} \sum_{i=1}^{n} v_i^T v_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} v_i^T s_i \right) (\beta_S - \hat{\beta}_{S_{OLS}}) \]

\[ + \left( \frac{1}{n} \sum_{i=1}^{n} s_i^T s_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} s_i^T v_i \right) (\beta_V - \hat{\beta}_{V_{OLS}}) \]

• 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 $X = \{X_1, X_2, \ldots, X_p\}$ could be partitioned into $k$ distinct groups $G_1, G_2, \ldots, G_k$. For $\forall i, j, i \neq j$ and $X_i, X_j \in G_l, l \in \{1, 2, \ldots, k\}$, we have $P_{X_iX_j}^e = P_{X_iX_j}$.
Differentiated Variable Decorrelation

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

  \[
  \text{Dis}(X_i, X_j) = \sqrt{\frac{1}{M-1} \sum_{l=1}^{M} (\text{Corr}(X_i^l, X_j^l) - \text{Ave_Corr}(X_i, X_j))^2},
  \]

- Remove the correlation between variables via sample reweighting:

  \[
  \min_W \sum_{i \neq j} I(i, j) \left\| (X_i^T \Sigma W X_j/n - X_i^T W/n \cdot X_j^T W/n) \right\|_2^2 \\
  \text{s.t.} \frac{1}{n} \sum_{i=1}^{n} W_i^2 < \gamma_1, \quad \left(\frac{1}{n} \sum_{i=1}^{n} W_i - 1\right)^2 < \gamma_2, \quad W \geq 0
  \]

## Experimental Results

### Scenario 1: varying sample size $n$

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\beta$ Error</th>
<th>Average Error</th>
<th>Stability Error</th>
<th>$\beta$ Error</th>
<th>Average Error</th>
<th>Stability Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>1.958</td>
<td>0.470</td>
<td>0.087</td>
<td>1.870</td>
<td>0.489</td>
<td>0.105</td>
</tr>
<tr>
<td>Lasso</td>
<td>2.021</td>
<td>0.476</td>
<td>0.092</td>
<td>1.905</td>
<td>0.494</td>
<td>0.110</td>
</tr>
<tr>
<td>Lasso</td>
<td>2.035</td>
<td>0.475</td>
<td>0.094</td>
<td>1.920</td>
<td>0.498</td>
<td>0.113</td>
</tr>
<tr>
<td>DWR</td>
<td>2.012</td>
<td>0.545</td>
<td>0.090</td>
<td>1.991</td>
<td>0.507</td>
<td>0.092</td>
</tr>
<tr>
<td>OUR</td>
<td><strong>1.892</strong></td>
<td><strong>0.469</strong></td>
<td><strong>0.040</strong></td>
<td><strong>1.741</strong></td>
<td><strong>0.489</strong></td>
<td><strong>0.050</strong></td>
</tr>
</tbody>
</table>

### Scenario 2: varying number of unstable variables $p_{uv}$

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\beta$ Error</th>
<th>Average Error</th>
<th>Stability Error</th>
<th>$\beta$ Error</th>
<th>Average Error</th>
<th>Stability Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>1.839</td>
<td>0.522</td>
<td>0.121</td>
<td>2.028</td>
<td>0.563</td>
<td>0.179</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.876</td>
<td>0.529</td>
<td>0.129</td>
<td>2.026</td>
<td>0.571</td>
<td>0.186</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.894</td>
<td>0.538</td>
<td>0.149</td>
<td>2.106</td>
<td>0.575</td>
<td>0.191</td>
</tr>
<tr>
<td>DWR</td>
<td>1.656</td>
<td>0.485</td>
<td>0.081</td>
<td>1.824</td>
<td>0.518</td>
<td>0.092</td>
</tr>
<tr>
<td>OUR</td>
<td><strong>1.369</strong></td>
<td><strong>0.476</strong></td>
<td><strong>0.042</strong></td>
<td><strong>1.641</strong></td>
<td><strong>0.460</strong></td>
<td><strong>0.064</strong></td>
</tr>
</tbody>
</table>

### Scenario 3: varying bias rate $r$ on training data

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\beta$ Error</th>
<th>Average Error</th>
<th>Stability Error</th>
<th>$\beta$ Error</th>
<th>Average Error</th>
<th>Stability Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>1.296</td>
<td><strong>0.452</strong></td>
<td>0.064</td>
<td>1.780</td>
<td>0.510</td>
<td>0.117</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.321</td>
<td>0.455</td>
<td>0.067</td>
<td>1.812</td>
<td>0.516</td>
<td>0.123</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.339</td>
<td>0.457</td>
<td>0.070</td>
<td>1.829</td>
<td>0.519</td>
<td>0.125</td>
</tr>
<tr>
<td>DWR</td>
<td><strong>1.153</strong></td>
<td><strong>0.457</strong></td>
<td><strong>0.023</strong></td>
<td><strong>1.562</strong></td>
<td><strong>0.458</strong></td>
<td><strong>0.035</strong></td>
</tr>
<tr>
<td>OUR</td>
<td><strong>1.236</strong></td>
<td><strong>0.463</strong></td>
<td><strong>0.021</strong></td>
<td><strong>1.236</strong></td>
<td><strong>0.450</strong></td>
<td><strong>0.023</strong></td>
</tr>
</tbody>
</table>

---

StableNet: From Linear Models to Deep Models

**Variable Decorrelation** by Sample Reweighting and RFF:

- Measure and eliminate the complex non-linear dependencies among features with RFF
- The computation cost is acceptable

\[ \mathcal{H}_{RFF} = \{ h : x \rightarrow \sqrt{2} \cos(\omega x + \phi) \mid \omega \sim N(0, 1), \phi \sim \text{Uniform}(0, 2\pi) \}, \]

\[ (f^{(t+1)}, g^{(t+1)}) = \arg \min_{f, g} \sum_{i=1}^{n} w_i^{(t)} L(g(f(X_i)), y_i), \]

\[ w^{(t+1)} = \arg \min_{w \in \Delta_n} \sum_{1 \leq i < j \leq m_z} \| \mathbf{Z}_i^{(t+1)} \mathbf{Z}_j^{(t+1)} w \|_F^2, \]

Learning sample weights globally

Optimize sample weights globally by saving and reloading all features and weights.

Learning sample weights globally

- Sample weights learning module is an independent module which can be easily assembled with current deep models.
- Sample weights and the classification model are trained iteratively.

Out-Of-Distribution Generalization

- The heterogeneity of training data is not significant nor known.
- The capacities of different domains can vary significantly.

NICO dataset
Flexible OOD Generalization

- The domains for different categories can be different.
- For instance, birds can be on trees but hardly in the water while fishes are the opposite.

<table>
<thead>
<tr>
<th></th>
<th>JiGen</th>
<th>M-ADA</th>
<th>DG-MMLD</th>
<th>RSC</th>
<th>ResNet-18</th>
<th>StableNet (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PACS</td>
<td>40.31</td>
<td>30.32</td>
<td>42.65</td>
<td>39.49</td>
<td>39.02</td>
<td><strong>45.14</strong></td>
</tr>
<tr>
<td>VLCS</td>
<td>76.75</td>
<td>69.58</td>
<td>78.96</td>
<td>74.81</td>
<td>73.77</td>
<td><strong>79.15</strong></td>
</tr>
<tr>
<td>NICO</td>
<td>54.42</td>
<td>40.78</td>
<td>47.18</td>
<td>57.59</td>
<td>51.71</td>
<td><strong>59.76</strong></td>
</tr>
</tbody>
</table>
Saliency maps of StableNet and other models

- The visualization of the gradient of the class score function with respect to the input pixels. The brighter the pixel is, the more contribution it makes to prediction.

OOD generalization: Model v.s. **Optimization**?

$$\theta_{ERM} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(\theta; X_i, Y_i)$$

Overall Good = **Majority Good** + **Minority Bad**
Overall Good = Majority Good + Minority Good

Problem I
Uncovering Heterogeneity

Problem II
Finding Invariance

To deal with the potential distributional shifts, one common assumption made in invariant learning is the **Invariance Assumption**.

**Assumption (Invariance Assumption)**

*There exists random variable $\Phi^*(X)$ such that the following properties hold:*

1. **Invariance property:** for all $e_1, e_2 \in \text{supp}(\mathcal{E})$, we have

   \[ P^{e_1}(Y|\Phi^*(X)) = P^{e_2}(Y|\Phi^*(X)) \]  

   (4)

2. **Sufficiency property:** $Y = f(\Phi^*) + \epsilon, \; \epsilon \perp X$.

Here we make some demonstrations on the Invariance Assumption:

- The first property assumes that the relationship between $\Phi^*(X)$ and $Y$ remains invariant across environments, which is also referred to as causal relationship.

- The second property assumes that $\Phi^*(X)$ can provide all information of the target label $Y$.

- $\Phi^*(X)$ is referred to as (Causally) Invariant Predictors.
To obtain the invariant predictor $\Phi^*(X)$, one can seeks for the **Maximal Invariant Predictor**\(^{12}\), which is defined as follows:

**Definition (Invariance Set & Maximal Invariant Predictor)**

The invariance set $\mathcal{I}$ with respect to $\mathcal{E}$ is defined as:

$$\mathcal{I}_\mathcal{E} = \{ \Phi(X) : Y \perp \mathcal{E} | \Phi(X) \} = \{ \Phi(X) : H[Y | \Phi(X)] = H[Y | \Phi(X), \mathcal{E}] \}$$

(5)

where $H[\cdot]$ is the Shannon entropy of a random variable. The corresponding maximal invariant predictor (MIP) of $\mathcal{I}_\mathcal{E}$ is defined as:

$$S = \arg \max_{\Phi \in \mathcal{I}_\mathcal{E}} I(Y; \Phi)$$

(6)

where $I(\cdot; \cdot)$ measures Shannon mutual information between two random variables.

**Remarks:**

- $\Phi^*(X)$ is MIP.
- Optimal for OOD is $\hat{Y} = \mathbb{E}[Y | \Phi^*(X)]$.
- "Find $\Phi^*(X)$" $\rightarrow$ "Find MIP"

---

\(^{1}\)Chang, S., Zhang, Y. et al. (2020, November). Invariant rationalization.

\(^{2}\)Koyama, M., & Yamaguchi, S. (2021). When is invariance useful in an Out-of-Distribution Generalization problem?
Quality of Training Environments

• The flow of Invariant Learning methods:

\[ \mathcal{E}_{tr} \rightarrow \text{Find MIP } \Phi^*_{tr} \text{ of } \mathcal{I}_{\mathcal{E}_{tr}} \rightarrow \text{Predict using } \Phi^*_{tr} \rightarrow \text{OOD "Optimal?"} \]

• Recall the definition of MIP:

\[
\arg \max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi) \tag{7}
\]

1. MIP relies on the invariance set \( \mathcal{I}_{\mathcal{E}} \)
2. Invariance set \( \mathcal{I}_{\mathcal{E}} \) relies on the given environments \( \mathcal{E} \).

• What happens when \( \mathcal{E} \) is replaced by \( \mathcal{E}_{tr} \)?

1. \( \text{supp}(\mathcal{E}_{tr}) \subset \text{supp}(\mathcal{E}) \)
2. \( \mathcal{I}_{\mathcal{E}} \subset \mathcal{I}_{\mathcal{E}_{tr}} \)
3. \( \Phi^*_{tr} \) NOT INVARIANT.

Remark: We need training environments where \( \mathcal{I}_{\mathcal{E}_{tr}} \rightarrow \mathcal{I}_{\mathcal{E}} \).
Modern datasets are frequently assembled by merging data from multiple sources without explicit source labels, which means there are not multiple environments but only one pooled dataset.
ERM → HRM (Heterogeneous Risk Minimization)

Theorem (Why using only $\Psi$?)

For $e_i, e_j \in \text{supp}(\mathcal{E}_T)$, assume that $X = [\Phi^*, \Psi^*]^T$ satisfying Invariance and Heterogeneity Assumption, where $\Phi^*$ is invariant and $\Psi^*$ variant. Then we have

$$D_{KL}(P^{e_i}(Y|X)||P^{e_j}(Y|X)) \leq D_{KL}(P^{e_i}(Y|\Psi^*)||P^{e_j}(Y|\Psi^*))$$

The Heterogeneity Identification Module $\mathcal{M}_c$

Recall that for $\mathcal{M}_c$,

$$\Psi(X) \rightarrow \mathcal{M}_c \rightarrow \mathcal{E}_{\text{learn}}$$

we implement it with a convex clustering method. Different from other clustering methods, we cluster the data according to the **relationship** between $\Psi(X)$ and $Y$.

- Assume the $j$-th cluster centre $P_{\Theta_j}(Y|\Psi)$ parameterized by $\Theta_j$ to be a Gaussian around $f_{\Theta_j}(\Psi)$ as $\mathcal{N}(f_{\Theta_j}(\Psi), \sigma^2)$:

  $$h_j(\Psi, Y) = P_{\Theta_j}(Y|\Psi) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(Y - f_{\Theta_j}(\Psi))^2}{2\sigma^2}\right)$$  \hspace{1cm} (8)

- The empirical data distribution is $\hat{P}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_i(\Psi, Y)$

- The target is to find a distribution in $Q = \{ Q | Q = \sum_{j \in [K]} q_j h_j(\Psi, Y), q \in \Delta_K \}$ to fit the empirical distribution best.

- The objective function of our heterogeneous clustering is:

  $$\min_{Q \in \mathcal{Q}} D_{KL}(\hat{P}_N || Q)$$  \hspace{1cm} (9)
The Invariant Prediction Module $\mathcal{M}_p$

Recall that for $\mathcal{M}_p$,

$$\mathcal{E}_{\text{learn}} \rightarrow \mathcal{M}_p \rightarrow \Phi(X) = M \odot X$$

The algorithm involves two parts, invariant prediction and feature selection.

- For invariant prediction, we adopt the regularizer\textsuperscript{4} as:

$$\mathcal{L}_p(M \odot X, Y; \theta) = \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e] + \lambda \text{trace}(\text{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e))$$  \hspace{1cm} (10)

  - Restrict the gradient across environments to be the same.
  - Only use invariant features.

- For feature selection, we adopt the continuous feature selection method that allows for continuous optimization of $M$:

$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e}\mathbb{E}_M[\ell(M \odot X^e, Y^e; \theta) + \alpha \|M\|_0]$$  \hspace{1cm} (11)

  - $\|M\|_0$ controls the number of selected features.
  - Conduct continuous optimization as \textsuperscript{5}.

\textsuperscript{4}Koyama, M., & Yamaguchi, S. (2021). When is invariance useful in an Out-of-Distribution Generalization problem?

\textsuperscript{5}Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates, in ICML2020
The Mutual Promotion

• Insight: We should only use $\Psi^*$ for Heterogeneity Identification.

Assumption (Heterogeneity Assumption from Information Theory)

Assume the pooled training data is made up of heterogeneous data sources: $P_{tr} = \sum_{e \in \text{supp} (E_{tr})} w_e P^e$. For any $e_i, e_j \in E_{tr}, e_i \neq e_j$, we assume

$$I_{i,j}^c(Y; \Phi^* | \Psi^*) \geq \max(I_i(Y; \Phi^* | \Psi^*), I_j(Y; \Phi^* | \Psi^*))$$ (12)

where $\Phi^*$ is invariant feature and $\Psi^*$ the variant. $I_i$ represents mutual information in $P^e_i$ and $I_{i,j}^c$ represents the cross mutual information between $P^e_i$ and $P^e_j$ takes the form of $I_{i,j}^c(Y; \Phi | \Psi) = H_{i,j}^c[Y | \Psi] - H_{i,j}^c[Y | \Phi, \Psi]$ and $H_{i,j}^c[Y] = -\int p^e_i(y) \log p^e_j(y) dy$.

• The mutual information $I_i(Y; \Phi^*) = H_i[Y] - H_i[Y | \Phi^*]$ can be viewed as the error reduction if we use $\Phi^*$ to predict $Y$ rather than predict by nothing.

• The cross mutual information $I_{i,j}^c(Y; \Phi^*)$ can be viewed as the error reduction if we use the predictor learned on $\Phi^*$ in environment $e_j$ to predict in environment $e_i$, rather than predict by nothing.

Theorem (Why using only $\Psi$?)

For $e_i, e_j \in \text{supp}(E_{tr})$, assume that $X = [\Phi^*, \Psi^*]^T$ satisfying Invariance and Heterogeneity Assumption, where $\Phi^*$ is invariant and $\Psi^*$ variant. Then we have

$$D_{KL}(P^e_i(Y|X) || P^e_j(Y|X)) \leq D_{KL}(P^e_i(Y|\Psi^*) || P^e_j(Y|\Psi^*))$$
**Results**

![Table](image)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Training environments</th>
<th>Testing environments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$e_1$</td>
<td>$e_2$</td>
</tr>
<tr>
<td>ERM</td>
<td>0.290</td>
<td>0.308</td>
</tr>
<tr>
<td>DRO</td>
<td>0.289</td>
<td>0.310</td>
</tr>
<tr>
<td>EIL</td>
<td>0.075</td>
<td>0.128</td>
</tr>
<tr>
<td>IRM (with $\varepsilon_{tr}$ label)</td>
<td>0.306</td>
<td>0.312</td>
</tr>
<tr>
<td>HRM$^5$</td>
<td>1.060</td>
<td>1.085</td>
</tr>
<tr>
<td>HRM</td>
<td>0.317</td>
<td>0.314</td>
</tr>
</tbody>
</table>

![Graphs](image)

Kernelized Heterogeneous Risk Minimization

• To solve the HRM problem beyond the raw feature level.

• Incorporate Neural Tangent Kernel.

• Perform the heterogeneity identification and invariant prediction in the Neural Tangent Feature Space.

Outline

➢ Brief introduction to causal inference
➢ Stable learning and its development
➢ Positioning stable learning in OOD generalization
➢ Benchmark and dataset
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*
  • May help for robustness
Domain Generalization

- 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 $1 - \delta$

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

$$\leq c_1 \cdot \mathbb{V}_{\mathcal{H}}(\mathbb{P}^1, \mathbb{P}^2, \ldots, \mathbb{P}^N) + c_2 \frac{N \cdot (\log \delta^{-1} + 2 \log N)}{n} + c_3 \frac{\log \delta^{-1}}{N} + \frac{c_4}{N}$$

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.

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

\[
Y^e = g(X^e_{S*}, \varepsilon^e), \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp X^e_{S*}
\]

- **Idea: Linking to causality**
  - Structural Causal Model (Pearl 2009):
    - The parent variables of \( Y \) in SCM satisfies Invariant Assumption
    - The causal variables lead to invariance w.r.t. “all” possible environments

Stable Learning

• Finding the common ground between causal inference and machine learning
Stable Learning

- One training distribution, multiple testing distributions
Outline

- Brief introduction to causal inference
- Stable learning and its development
- Positioning stable learning in OOD generalization
- Benchmark and dataset
Image Dataset —— Synthetic Transformation

Colored MNIST\(^1\)

Waterbirds\(^2\)


Image Dataset ——— Fixed Wild Data

<table>
<thead>
<tr>
<th>Train</th>
<th>Test (OOD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d =$ Location 1</td>
<td>$d =$ Location 2</td>
</tr>
<tr>
<td>Vulturine Guineafowl</td>
<td>African Bush Elephant</td>
</tr>
<tr>
<td>Cow</td>
<td>unknown</td>
</tr>
<tr>
<td>$d =$ Location 245</td>
<td>$d =$ Location 246</td>
</tr>
<tr>
<td>Wild Horse</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test (ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d =$ Location 1</td>
</tr>
<tr>
<td>Giraffe</td>
</tr>
</tbody>
</table>

*Image Dataset —— Controllable Wild Data*

**NICO**\(^{[1]}\) (Non-I.I.D. Image Dataset with Contexts)

---

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

• Contextual labels (Contexts)
  • the attributes or actions of a category
    • e.g. white bear, double decker
  • the background or scene of a category
    • e.g. cat on snow, airplane in sunrise

• Structure of NICO

Diagram:
- 2 Superclass per
- 10 or 9 Class per
- 10 or 9 Contexts

- Animal
  - Bird
    - flying
  - ...
- Vehicle
  - Train
    - on bridge
  - ...

Overlapping
Diverse & Meaningful
NICO—Non-I.I.D. Image Dataset with Contexts

- Data size of each class in NICO

- Samples with contexts in NICO

<table>
<thead>
<tr>
<th>Animal</th>
<th>Data Size</th>
<th>Animal</th>
<th>Data Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bear</td>
<td>1609</td>
<td>Airplane</td>
<td>930</td>
</tr>
<tr>
<td>Bird</td>
<td>1590</td>
<td>Bicycle</td>
<td>1639</td>
</tr>
<tr>
<td>Cat</td>
<td>1479</td>
<td>Boat</td>
<td>2156</td>
</tr>
<tr>
<td>Cow</td>
<td>1192</td>
<td>Bus</td>
<td>1069</td>
</tr>
<tr>
<td>Dog</td>
<td>1624</td>
<td>Car</td>
<td>1026</td>
</tr>
<tr>
<td>Elephant</td>
<td>1178</td>
<td>Helicopter</td>
<td>1351</td>
</tr>
<tr>
<td>Horse</td>
<td>1258</td>
<td>Motorcycle</td>
<td>1542</td>
</tr>
<tr>
<td>Monkey</td>
<td>1117</td>
<td>Train</td>
<td>750</td>
</tr>
<tr>
<td>Rat</td>
<td>846</td>
<td>Truck</td>
<td>1000</td>
</tr>
<tr>
<td>Sheep</td>
<td>918</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- Range of average NI over Animal superclass for different settings supported in NICO.

Other Data Type

Graph Data (OGB-LSC[1])

Text Data (Amazon Review[2])

<table>
<thead>
<tr>
<th>Reviewer ID (d)</th>
<th>Review Text (x)</th>
<th>Stars (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer 1</td>
<td>They are decent shoes. Material quality is good but the color fades very quickly. Not as black in person as shown. Super easy to put together. Very well built.</td>
<td>5</td>
</tr>
<tr>
<td>Reviewer 2</td>
<td>This works well and was easy to install. The only thing I don’t like is that it tilts forward a little bit and I can’t figure out how to stop it. Perfect for the trail camera</td>
<td>4</td>
</tr>
<tr>
<td>Reviewer 10,000</td>
<td>I am disappointed in the quality of these. They have significantly deteriorated in just a few uses. I am going to stick with using foil. Very sturdy especially at this price point. I have a memory foam mattress on it with nothing underneath and the slats perform well.</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer 10,001</td>
<td>Solidly built plug. I have had 4 devices plugged in and all charge just fine. Works perfectly on the wall to hang our wreath without having to do any permanent damage.</td>
<td>5</td>
</tr>
</tbody>
</table>


OOD Evaluation Metric

**Average Accuracy**

\[
\overline{Acc} = \frac{1}{K} \sum_{k=1}^{K} acc_k
\]

**Standard Deviation (STD)**

\[
ACC_{std} = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (acc_k - \overline{Acc})^2}
\]

**Worst-Case Accuracy**

\[
ACC_{worst} = \min_{k \in [K]} acc_k
\]

performance in \( k_{th} \) environment
Conclusions

• Explainability, Stability, Fairness, Verifiability problems are becoming more critical
• They are not independent!
• Stable Learning: finding the common ground between causal inference and machine learning
  • Theoretical problems
  • Sample efficiency problems
  • Application problems
A survey on OOD generalization

Towards Out-Of-Distribution Generalization: A Survey

Zheyan Shen*, Jiashuo Liu*, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, Peng Cui†, Senior Member, IEEE

Abstract—Classic machine learning methods are built on the i.i.d. assumption that training and testing data are independent and identically distributed. However, in real scenarios, the i.i.d. assumption can hardly be satisfied, rendering the sharp drop of classic machine learning algorithms’ performances under distributional shifts, which indicates the significance of investigating the Out-of-Distribution generalization problem. Out-of-Distribution (OOD) generalization problem addresses the challenging setting where the testing distribution is unknown and different from the training. This paper serves as the first effort to systematically and comprehensively discuss the OOD generalization problem, from the definition, methodology, evaluation to the implications and future directions. Firstly, we provide the formal definition of the OOD generalization problem. Secondly, existing methods are categorized into three parts based on their positions in the whole learning pipeline, namely unsupervised representation learning, supervised model learning and optimization, and typical methods for each category are discussed in detail. We then demonstrate the theoretical connections of different categories, and introduce the commonly used datasets and evaluation metrics. Finally, we summarize the whole literature and raise some future directions for OOD generalization problem. The summary of OOD generalization methods reviewed in this survey can be found at http://out-of-distribution-generalization.com.

Index Terms—Out-of-Distribution Generalization, Causal Inference, Invariant Learning, Stable Learning, Representation Learning, Distributionally Robust Optimization

Reference

➢ Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen. Deep Stable Learning for Out-Of-Distribution Generalization. CVPR, 2021
➢ Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, Bo Li. Stable Prediction with Model Misspecification and Agnostic Distribution Shift. AAAI, 2020.
➢ Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. KDD, 2018.
➢ Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ACM Multimedia, 2018.
Thanks!

Peng Cui
cuip@tsinghua.edu.cn
http://pengcui.thumedialab.com