Stable Learning:
Finding the Common Ground between Causal Inference and Machine Learning

Peng Cui
Tsinghua University
Now AI is stepping into risk-sensitive areas

Healthcare

Law

Transportation

Fintech

Shifting from Performance Driven to Risk Sensitive
Risks of Today’s AI Algorithms

Unexplainable

Medical

Military

Finance

Explainability

Human in the loop
Most ML methods are developed under I.I.D hypothesis
Risks of Today’s AI Algorithms

Yes

Maybe

No
Risks of Today’s AI Algorithms

• Cancer survival rate prediction

Features:
• Body status
• Income
• Treatments
• Medications

City Hospital
Higher income, higher survival rate.

City Hospital
Higher income, higher survival rate.

Testing Data

Predictive Model

University Hospital
Survival rate is not so correlated with income.
The Current Condition

- **Explainability**: We cannot *understand* AI
- **Stability**: We don’t *trust* AI

**Dilemma**
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
Correlation is ‘unstable’
It’s not the fault of correlation, but the way we use it

• Three sources of correlation:
  - Causation
    • Causal mechanism
    • Stable and explainable
  - Confounding
    • 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**

- **T**: grass
- **X**: dog nose
- **Y**: label

**Grass**—Label: Strong correlation
  - Weak causation

**Dog nose**—Label: Strong correlation
  - Strong causation

More *Explainable* and More *Stable*
Stable Learning

Training

Distribution 1

Distribution 2

Distribution 3

Distribution n

Model

Testing

Accuracy 1

Accuracy 2

Accuracy 3

Accuracy n

I.I.D. Learning

VAR (Acc)

Stable Learning

Transfer Learning
Revisit Directly Balancing for causal inference

Typical Causal Framework

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)

Sample reweighting can make a variable independent of other variables.
Global Balancing: making all variables independent

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

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

\[\triangleright\]

$0$

Causal Regularizer

Set feature $j$ as treatment variable

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

- All features excluding treatment $j$
- Sample Weights
- Indicator of treatment status

Causally Regularized Logistic Regression

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

Sample reweighted logistic loss

Causal Contribution

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
Experimental Result - insights
Experimental Result - insights
Stable Learning with **Continuous** Variables

**Variable Decorrelation** by Sample Reweighting:

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

**Decorrelated Weighted Regression:**

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

subject to:

$$\sum_{j=1}^{p} \left\| X_{i,j}^T \Sigma_W X_{i,-j}/n - X_{i,j}^T W/n \cdot X_{i,-j}^T W/n \right\|_2^2 < \lambda_2$$

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

$$(\frac{1}{n} \sum_{i=1}^{n} W_i - 1)^2 < \lambda_4, \quad W \geq 0,$$
Stable Learning with *Continuous* Variables

Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, Bo Li. Stable Prediction with Model Misspecification and Agnostic Distribution Shift. AAAI, 2020.
Stable Learning with **Differentiated** Variables

- More detailed analysis:

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

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

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

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

- We can focus on only the **spurious part** of correlation
- But how?
- Leveraging the abundant sources of unlabeled data!
Stable Learning with *Differentiated* Variables

**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 $P_{X_iX_j}^e = P_{X_iX_j}$. 

![Clustering?](image)
Stable Learning with *Differentiated* Variables

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

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

- Remove the correlation between variables via sample reweighting:

\[
\min_W \sum_{i \neq j} \mathbb{1}(i, j) \left\| \left( X_i^T \Sigma_W X_j / n - X_i^T W / n \cdot X_j^T W / n \right) \right\|_2^2
\]

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

## Stable Learning with **Differentiated** Variables

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

<table>
<thead>
<tr>
<th>$n$, $p_{cb}$, $r$</th>
<th>$n = 120$, $p_{cb} = p \times 0.2$, $r = 1.9$</th>
<th>$n = 160$, $p_{cb} = p \times 0.2$, $r = 1.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>$\beta$ Error</td>
<td>Average Error</td>
</tr>
<tr>
<td>OLS</td>
<td>1.958</td>
<td>0.470</td>
</tr>
<tr>
<td>Lasso</td>
<td>2.021</td>
<td>0.476</td>
</tr>
<tr>
<td>IIIlasso</td>
<td>2.035</td>
<td>0.475</td>
</tr>
<tr>
<td>DWR</td>
<td>2.012</td>
<td>0.545</td>
</tr>
<tr>
<td><strong>Our</strong></td>
<td><strong>1.892</strong></td>
<td><strong>0.469</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$n$, $p_{cb}$, $r$</th>
<th>$n = 200$, $p_{cb} = p \times 0.2$, $r = 1.9$</th>
<th>$n = 200$, $p_{cb} = p \times 0.3$, $r = 1.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>$\beta$ Error</td>
<td>Average Error</td>
</tr>
<tr>
<td>OLS</td>
<td>1.839</td>
<td>0.522</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.876</td>
<td>0.529</td>
</tr>
<tr>
<td>IIIlasso</td>
<td>1.894</td>
<td>0.538</td>
</tr>
<tr>
<td>DWR</td>
<td>1.656</td>
<td>0.485</td>
</tr>
<tr>
<td><strong>Our</strong></td>
<td><strong>1.369</strong></td>
<td><strong>0.476</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$n$, $p_{cb}$, $r$</th>
<th>$n = 200$, $p_{cb} = p \times 0.2$, $r = 1.6$</th>
<th>$n = 200$, $p_{cb} = p \times 0.2$, $r = 1.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>$\beta$ Error</td>
<td>Average Error</td>
</tr>
<tr>
<td>OLS</td>
<td>1.296</td>
<td><strong>0.452</strong></td>
</tr>
<tr>
<td>Lasso</td>
<td>1.321</td>
<td>0.455</td>
</tr>
<tr>
<td>IIIlasso</td>
<td>1.339</td>
<td>0.457</td>
</tr>
<tr>
<td>DWR</td>
<td>1.159</td>
<td>0.457</td>
</tr>
<tr>
<td><strong>Our</strong></td>
<td><strong>1.136</strong></td>
<td><strong>0.463</strong></td>
</tr>
</tbody>
</table>

---

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
Interpretation from Statistical Learning perspective

• Consider the linear regression with misspecification bias

\[ y = x^\top \bar{\beta}_{1:p} + \bar{\beta}_0 + b(x) + \epsilon \]

Goes to infinity when perfect collinearity exists!

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

• By accurately estimating \( \bar{\beta} \) 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 \( \| \hat{\beta} - \bar{\beta} \|_2 \leq 2(\delta/\gamma) + \delta \), where \( \gamma^2 \) is the smallest eigenvalue of centered covariance matrix.

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

<table>
<thead>
<tr>
<th>Animal</th>
<th>Data Size</th>
<th>Vehicle</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>1009</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>

Examples:
- Dog: At home, on beach, eating, in cage, in water, lying, on grass, in street, running, on snow
- Horse: on beach, in forest, at home, in river, lying, on grass, in street, aside people, running, on snow
- Boat: on beach, cross bridge, in city, with people, in river, sailboat, in sunset, at wharf, wooden, yacht
NICO - Non-I.I.D. Image Dataset with Contexts

http://nico.thumedialab.com/

Conclusions

- Why can’t the current AI generalize well to unknown environments?

Know **What**, but don’t know **Why**

知其 **然**，但不知其 **所以然**

Correlation  Causality

Stable Learning: Try to promote the convergence of causal inference and machine learning.
Reference

- Yue He, Peng Cui, Jianxin Ma, Zou Hao, Xiaowei Wang, Hongxia Yang and Philip S. Yu. Learning Stable Graphs from Multiple Environments with Selection Bias. KDD, 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