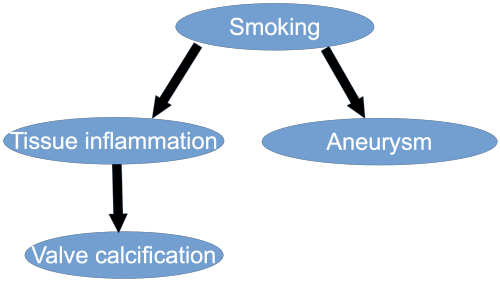
Efficient Nonlinear DAG Learning under Projection Framework

Naiyu Yin[1] Yue Yu[2] Tian Gao[3] Qiang Ji[1]

[1]Rensselaer Polytechnic Institute
[2]Lehigh University
[3]IBM Research

Introduction - DAG and Causality

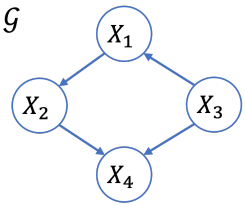
- Direct acyclic graphs (DAGs) are ideal for representing causal relations.



- DAG Learning:
  - Learn and identify the cause-effect relationships among a set of variables from the observational data.

Introduction - Structural Causal Model

- Structural Causal Model (SCM)
  - Capture the causal relations using DAGs.
  - Quantifies the causal relations with structural equation models (SEMs).



$$X_1 := f_1(X_3) + E_1$$

$$X_2 := f_2(X_1) + E_2$$

$$X_3 := f_3 + E_3, f_3 = 0$$

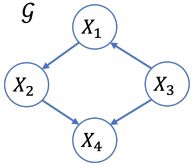
$$X_4 := f_4(X_2, X_3) + E_4$$

$f_n(\cdot)$s are causal functions; E_ns are exogeneous noises.

Introduction - Continuous DAG Learning Methods

- Represent the DAG with continuous parameters, such as weighted adjacency matrix W for linear SEM.
- Employ a continuous acyclicity constraint from Zheng et al.(2018).
- Formulate the DAG learning as continuous optimization problem and search in a continuous DAG space.

Linear SEM Example



$$X_1 := w_{31}X_3 + E_1$$
$$X_2 := w_{12}X_1 + E_2$$
$$X_3 := E_3$$
$$X_4 := w_{24}X_2 + w_{34}X_3 + E_4$$

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} 0 & w_{12} & 0 & 0 \\ 0 & 0 & 0 & w_{24} \\ w_{31} & 0 & 0 & w_{34} \\ 0 & 0 & 0 & 0 \end{bmatrix}^T \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \end{bmatrix}$$

Weighted adjacency matrix W

Related Works

- Challenges in existing methods:
  - Experience high computational costs due to iterative optimization for enforcing the acyclicity constraint.

- Recent efforts that do not require the iterative optimization process:
  - Ng et al., (2020) proposed to softly enforce the continuous acyclicity constraint.
  - Yu et al., (2021) decomposed the weighted adjacency matrix of a DAG into the product of smaller matrices with relaxed constraints.
  - ...

Although achieving efficiency gains, the above methods are built upon linear SEMs. We propose an approach that learns DAG under nonlinear SEMs without any explicit acyclicity constraint.

Problem Statement

- We adopt a standard form of SCM with nonlinear causal functions and additive homoscedastic noises.

- Let $X = [X_1, X_2, \cdots, X_N] \in \mathbb{R}^N$ denotes a set of N random variables. The causal relations between a variable $X_n \in X$ and its parents $\pi(X_n)$ can be modeled via following equation:

$$X_n = f_n\big(\pi(X_n)\big) + E_n, n = 1, 2, \cdots, N$$

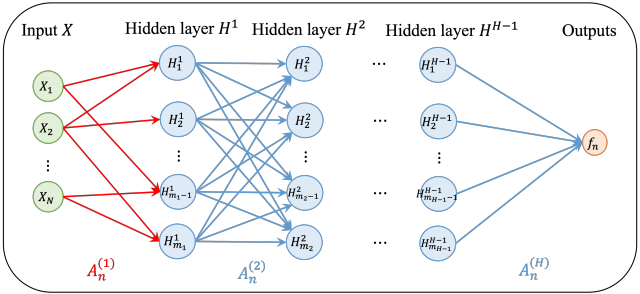
  - $f_n(\cdot)$ is the structural causal function for variable X_n.
  - E_n is the exogenous noise variable corresponding to X_n.

Problem Statement (cont'd)

- We implement the causal functions with MLPs.

$$f_n(X; A_n) = A_n^{(H)} \sigma\big(\cdots A_n^{(2)} \sigma(A_n^{(1)} X) \cdots \big)$$

  - $\sigma(\cdot)$ is the sigmoid activation function.
  - $A_n^{(h)}$ is the h^{th} layer weight matrix of the MLP for variable X_n.



- The structural information is encoded in $A^{(1)} = [A_1^{(1)}, A_2^{(1)}, \cdots, A_N^{(1)}]$.
  - If there is no causal relations $X_k \to X_n$, $A_n^{(1)}[:, k] = \mathbf{0}$.

Problem Statement (cont'd)

- The DAG learning problem becomes:

$$A = \arg\min_{A} \sum_{n=1}^{N} \left(X_n - A_n^{(H)} \sigma\left(\cdots A_n^{(2)} \sigma(A_n^{(1)} X) \cdots \right) \right)^2$$
$$\text{subject to } h\left(W(A^{(1)})\right) = 0$$

  - We can convert the first layer weights $A^{(1)}$ into the weighted adjacency matrix W with Eq. (1).

$$W(A^{(1)})[k, n] = \sqrt{\sum_{b=1}^{m_1} (A_n^{(1)}[b, k])^2} \tag{1}$$

  - We apply the acyclicity constraint on $W(A^{(1)})$:

$$h(W(A^{(1)})) = tr(e^{W(A^{(1)}) \circ W(A^{(1)})}) - N = 0$$

DAG Projection Framework

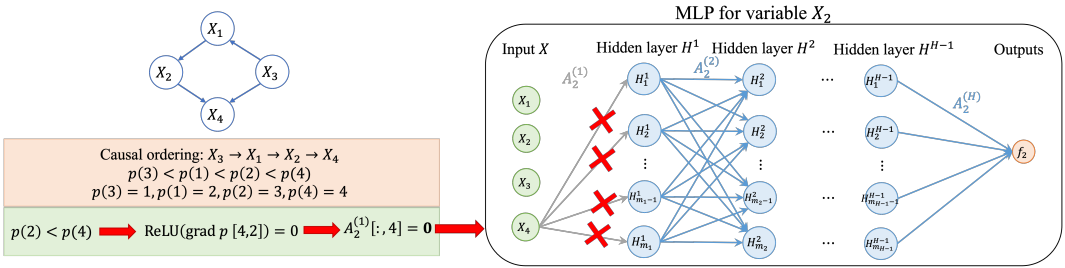
- We discover that if $W(A^{(1)})$ satisfies the acyclicity constraint, then $A_n^{(1)} \in \mathbb{R}^{m_1 \times N}$ can be represented in the format of a weight matrix $S_n \in \mathbb{R}^{m_1 \times N}$ and a potential vector $p \in \mathbb{R}^N$.

$$A_n^{(1)} = S_n \circ \mathrm{ReLU}\big((\mathrm{grad}p)[:, n]^T\big) \tag{2}$$

  - where grad is an operation on p: $(\mathrm{grad}p)[k, n] = p(n) - p(k)$.

- Intuitively, p can be seen as reserving the topological order of a DAG, where variables with higher orders correspond to smaller values.

DAG Projection Framework - (cont')



- Eq. (2) ensures acyclicity for $W(A^{(1)})$ by preventing variables with lower orders from becoming parent variables of those with higher orders.
- We propose to optimize over the weight matrices $S = [S_1, S_2, \cdots, S_N]$ and p instead of constrained $A^{(1)}$, and given (S, p), construct $A^{(1)}$ using Eq. (2).

Proposed Method - DAG-NCMLP

- With fixed number of iterations, solve for an initialization $\hat{A}$.

$$\hat{A} = \arg\min_{A} \sum_{n=1}^{N} \Big(X_n - A_n^{(H)} \sigma\big(\cdots A_n^{(2)} \sigma(A_n^{(1)} X) \cdots \big) \Big)^2 + \lambda \|A_n^{(1)}\|_1 \\ \alpha h\big(W(A^{(1)})\big) + \frac{\rho}{2} \|h\big(W(A^{(1)})\big)\|^2 \tag{3}$$

  Obtain an initial guess of potential vector p^{pre} from $\hat{A}$.

- We use p^{pre} as initialization for p and solve for the optimal solution (S^*, p^*) by minimizing the objective in Eq. (4).

$$\sum_{n=1}^{N} \Big(X_n - A_n^{(H)} \sigma\big(\cdots A_n^{(2)} \sigma(S_n \circ \mathrm{ReLU}((\mathrm{grad}p)[:, n]) X^T) \cdots \big) \Big)^2 + \lambda \|S_n \circ \mathrm{ReLU}((\mathrm{grad}p)[:, n])\|_1 \tag{4}$$

- Reconstruct $A^{(1)}$ using (S^*, p^*).

Empirical Results - Synthetic Data

ER2: SHD					
d	GraN-DAG	DAG-GNN	GS-GES	NOTEARS-MLP	DAG-NCMLP
10	$2.6e3 \pm 5.2e3$	$6.6e2 \pm 2.3e2$	$3.1e2 \pm 5.3e1$	$3.1e2 \pm 1.1e2$	$\mathbf{1.2e2 \pm 6.0e1}$
20	$2.5e3 \pm 5.8e2$	$1.5e3 \pm 2.3e2$	$1.5e3 \pm 1.2e2$	$7.6e2 \pm 1.4e2$	$\mathbf{3.9e2 \pm 8.6e1}$
40	$8.6e3 \pm 1.8e3$	$8.0e3 \pm 1.9e2$	$6.8e3 \pm 1.4e3$	$1.6e3 \pm 2.2e2$	$\mathbf{1.1e3 \pm 2.1e2}$
50	$1.4e4 \pm 1.5e3$	$1.2e4 \pm 1.1e2$	$1.0e4 \pm 9.4e2$	$8.5e3 \pm 2.3e3$	$\mathbf{1.6e3 \pm 3.9e2}$
100	$> 60h$	$2.0e4 \pm 6.4e2$	$> 60h$	limit to $60h$	$\mathbf{6.5e3 \pm 1.3e3}$

ER2: SHD					
d	GraN-DAG	DAG-GNN	GS-GES	NOTEARS-MLP	DAG-NCMLP
10	15.0 ± 6.0	13.3 ± 5.5	10.5 ± 3.9	5.7 ± 3.2	$\mathbf{5.5 \pm 2.5}$
20	22.7 ± 1.8	25.7 ± 3.6	19.4 ± 5.6	$\mathbf{13.0 \pm 3.8}$	13.5 ± 4.0
40	57.5 ± 8.6	56.1 ± 6.7	40.5 ± 9.5	$\mathbf{27.7 \pm 5.1}$	27.8 ± 5.8
50	68.9 ± 13.3	65.8 ± 7.8	50.6 ± 8.4	$\mathbf{36.0 \pm 9.7}$	36.9 ± 10.3
100	$> 60h$	144.8 ± 7.1	$> 60h$	$\mathbf{77.3 \pm 4.0}$	80.5 ± 6.0

- Across all graph and causal function settings, our method completes computations in approximately $\frac{1}{2}$ to $\frac{1}{10}$ of the time required by the second-best baseline.
- Our method achieves accuracy comparable to that of state-of-the-art methods.

Empirical Results - Real Data

Dataset	SHD	# Correct Edges	Ratio of Correct Edges	Runtime
GraN-DAG	**13**	6/17	0.353	6.1*e*2
DAG-GNN	19	**8**/**17**	**0.471**	5.3*e*2
GS-GES	17	6/17	0.353	5.0*e*2
NOTEARS-MLP	15	7/17	0.412	4.4*e*2
DAG-NCMLP	15	7/17	0.412	**1.5e2**

- Our method, DAG-NCMLP, significantly enhances computational efficiency while maintaining decent accuracy.

Conclusion

- To address the low computational efficiency issue, we propose an efficient nonlinear continuous learning method that avoids explicitly imposing the acyclicity constraint.

- By performing optimization in a DAG projection space that satisfies the acyclicity constraint, our approach significantly decreases runtime while maintaining accuracy comparable to state-of-the-art methods.

- Future work directions:
  - Enhance our algorithm by eliminating its reliance on heuristic and *ad hoc* strategies for estimating the potential vector p.
  - Apply the nonlinear DAG projection framework to other DAG learning methods, such as Bayesian DAG learning.

Questions?

Thank you!

Contact: naiyu.yin@gmail.com