Differentiable Structure Learning with Ancestral Constraints

Taiyu Ban, Changxin Rong, Xiangyu Wang, Lyuzhou Chen, Xin Wang, Derui Lyu, Qinrui Zhu, Huanhuan Chen

ICML 2025

Reporter: Taiyu Ban

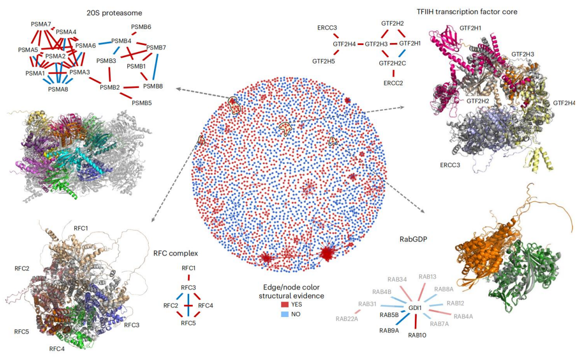
中国科学技术大学
University of Science and Technology of China

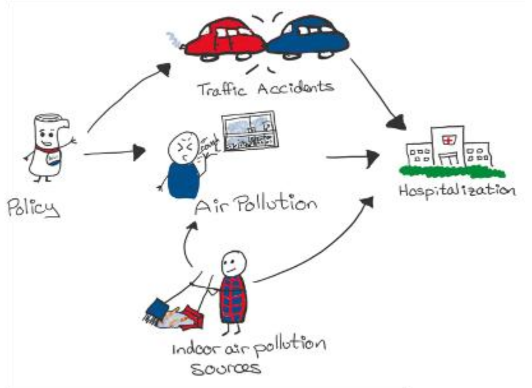
Outline

- **Background**
- Preliminary
- Method and Analyses

Differentiable Structure Learning for Causal Discovery

- Structure learning aims to recover the structure of the causal grahical model, a directed acyclic graph (DAG), that represents causal mechanisms underlying the observational data.
  - **Biology**
  - **Advertising**
  - **Public policy**
  - …





Differentiable Structure Learning for Causal Discovery

- Traditional structure learning is a combinatorial optimization problem, searching for the DAG with the optimal data approximation score.
- Zheng et al. [2018] reformulates structure learning as a continuous optimization problem by proposing a smooth function to characterize the ayclicity property of a graph.

Lyuzhou Chen,

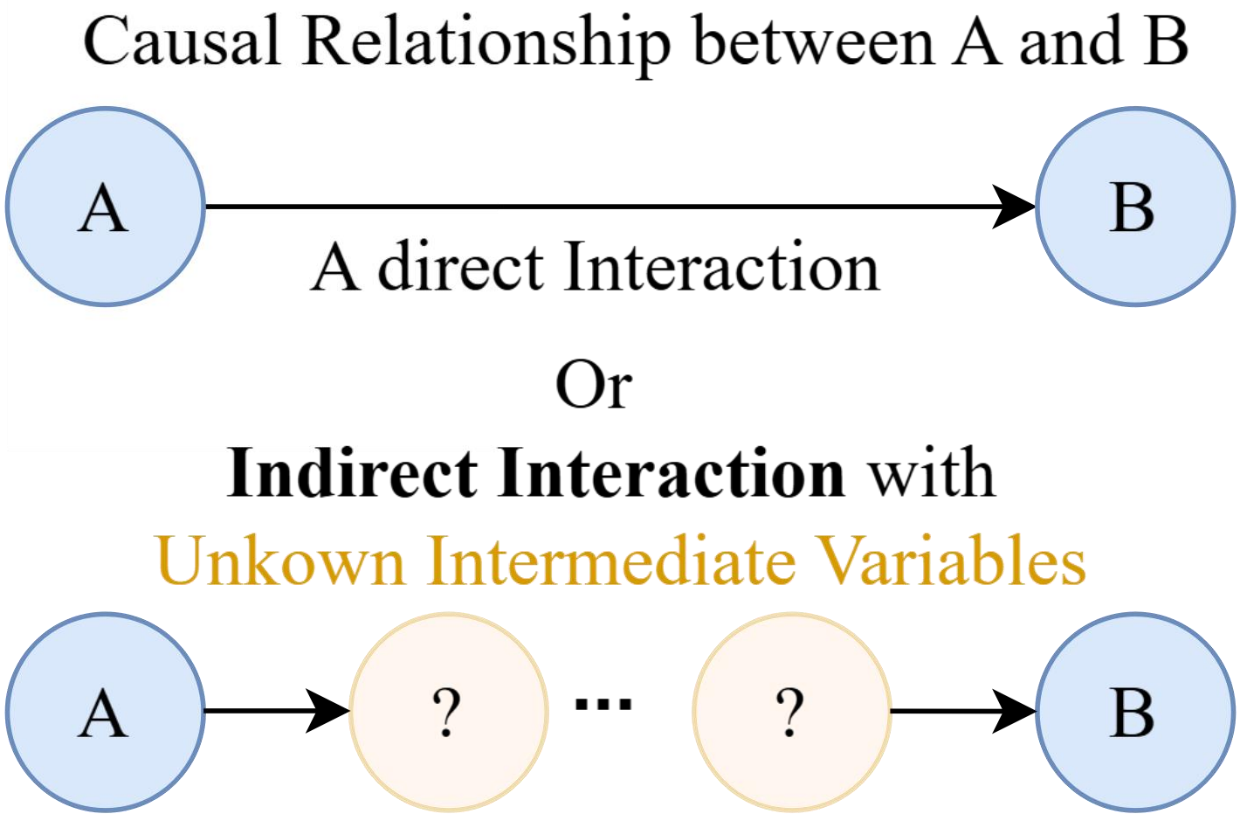
$$\min_{W \in \mathbb{R}^{d \times d}} \quad F(W) \qquad\qquad \min_{W \in \mathbb{R}^{d \times d}} \quad F(W)$$

$$\text{subject to} \;\; \mathsf{G}(W) \in \mathsf{DAGs} \quad \Longleftrightarrow \quad \text{subject to} \;\; h(W) = 0,$$

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. P. (2018, December). DAGs with NO TEARS: continuous optimization for structure learning. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 9492-9503).

Encoding Broad-Grained Prior Knowledge as Paths

Prior Knowledge Corresponds to **A Directed Path (Including Edge)** Between Variables in the Structure



Outline

- Background
- **Preliminary**
- Method and Analyses

Structural Equation Model

Structural equation model Let G denote a directed acyclic graph (DAG) with d nodes, where the vertex set V corresponds to a set of random variables $X = \{X_1, X_2, \ldots, X_d\}$, and the edge set $E(G) \subset V \times V$ defines the causal relationships among the variables. The structural equation model (SEM) specifies that the value of each variable is determined by a function of its parent variables in G and an independent noise component:

$$X_j = f_j(\mathrm{Pa}_j^G, z_j) \tag{1}$$

where $\mathrm{Pa}_j^G = \{X_i \mid X_i \in X, (X_i, X_j) \in E\}$ denotes the set of parent variables of X_j in G, and z_j represents noise that is independent across different j. Denoting the structure of G as a weighted adjacent matrix $W \in \mathbb{R}^{d \times d}$, where $W_{i,j} \neq 0$ equals that $(X_i, X_j) \in E(G)$, we have:

$$X_j = f_j(W_{:,j}, X, z_j) \tag{2}$$

Task Definition of Differentiable Structure Learning

$$\min_{W \in \mathbb{R}^{d\times d}} \mathcal{F}(W) \quad \text{subject to } h(W) = 0$$

$$h(W) = \boxed{\text{Trace}\left(\sum_{i=1}^{d} c_i (W \circ W)^i\right),} \quad c_i > 0$$

→ **For all i =1,…,d, forbid i-length path from a node to itself.**

⇩ **Path Absence Characterization**

$$\sum_{k=1}^{d} (|W|^k)_{i,j} = 0$$

Forbid k-length path from i to j for all k=1,…,d

⇔ Absence of Path (i,j)

Some designs of the Acyclicity Constraint:

$$h(W) = \text{Trace}(e^{W \circ W}) - d$$

$$h(W) = \text{Trace}\left(\left(I + \frac{1}{d} W \circ W\right)^d - I\right)$$

$$h(W) = -\log\det(sI - W \circ W) + d \log s$$

Outline

- Background
- Preliminary
- **Method and Analyses**

Differentiable Structure Learning with Ancestral Constraints

- Differentiable structure learning with ancestral constraints (mainly path existence here) can be formulated as:

$$\min_{W} F(W) \quad \text{subject to } h(W) = 0, \ \text{path } i \rightsquigarrow j \in G(W)$$

HOW TO CHARACTERIZE PATH EXISTENCE

DIFFERENTIALBLY AND EQUIVALENTLY?

Path Existence Characterization with Relaxation

Path Absence Characterization

$$\sum_{k=1}^{d}(|W|^{k})_{i,j} = 0$$

⇩ **Relaxation**

$$\sum_{k=1}^{d}(|W|^{k})_{i,j} \geq \epsilon$$

Path Existence Characterization

Issue of In-Equivalence to Path Existence

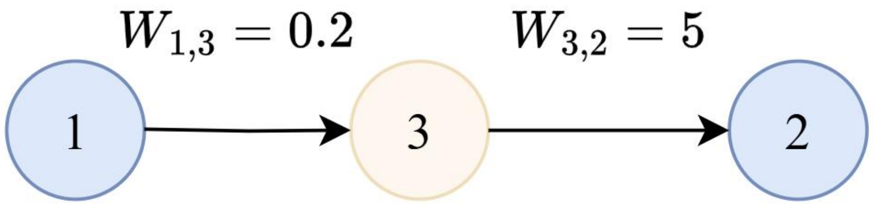
Consider $\bar{p}(W) = \text{ReLU}(\epsilon - p(W)), \quad p(W) = \sum_{k=1}^{d} |W|^k$

$$(\bar{p}(W))_{i,j} = 0 \;\not\Longleftrightarrow\; i \rightsquigarrow j \in G(W)$$

Under the edge thresholding process:

$$\text{Edge } (i, j) \in G(W) \iff |W_{i,j}| \geq \epsilon_0$$

Issue of In-Equivalence to Path Existence



Edge threshold $\epsilon_0 = 0.3$
Path threshold $\epsilon = 0.9$

$(p(W))_{1,2} \geq |W|_{1,3}|W|_{3,2} = 1.0$

$(\bar{p}(W))_{1,2} = \text{ReLU}(0.9 - (p(W))_{1,2}) = 0$

$|W|_{1,3} < 0.3 \Rightarrow \text{Edge } (1,3) \notin G(W)$

$\Rightarrow \text{Path } (1,3,2) \notin G(W)$

$(\bar{p}(W))_{1,2} = 0$ but Path $1 \rightsquigarrow 2 \notin G(W)$

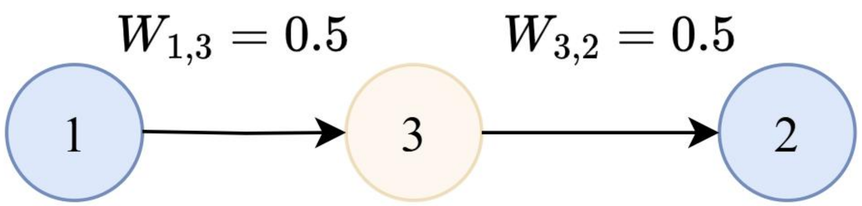
Lemma 1. *(Sufficient Condition) There exists a finite threshold* $f(\epsilon_0, \sigma) > \epsilon_0$ *such that* $(\bar{p}(W))_{i,j} = 0$ *is sufficient to guarantee path existence* $x_i \rightsquigarrow x_j \in G(W)$ *under edge relaxation in Equation* (14) *if and only if* $\epsilon \geq f(\epsilon_0, \sigma)$.

In-Sufficiency If $\epsilon < f(\epsilon_0, \sigma)$

$$(\bar{p}(W))_{i,j} = 0 \not\Rightarrow i \rightsquigarrow j \in G(W)$$

if without sufficiently large ϵ

Issue of In-Equivalence to Path Existence



Edge threshold $\epsilon_0 = 0.3$
Path threshold $\epsilon = 0.9$

$(\bar{p}(W))_{1,2} = \text{ReLU}(0.9 - 0.5 \times 0.5) = 0.65$

Edges $(1,3),(3,2) \in G(W)$
$\Rightarrow$ Path $(1,3,2) \in G(W)$

Path $1 \rightsquigarrow 2 \in G(W)$ but $(\bar{p}(W))_{1,2} \neq 0$

Lemma 2. *(Necessary Condition) The continuous equality* $(\bar{p}(W))_{i,j} = 0$ *is necessary for the path existence* $x_i \rightsquigarrow x_j \in G(W)$ *under edge relaxation in Equation* (14) *if and only if* $\epsilon \leq \min(\epsilon_0, \epsilon_0^d)$.

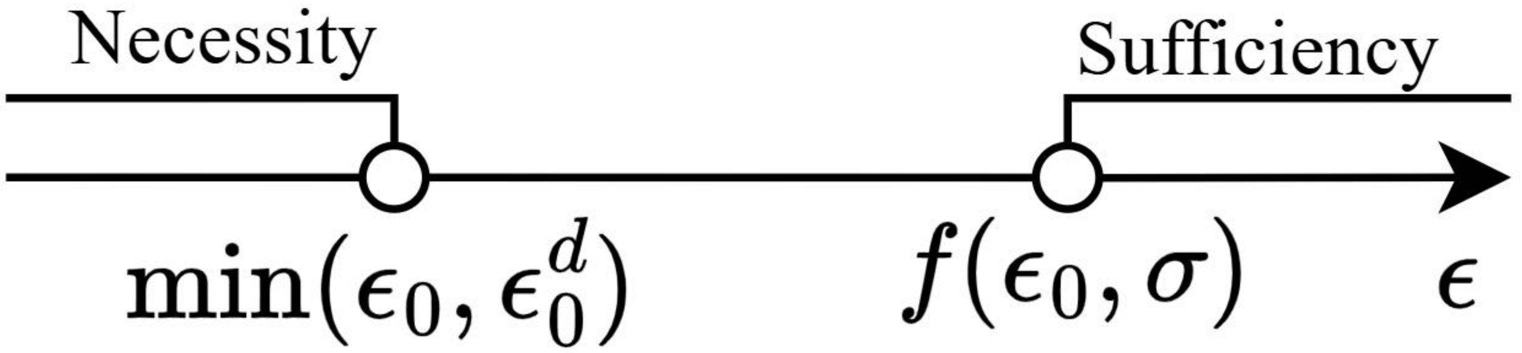
Un-Necessity If $\epsilon > \min(\epsilon_0, \epsilon_0^d)$

$$i \rightsquigarrow j \in G(W) \nRightarrow (\bar{p}(W))_{i,j} = 0$$

if without sufficiently small ϵ

Issue of In-Equivalence to Path Existence

$$\bar{p}(W) = \mathrm{ReLU}(\epsilon - p(W)) = 0 \text{ for Path Existence}$$



NO ϵ TO SATISFY BOTH NECESSITY AND SUFFICENCY

Equivalent Path Existence Characterization

$$\hat{p}(W) = \bar{p}(W) \circ b(W)$$

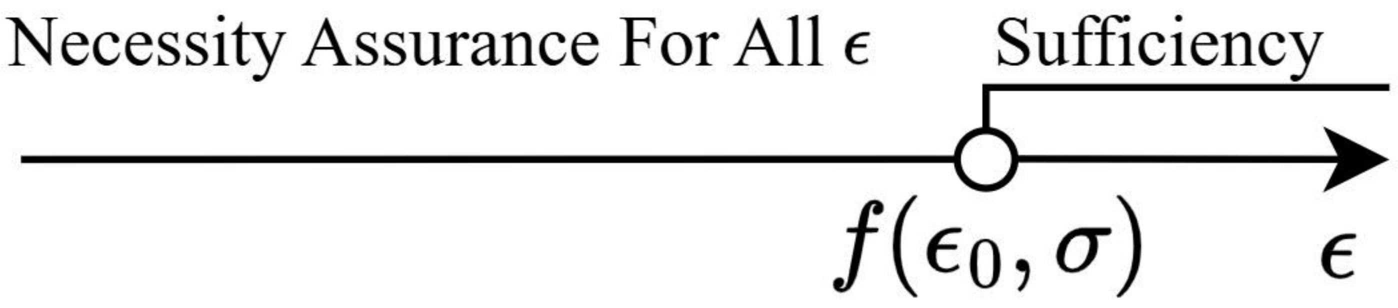
$$b(W) = \mathbb{I}\left(\sum_{k=1}^{d} (\mathbb{I}(|W| \geq \epsilon_0))^k = 0\right)$$

Binary $b(W) \in \{0,1\}^{d \times d}$ 1 For Path Absence and 0 For Path Existence

Neccesity Assurance: Path $i \rightsquigarrow j \in G(W) \Rightarrow (b(W))_{i,j} = 0 \Rightarrow (\hat{p}(W))_{i,j} = 0$

Equivalent Path Existence Characterization

$\hat{p}(W) = \bar{p}(W) \circ b(W) = 0$ for Path Existence



SUFFICINETLY LARGE ϵ TO SATISFY
BOTH NECESSITY AND SUFFICENCY

Binary $b(W) \in \{0, 1\}^{d \times d}$ 1 For Path Absence and 0 For Path Existence

Neccesity Assurance: Path $i \rightsquigarrow j \in G(W) \Rightarrow (b(W))_{i,j} = 0 \Rightarrow (\hat{p}(W))_{i,j} = 0$

Equivalent Path Existence Characterization

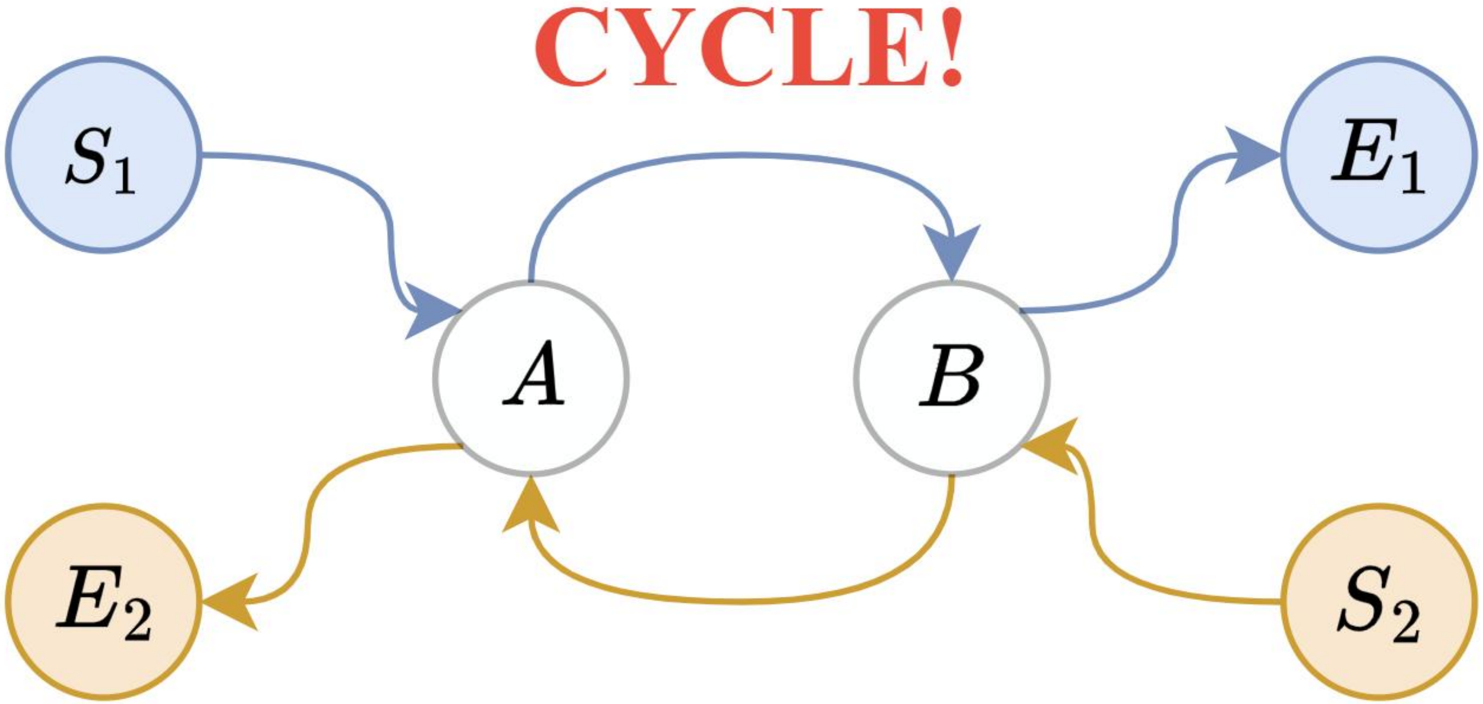
Theorem 1. *There exists at least one directed path from* x_i *to* x_j *in* $G(W)$ *constructed by Equation* (14) *if and only if* $(\hat{p}(W))_{i,j} = 0$, *where* $\hat{p}(W)$ *is defined by Equation* (16) *with* $\epsilon \geq f(\epsilon_0, \sigma)$ *for some finite* $f(\epsilon_0, \sigma)$.

$$\hat{p}(W) = \bar{p}(W) \circ b(W)$$

$$b(W) = \mathbb{I}\left(\sum_{k=1}^{d} (\mathbb{I}(|W| \geq \epsilon_0))^k = 0\right)$$

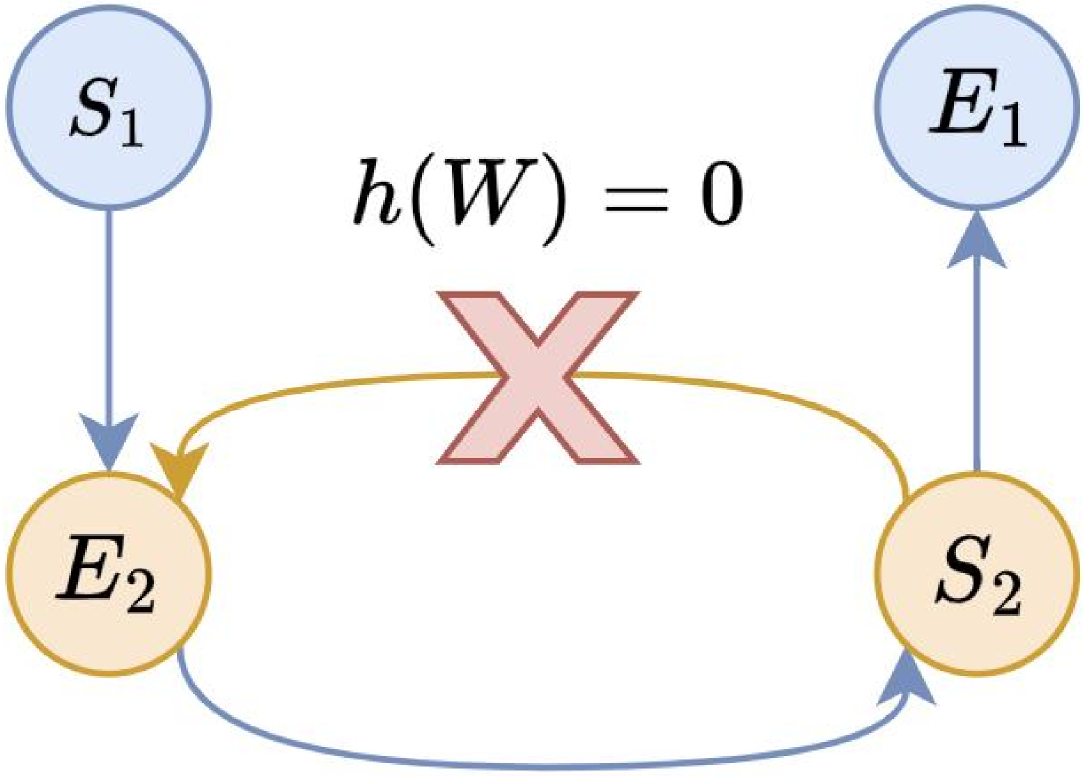
$$\bar{p}(W) = \mathrm{ReLU}(\epsilon - p(W)), \quad p(W) = \sum_{k=1}^{d} |W|^k$$

Order Violation Among Paths



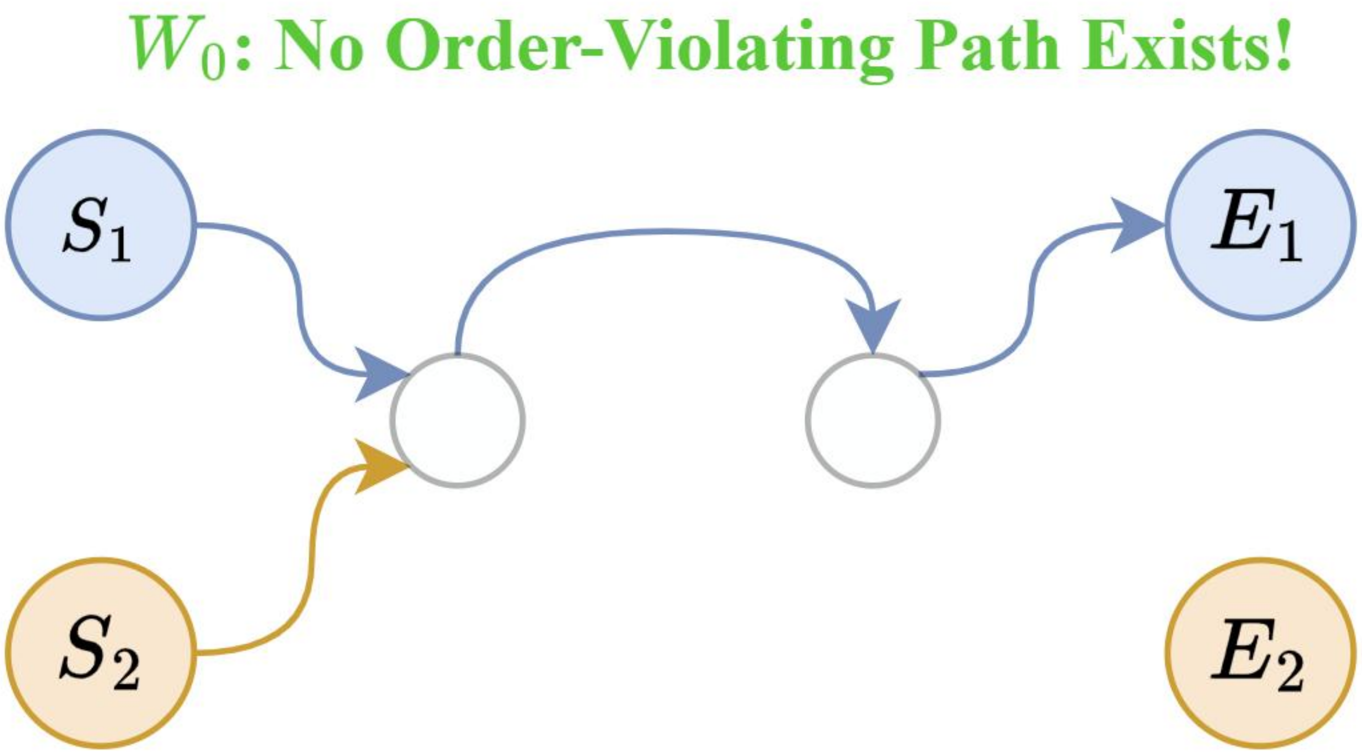
Order Violation Between Path $S_1 \rightsquigarrow E_1$ and $S_2 \rightsquigarrow E_2$

Order Violation Among Paths



Fail the recovery of $S_2 \rightsquigarrow E_2$ due to Acyclicity

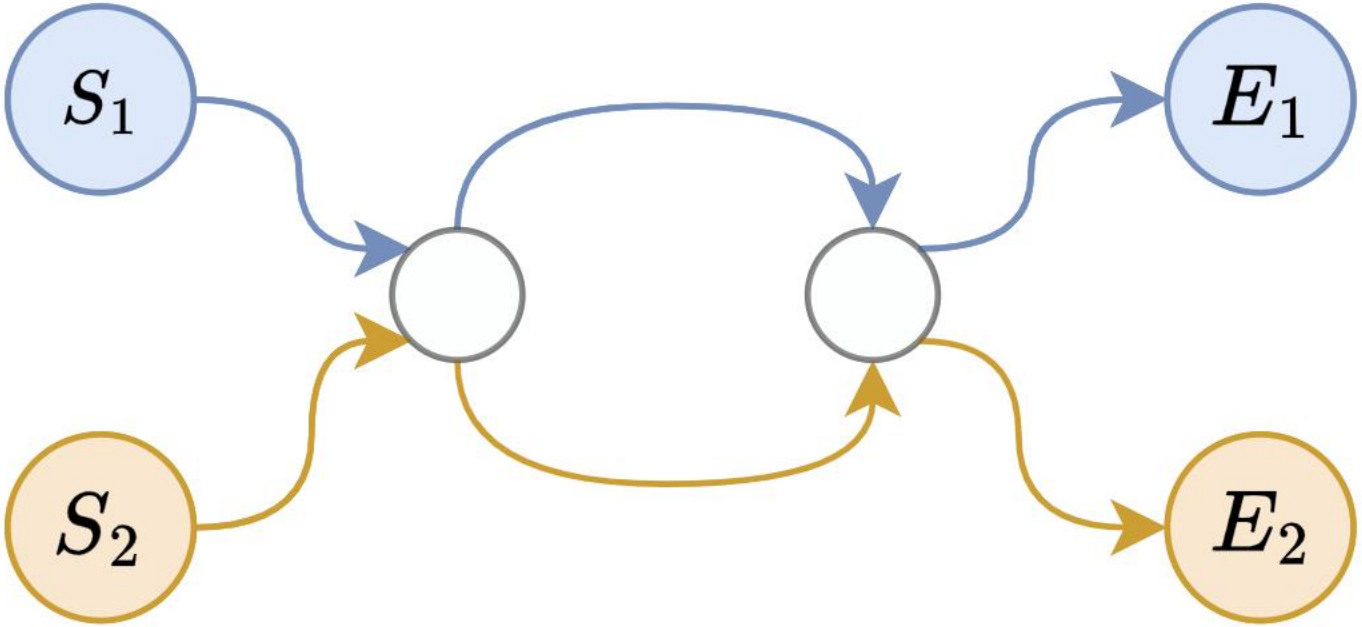
Solution: Order-Guided Optimization



Impose Partial Order Constraint
$S_1 \prec E_1, S_2 \prec E_2$ Implied by Paths

Solution: Order-Guided Optimization

Avoid Order Violation with Initialization W_0



Solving the Path Existence-Constrained Issue
Starting from Order-Guided Optima W_0

Overall Alagrithm

Algorithm 1 Differentiable Structure Learning with Path Existence Constraints

Require: Data D, binary mask $A \in \{0,1\}^{d\times d}$ of path-existence constraints, edge threshold ϵ_0

1: **Define** backbone model: $M\langle L, h, W_\theta\rangle$, with data-fit loss L, acyclicity loss h, and structure parameters W_θ.

2: **Define** path existence loss:

$$L' = L + \sum(\hat{p}(W) \circ A)$$

3: **Define** order-based acyclicity loss:

$$h_o = h + \sum(p(W) \circ (A^+)^T)$$

4: Solve order-based optimization (initializing from zero):

$$W_o \leftarrow M_o(D, 0), \quad \text{where} \quad M_o\langle L, h_o, W_\theta\rangle$$

5: Solve path existence-based optimization using the order-based optimization result W_o as initialization:

$$W_p \leftarrow M_p(D, W_o), \quad \text{where} \quad M_p\langle L', h, W_\theta\rangle$$

6: Threshold learned structure:

$$\bar{W}_p \leftarrow \mathbb{I}(|W_p| > \epsilon_0)$$

7: **Return** Final learned structure $\bar{W}_p$

GOAL: Solve structure learning with path existence constraints $A \in \{0,1\}^{d\times d}$

STEP 1: Solve structure learning with partial order constraints A and derive W_0.

This task has been addressed by a previous work "Differentiable structure learning with partial orders."

STEP 2: Solve structure learning with path existence constraints A with init point W_0.

Thank you