Differentiable Structure Learning with Partial Orders

Taiyu Ban, Lyuzhou Chen, Xiangyu Wang, Xin Wang, Derui Lyu, Huanhuan Chen

NeurIPS 2024

Reporter: Taiyu Ban

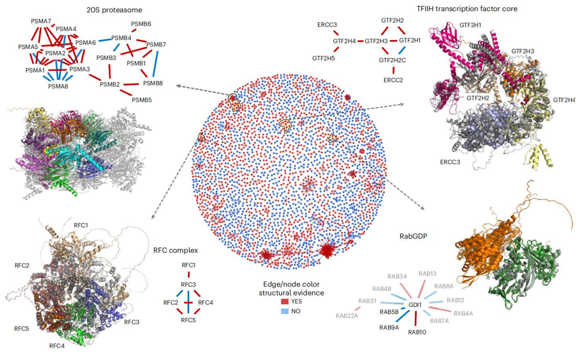
中国科学技术大学
University of Science and Technology of China

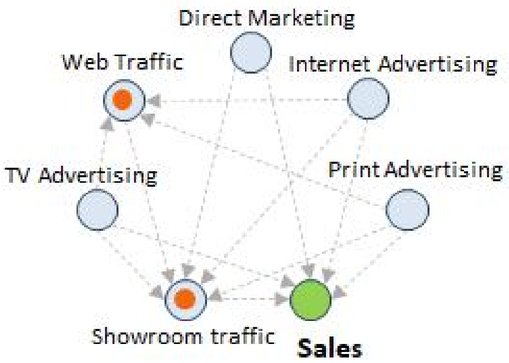
Outline

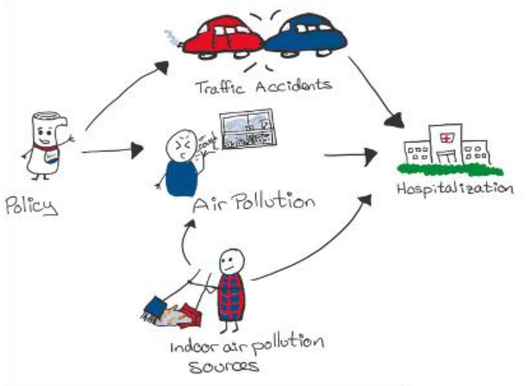
- **Background**
- Preliminary
- Method and Theoretical Analyses
- Conclusion

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.

$$\begin{array}{ll} \min\limits_{W \in \mathbb{R}^{d \times d}} & F(W) \\ \text{subject to} & \mathsf{G}(W) \in \mathsf{DAGs} \end{array} \quad \Longleftrightarrow \quad \begin{array}{ll} \min\limits_{W \in \mathbb{R}^{d \times d}} & F(W) \\ \text{subject to} & h(W) = 0, \end{array}$$

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).

Differentiable Structure Learning with Partial Orders

- The order relationship of variables, characterized by partial orders, is a common and important prior type in real-world, and **can be easily incorporated** in order-based search in combinatorial structure learning.
- However, partial order constraints **can not be easily integrated** in differentiable structure learning as it is modeled in the graph space instead of order space.

$$\min_{\Pi, W} F(\mathrm{Triu}(\Pi W \Pi^T); X) \quad \text{subject to } \Pi \models \mathcal{O} \qquad \Longleftrightarrow \qquad \min_{W \in \mathbb{R}^{d \times d}} F(W) \quad \text{subject to } h(W) = 0, \Pi(W) \models \mathcal{O}$$

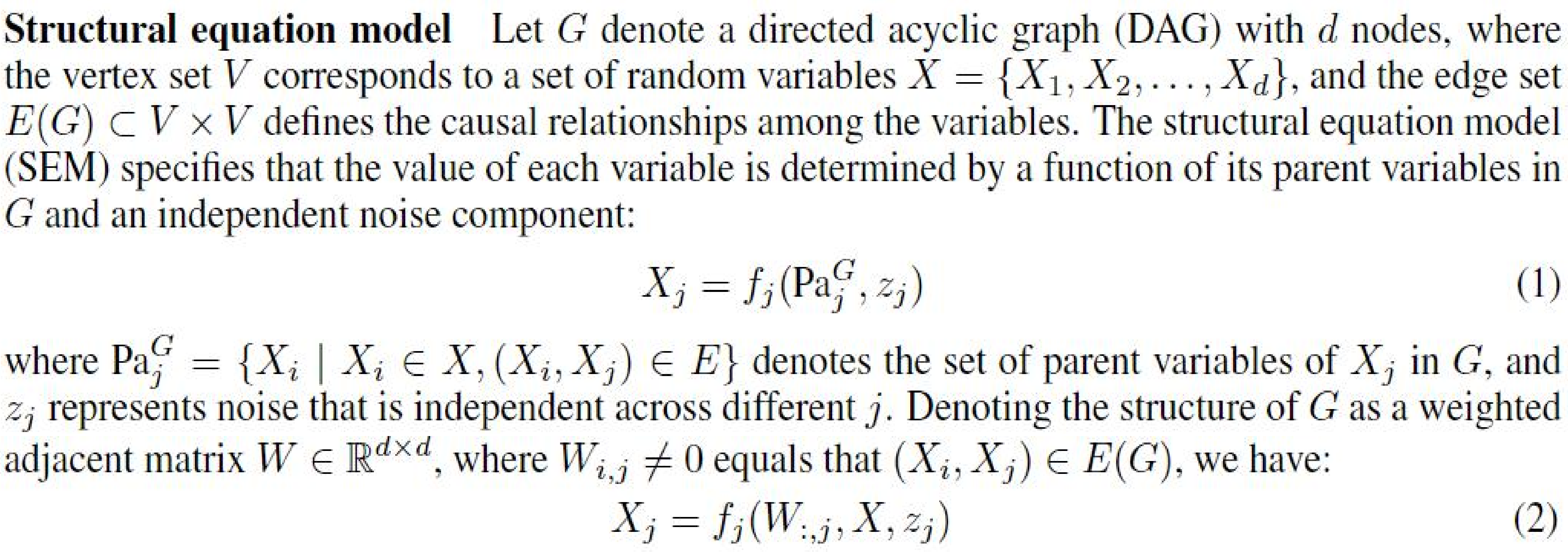
HOW TO CHARACTERIZE $\Pi(W) \models \mathcal{O}$ DIFFERENTIALBLY?

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).

Outline

- ➢Background
- ➢ **Preliminary**
- ➢Method and Theoretical Analyses
- ➢Conclusion

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:

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

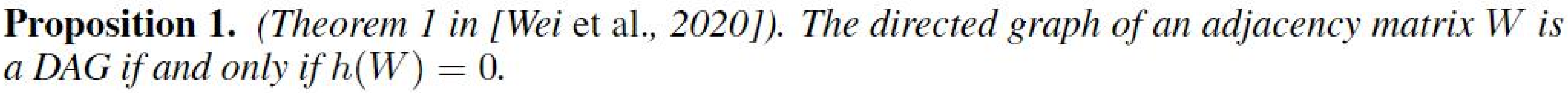
where $\mathrm{Pa}_j^G = \{X_i \mid X_i \in X, (X_i, X_j) \in E\}$ denotes the set of parent variables of X_j in G, and z_j represents noise that is independent across different j. Denoting the structure of G as a weighted adjacent matrix $W \in \mathbb{R}^{d \times d}$, where $W_{i,j} \neq 0$ equals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

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

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


Proposition 1. *(Theorem 1 in [Wei* et al.*, 2020]). The directed graph of an adjacency matrix W is a DAG if and only if $h(W) = 0$.*

Some designs of the Acyclicity Constraint:

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

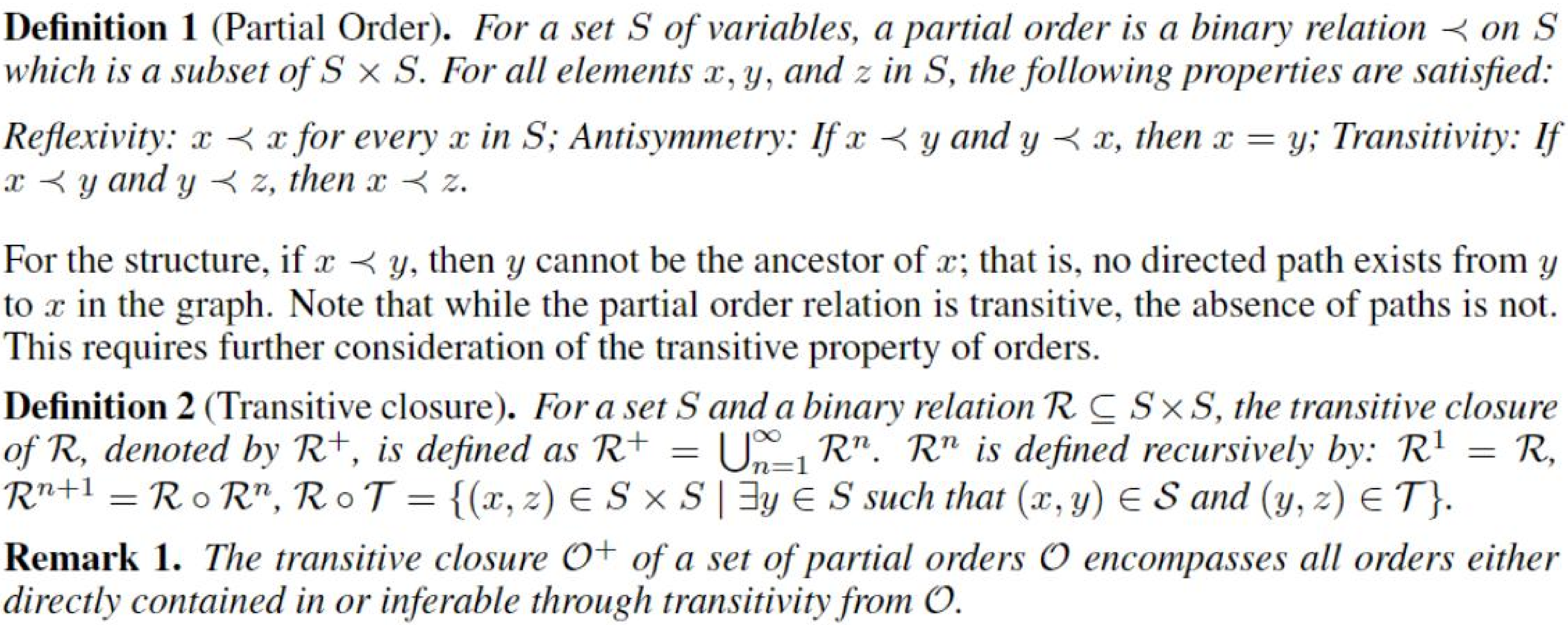
$$h(W) = \operatorname{Trace}\left(\left(I + \frac{1}{d} W \circ W\right)^d - I\right)$$

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

Outline

- Background
- Preliminary
- **Method and Theoretical Analyses**
- Conclusion

Partial Orders to Equivalent Path Absences



Definition 1 (Partial Order). *For a set S of variables, a partial order is a binary relation $\prec$ on S which is a subset of $S \times S$. For all elements x, y, and z in S, the following properties are satisfied:*

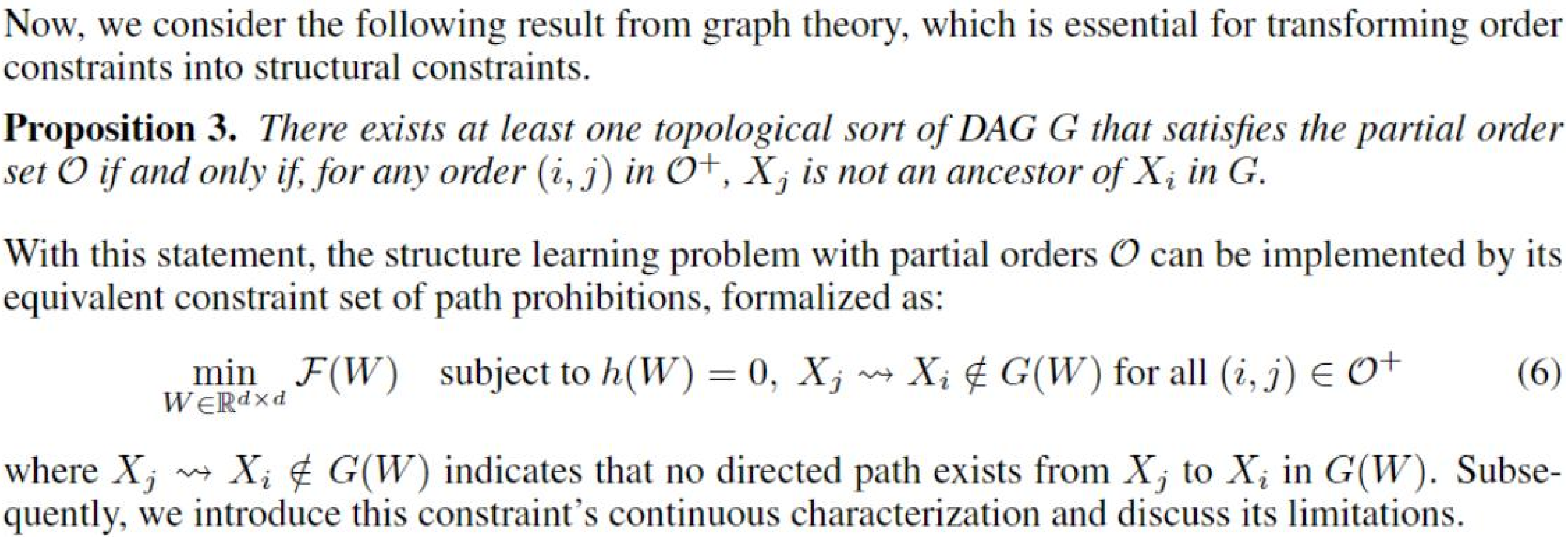
Reflexivity: $x \prec x$ for every x in S; Antisymmetry: If $x \prec y$ and $y \prec x$, then $x = y$; Transitivity: If $x \prec y$ and $y \prec z$, then $x \prec z$.

For the structure, if $x \prec y$, then y cannot be the ancestor of x; that is, no directed path exists from y to x in the graph. Note that while the partial order relation is transitive, the absence of paths is not. This requires further consideration of the transitive property of orders.

Definition 2 (Transitive closure). *For a set S and a binary relation $\mathcal{R} \subseteq S \times S$, the transitive closure of $\mathcal{R}$, denoted by $\mathcal{R}^+$, is defined as $\mathcal{R}^+ = \bigcup_{n=1}^{\infty} \mathcal{R}^n$. $\mathcal{R}^n$ is defined recursively by: $\mathcal{R}^1 = \mathcal{R}$, $\mathcal{R}^{n+1} = \mathcal{R} \circ \mathcal{R}^n$, $\mathcal{R} \circ \mathcal{T} = \{(x, z) \in S \times S \mid \exists y \in S$ such that $(x, y) \in \mathcal{S}$ and $(y, z) \in \mathcal{T}\}$.*

Remark 1. *The transitive closure $\mathcal{O}^+$ of a set of partial orders $\mathcal{O}$ encompasses all orders either directly contained in or inferable through transitivity from $\mathcal{O}$.*

Partial Orders to Equivalent Path Absences



Now, we consider the following result from graph theory, which is essential for transforming order constraints into structural constraints.

Proposition 3. *There exists at least one topological sort of DAG G that satisfies the partial order set $\mathcal{O}$ if and only if, for any order (i, j) in $\mathcal{O}^+$, X_j is not an ancestor of X_i in G.*

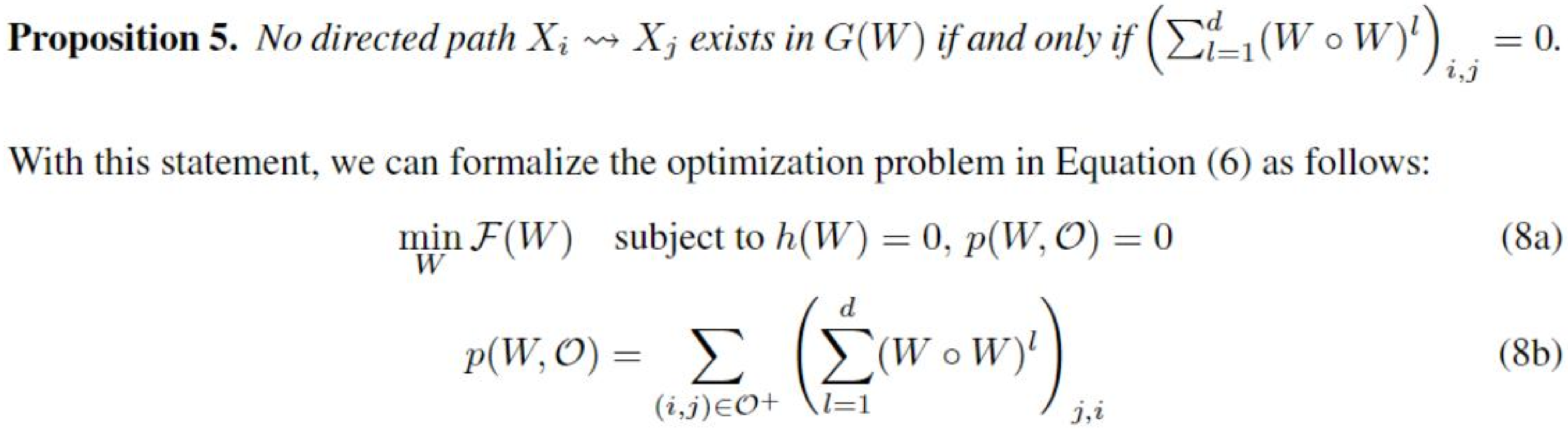
With this statement, the structure learning problem with partial orders $\mathcal{O}$ can be implemented by its equivalent constraint set of path prohibitions, formalized as:

$$\min_{W \in \mathbb{R}^{d \times d}} \mathcal{F}(W) \quad \text{subject to } h(W) = 0,\ X_j \rightsquigarrow X_i \notin G(W) \text{ for all } (i, j) \in \mathcal{O}^+ \tag{6}$$

where $X_j \rightsquigarrow X_i \notin G(W)$ indicates that no directed path exists from X_j to X_i in $G(W)$. Subsequently, we introduce this constraint's continuous characterization and discuss its limitations.

Continuous Characterization of Path Absences

- Given that the (i, j)th element of $(W \circ W)^k$ represents the existence of k-length paths from node i to j, we have the following formula to characterize path absence constriants:



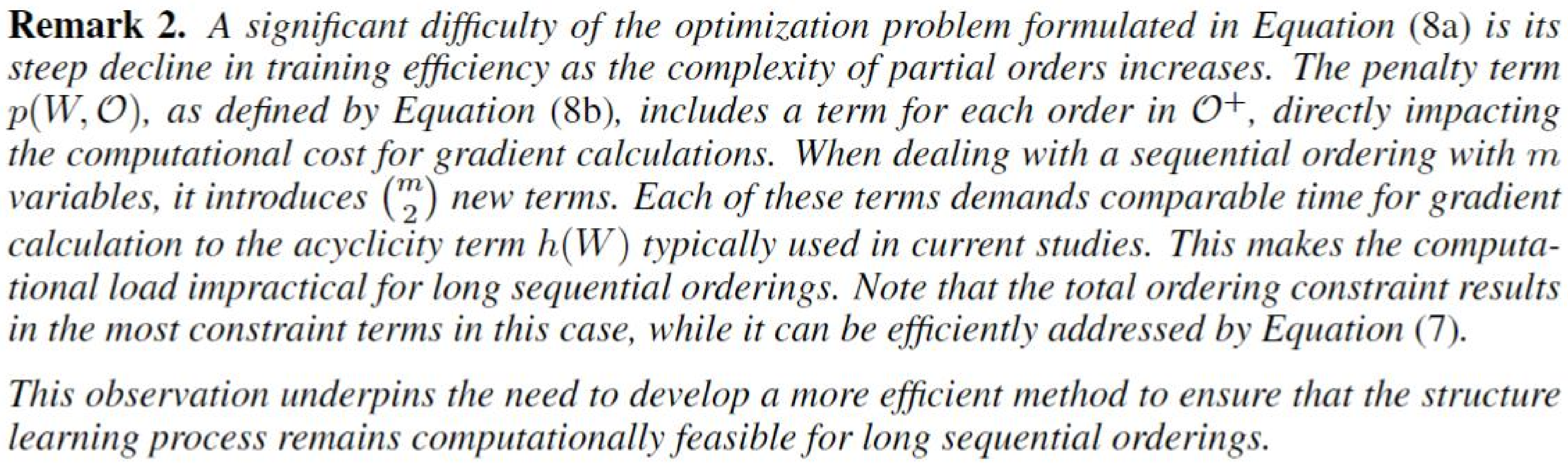
Proposition 5. *No directed path $X_i \rightsquigarrow X_j$ exists in $G(W)$ if and only if* $\left(\sum_{l=1}^{d}(W \circ W)^l\right)_{i,j} = 0$.

With this statement, we can formalize the optimization problem in Equation (6) as follows:

$$\min_{W} \mathcal{F}(W) \quad \text{subject to } h(W) = 0,\ p(W, \mathcal{O}) = 0 \tag{8a}$$

$$p(W, \mathcal{O}) = \sum_{(i,j) \in \mathcal{O}+} \left(\sum_{l=1}^{d}(W \circ W)^l\right)_{j,i} \tag{8b}$$

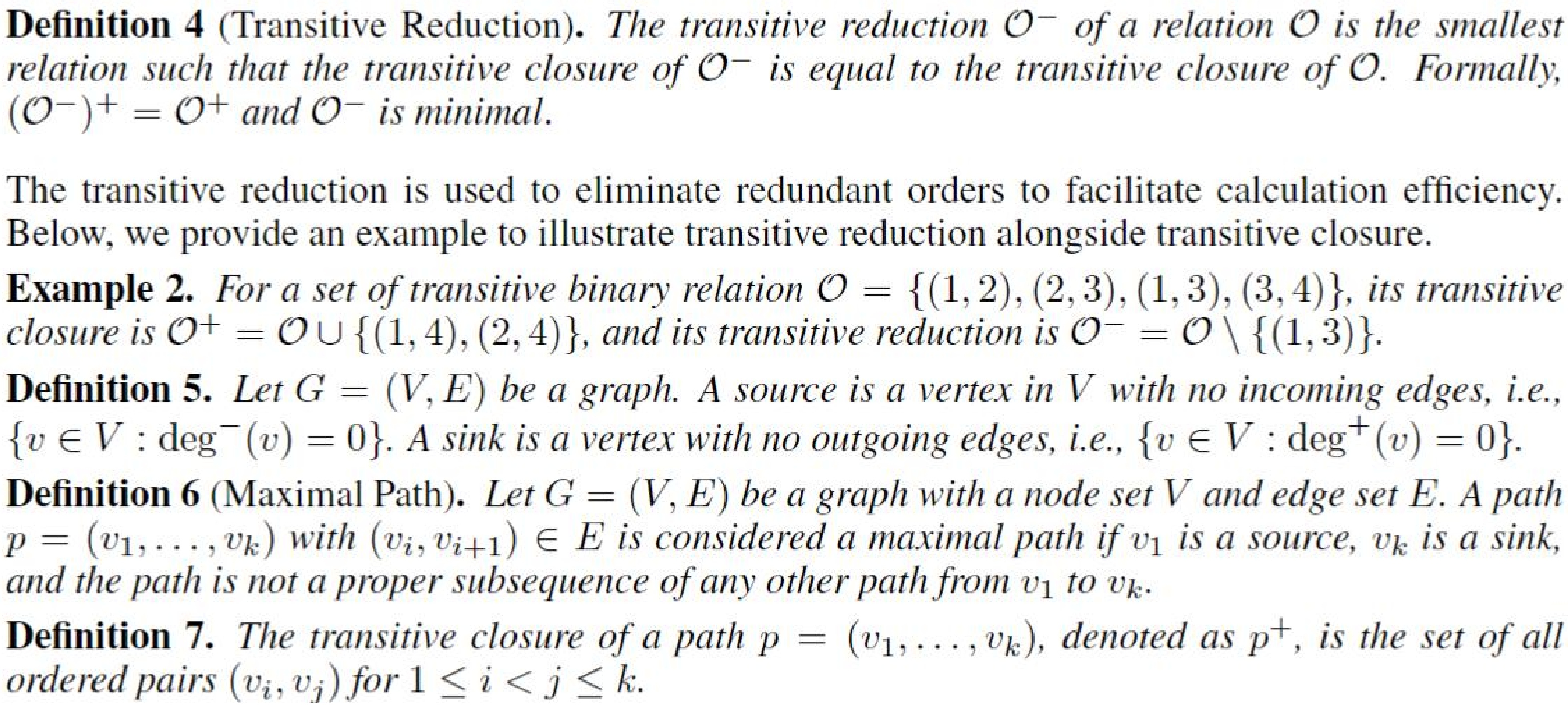
Continuous Characterization of Path Absences



Remark 2. *A significant difficulty of the optimization problem formulated in Equation* (8a) *is its steep decline in training efficiency as the complexity of partial orders increases. The penalty term* $p(W, \mathcal{O})$, *as defined by Equation* (8b), *includes a term for each order in* $\mathcal{O}^+$, *directly impacting the computational cost for gradient calculations. When dealing with a sequential ordering with* m *variables, it introduces* $\binom{m}{2}$ *new terms. Each of these terms demands comparable time for gradient calculation to the acyclicity term* $h(W)$ *typically used in current studies. This makes the computational load impractical for long sequential orderings. Note that the total ordering constraint results in the most constraint terms in this case, while it can be efficiently addressed by Equation* (7).

This observation underpins the need to develop a more efficient method to ensure that the structure learning process remains computationally feasible for long sequential orderings.

Augmented Acyclicity for Partial Orders



Definition 4 (Transitive Reduction). *The transitive reduction $\mathcal{O}^-$ of a relation $\mathcal{O}$ is the smallest relation such that the transitive closure of $\mathcal{O}^-$ is equal to the transitive closure of $\mathcal{O}$. Formally, $(\mathcal{O}^-)^+ = \mathcal{O}^+$ and $\mathcal{O}^-$ is minimal.*

The transitive reduction is used to eliminate redundant orders to facilitate calculation efficiency. Below, we provide an example to illustrate transitive reduction alongside transitive closure.

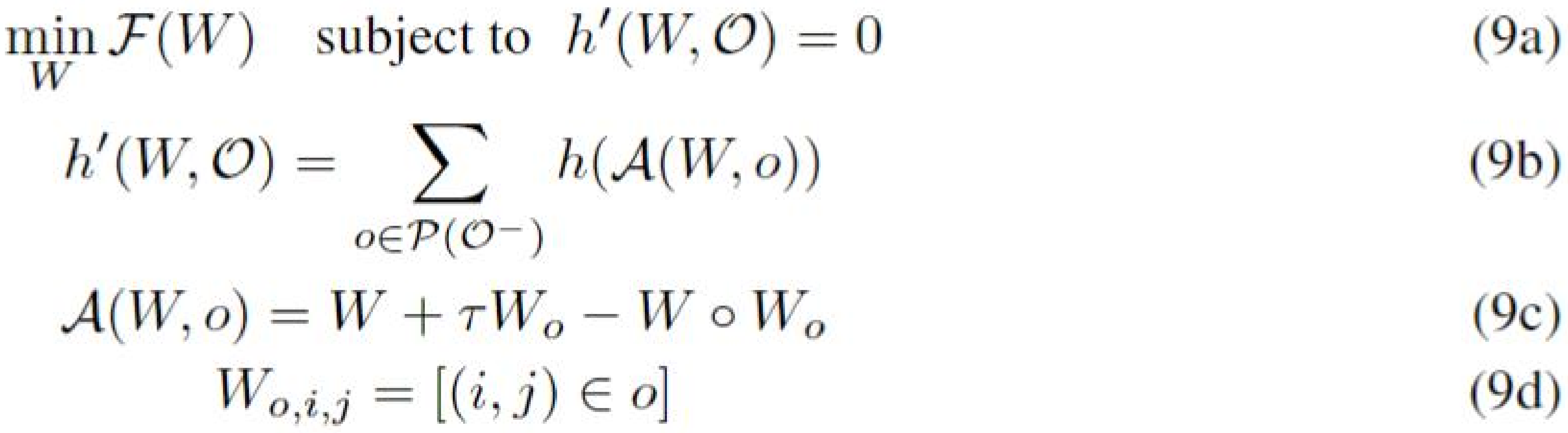
Example 2. *For a set of transitive binary relation $\mathcal{O} = \{(1,2),(2,3),(1,3),(3,4)\}$, its transitive closure is $\mathcal{O}^+ = \mathcal{O} \cup \{(1,4),(2,4)\}$, and its transitive reduction is $\mathcal{O}^- = \mathcal{O} \setminus \{(1,3)\}$.*

Definition 5. *Let $G = (V,E)$ be a graph. A source is a vertex in V with no incoming edges, i.e., $\{v \in V : \deg^-(v) = 0\}$. A sink is a vertex with no outgoing edges, i.e., $\{v \in V : \deg^+(v) = 0\}$.*

Definition 6 (Maximal Path). *Let $G = (V,E)$ be a graph with a node set V and edge set E. A path $p = (v_1,\ldots,v_k)$ with $(v_i, v_{i+1}) \in E$ is considered a maximal path if v_1 is a source, v_k is a sink, and the path is not a proper subsequence of any other path from v_1 to v_k.*

Definition 7. *The transitive closure of a path $p = (v_1,\ldots,v_k)$, denoted as p^+, is the set of all ordered pairs (v_i, v_j) for $1 \le i < j \le k$.*

Augmented Acyclicity for Partial Orders

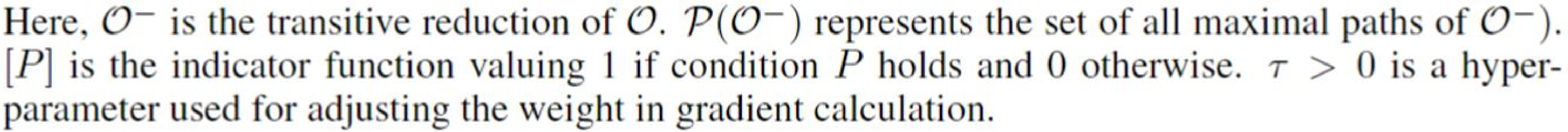


$$\min_{W} \mathcal{F}(W) \quad \text{subject to} \quad h'(W, \mathcal{O}) = 0 \tag{9a}$$

$$h'(W, \mathcal{O}) = \sum_{o \in \mathcal{P}(\mathcal{O}^-)} h(\mathcal{A}(W, o)) \tag{9b}$$

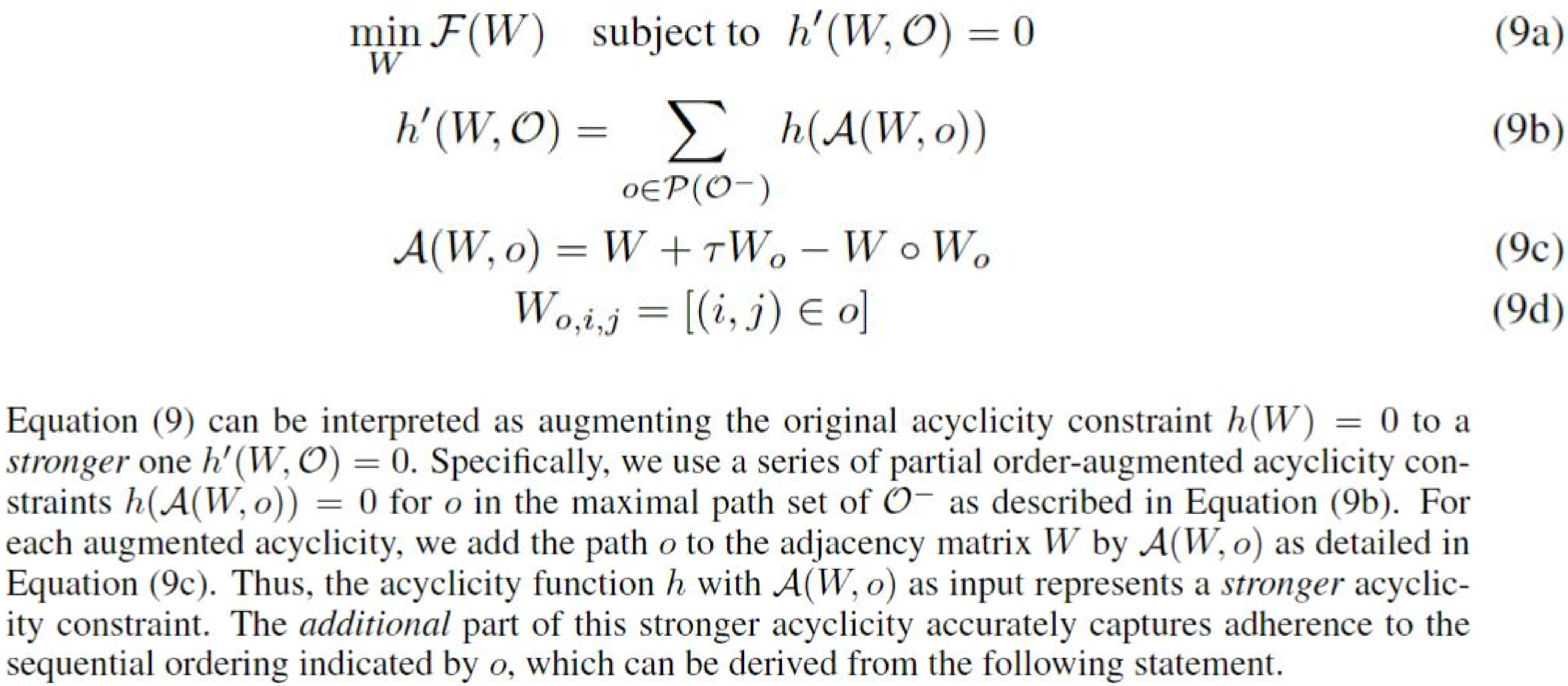
$$\mathcal{A}(W, o) = W + \tau W_o - W \circ W_o \tag{9c}$$

$$W_{o,i,j} = [(i, j) \in o] \tag{9d}$$



Here, $\mathcal{O}^-$ is the transitive reduction of $\mathcal{O}$. $\mathcal{P}(\mathcal{O}^-)$ represents the set of all maximal paths of $\mathcal{O}^-$). $[P]$ is the indicator function valuing 1 if condition P holds and 0 otherwise. $\tau > 0$ is a hyperparameter used for adjusting the weight in gradient calculation.

Augmented Acyclicity for Partial Orders



$$\min_{W} \mathcal{F}(W) \quad \text{subject to} \quad h'(W, \mathcal{O}) = 0 \tag{9a}$$

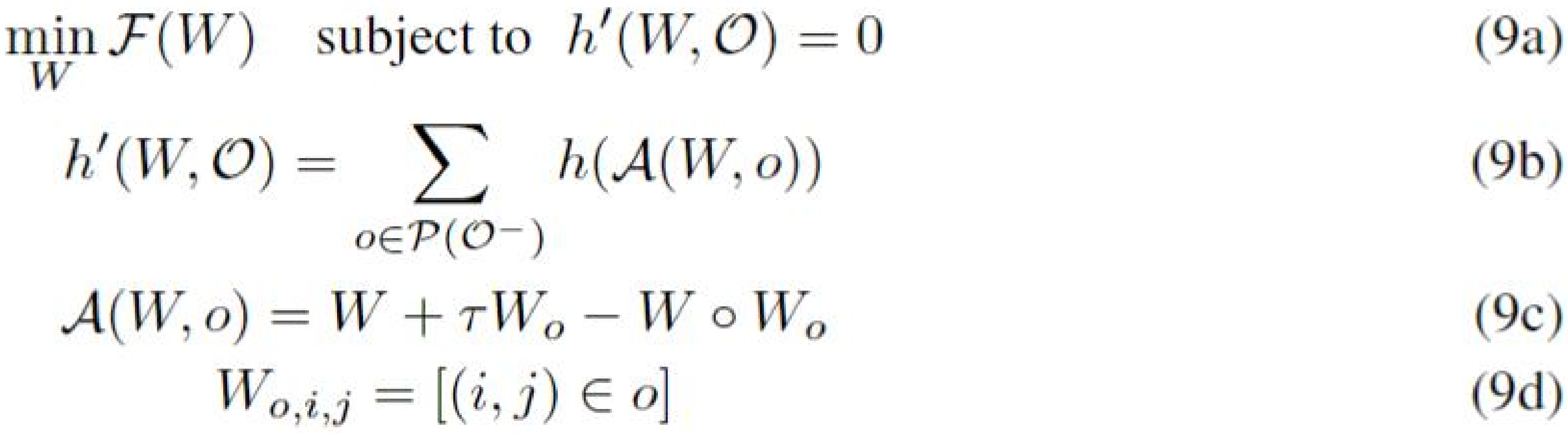
$$h'(W, \mathcal{O}) = \sum_{o \in \mathcal{P}(\mathcal{O}^{-})} h(\mathcal{A}(W, o)) \tag{9b}$$

$$\mathcal{A}(W, o) = W + \tau W_o - W \circ W_o \tag{9c}$$

$$W_{o,i,j} = [(i, j) \in o] \tag{9d}$$

Equation (9) can be interpreted as augmenting the original acyclicity constraint $h(W) = 0$ to a *stronger* one $h'(W, \mathcal{O}) = 0$. Specifically, we use a series of partial order-augmented acyclicity constraints $h(\mathcal{A}(W, o)) = 0$ for o in the maximal path set of $\mathcal{O}^{-}$ as described in Equation (9b). For each augmented acyclicity, we add the path o to the adjacency matrix W by $\mathcal{A}(W, o)$ as detailed in Equation (9c). Thus, the acyclicity function h with $\mathcal{A}(W, o)$ as input represents a *stronger* acyclicity constraint. The *additional* part of this stronger acyclicity accurately captures adherence to the sequential ordering indicated by o, which can be derived from the following statement.

Augmented Acyclicity for Partial Orders

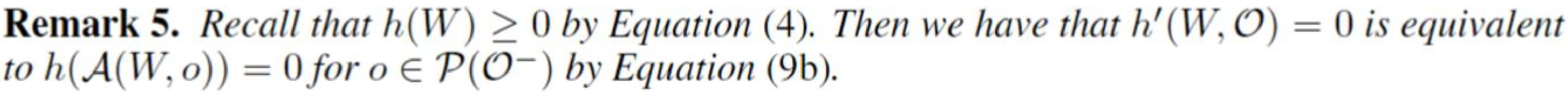


$$\min_{W} \mathcal{F}(W) \quad \text{subject to} \quad h'(W, \mathcal{O}) = 0 \tag{9a}$$

$$h'(W, \mathcal{O}) = \sum_{o \in \mathcal{P}(\mathcal{O}^-)} h(\mathcal{A}(W, o)) \tag{9b}$$

$$\mathcal{A}(W, o) = W + \tau W_o - W \circ W_o \tag{9c}$$

$$W_{o,i,j} = [(i, j) \in o] \tag{9d}$$

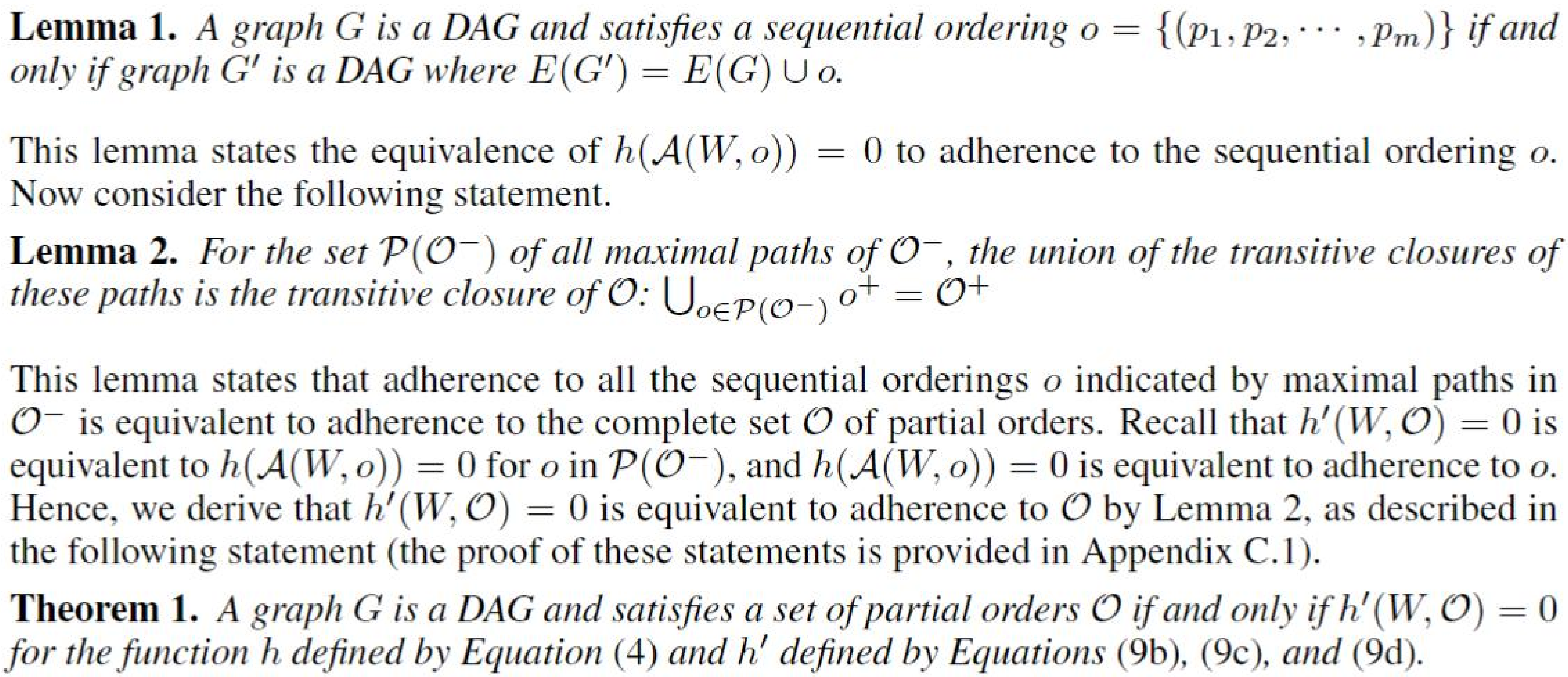


Remark 5. *Recall that $h(W) \geq 0$ by Equation* (4)*. Then we have that $h'(W, \mathcal{O}) = 0$ is equivalent to $h(\mathcal{A}(W, o)) = 0$ for $o \in \mathcal{P}(\mathcal{O}^-)$ by Equation* (9b).

$$h'(W, \mathcal{O}) = 0 \Longleftrightarrow \forall o \in \mathcal{P}(\mathcal{O}^-),\ h(\mathcal{A}(W, o)) = 0$$

For each maximal path in $\mathcal{O}^-$, we constraint that adding the edges from the path to the graph preserves it as a DAG.

Augmented Acyclicity for Partial Orders



Lemma 1. *A graph G is a DAG and satisfies a sequential ordering $o = \{(p_1, p_2, \cdots, p_m)\}$ if and only if graph G' is a DAG where $E(G') = E(G) \cup o$.*

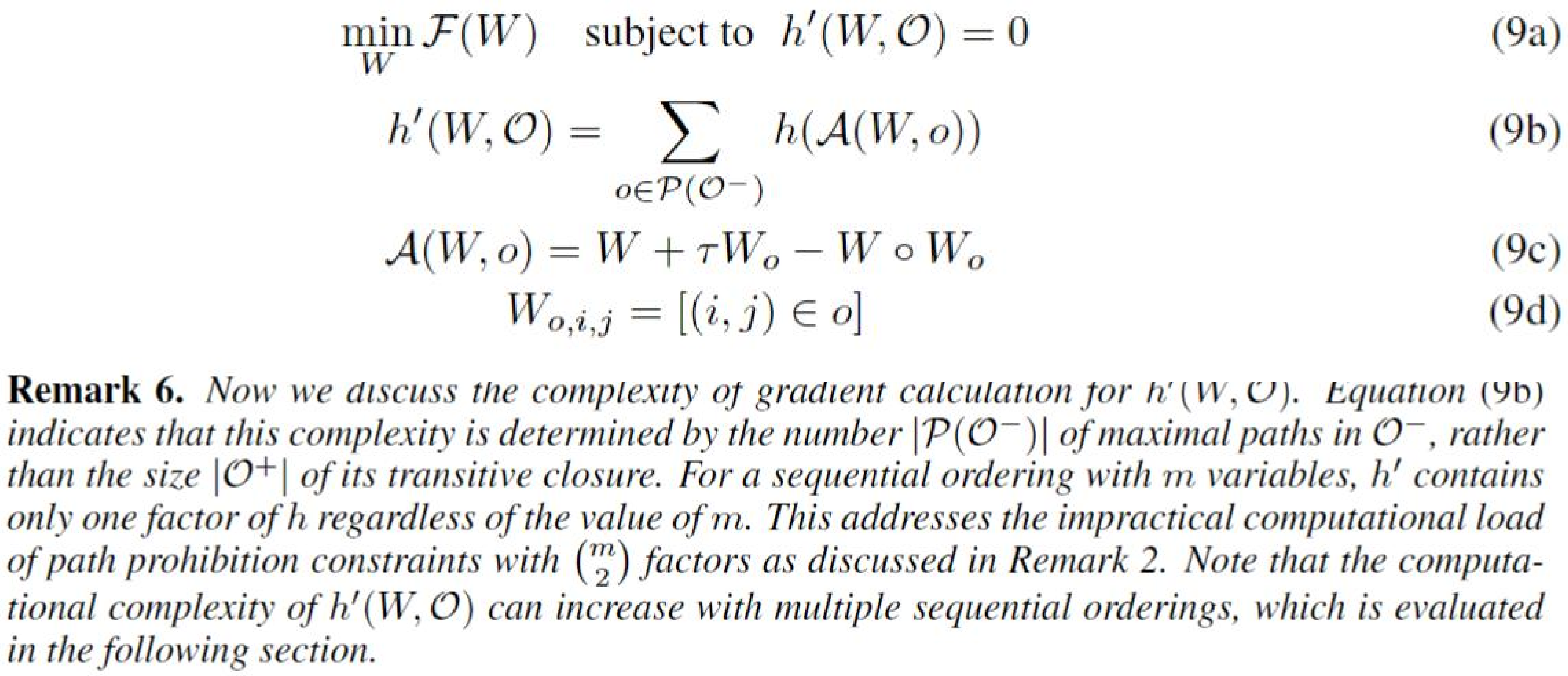
This lemma states the equivalence of $h(\mathcal{A}(W, o)) = 0$ to adherence to the sequential ordering o. Now consider the following statement.

Lemma 2. *For the set $\mathcal{P}(\mathcal{O}^-)$ of all maximal paths of $\mathcal{O}^-$, the union of the transitive closures of these paths is the transitive closure of $\mathcal{O}$: $\bigcup_{o\in\mathcal{P}(\mathcal{O}^-)} o^+ = \mathcal{O}^+$*

This lemma states that adherence to all the sequential orderings o indicated by maximal paths in $\mathcal{O}^-$ is equivalent to adherence to the complete set $\mathcal{O}$ of partial orders. Recall that $h'(W, \mathcal{O}) = 0$ is equivalent to $h(\mathcal{A}(W, o)) = 0$ for o in $\mathcal{P}(\mathcal{O}^-)$, and $h(\mathcal{A}(W, o)) = 0$ is equivalent to adherence to o. Hence, we derive that $h'(W, \mathcal{O}) = 0$ is equivalent to adherence to $\mathcal{O}$ by Lemma 2, as described in the following statement (the proof of these statements is provided in Appendix C.1).

Theorem 1. *A graph G is a DAG and satisfies a set of partial orders $\mathcal{O}$ if and only if $h'(W, \mathcal{O}) = 0$ for the function h defined by Equation* (4) *and h' defined by Equations* (9b), (9c), *and* (9d).

Augmented Acyclicity for Partial Orders



$$\min_{W} \mathcal{F}(W) \quad \text{subject to} \quad h'(W, \mathcal{O}) = 0 \tag{9a}$$

$$h'(W, \mathcal{O}) = \sum_{o \in \mathcal{P}(\mathcal{O}^-)} h(\mathcal{A}(W, o)) \tag{9b}$$

$$\mathcal{A}(W, o) = W + \tau W_o - W \circ W_o \tag{9c}$$

$$W_{o,i,j} = [(i, j) \in o] \tag{9d}$$

Remark 6. *Now we discuss the complexity of gradient calculation for $h'(W, \mathcal{O})$. Equation (9b) indicates that this complexity is determined by the number $|\mathcal{P}(\mathcal{O}^-)|$ of maximal paths in $\mathcal{O}^-$, rather than the size $|\mathcal{O}^+|$ of its transitive closure. For a sequential ordering with m variables, h' contains only one factor of h regardless of the value of m. This addresses the impractical computational load of path prohibition constraints with $\binom{m}{2}$ factors as discussed in Remark 2. Note that the computational complexity of $h'(W, \mathcal{O})$ can increase with multiple sequential orderings, which is evaluated in the following section.*

Outline

- Background
- Preliminary
- Method and Theoretical Analyses
- **Conclusion**

Conclusion

- This paper enhances the field of differentiable structure learning by enabling this framework to apply priors of partial order constraints.

- We systematically analyze the related challenges of applying flexible order constraints and propose a novel and effective strategy to address them by augmenting the acyclicity constraint, with a theoretical proof confirming the correctness and completeness of our strategy.

Thank you