Non-negative Weighted DAG Structure Learning

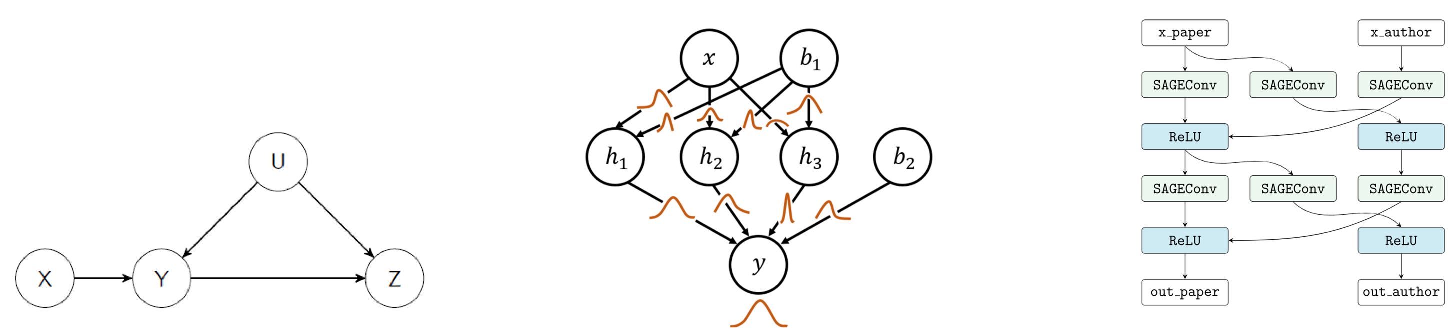
Samuel Rey*, Seyed Saman Saboksayr† and Gonzalo Mateos†

*Dept. of Signal Theory and Communications, Rey Juan Carlos University, Madrid, Spain

†Dept. of Electrical and Computer Eng., University of Rochester, Rochester, USA

Motivation, context, and goal

- **Directed acyclic graphs (DAGs)** have become prominent models in ML applications.
  - ⇒ DAG edges may have causal interpretations [Peters17].
  - ⇒ Conditional independencies exist among variables in Bayesian networks.
- DAGs appear in a gamut of applications: biology, genetics, and finance [Sachs05].
  - ⇒ The structure of the DAG is often unknown or unavailable.



Causal inference | Bayesian networks | Neural networks

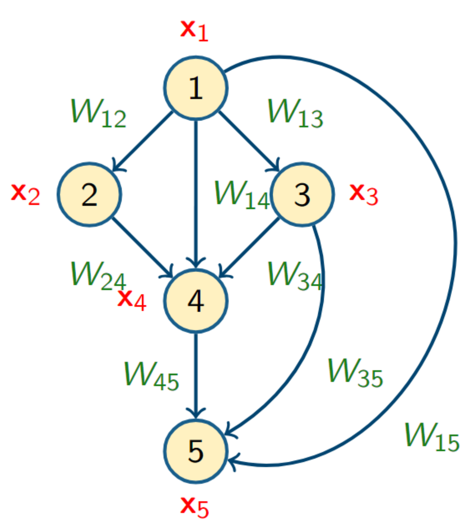
- Learning graphs with cycles from nodal observations is a well-studied problem.
  - ⇒ Imposing acyclicity is a challenge due to its **combinatorial** nature.
  - ⇒ Initial methods based on combinatorial/greedy search faced scalability issues.
  - ⇒ Recent work introduced non-convex continuous acyclicity functions [Zheng18].
- **Contribution**: Learning DAG structure based on a **convex acyclicity function**.
  - ⇒ Recovery guarantees under the simplifying assumption of non-negative weights.

Preliminaries: DAGs and linear SEM

- A DAG $\mathcal{D}=(\mathcal{V},\mathcal{E})$ is a set $\mathcal{V}$ of d nodes and a set of edges $\mathcal{E}$.
  - ⇒ The adjacency matrix $\mathbf{W}\in\mathbb{R}^{d\times d}$ encodes its connectivity.
  - ⇒ The entry $W_{ij}\neq 0$ indicates a directed link $i\to j$.
- Define a graph signal $\mathbf{x}\in\mathbb{R}^d$ whose properties depend on $\mathcal{D}$.
  - ⇒ x_i depends on its parents $PA_i=\{j\in\mathcal{V}: W_{ij}\neq 0\}$.
- Structural equation model (SEM) widely used in causal inference.
  - ⇒ A linear SEM generates the signals $\mathbf{X}\in\mathbb{R}^{d\times n}$ according to

$$\mathbf{X}=\mathbf{W}^\top\mathbf{X}+\mathbf{Z}.$$

  - ⇒ Exogenous input $\mathbf{Z}$ is a random variable with diagonal covariance.

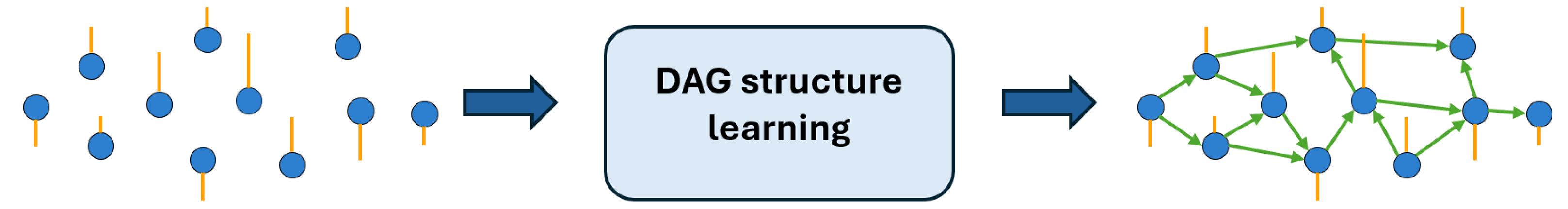


DAG structure learning

- Given data $\mathbf{X}=[\mathbf{x}_1,...,\mathbf{x}_n]\in\mathbb{R}^{d\times n}$, adhering to a linear SEM determined by the DAG $\mathcal{D}$,
  - ⇒ learn the adjacency matrix $\mathbf{W}$ by solving a score-minimization problem.

$$\min_{\mathbf{W}} F(\mathbf{W},\mathbf{X}) \text{ subject to } \mathbf{W}\in\mathbb{D}.$$

  - ⇒ With $F(\mathbf{W},\mathbf{X})$ being a score function of interest, such as least squares.



Challenges

- Learning a DAG solely from observational data $\mathbf{X}$ is NP-hard.
  - ⇒ The combinatorial acyclicity constraint $\mathbf{W}\in\mathbb{D}$ is difficult to enforce.
  - ⇒ The optimization problem may not be identifiable.

Non-convex acyclicity functions

- The pioneering work in [Zheng18] characterizes acyclicity via a **smooth function** $h(\mathbf{W})$.
  - ⇒ **Key**: The zero-level set corresponds to DAGs: $h(\mathbf{W})=0 \iff \mathbf{W}\in\mathbb{D}$.

Continuous acyclicity functions

Allow us to move from a combinatorial search to non-convex continuous optimization.

$$\min_{\mathbf{W}} F(\mathbf{W},\mathbf{X}) \text{ s. to } \mathbf{W}\in\mathbb{D} \iff \min_{\mathbf{W}} F(\mathbf{W},\mathbf{X}) \text{ s. to } h(\mathbf{W})=0$$

- Examples of continuous acyclicity functions include NoTears [Zheng18] and DAGMA [Bello22].

$$h_{\text{notears}}(\mathbf{W})=\operatorname{Tr}\left(e^{\mathbf{W}\circ\mathbf{W}}\right)-d, \qquad h^s_{\text{dagma}}(\mathbf{W})=d\log(s)-\log\det(s\mathbf{I}-\mathbf{W}\circ\mathbf{W}).$$

- **Limitation**: The product $\mathbf{W}\circ\mathbf{W}$ renders the acyclicity functions non-convex.

Learning non-negative DAGs

- **Idea**: assume non-negative weights and harness additional structure to achieve convexity.
  - ⇒ We learn a sparse DAG by minimizing the least squares score function.

$$\hat{\mathbf{W}}=\arg\min_{\mathbf{W}}\left\{\frac{1}{2n}\|\mathbf{X}-\mathbf{W}^\top\mathbf{X}\|_F^2+\alpha\sum_{i,j=1}^{d}W_{ij}\right\} \text{ s. to: } \mathbf{W}\geq 0,\ \ h(\mathbf{W})=0. \tag{1}$$

- We demonstrate that the non-negativity of $\mathbf{W}$ leads to a convex acyclicity function.

Convex acyclicity function

For any matrix $\mathbf{W}\in\mathbb{R}_+^{d\times d}$ whose spectral radius is bounded by $\rho(\mathbf{W})<s\in\mathbb{R}_+$, define

$$h_{ldet}(\mathbf{W}):=d\log(s)-\log\det(s\mathbf{I}-\mathbf{W}), \tag{2}$$

Then, $h_{ldet}(\mathbf{W})\geq 0$ for every $\mathbf{W}$ such that $\rho(\mathbf{W})<s$, and $h_{ldet}(\mathbf{W})=0$ if and only if $\mathbf{W}\in\mathbb{D}$.

- Using the convex acyclicity $h_{ldet}(\mathbf{W})$ in (1) leads to an **abstract convex optimization**.
  - ⇒ Enables finding the global minimum at the expense of additional structure.

DAG learning algorithm

- Estimate the non-negative DAG structure using the method of multipliers.
  - ⇒ Iterative method for constrained optimization with convergence guarantees.
- Let the augmented Lagrangian of (1) be given by

$$L_c(\mathbf{W},\lambda)=\frac{1}{2n}\|\mathbf{X}-\mathbf{W}^\top\mathbf{X}\|_F^2+\alpha\sum_{i,j=1}^{d}W_{ij}+\lambda h(\mathbf{W})+\frac{c}{2}h(\mathbf{W})^2.$$

Method of multipliers for non-negative DAG learning

Perform the following sequence of steps with positive constants $0<\gamma<1$ and $\beta>1$

Step 1. Update the adjacency matrix $\mathbf{W}^{(k+1)}=\arg\min_{\mathbf{W}\geq 0} L_{c^{(k)}}(\mathbf{W},\lambda^{(k)})$.

Step 2. Update the Lagrange multiplier $\lambda^{(k+1)}=\lambda^{(k)}+c^{(k)}h(\mathbf{W}^{(k+1)})$.

Step 3. Update the penalty parameter $c^{(k+1)}=\begin{cases}\beta c^{(k)} & \text{if } h(\mathbf{W}^{(k+1)})>\gamma h(\mathbf{W}^{(k)})\\ c^{(k)} & \text{otherwise.}\end{cases}$

- Convergence to the global optimum of (1) due to the convexity of $L_c(\mathbf{W},\lambda)$.
  - ⇒ Optimization problem in Step 1 solved via gradient descent.

Recovering the true DAG structure

- Our proposed algorithm recovers the true DAG structure $\mathbf{W}^*$ in the infinite sample regime.
  - ⇒ Assume the distribution of $\mathbf{x}$ is known and consider the following score function.

$$\bar{F}(\mathbf{W},\mathbf{x})=\mathbb{E}_{\mathbf{x}}\left[\left\|\boldsymbol{\Sigma}_{\mathbf{z}}^{-\frac{1}{2}}\left(\mathbf{I}-\mathbf{W}^\top\right)\mathbf{x}\right\|_2^2\right].$$

Theorem

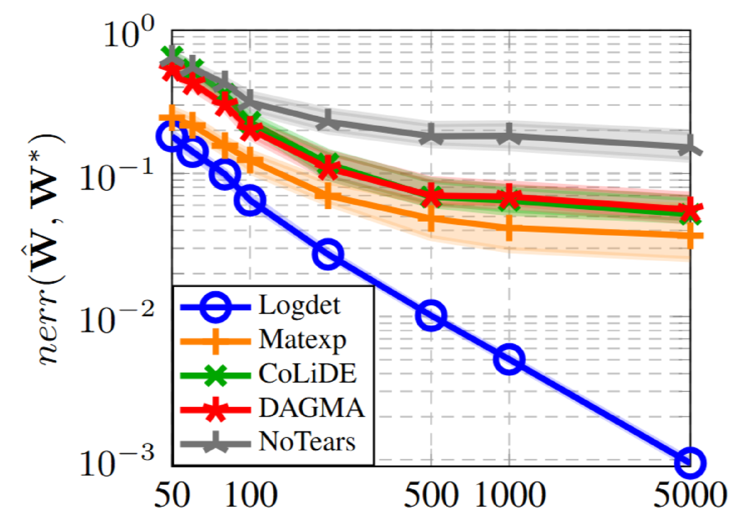
Let $\mathbf{x}\in\mathbb{R}^d$ be a random vector following a linear SEM with non-negative DAG $\mathbf{W}^*\geq 0$ and exogenous input $\mathbf{z}$ with covariance $\boldsymbol{\Sigma}_{\mathbf{z}}$ known up to a scaling factor. Then, the estimate $\hat{\mathbf{W}}$ from solving

$$\min_{\mathbf{W}} \bar{F}(\mathbf{W},\mathbf{x}) \quad \text{s.to} \quad \mathbf{W}\geq 0,\ h_{ldet}(\mathbf{W})=0,$$

with the iterates from Step 1 to Step 3, satisfies $\hat{\mathbf{W}}=\mathbf{W}^*$.

Test case I - Number of samples

- Non-negative ER graphs with $d=100$ nodes and average degree 4.
  - ⇒ Signals sampled from linear SEM with $\mathbf{z}\sim\mathcal{N}(\mathbf{0},\sigma\mathbf{I})$.



(a) Number of samples n

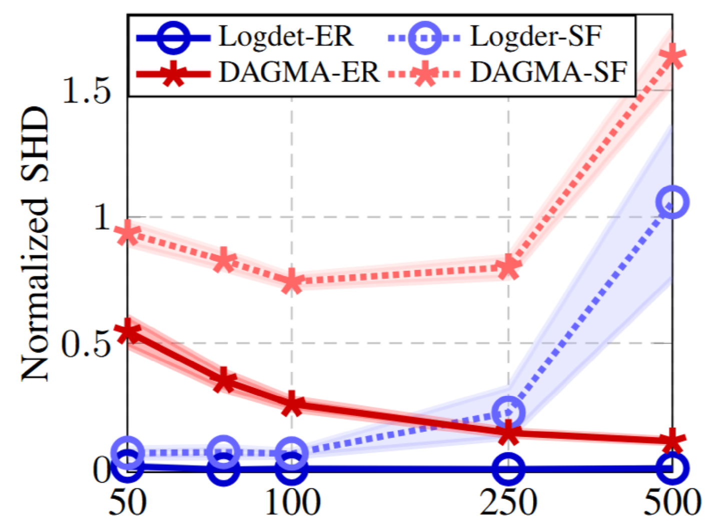
- We report the normalized error

$$nerr(\hat{\mathbf{W}},\mathbf{W}^*)=\|\mathbf{W}^*-\hat{\mathbf{W}}\|_F^2/\|\mathbf{W}^*\|_F^2.$$

- Convex acyclicity constraints outperform alternatives.
- Error of convex method goes to 0 as the number of samples grows.
  - ⇒ Aligned with theoretical result.

Test case II - Number of ndoes

- We sample $n=1000$ signals and consider ER and SF graphs.



(b) Number of nodes d

- We report the normalized Structural Hamming Distance (SHD).
- Convex logdet constraint consistently outperforms the alternative.
- Convex constraint achieves a SHD of 0 even with moderately large ER graphs.
  - ⇒ Recovers the true support even in the small-sample regime.

References

Peters17 J. Peters, D. Janzing, and B. Scholkopf. "Elements of Causal Inference: Foundations and Learning Algorithms". MIT Press, 2017.

Zheng18 X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. "DAGs with NO TEARS: Continuous optimization for structure learning". Neurips, 2018.

Sachs05 K. Sachs, O. Perez, D. Pe'er, D. A Lauffenburger, and G. P. Nolan. "Causal protein-signaling networks derived from multiparameter singlecell data". Science, 2005.

Bello22 K. Bello, B. Aragam, and P. Ravikumar. "DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization". Neurips, 2022.