ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888302

Non-negative Weighted DAG Structure Learning

Samuel Rey*, Seyed Saman Saboksayr†, Gonzalo Mateos†,
*Dept. of Signal Theory and Communications, Rey Juan Carlos University, Madrid, Spain
†Dept. of Electrical and Computer Eng., University of Rochester, Rochester, NY, United States

***Abstract*—We address the problem of learning the topology of directed acyclic graphs (DAGs) from nodal observations, which adhere to a linear structural equation model. Recent advances framed the combinatorial DAG structure learning task as a continuous optimization problem, yet existing methods must contend with the complexities of non-convex optimization. To overcome this limitation, we assume that the latent DAG contains only non-negative edge weights. Leveraging this additional structure, we argue that cycles can be effectively characterized (and prevented) using a convex acyclicity function based on the log-determinant of the adjacency matrix. This convexity allows us to relax the task of learning the non-negative weighted DAG as an abstract convex optimization problem. We propose a DAG recovery algorithm based on the method of multipliers, that is guaranteed to return a global minimizer. Furthermore, we prove that in the infinite sample size regime, the convexity of our approach ensures the recovery of the true DAG structure. We empirically validate the performance of our algorithm in several reproducible synthetic-data test cases, showing that it outperforms state-of-the-art alternatives.**

***Index Terms*—DAG learning, network topology inference, causal discovery, graph signal processing, convex relaxation.**

I. Introduction

Directed acyclic graphs (DAGs) are crucial tools for modeling complex systems where directionality plays a key role [1], and they are widely recognized for their ability to represent causal relationships [2], [3]. Consequently, DAGs and associated Bayesian networks have become increasingly common tools in biology [4], [5], genetics [6], machine learning [7]–[9], signal processing [3], [10], and causal inference [11], [12]. Despite their widespread adoption, often the DAG structure is not known in advance and must be inferred from data.

Learning a graph from nodal observations is a prominent problem rooted in the relation between the properties of the observed data and the graph topology [13]. Noteworthy approaches for undirected graphs include Gaussian graphical models [14]–[16], smoothness models [17]–[19], or graph stationary models [20]–[24], among others. When the graph of interest is a DAG, structural equation models (SEMs) are often the method of choice [25]–[28]. Accounting for the acyclicity of the graph renders the DAG structure learning a challenging combinatorial, in fact NP-hard, endeavor [29], [30].

Recent works managed to circumvent the combinatorial nature of DAG structure learning by introducing a continuous relaxation that allows for efficient exploration of the DAG space. A breakthrough in [25] advocated a continuous non-convex acyclicity constraint based on the matrix exponential. This inspired further developments, including acyclicity functions based on powers of the adjacency matrix [26], [31] and the log determinant [27], [28], [32]. Despite significant progress, learning the DAG structure remains a challenging task involving a non-convex optimization problem. Consequently, current methods rely on heuristics [27] and are content with estimates corresponding to local minima; see also [33].

Contributions. To circumvent these non-convexity issues, we focus on the class of DAGs with non-negative edge weights and propose a *convex acyclicity function* that enables recovering the global minimizer. To the best of our knowledge, this is the first work proposing a convex relaxation for DAG estimation. We contribute the following technical innovations:

- By leveraging the non-negativity of the DAG edge weights, we propose a convex log-determinant function to characterize the acyclicity of the graph.
- We cast the DAG learning task as an abstract convex optimization problem and propose an iterative algorithm based on the method of multipliers. This approach warrants the recovery of the global minimum.
- We prove that the proposed method recovers the true DAG structure when infinite observations are available.

II. Fundamentals of DAG structure learning

Let $\mathcal{D} = (\mathcal{V}, \mathcal{E})$ denote a DAG, where $\mathcal{V}$ is a set of d nodes, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of directed edges. An edge $(i, j) \in \mathcal{E}$ exists if and only if there is a directed link from node i to node j. The connectivity of the DAG is captured by the weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$, where $W_{ij} \neq 0$ if and only if $(i, j) \in \mathcal{E}$. Then, a graph signal is defined on the nodes of the DAG and is represented as a vector $\mathbf{x} \in \mathbb{R}^d$, with x_i denoting the signal value at node i.

The task of DAG structure learning involves inferring the topology of a DAG from a set of nodal observations. Collecting the n observed signals in the matrix $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, we suppose $\mathbf{X}$ adheres to a linear SEM given by

$$\mathbf{X} = \mathbf{W}^\top \mathbf{X} + \mathbf{Z}, \tag{1}$$

where $\mathbf{Z} \in \mathbb{R}^{d\times n}$ collects zero-mean exogenous noises whose columns are i.i.d. random vectors with covariance matrix $\mathbf{\Sigma_z} = \sigma^2\mathbf{I}$. Mutual independence of the noise variables is crucial [2, pp. 83-84].

This work was supported in part by the Spanish AEI Grants PID2022-136887NB-I00, TED2021-130347B-I00, PID2023-149457OB-I00, and the Community of Madrid (F1180-AdvGSP4BIO Doctores emergentes, and Madrid ELLIS Unit). Emails: samuel.rey.escudero@urjc.es, ssaboksa@ur.rochester.edu, gmateosb@ur.rochester.edu,

With the previous definitions in place, the DAG structure encoded in $\mathbf{W}$ can be inferred from $\mathbf{X}$ by solving the optimization problem

$$\min_{\mathbf{W}} F(\mathbf{W}, \mathbf{X}) \quad \text{s.to} \quad \mathbf{W} \in \mathbb{D}, \tag{2}$$

where $F(\mathbf{W}, \mathbf{X})$ denotes a data-dependent score function that captures the relation between $\mathbf{W}$ and the signals, and $\mathbb{D}$ denotes the set of adjacency matrices corresponding to a DAG.

The optimization problem in (2) is challenging to solve due to the non-convex and combinatorial nature of the constraint $\mathbf{W} \in \mathbb{D}$. Recent advances advocate recovery of DAG structure by replacing this constraint with an *acyclicity condition* of the form $h(\mathbf{W}) = 0$, where $h : \mathbb{R}^{d\times d} \mapsto \mathbb{R}$ is a differentiable function whose zero level set corresponds to $\mathbb{D}$. This approach was pioneered in [25] via the acyclicity function

$$h_{\text{notears}}(\mathbf{W}) = \text{tr}\left(e^{\mathbf{W}\circ\mathbf{W}}\right) - d, \tag{3}$$

where $\circ$ denotes the Hadamard (entry-wise) product. More recently, [27] proposed an alternative acyclicity characterization

$$h^s_{\text{dagma}}(\mathbf{W}) = d\log(s) - \log\det\left(s\mathbf{I} - \mathbf{W}\circ\mathbf{W}\right), \tag{4}$$

with $s \in \mathbb{R}_+$. The log-determinant function alleviates numerical issues and has been shown to outperform prior relaxations.

Replacing the combinatorial constraint $\mathbf{W} \in \mathbb{D}$ with the acyclicity condition $h(\mathbf{W}) = 0$ constitutes a significant advancement, enabling the use of standard continuous optimization methods to learn the DAG structure. However, the presence of the term $\mathbf{W} \circ \mathbf{W}$ renders these functions non-convex, still posing important challenges to recovering the true DAG. To overcome this limitation, we henceforth assume $\mathbf{W}$ *has non-negative weights* and propose: (i) a convex alternative to (4); and (ii) a method to obtain the global minimizer of an optimization problem equivalent to (2).

III. Non-negative DAG structure learning

We tackle the problem of learning the DAG structure by assuming that the entries of $\mathbf{W}$ are non-negative. The restriction is still relevant to binary DAGs and other pragmatic settings; see e.g., [3]. This additional structure is crucial for simplifying the optimization problem in (2), allowing us to characterize the acyclicity of the graph using a convex function.

When $\mathbf{X}$ adheres to a linear SEM, a common choice for the score function is $F(\mathbf{W}, \mathbf{X}) = \frac{1}{2n}\|\mathbf{X} - \mathbf{W}^\top\mathbf{X}\|_F^2 + \alpha\|\mathbf{W}\|_1$, which balances a data-fidelity term with ℓ_1 norm regularization to encourage sparse solutions. This trade-off is controlled by the tunable weight $\alpha \in \mathbb{R}_+$. Using this convex score function, a continuous acyclicity constraint, and the entrywise non-negativity of $\mathbf{W}$, the DAG structure learning problem can be alternatively formulated as

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \left\{ \frac{1}{2n}\|\mathbf{X} - \mathbf{W}^\top\mathbf{X}\|_F^2 + \alpha \sum_{i,j=1}^{d} W_{ij} \right\}$$
$$\text{s.to} \qquad \mathbf{W} \geq 0, \;\; h(\mathbf{W}) = 0, \tag{5}$$

where $h(\mathbf{W})$ denotes an acyclicity function of interest, and the ℓ_1 norm is replaced by $\sum_{i,j=1}^{d} W_{ij}$ due to the non-negativity of $\mathbf{W}$. The acyclicity constraint renders the optimization problem in (5) non-convex, an issue that is dealt with next.

A. Convex acyclicity functions via non-negative matrices

Relying on a smooth acyclicity constraint to ensure that $\mathbf{W}$ is cycle-free is central to modern DAG learning methods. As discussed in [25] an effective acyclicity function $h(\mathbf{W})$ should be smooth, have an easy to compute gradient, and satisfy $h(\mathbf{W}) = 0$ if and only if $\mathbf{W} \in \mathbb{D}$. Harnessing the non-negativity of $\mathbf{W}$ and inspired by (4), we introduce a *convex* acyclicity function that meets these criteria.

Proposition 1. *For any matrix $\mathbf{W} \in \mathbb{R}_+^{d\times d}$ whose spectral radius is bounded by $\rho(\mathbf{W}) < s$ with $s \in \mathbb{R}_+$, define*

$$h_{ldet}(\mathbf{W}) := d\log(s) - \log\det(s\mathbf{I} - \mathbf{W}), \tag{6}$$

with gradient $\nabla h_{ldet}(\mathbf{W}) = (s\mathbf{I} - \mathbf{W})^{-\top}$. 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 $\mathbf{W} \in \mathbb{D}$.

Proof. We start by rewriting h_{ldet} in the equivalent form

$$\begin{aligned} h_{ldet}(\mathbf{W}) &= d\log(s) - \log(s^d) - \log\det(\mathbf{I} - s^{-1}\mathbf{W}) \\ &= -\log\det(\mathbf{I} - s^{-1}\mathbf{W}). \end{aligned}$$

From the bounded spectral radius $\rho(\mathbf{W}) < s$, it follows that $\log\det(\mathbf{I} - s^{-1}\mathbf{W})$ is well defined for every $\mathbf{W}$, so for the rest of the proof we set $s = 1$ without loss of generality.

First, we show that $h_{ldet}(\mathbf{W}) \geq 0$ for every non-negative $\mathbf{W}$, and then establish $h_{ldet}(\mathbf{W}) = 0 \Leftrightarrow \mathbf{W} \in \mathbb{D}$. To that end, we start by showing that $\log\det(\mathbf{I} - \mathbf{W}) \leq 0$. With $\lambda_i(\mathbf{B})$ denoting the i-th eigenvalue of some matrix $\mathbf{B}$, we have

$$\begin{aligned} \log\det(\mathbf{I}-\mathbf{W}) &= \sum_{i=1}^{d} \log\left(\lambda_i(\mathbf{I}-\mathbf{W})\right) = d\sum_{i=1}^{d} \frac{\log\left(\lambda_i(\mathbf{I}-\mathbf{W})\right)}{d} \\ &\leq d\log\left(\sum_{i=1}^{d} \frac{\lambda_i(\mathbf{I}-\mathbf{W})}{d}\right) = d\log\left(\frac{\text{tr}(\mathbf{I}-\mathbf{W})}{d}\right), \end{aligned}$$

where the inequality follows from Jensen's inequality for concave functions.

Next, we leverage that $\mathbf{W} \geq 0$, and hence, $\text{tr}(\mathbf{W}) \geq 0$, to obtain the bound $\text{tr}(\mathbf{I} - \mathbf{W}) = \text{tr}(\mathbf{I}) - \text{tr}(\mathbf{W}) \leq \text{tr}(\mathbf{I}) = d$. Combining this with the monotonicity of the logarithm renders

$$\log\det(\mathbf{I} - \mathbf{W}) \leq d\log\left(\frac{\text{tr}(\mathbf{I}-\mathbf{W})}{d}\right) \leq d\log\left(\frac{\text{tr}(\mathbf{I})}{d}\right) = 0.$$

Therefore, $h_{ldet}(\mathbf{W}) = -\log\det(\mathbf{I} - \mathbf{W}) \geq 0$, as intended.

Finally, note that $h_{ldet}(\mathbf{W}) = -\log\det(\mathbf{I}-\mathbf{W}) = 0$ if and only if all the eigenvalues of $\mathbf{W}$ are zero, which means that $\mathbf{W}$ is a nilpotent matrix, equivalent to $\mathbf{W}$ being a DAG [27]. □

Ensuring the acyclicity of $\mathbf{W}$ using a convex function such as h_{ldet} offers significant advantages. First, the convexity of h_{ldet} renders (5) an *abstract convex optimization problem [34]*, enabling us to reliably recover the global minimum. To see this, note that under the conditions of Proposition 1, the feasible set defined by $h_{ldet}(\mathbf{W}) = 0$ is a convex set and, in fact, is equivalent to the feasible set of the convex constraint

$h_{ldet}(\mathbf{W}) \leq 0$. In turn, recovering the global minimum provides new opportunities to characterize the estimate $\hat{\mathbf{W}}$, a promising research direction that we start pursuing next, leaving a more in-depth analysis as future work. Moreover, unlike the acyclicity functions in (3) and (4), h_{ldet} does not have stationary points at DAGs, meaning $\nabla h_{ldet}(\mathbf{W}^*) \neq 0$ for $\mathbf{W}^* \in \mathbb{D}$. This avoids algorithmic issues highlighted in [26].

Similar to Proposition 1, when $\mathbf{W}$ is non-negative, a convex alternative to (3) is given by

$$h_{\text{mexp}}(\mathbf{W}) = \text{tr}\left(e^{\mathbf{W}}\right) - d, \tag{7}$$

where $h_{\text{mexp}}(\mathbf{W}) = 0$ if and only if $\mathbf{W} \in \mathbb{D}$ [26]. While (7) is also a convex function, acyclicity constraints based on the log-determinant have demonstrated superior performance [27], [28]. Indeed, we further examine how different acyclicity functions impact the DAG recovery in Section IV.

B. DAG structure learning via method of multipliers

Considering a convex acyclicity function $h(\mathbf{W})$, we solve the optimization problem in (5) using the method of multipliers [35, Ch. 4.2], an iterative approach based on the augmented Lagrangian tailored to constrained optimization problems.

Let the augmented Lagrangian of (5) be given by

$$L_c(\mathbf{W}, \lambda) = \frac{1}{2n}\|\mathbf{X} - \mathbf{W}^\top \mathbf{X}\|_F^2 + \alpha \sum_{i,j=1}^{d} W_{ij} + \lambda h(\mathbf{W}) + \frac{c}{2} h(\mathbf{W})^2, \tag{8}$$

where $\lambda \in \mathbb{R}_+$ is the Lagrange multiplier, and $c \in \mathbb{R}_+$ is a penalty parameter. Note that the term $h(\mathbf{W})^2$ is convex since it is a composition of a convex function and a convex and non-decreasing function [34], and hence, the augmented Lagrangian is convex. Note that no term is related to the constraint $\mathbf{W} \geq 0$ since it can be enforced through a simple projection. Then, at each iteration $k = 1, \ldots, \kappa_{max}$, we perform the following sequence of steps.

Step 1. We update $\mathbf{W}^{(k+1)}$ by minimizing

$$\mathbf{W}^{(k+1)} = \arg\min_{\mathbf{W} \geq 0} L_{c^{(k)}}(\mathbf{W}, \lambda^{(k)}). \tag{9}$$

Thanks to the convexity of $L_{c^{(k)}}$, we can recover the global minimum $\mathbf{W}^{(k+1)}$ with standard convex optimization methods such as projected gradient descent.

Step 2. The update of the Langrange multiplier $\lambda^{(k+1)}$ depends on the degree of constraint violation, given by

$$\lambda^{(k+1)} = \lambda^{(k)} + c^{(k)} h(\mathbf{W}^{(k+1)}). \tag{10}$$

This update can also be interpreted as a gradient ascent step since the constraint violation corresponds to the gradient of $L_{c^{(k)}}(\mathbf{W}^{(k+1)}, \lambda)$ with respect to λ.

Step 3. The penalty parameter $c^{(k)}$ needs to be progressively increased so the constraint $h(\mathbf{W}) = 0$ is satisfied as $\kappa_{max} \to \infty$. A typical update scheme is given by

$$c^{(k+1)} = \begin{cases} \beta c^{(k)} & \text{if } \; h(\mathbf{W}^{(k+1)}) > \gamma h(\mathbf{W}^{(k)}) \\ c^{(k)} & \text{otherwise,} \end{cases} \tag{11}$$

where $0 < \gamma < 1$ and $\beta > 1$ are positive constants. Intuitively, $c^{(k)}$ is increased only if the constraint violation is not decreased by a factor of γ.

When the sequence of iterations is completed, the estimated DAG structure is given by $\hat{\mathbf{W}} = \mathbf{W}^{(\kappa_{max})}$. The convexity of the augmented Lagrangian guarantees that $\mathbf{W}^{(k)}$ corresponds to the global minimum of (9) for every k. Therefore, [35, Prop. 4.2.1] guarantees that every limit point of the sequence $\mathbf{W}^{(k)}$ is a global minimum of the constrained problem in (5).

Finally, we demonstrate that the above algorithm can recover the true DAG structure $\mathbf{W}^*$ in the infinite sample regime, i.e., when the distribution of the random vector $\mathbf{x}$ is known. To that end, replace the score function in (5) with

$$\bar{F}(\mathbf{W}, \mathbf{x}) = \mathbb{E}_{\mathbf{x}}\left[\left\|\boldsymbol{\Sigma}_{\mathbf{z}}^{-\frac{1}{2}}\left(\mathbf{I} - \mathbf{W}^\top\right)\mathbf{x}\right\|_2^2\right]. \tag{12}$$

Then, the next theorem guarantees the recovery of $\mathbf{W}^*$.

Theorem 1. *Consider the score function $\bar{F}(\mathbf{W}, \mathbf{x})$ in* (12) *and the convex acyclicity function $h_{ldet}(\mathbf{W})$ from* (6)*. Let $\mathbf{x}$ 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*

$$\min_{\mathbf{W}} \; \bar{F}(\mathbf{W}, \mathbf{x}) \quad \text{s.to} \quad \mathbf{W} \geq 0, \; h_{ldet}(\mathbf{W}) = 0, \tag{13}$$

with the iterates (9)-(11)*, satisfies $\hat{\mathbf{W}} = \mathbf{W}^*$.*

Proof. Given the convexity of the optimization problem (9) and the updates in (10) and (11), from [35, Prop. 4.2.1] it follows that $\hat{\mathbf{W}}$ is the global minimizer of (5). Then, from Proposition 1, we have that $h_{ldet}(\mathbf{W}) = 0$ if and only if $\mathbf{W} \in \mathbb{D}$. Consequently, minimizing (13) is equivalent to solving

$$\tilde{\mathbf{W}} \; = \; \arg\min_{\mathbf{W}} \; \bar{F}(\mathbf{W}, \mathbf{x}) \quad \text{s.to} \quad \mathbf{W} \geq 0, \; \mathbf{W} \in \mathbb{D}, \tag{14}$$

so $\hat{\mathbf{W}} = \tilde{\mathbf{W}}$.

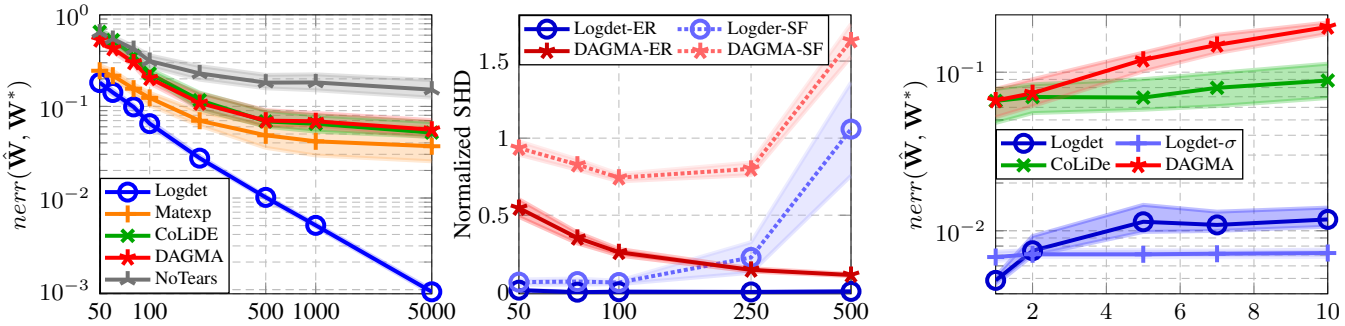
Finally, from [36, Thm. 7] it follows that the global minimizer of (14), $\tilde{\mathbf{W}}$, corresponds to the true DAG structure, $\mathbf{W}^*$. Therefore, the proof is concluded since $\hat{\mathbf{W}} = \tilde{\mathbf{W}} = \mathbf{W}^*$. □

Theorem 1 guarantees that our proposed method recovers the true DAG $\mathbf{W}^*$ when the distribution of $\mathbf{x}$ is known, which is tantamount to assuming access to infinite observations n. While such an assumption is unlikely to be satisfied in practical settings, this preliminary recoverability guarantee brings to light the potential benefits of relying on convex acyclicity functions to learn the DAG structure. In the following section, we complete the assessment of our method by empirically evaluating its performance in the finite sample regime.

IV. Numerical experiments

We now evaluate the performance of the proposed method across different scenarios and compare it with state-of-the-art alternatives. The code with the proposed method and all implementation details is publicly available on GitHub[1].

[1]https://github.com/reysam93/cvx_dag_learning



(a) Number of samples n (b) Number of nodes d (c) Exogenous noise variance σ^2

Fig. 1: Evaluation of the proposed DAG structure learning method across different scenarios. a) reports the error of $\hat{\mathbf{W}}$ as the number of samples increases. b) presents the normalized SHD between $\hat{\mathbf{W}}$ and the true DAG structure as the number of nodes increases. c) illustrates the error of $\hat{\mathbf{W}}$ for different values of the variance of the exogenous input $\mathbf{Z}$.

We measure the performance in terms of the normalized Frobenius error, calculated as

$$nerr(\hat{\mathbf{W}}, \mathbf{W}^*) = \|\mathbf{W}^* - \hat{\mathbf{W}}\|_F^2 / \|\mathbf{W}^*\|_F^2, \qquad (15)$$

and the structural Hamming distance (SHD) normalized by the number of nodes [25, App. D.2], which counts the number of edges in $\hat{\mathbf{W}}$ that need to be changed to match the support of $\mathbf{W}^*$. As baselines, we consider the following relevant non-convex approaches: NO TEARS [25], DAGMA [27], and CoLiDE [28]. In addition, we consider the convex constraints (6) and (7), respectively denoted as "Logdet" and "Matexp", in Fig. 1. Unless otherwise stated, we simulate Erdős-Rényi (ER) DAGs with $d = 100$ nodes and an average degree of 4 by sampling triangular adjacency matrices and permuting their rows and columns. We create 1000 samples following a linear SEM with $\mathbf{z}$ sampled from a standard Gaussian distribution. We report the median and the 25th and 75th percentiles of 100 independent realizations.

Test case 1. The first experiment examines the error $nerr(\hat{\mathbf{W}}, \mathbf{W}^*)$ of different methods as the number of samples n increases, as indicated on the x-axis. From the results in Fig. 1 (a), it is evident that leveraging a convex acyclicity constraint consistently leads to superior performance. Moreover, while the error for the considered baselines saturates, the error associated with our convex constraint "Logdet" approaches 0 as the number of samples increases. This behavior aligns with Theorem 1, showcasing the potential of our convex method for recovering the true DAG structure. Additionally, although both "Matexp" and "Logdet" use convex constraints, the latter outperforms the former. This highlights the benefits of using the log determinant to ensure acyclicity, which coincides with the conclusions drawn for non-convex approaches [27].

Test case 2. Next, Fig. 1 (b) depicts the normalized SHD as the number of nodes d increases, with the number of samples fixed at $n = 1000$. This experiment considers both ER graphs and scale-free (SF) graphs. Our results demonstrate that "Logdet" consistently outperforms "DAGMA". Since both methods utilize an acyclicity constraint based on the log determinant [cf. (4) and (6)], the difference in performance underscores the advantages of exploiting the non-negativity and convexity of $h_{ldet}(\mathbf{W})$. Specifically, for ER graphs, our method achieves a normalized SHD of zero, showcasing that it accurately recovers the support of the true DAG even in the small-sample regime. Regarding the SF graphs, the performance of the non-convex "DAGMA" method deteriorates significantly more than that of "Logdet" as the number of nodes increases. Overall, the experiment also suggests that estimating non-negative SF DAGs is more challenging than estimating ER.

Test case 3. To conclude the numerical evaluation, we assume the covariance of the exogenous input is given by $\mathbf{\Sigma_z} = \sigma^2\mathbf{I}$, and evaluate the performance as σ^2 increases. We also include a variant of (5), which leverages the knowledge of $\mathbf{\Sigma_z}$, referred to as "Logdet-σ". Consistent with previous results, Fig. 1 (c) demonstrates that our proposed method based on the convex log determinant consistently outperforms the non-convex approaches. Moreover, while the performance of "DAGMA" and "Logdet" deteriorates for larger values of σ^2, the error of "CoLiDE" and "Logdet-σ" remains stable, highlighting the advantages of leveraging information about $\mathbf{\Sigma_z}$ or, if unavailable, estimating it as in "CoLiDE" [28].

V. Conclusion

This paper studied the prominent task of learning the structure of a DAG from a set of nodal observations using a linear SEM. Leveraging the assumed non-negativity of $\mathbf{W}$, we framed DAG learning as a continuous optimization problem and introduced the first method that guarantees the estimated DAG corresponds to a global minimizer. Specifically, we demonstrated that a convex constraint based on the log determinant ensures acyclicity when the graph does not contain negative edges. The convexity of the acyclicity function enables us to formulate an abstract convex optimization problem, which we solve with an iterative algorithm based on the method of multipliers, recovering the global minimum. After establishing preliminary recovery guarantees by demonstrating that our method recovers the ground truth DAG in the infinite sample size regime, we validated its performance through reproducible numerical experiments on synthetic data, where it outperformed state-of-the-art methods.

REFERENCES

[1] A. G. Marques, S. Segarra, and G. Mateos, "Signal processing on directed graphs: The role of edge directionality when processing and learning from network data," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 99–116, 2020.

[2] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press, 2017.

[3] B. Seifert, C. Wendler, and M. Püschel, "Causal Fourier analysis on directed acyclic graphs and posets," *IEEE Trans. Signal Process.*, vol. 71, pp. 3805–3820, 2023.

[4] K. Sachs, O. Perez, D. Pe'er, D. A Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[5] P. J. F. Lucas, L. C. Van der Gaag, and A. Abu-Hanna, "Bayesian networks in biomedicine and health-care," *Artif Intell Med*, vol. 30, no. 3, pp. 201–214, 2004.

[6] B. Zhang, C. Gaiteri, L.-G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, et al., "Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease," *Cell*, vol. 153, no. 3, pp. 707–720, 2013.

[7] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT press, 2009.

[8] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," in *Intl. Conf. Mach. Learn. (ICML)*. PMLR, 2019, pp. 7154–7163.

[9] S. Rey, H. Ajorlou, and G. Mateos, "Convolutional learning on directed acyclic graphs," *arXiv preprint arXiv:2405.03056*, 2024.

[10] P. Misiakos, V. Mihal, and M. Püschel, "Learning signals and graphs from time-series graph data with few causes," in *IEEE Intl. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2024, pp. 9681–9685.

[11] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, MIT Press, 2001.

[12] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Trans. Knowl. Discovery Data*, vol. 15, no. 5, pp. 1–46, 2021.

[13] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, 2019.

[14] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[15] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE J. Select. Topics Signal Process.*, vol. 11, no. 6, pp. 825–841, 2017.

[16] S. Rey, T. M. Roddenberry, S. Segarra, and A. G. Marques, "Enhanced graph-learning schemes driven by similar distributions of motifs," *IEEE Trans. Signal Process.*, vol. 71, pp. 3014–3027, 2023.

[17] V. Kalofolias, "How to learn a graph from smooth signals," in *Intl. Conf. Artif. Intel. Statist. (AISTATS)*, 2016, pp. 920–929.

[18] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, 2016.

[19] S. S. Saboksayr and G. Mateos, "Accelerated graph learning from smooth signals," *IEEE Signal Process. Lett.*, vol. 28, pp. 2192–2196, 2021.

[20] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 467–483, Sep. 2017.

[21] R. Shafipour and G. Mateos, "Online topology inference from streaming stationary graph signals with partial connectivity information," *Algorithms*, vol. 13, no. 9, pp. 228, 2020.

[22] T. M. Roddenberry, M. Navarro, and S. Segarra, "Network topology inference with graphon spectral penalties," in *IEEE Intl. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2021, pp. 5390–5394.

[23] A. Buciulea, S. Rey, and A. G. Marques, "Learning graphs from smooth and graph-stationary signals with hidden variables," *IEEE Trans. Signal, Inform. Process. Networks*, vol. 8, pp. 273–287, 2022.

[24] M. Navarro, S. Rey, A. Buciulea, A. G. Marques, and S. Segarra, "Joint network topology inference in the presence of hidden nodes," *IEEE Trans. Signal Process.*, vol. 72, pp. 2710–2725, 2024.

[25] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," *Advances Neural Inf. Process. Syst.*, vol. 31, 2018.

[26] D. Wei, T. Gao, and Y. Yu, "DAGs with No Fears: A closer look at continuous optimization for learning Bayesian networks," *Advances Neural Inf. Process. Syst.*, vol. 33, pp. 3895–3906, 2020.

[27] K. Bello, B. Aragam, and P. Ravikumar, "DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization," *Advances Neural Inf. Process. Syst.*, vol. 35, pp. 8226–8239, 2022.

[28] S. S. Saboksayr, G. Mateos, and M. Tepper, "CoLiDE: Concomitant linear DAG estimation," *Intl. Conf. Learn. Repr. (ICLR)*, 2024.

[29] D. M. Chickering and D. Heckerman, "Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables," *Mach. Learn.*, vol. 29, pp. 181–212, 1997.

[30] D. M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," *J. Mach. Learn. Res.*, vol. 5, pp. 1287–1330, 2004.

[31] R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, Konstantinos G., Paul B., and B. Aragam, "DYNOTEARS: Structure learning from time-series data," in *Intl. Conf. Artif. Intel. Statist. (AISTATS)*, 2020, pp. 1595–1605.

[32] I. Ng, A. Ghassami, and K. Zhang, "On the role of sparsity and DAG constraints for learning linear DAGs," *Advances Neural Inf. Process. Syst.*, vol. 33, pp. 17943–17954, 2020.

[33] C. Deng, K. Bello, P. K. Ravikumar, and B. Aragam, "Global optimality in bivariate gradient-based DAG learning," in *Intl. Conf. Mach. Learn. (ICML)*, 2023.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[35] D. P. Bertsekas, *Dynamic Programming*, vol. 4, Athena Scientific, 1997.

[36] P.-L. Loh and P. Bühlmann, "High-dimensional learning of linear causal networks via inverse covariance estimation," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3065–3105, 2014.