Causal Structural Learning from Time Series: A Convex Optimization Approach

Song Wei, **Yao Xie**
School of Industrial and Systems Engineering, Georgia Institute of Technology
song.wei@gatech.edu yao.xie@isye.gatech.edu

Abstract

Structural learning, which aims to learn directed acyclic graphs (DAGs) from observational data, is foundational to causal reasoning and scientific discovery. Recent advancements formulate structural learning into a continuous optimization problem; however, DAG learning remains a highly non-convex problem, and there has not been much work on leveraging well-developed convex optimization techniques for causal structural learning. We fill this gap by proposing a data-adaptive linear approach for causal structural learning from time series data, which can be conveniently cast into a convex optimization problem using a recently developed monotone operator variational inequality (VI) formulation. Furthermore, we establish non-asymptotic recovery guarantee of the VI-based approach and show the superior performance of our proposed method on structure recovery over existing methods via extensive numerical experiments.

1 Introduction

Causal discovery, which aims to capture the interactions among events of interest using directed acyclic graphs (or Bayesian networks), is a crucial part of scientific discovery [Pearl, 2009] and has drawn much attention recently. With advanced data acquisition techniques, we usually observe time series data in many modern applications, posing both opportunities to learn a dynamic Bayesian network and challenges in finding an efficient approach for learning a directed acyclic graph (DAG) from serially correlated data [Pamfil *et al.*, 2020].

However, learning DAGs from observational data, i.e., the structural learning problem, is NP-hard due to the combinatorial acyclicity constraint [Chickering *et al.*, 2004], motivating many research efforts in finding efficient approaches for learning DAGs. Recently, [Zheng *et al.*, 2018] proposed a continuous differentiable characterization of DAG, which formulates the DAG learning problem into a constrained continuous optimization problem; they applied augmented Lagrangian method to transfer constraint into penalty and achieved efficient DAG learning. Later on, [Ng *et al.*, 2020] proposed to treat the non-convex DAG characterization as penalty and proved asymptotic recovery guarantee for linear Gaussian models.

On the other hand, recently much work has been done on causal discovery from time series; notable contributions include Fourier-transform based time series approach for continuous-time Hawkes process models [Etesami *et al.*, 2016]. However, existing works have been mostly focusing on Granger causality, which has been deemed less useful due to the lack of DAG structure in the estimated causal graph. To fix this issue, [Pamfil *et al.*, 2020] leveraged the continuous DAG characterization as the constraint in structural vector autoregressive models for Granger causal discovery and solved the constrained optimization problem via augmented Lagrangian method as [Zheng *et al.*, 2018] did. Despite those recent advancements, DAG learning remains a non-convex problem. Thus, how to leverage the well-developed convex optimization techniques to learn a DAG largely remains an open problem.

In this work, we present a generalized linear model (GLM) based approach for causal discovery from time series data, while seeking the DAG structure via a novel data-adaptive linear regularizer. Furthermore, we cast the DAG structural learning problem into a convex optimization program by a monotone operator variational inequality (VI) formulation. The convex formulation enables us to establish non-asymptotic performance guarantee for a wide range of non-linear link functions via recent advances in VI-based signal recovery [Juditsky and Nemirovski, 2019; Juditsky *et al.*, 2020]. We provide extensive numerical experiments to show the competitive performance of the proposed method and observe that our approach achieves more performance gain in the presence of limited data (see Figure 1 for illustration).

1.1 Literature

Efficient structural learning of a DAG is the heart of scientific discovery in many fields, e.g., biology [Sachs *et al.*, 2005], genetics [Zhang *et al.*, 2013], and so on. In particular, in causal reasoning, structural causal model based causal discovery methods oftentimes boil down to maximizing a score function within the DAG family [Glymour *et al.*, 2019]. There is rich literature in DAG learning: [Yuan *et al.*, 2019] proposed to use indicator function to enumerate and eliminate all possible directed cycles; to efficiently solve such problem, they used truncate ℓ_1-function as a continuous surrogate of indicator function and proposed to use alternating direc-

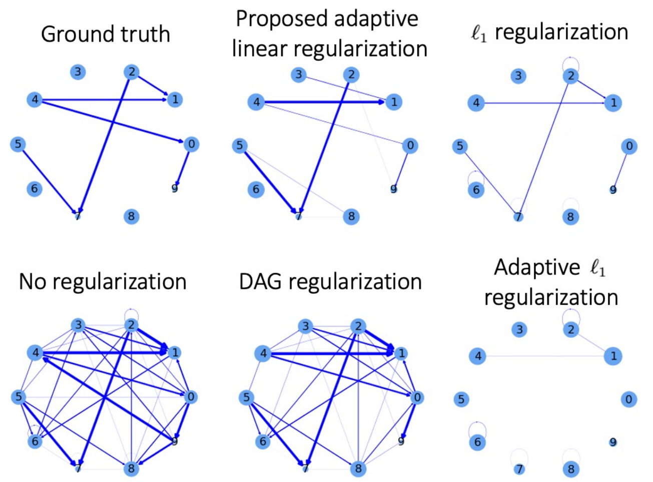

Figure 1: Visualization of estimated graphs, where the size of the node is proportional to the background intensity and the width of the edge is proportional to the exciting coefficient magnitude. We consider a graph with $d_1 = 10$ nodes and time horizon $T = 500$; the data is generated via our GLM with exponential link. We compare various types of regularization (specified on top of each panel). The Structural Hamming Distances between the estimated graph and the ground truth are 39 (no regularization), *7 (proposed)*, 21 (DAG regularization), 25 (ℓ_1 regularization) and 12 (adaptive ℓ_1 regularization), respectively. Our proposed data-adaptive linear regularization achieves the best graph structure recovery.

tion method of multipliers to numerically solve it. [Manzour *et al.*, 2021] transferred indicators into binary variables and leveraged mixed integer programming to solve it. There are also dynamic programming based approaches, e.g., [Loh and Bühlmann, 2014], but they are not scalable in high dimensions unless coupled sparse structure, e.g., $A*$ Lasso [Xiang and Kim, 2013]. Another line of research follows the continuous DAG characterization by [Zheng *et al.*, 2018]; in addition to aforementioned developments, notable extensions along this direction includes a discrete backpropagation method, exploration of low-rank structure [Fang *et al.*, 2020] and neural DAG learning [Yu *et al.*, 2019; Ke *et al.*, 2019; Lachapelle *et al.*, 2019]. We refer readers to [Scanagatta *et al.*, 2019; Vowels *et al.*, 2022] for systematic surveys on structural learning and causal discovery.

2 Background

2.1 Problem Set-Up

Consider observing d_1 binary time series over time horizon T, among which there exist lagged mutual-exciting effects and such effects have a finite memory depth $\tau \geq 1$. Specifically, we are given history data $\{y_t^{(i)} : t = 1 - \tau \dots, 0\}$ and observations $\{y_t^{(i)} : t = 1 \dots, T\}$ for $i \in \{1, \dots, d_1\}$, where $y_t^{(i)} = 1$ (or 0) represents type-i event occurrence (or not) at time t. We adopt the discrete-time Bernoulli process [Juditsky *et al.*, 2020] and model the probability of i-th event's occurrence at time step $t \in \{1, \dots, T\}$ via the following generalized linear model:

$$\mathbb{P}\left(y_t^{(i)} = 1 | \mathcal{H}_{t-1}\right) = g\left(\nu_i + \sum_{j=1}^{d_1} \sum_{k=1}^{\tau} \alpha_{ijk} y_{t-k}^{(j)}\right), \qquad (1)$$

where $\mathcal{H}_{t-1}$ denotes all observations up to time $t-1$. In the following, we will refer to those events as node variables. For i-th node, $\nu_i \geq 0$ reflects the deterministic background intensity, and $\alpha_{ijk} \geq 0$ represents the magnitude of triggering effect from the j-th node variable at time lag k. Link function $g : \mathbb{R} \to [0, 1]$ can be non-linear, such as sigmoid link function $g(x) = 1/(1 + e^{-x})$ on domain $x \in \mathbb{R}$ and $g(x) = 1 - e^{-x}$ on domain $x \in [0, \infty)$; also, it can be linear $g(x) = x$ on domain $x \in [0, 1]$, which reduces our GLM to the simple linear model. The major goal is to recover the mutual-excitation graphs (which is induced by mutual-excitation matrices $A_k = (\alpha_{ijk}) \in \mathbb{R}^{d_1 \times d_1}, k \in \{1, \dots, \tau\}$) over those d_1 nodes.

For brevity, we use $w_{t-\tau:t-1}$ to denote the observations from time $t - \tau$ to $t - 1$ and $\theta_i \in \mathbb{R}^d$ (where $d = 1 + \tau d_1$ denotes the dimensionality) to denote the problem parameter:

$$w_{t-\tau:t-1} = \left(1, y_{t-1}^{(1)}, \dots, y_{t-\tau}^{(1)}, \dots, y_{t-1}^{(d_1)}, \dots, y_{t-\tau}^{(d_1)}\right)^{\mathrm{T}},$$
$$\theta_i = (\nu_i, \alpha_{i11}, \dots, \alpha_{i1\tau}, \dots, \alpha_{id_11}, \dots, \alpha_{id_1\tau})^{\mathrm{T}},$$

where superscript $^{\mathrm{T}}$ denotes vector/matrix transpose. Parameter θ_i summarizes the influence from all nodes to node i. Now, we can rewrite (1) into the following compact form:

$$\mathbb{P}\left(y_t^{(i)} = 1 \middle| w_{t-\tau:t-1}\right) = g\left(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\right), \quad \theta_i \in \Theta, \qquad (2)$$

where $\Theta \subset \mathbb{R}_+^d = [0, \infty)^d$ is the feasible region and depends on the link function. For example, in the linear link case where $g(x) = x$, the feasible region is

$$\Theta = \{\theta \in \mathbb{R}_+^d : 0 \leq w_{t-\tau:t-1}^{\mathrm{T}} \theta \leq 1,\ t = 1, \dots, T\}.$$

2.2 Decoupled Estimation with Variational Inequality

In this section, we introduce a recently developed technique [Juditsky and Nemirovski, 2019; Juditsky *et al.*, 2020] to estimate the parameters of the GLM by solving stochastic monotone variational inequality. For $i \in \{1, \dots, d_1\}$. we assume the feasible region Θ is convex and compact and use the weak solution to the following variational inequality as the estimator $\hat{\theta}_i$ (which we will refer to as VI estimator; see [Juditsky and Nemirovski, 2019] for detailed background on solving VIs):

$$\text{find } \hat{\theta}_i \in \Theta : \langle F_T^{(i)}(\theta_i), \theta_i - \hat{\theta}_i \rangle \geq 0,\ \forall \theta_i \in \Theta,\ \ \mathrm{VI}[F_T^{(i)}, \Theta]$$

where $\langle \cdot \rangle$ represents the standard inner product in Euclidean space and $F_T^{(i)}(\theta_i)$ is the empirical vector field and defined as:

$$F_T^{(i)}(\theta_i) = \frac{1}{T} \sum_{t=1}^{T} w_{t-\tau:t-1} \left(g\left(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\right) - y_t^{(i)}\right). \qquad (3)$$

As we can see, the statistical inference for each node can be *decoupled* and therefore we can perform the computation in parallel and simplify the analysis.

The intuition behind this method is straightforward. Let us consider the global counterpart of the above vector field, whose root is the unknown ground truth $\theta_i^\star$,

$$\begin{aligned} F^{(i)}(\theta_i) &= \mathbb{E}_{(w,y^{(i)})}\left[w\left(g\left(w^{\mathrm{T}}\theta_i\right)-y^{(i)}\right)\right] \\ &= \mathbb{E}_{(w,y^{(i)})}\left[w\left(g\left(w^{\mathrm{T}}\theta_i\right)-g\left(w^{\mathrm{T}}\theta_i^\star\right)\right)\right]. \end{aligned}$$

Although we cannot access this global counterpart, by solving the empirical one $\mathrm{VI}[F_T^{(i)},\Theta]$ we could approximate the ground truth very well. We will show how well this approximation can be by generalizing the parameter recovery guarantee in [Juditsky *et al.*, 2020] to handle general non-linear link functions in Section 4.

3 Proposed Method

In our prior work [Wei *et al.*, 2023], evidence from synthetic and real data experiments shows that there could exist a DAG structure on the graphs induced by the (lagged) mutual-excitation matrices. However, as illustrated in Figure 1, it is difficult to recover the true graph structure in the presence of limited data without the help of DAG-inducing regularization. In this section, we will present our proposed data-adaptive linear regularization to encourage the DAG structure and show how to leverage such constraint (or rather, penalty) in the VI estimator.

3.1 Data-Adaptive Linear Cycle Elimination Regularization

Consider the graphs induced by the estimated adjacency matrices $\hat{A}_\ell = (\hat{\alpha}_{ij\ell}) \in \mathbb{R}^{d_1\times d_1}, \ell \in \{1,\dots,\tau\}$, using estimator $\mathrm{VI}[F_T^{(i)},\Theta]$. In DAG recovery problem, cycles in those estimated graphs are undesirable and should be removed.

First, let us formally define cycles: for positive integer $L \geq 2$, if there exist $\ell \in \{1,\dots,\tau\}$ and mutually different indices $i_1,\dots,i_L \in \{1,\dots,d_1\}$ such that

$$\hat{\alpha}_{i_1 i_L \ell} > 0, \quad \hat{\alpha}_{i_{k+1} i_k \ell} > 0, \quad k \in \{1,\dots,L-1\},$$

then we say there exists a *length-L (directed) cycle* in the directed graphs induced by $\hat{A}_\ell$'s. In particular, for $L=1$ case, we say there is a *length-*1 *cycle (or lagged self-exciting component)* if there exist $\ell \in \{1,\dots,\tau\}$ and index $i \in \{1,\dots,d_1\}$ such that $\hat{\alpha}_{ii\ell} > 0$.

To remove those cycles, we consider all possible length-1, 2 and 3 cycles in the estimated graphs, whose indices are denoted as follows: for all $\ell \in \{1,\dots,\tau\}$,

$$\begin{aligned} &I_{1,\ell} = \big\{i:\ \hat{\alpha}_{ii\ell} > 0\big\},\ I_{2,\ell} = \big\{(i,j): i \neq j,\ \hat{\alpha}_{ij\ell}, \hat{\alpha}_{ji\ell} > 0\big\}, \\ &I_{3,\ell} = \big\{(i,j,k): i,j,k \text{ mutually different}, \\ &\qquad\qquad\qquad\qquad \hat{\alpha}_{ij\ell}, \hat{\alpha}_{jk\ell}, \hat{\alpha}_{ki\ell} > 0\big\}. \end{aligned}$$

Intuitively, in each length-2 (or 3) cycle of those estimated graphs, the edge with the smallest weight could be caused by noisy observation, meaning that we should remove such edge to eliminate the corresponding cycle. To do so, we impose the following *data-adaptive linear cycle elimination constraints*, aiming to shrink the weight of those "least important edges" in the cycle: for all $\ell \in \{1,\dots,\tau\}$,

$$\begin{aligned} \alpha_{ij\ell} + \alpha_{ji\ell} &\leq \delta_{2,\ell}(i,j), \quad (i,j) \in I_{2,\ell}, \\ \alpha_{ij\ell} + \alpha_{jk\ell} + \alpha_{ki\ell} &\leq \delta_{3,\ell}(i,j,k), \quad (i,j,k) \in I_{3,\ell}, \end{aligned} \tag{4}$$

where the *adaptive regularization strength parameters* are

$$\begin{aligned} \delta_{2,\ell}(i,j) &= \hat{\alpha}_{ij\ell} + \hat{\alpha}_{ji\ell} - \min\{\hat{\alpha}_{ij\ell}, \hat{\alpha}_{ji\ell}\}, \\ \delta_{3,\ell}(i,j,k) &= \hat{\alpha}_{ij\ell} + \hat{\alpha}_{jk\ell} + \hat{\alpha}_{ki\ell} - \min\{\hat{\alpha}_{ij\ell}, \hat{\alpha}_{jk\ell}, \hat{\alpha}_{ki\ell}\}. \end{aligned}$$

3.2 Joint VI Estimator with Penalty

Different from the decoupled learning approach in Section 2.2, parameters $\theta_1,\dots,\theta_{d_1}$ should be estimated jointly to account for the desired DAG structure. We concatenate the parameter and response vectors into matrices as follows:

$$\theta = (\theta_1,\dots,\theta_{d_1}) \in \mathbb{R}^{d\times d_1},\ Y = (Y_{1:T}^{(1)},\dots,Y_{1:T}^{(d_1)}) \in \mathbb{R}^{T\times d_1},$$

where $Y_{1:T}^{(i)} = (y_1^{(i)},\dots,y_T^{(i)})^{\mathrm{T}}$. The feasible region of the concatenated parameter $\tilde{\Theta}$ is then defined as follows:

$$\tilde{\Theta} = \{\theta = (\theta_1,\dots,\theta_{d_1}) : \theta_i \in \Theta,\ i=1,\dots,d_1\}.$$

One natural idea to incorporate the data-adaptive linear constraints (4) is to directly include them into the feasible region $\tilde{\Theta}$. Since adding linear constraints into the original convex feasible region will ensure the new feasible region is still convex (intersection of convex sets remains convex), solving VI in the new feasible region remains a convex problem.

In practice, we typically treat the empirical vector field as the gradient field and perform projected gradient descent (PGD) to numerically solve for the VI estimator [Juditsky and Nemirovski, 2019]. Thus, adding more constraints to feasible region will make the projection harder to implement; one can see a special case on how to use PGD to solve for VI estimator in Appendix A.2. Alternatively, we propose a *data-adaptive linear penalized VI estimator*, which is the weak solution to the following Variational Inequality:

$$\text{find } \hat{\theta} \in \tilde{\Theta} : \langle \mathrm{vec}(F_T^{\mathrm{AL}}(\theta)), \mathrm{vec}(\theta - \hat{\theta})\rangle \geq 0, \quad \forall \theta \in \tilde{\Theta},$$

where $\mathrm{vec}(A)$ is the vector of columns of A stacked one under the other. The data-adaptive linear penalized vector field $F_T^{\mathrm{AL}}(\theta)$ is defined as follows:

$$\begin{aligned} F_T^{\mathrm{AL}}(\theta) = F_T(\theta) + \lambda \sum_{\ell=1}^{\tau} \Bigg(&\sum_{i\in I_{1,\ell}} \frac{e_{f_{i,\ell},d} e_{i,d_1}^{\mathrm{T}}}{\hat{\alpha}_{ii\ell}} \\ &+ \sum_{i\notin I_{1,\ell}} \frac{e_{f_{i,\ell},d} e_{i,d_1}^{\mathrm{T}}}{\Lambda} + \sum_{(i,j)\in I_{2,\ell}} \frac{e_{f_{j,\ell},d} e_{i,d_1}^{\mathrm{T}} + e_{f_{i,\ell},d} e_{j,d_1}^{\mathrm{T}}}{\delta_{2,\ell}(i,j)} \\ &+ \sum_{(i,j,k)\in I_{3,\ell}} \frac{e_{f_{j,\ell},d} e_{i,d_1}^{\mathrm{T}} + e_{f_{k,\ell},d} e_{j,d_1}^{\mathrm{T}} + e_{f_{i,\ell},d} e_{k,d_1}^{\mathrm{T}}}{\delta_{3,\ell}(i,j,k)} \Bigg), \end{aligned} \tag{5}$$

where the "concatenated empirical vector field" $F_T(\theta)$ is

$$F_T(\theta) = (F_T^{(1)}(\theta_1),\dots,F_T^{(d_1)}(\theta_{d_1})) \in \mathbb{R}^{d\times d_1}, \tag{6}$$

and the empirical vector field $F_T^{(i)}(\theta_i)$ is defined in (3). Since vector field is treated as gradient field, we add the derivative of the linear penalty term to the vector field in (5).

Interpretation of the penalized vector field. In (5), $e_{i,d} \in \mathbb{R}^d$ is the standard basis vector with its i-th element being one and $f_{j,\ell} = 1 + (j-1)\tau + \ell$, which gives us

$$e^{\mathrm{T}}_{f_{j,\ell},d}\theta e_{i,d_1} = \alpha_{ij\ell}, \quad \nabla_\theta(e^{\mathrm{T}}_{f_{j,\ell},d}\theta e_{i,d_1}) = e_{f_{j,\ell},d}e^{\mathrm{T}}_{i,d_1}.$$

The penalties at the end of the first line and the beginning of the second in (5) are very similar to adaptive Lasso [Zou, 2006], aiming to remove all lagged self-exiting components. Intuitively, the smaller the adaptive regularization strength parameters are, the stronger penalties should be applied, which explains why those regularization strength parameters appear in the denominator.

Selection of regularization strength. Hyperparameters λ and Λ are tunable and control the penalty strength. In practice, hyperparameter Λ is usually set to be a small number, such as 10^{-3}, to ensure there will not exist self-exciting components, whereas λ is selected based on the continuous DAG characterization [Zheng *et al.*, 2018]. To be precise, let us consider the $\tau = 1$ special case in the illustrative example in Figure 1, and use $A = (\alpha_{ij})$ to denote $A_1 = (\alpha_{ij1})$ for brevity. The DAG characterization of a graph induced by adjacency matrix A is:

$$h(A) = \mathrm{tr}(e^A) - d, \tag{7}$$

where $\mathrm{tr}(e^A)$ is the trace of matrix exponential of A. For $A \in \mathbb{R}_+^{d_1 \times d_1}$, we have $h(A) \geq 0$ and $h(A) = 0$ if and only if the directed graph induced by adjacency matrix A is a DAG. Therefore, $h(A)$ can measure the "DAG-ness" of A. We study how the performances vary with respect to (w.r.t.) hyperparameter λ in Figure 2; the performance evaluation metrics are: (i) matrix F-norm of the mutual-exciting matrix estimation error (A err.), (ii) the ℓ_2 norm of the background intensity estimation error (ν err.), (iii) "DAG-ness" of estimated adjacency matrix $h(A)$, and (iv) Structural Hamming Distance (SHD) between the estimated and the true adjacency matrices.

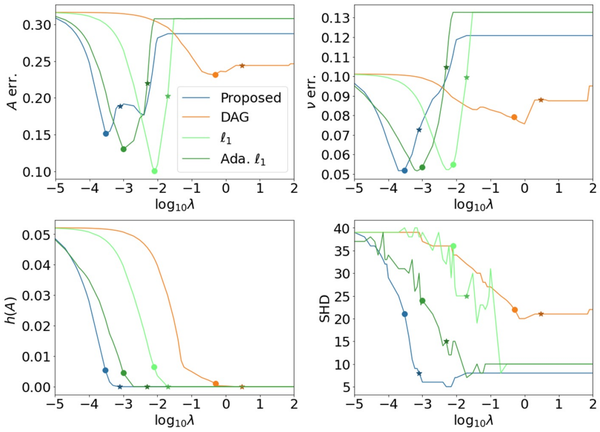

Figure 2: Illustration of the hyperparameter selection. We plot the trajectories of four performance metrics w.r.t. hyperparameter λ for the example in Figure 1. In our numerical simulation, λ is selected to be the smallest one which satisfies $h(A) \leq 10^{-8}$. The selected λ is marked with a star; in addition, we mark the λ which minimizes A err. with a dot.

From Figure 2, we can observe that the λ which minimizes the A err. (marked with a dot) typically does not give the best structural recovery (i.e., the smallest SHD); A err. cannot be used to select hyperparameter anyways since its calculation requires knowledge on the ground truth. Fortunately, we observe that the SHD converges (to its near optimal value) almost the same time when the "DAG-ness" measure $h(A)$ converges to zero. Therefore, we propose to select λ as *the smallest one which satisfies that* $h(A) \leq \mathrm{thres.}$, where thres. is again user-specified. Later in our numerical experiments, we will show how hyperparameter thres. controls the balance between structural recovery (SHD) and weight recovery (A err.).

4 Theoretical Analysis

In this section, we extend the non-asymptotic recovery guarantee of VI$[F_T^{(i)}, \Theta]$ for linear link function case in [Juditsky *et al.*, 2020] to general non-linear link function case by imposing the following assumption:

Assumption 1. The link function $g(\cdot)$ is continuous and monotone, and the vector field $G(\theta) = \mathbb{E}_w[wg(w^{\mathrm{T}}\theta)]$ is well defined (and therefore monotone along with g). Moreover, g is differentiable and has uniformly bounded first order derivative $m_g \leq |g'| \leq M_g$ for $0 < m_g \leq M_g$.

Then, the non-asymptotic upper bound on estimation error is given as follows:

Theorem 1. *Under Assumption 1, for $i \in \{1, \ldots, d_1\}$ and any $\varepsilon \in (0,1)$, with probability at least $1-\varepsilon$, the ℓ_2 estimation error of VI$[F_T^{(i)}, \Theta]$ can be upper bounded as follows:*

$$\|\hat{\theta}_i - \theta_i^\star\|_2 \leq \frac{1}{m_g\lambda_1}\sqrt{\frac{d\log(2d/\varepsilon)}{T}},$$

where $\theta_i^\star$ is the unknown ground truth parameter, and λ_1 is the smallest eigenvalue of $\mathbb{W}_{1:T} = \sum_{t=1}^T w_{t-\tau:t-1}w_{t-\tau:t-1}^{\mathrm{T}}/T$.

As pointed out in [Juditsky *et al.*, 2020], $\mathbb{W}_{1:T} \in \mathbb{R}^{d \times d}$ will be full rank when T is sufficiently large, i.e., with high probability, λ_1 will be a positive constant. The complete proof of the above theorem can be found in Appendix A.1. One pitfall of the theoretical analysis is the lack of guarantee for the proposed data-adaptive linear regularizer and we leave this part for future discussion. In the following, we will use numerical experiments to show the good performance of our method.

Identifiablility of our proposed estimator. In addition, we can show the uniqueness, or rather, the identifiablility of the VI estimator VI$[F_T^{(i)}, \Theta]$, which comes from the nice property of the underlying vector field. To be precise, in the proof of the above theorem, we have shown the vector field $F_T^{(i)}(\theta_i)$ is monotone modulus $m_g\lambda_1$ under Assumption 1. Then, the following lemma tells us that our proposed estimator is unique:

Lemma 1 (Lemma 3.1 [Juditsky and Nemirovski, 2019])**.** Let Θ be a convex compact set and H be a monotone vector field on Θ with monotonicity modulus $\kappa > 0$, i.e.,

$$\forall\, z, z' \in \Theta, [H(z) - H(z')]^{\mathrm{T}}(z - z') \geq \kappa\|z - z'\|_2^2.$$

Then, the weak solution $\bar{z}$ to VI$[H, \Theta]$ exists and is unique. It satisfies:

$$H(z)^{\mathrm{T}}(z - \bar{z}) \geq \kappa\|z - \bar{z}\|_2^2.$$

5 Numerical Experiments

In this section, we provide more numerical experiments to show the effectiveness of our proposed method. We will 1) show its competitive performance under various settings and 2) study the effect regularization strength hyperparameter. In our numerical simulation, we consider $\tau = 1$ case for simplicity and choose SHD (for structural recovery) and A err. (for weight recovery) as the primary performance metrics. We report the mean and standard deviation of those metrics over 200 independent trials. Complete details, such as random DAG generation, can be found in Appendix B.

Let us begin with presenting benchmark methods. The idea of transferring constraint into penalty by adding the penalty's derivative to the vector field opens up possibilities to consider various type of DAG-inducing penalties when using the VI estimator, e.g., the continuous DAG penalty [Zheng *et al.*, 2018] and the adaptive Lasso [Zou, 2006]. As mentioned earlier, we use $A = (\alpha_{ij})$ to denote $A_1 = (\alpha_{ij1})$ for brevity; in addition, we denote $J = (\mathbf{0}_{d_1}, I_{d_1}) \in \mathbb{R}^{d_1 \times d}$ such that we have $J\theta = A^{\mathrm{T}}$.

Continuous DAG regularization. The DAG characterization (7) has closed-from derivative as follows:

$$\nabla h(A) = \left(e^A\right)^{\mathrm{T}}.$$

Inspired by [Ng *et al.*, 2020] who treated the DAG characterization directly as a penalty, we take advantage of the differentiability of the DAG penalty and add its derivative to the concatenated field $F_T(\theta)$ (6), which will later be treated as the gradient field when we use PGD to solve for the estimator. More precisely, the DAG-penalized vector field $F_T^{\mathrm{DAG}}(\cdot)$ is defined as follows:

$$F_T^{\mathrm{DAG}}(\theta) = F_T(\theta) \ + \lambda J^{\mathrm{T}} \nabla h(J\theta) = F_T(\theta) \ + \lambda J^{\mathrm{T}} e^A.$$

ℓ_1 regularization. We adopt the ℓ_1 penalty as another benchmark method, which will encourage a sparse structure on the adjacency matrix A and in turn eliminates cycles. To be precise, the ℓ_1 penalized vector field is defined as follows:

$$F_T^{\ell_1}(\theta) = F_T(\theta) + \lambda J^{\mathrm{T}} \nabla(|J\theta|_1), \tag{8}$$

where $|\cdot|_1$ is the summation of all entries' absolute values.

Adaptive Lasso. As a variant of ℓ_1 regularization, adaptive ℓ_1 regularization, or adaptive Lasso [Zou, 2006], replaces $\lambda|\alpha_{ij}|$ with $\frac{\lambda}{\hat{\alpha}_{ij}}|\alpha_{ij}|$ in (8). In addition, for $\hat{\alpha}_{ij} = 0$ case, we use a simple remedy by adding penalty term $\frac{\lambda}{\Lambda}|\alpha_{ij}|$ as in (5) to restrict α_{ij} to be zero.

As shown in Figure 1, our proposed data-adaptive linear approach has superior performance compared with the aforementioned DAG-inducing penalties. We will give more numerical evidence to support this in the following.

Experiment 1. First, we show the superior performance of our proposed method under settings $(d_1, T) \in \{(10, 500), (20, 1000), (300, 1500)\}$; results are reported in Figure 3. We observe that our proposed method achieves the best structural recovery among all methods, especially in higher dimensions. Besides, ℓ_1 regularization does well in weight recovery but poorly in structural recovery. As a comparison, our proposed method achieves comparable weight recovery accuracy with ℓ_1 regularization but much better structural recovery accuracy. On the contrary, DAG regularization is completely dominated by our proposed method, potential due to the non-convexity incurred by the DAG characterization (7); adaptive ℓ_1 regularization achieves improved structural recovery accuracy compared with ℓ_1 regularization, but is again dominated by our proposed method in most cases. As a sanity check, we observe the A err.'s are all on the same scale for different dimension cases — this is because we normalize each row of A to sum to one to ensure it stays within the feasible region for linear link function case. For completeness, we also report the ν err. and the "DAG-ness" measure $h(A)$ in Table 1 in Appendix B. Those results do not only further validate our aforementioned observations, but also show ℓ_1 regularization does the best in returning a DAG (even better than DAG regularization) but cannot return an accurate graph structure. This agrees with our illustration in Figure 1 — it does very well in encouraging sparse structure, but may shrink some important edges' weights to zeros.

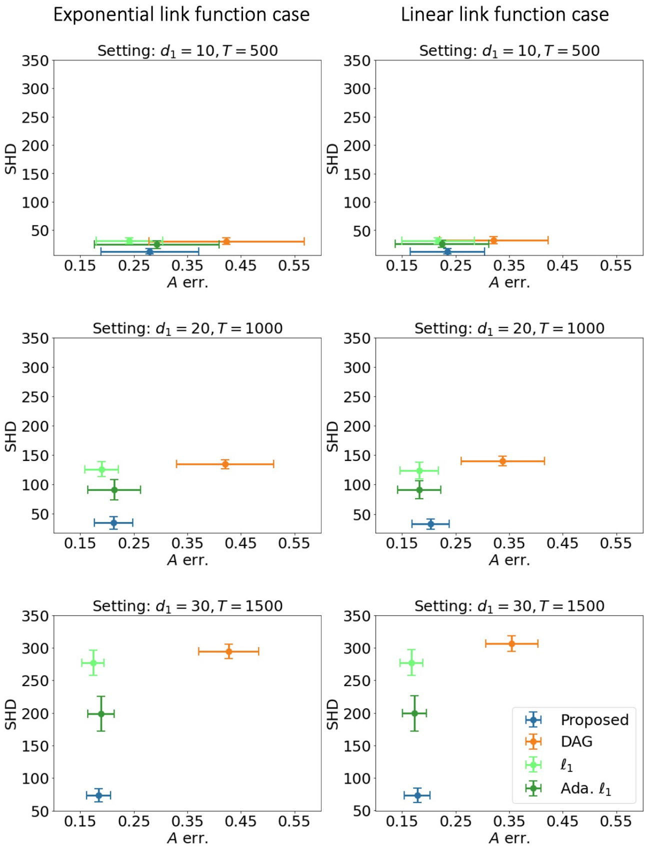

Figure 3: Comparison among different types of regularization in DAG recovery. We plot the mean (dot) and standard deviation (error bar) of matrix F-norm of the mutual-exciting matrix estimation error (A err.) and Structural Hamming Distance over 200 independent trials for various types of regularization. Hyperparameter λ is selected to be the smallest one which satisfies $h(A) \leq 10^{-4}$. For each regularization, the closer it is to the origin, the better it is. We can observe that our proposed data-adaptive linear regularization performs the best (especially in higher dimensional case).

Experiment 2. We now study the effect of the hyperparameter $\mathrm{thres.}$ introduced in Section 3.2. We plot the SHD and A err. in Figure 4 for the exponential link function case; for completeness, we report the result for linear link function in Figure 5 in Appendix B. From both figures, we can observe that: (i) On one hand, smaller $\mathrm{thres.}$ does give better SHD. (ii) On the other hand, A err. exhibits a U-shape property w.r.t. $\mathrm{thres.}$, which agrees with the U-shape curves for both A err. and ν err. w.r.t. λ in Figure 2 and suggests that there could exist one optimal hyperparameter in outputting the smallest A err.; however, it is an open problem on how to select it to minimize A err. — one possible approach is through the norm of empirical vector field, since it is treated as the gradient field in PGD. Nevertheless, we mainly focus on the structural recovery (i.e., SHD), and it is safe to choose a sufficient small $\mathrm{thres.}$ (e.g., 10^{-4}) in practice.

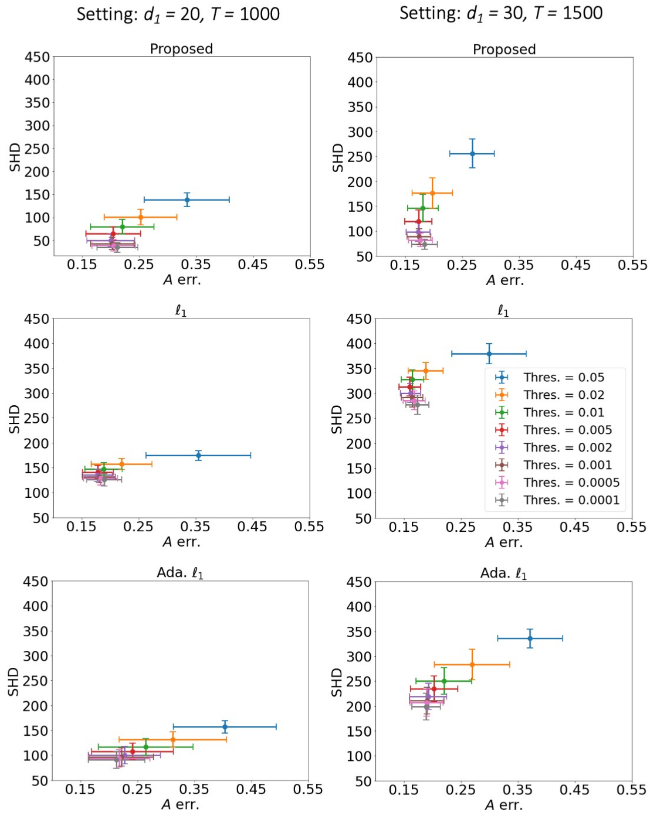

Figure 4: Effect of hyperparameter. We consider the VI estimator with exponential link function. Regularization strength hyperparameter λ is selected to be the smallest one which satisfies that $h(A)$ is smaller than a given threshold ($\mathrm{thres.}$). We plot the mean (dot) and standard deviation (error bar) of A err. and SHD over 200 independent trials for different choices of this threshold. We can observe that smaller $\mathrm{thres.}$ typically leads to better SHD.

6 Conclusion

In this work, we go beyond the continuous but non-convex optimization approach for structural learning [Zheng *et al.*, 2018] and formulate the DAG learning problem as a general convex optimization problem. Our theoretical analysis for the VI estimator extends the recovery guarantee in [Juditsky *et al.*, 2020] to the general non-linear monotone ink function cases and our numerical experiments show our method's superior performance over existing methods in structural learning, opening up possibility for future work to adopt this method in a wide range of applications.

Acknowledgments

This work is partially supported by NSF CAREER CCF-1650913, and NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, DMS-1830210.

References

[ApS, 2019] MOSEK ApS. *The MOSEK optimization toolbox for Python manual. Version 10.0.*, 2019.

[Chickering *et al.*, 2004] Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

[Etesami *et al.*, 2016] Jalal Etesami, Negar Kiyavash, Kun Zhang, and Kushagra Singhal. Learning network of multivariate hawkes processes: a time series approach. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 162–171, 2016.

[Fang *et al.*, 2020] Zhuangyan Fang, Shengyu Zhu, Jiji Zhang, Yue Liu, Zhitang Chen, and Yangbo He. Low rank directed acyclic graphs and causal structure learning. *arXiv preprint arXiv:2006.05691*, 2020.

[Glymour *et al.*, 2019] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[Juditsky and Nemirovski, 2019] Anatoli B Juditsky and AS Nemirovski. Signal recovery by stochastic optimization. *Automation and Remote Control*, 80(10):1878–1893, 2019.

[Juditsky *et al.*, 2020] Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, and Yao Xie. Convex parameter recovery for interacting marked processes. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[Ke *et al.*, 2019] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[Lachapelle *et al.*, 2019] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2019.

[Loh and Bühlmann, 2014] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[Manzour *et al.*, 2021] Hasan Manzour, Simge Küçükyavuz, Hao-Hsiang Wu, and Ali Shojaie. Integer programming for learning directed acyclic graphs from continuous data. *INFORMS Journal on Optimization*, 3(1):46–73, 2021.

[Ng *et al.*, 2020] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

[Pamfil *et al.*, 2020] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Sachs *et al.*, 2005] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[Scanagatta *et al.*, 2019] Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. A survey on bayesian network structure learning from data. *Progress in Artificial Intelligence*, 8(4):425–439, 2019.

[Vowels *et al.*, 2022] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

[Wei *et al.*, 2023] Song Wei, Yao Xie, Christopher S Josef, and Rishikesan Kamaleswaran. Causal graph discovery from self and mutually exciting time series. *arXiv preprint arXiv:2106.02600*, 2023.

[Xiang and Kim, 2013] Jing Xiang and Seyoung Kim. A* lasso for learning a sparse bayesian network structure for continuous variables. *Advances in neural information processing systems*, 26, 2013.

[Yu *et al.*, 2019] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

[Yuan *et al.*, 2019] Yiping Yuan, Xiaotong Shen, Wei Pan, and Zizhuo Wang. Constrained likelihood for reconstructing a directed acyclic gaussian graph. *Biometrika*, 106(1):109–125, 2019.

[Zhang *et al.*, 2013] Bin Zhang, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, et al. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707–720, 2013.

[Zheng *et al.*, 2018] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[Zou, 2006] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

A Additional Technical Details

A.1 Proofs

We begin with defining an auxiliary vector field

$$\tilde{F}_T^{(i)}(\theta_i) = \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1}(g(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i) - g(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i^\star)),$$

where $\theta_i^\star$ is the unknown ground truth. This vector field has a nice property that its unique root/weak solution to corresponding VI is $\theta_i^\star$, whereas the VI estimator $\hat{\theta}_i$ is the root of $F_T^{(i)}(\theta_i)$. Next, we bound the difference between $\hat{\theta}_i$ and $\theta_i^\star$ by bounding the difference between the empirical vector field $F_T^{(i)}(\theta_i)$ and the auxiliary vector field $\tilde{F}_T^{(i)}(\theta_i)$, i.e.,

$$\Delta^{(i)} = F_T^{(i)}(\theta_i) - \tilde{F}_T^{(i)}(\theta_i) = F_T^{(i)}(\theta_i^\star).$$

Proposition 1. Under Assumption 1, for $i \in \{1, \dots, d_1\}$ and any $\varepsilon \in (0, 1)$, with probability at least $1 - \varepsilon$, the following holds:

$$\|\Delta^{(i)}\|_\infty \le \sqrt{\log(2d/\varepsilon)/T}. \tag{9}$$

Moreover, this implies

$$\|\Delta^{(i)}\|_2 \le \sqrt{d\log(2d/\varepsilon)/T}. \tag{10}$$

Proof. Denote random vector

$$\xi_t = w_{t-\tau:t-1}\left(g\left(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i^\star\right) - y_t^{(i)}\right).$$

We can re-write $\Delta^{(i)} = \sum_{t=1}^{T}\xi_t/T$. Define σ-field $\mathcal{F}_t = \sigma(\mathcal{H}_t)$, and $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \mathcal{F}_T$ form a filtration. We can show $\mathbb{E}[(\xi_t)_k|\mathcal{F}_{t-1}] = 0$, and

$$\begin{aligned}\mathrm{Var}((\xi_t)_k|\mathcal{F}_{t-1}) &\le g\left(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i\right)\left(1 - g\left(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i\right)\right)\\ &\le 1/4,\end{aligned}$$

where the subscript $_k$ represents the corresponding k-th entry of the vector, and the bound on the variance comes from the property of a Bernoulli distribution. This means $\xi_t, t \in \{1, \dots, T\}$, is a Martingale Difference Sequence; additionally, its infinity norm is upper bounded by one since it only consists of binary elements. Therefore, Azuma's inequality gives us:

$$\mathbb{P}\left(|\Delta_k^{(i)}| > u\right) \le 2\exp\left\{-\frac{Tu^2}{2}\right\},\ k = 1, \dots, d,\ \forall\, u > 0,$$

where $\Delta_k^{(i)}$ is the k-th entry of vector $\Delta^{(i)}$. By union bound,

$$\mathbb{P}\left(|\Delta_k^{(i)}| > u,\ k = 1, \dots, d\right) \le 2d\exp\left\{-\frac{Tu^2}{2}\right\}, \forall\, u > 0.$$

Setting the RHS of above inequality to ε and solving for u, we prove (9); notice that $\|\Delta\|_2 \le \sqrt{d}\|\Delta\|_\infty$, we prove (10). □

The proof of Proposition 1 leverages the concentration property of martingales. By this proposition, we can now prove the non-asymptotic estimation error bound as follows:

Proof of Theorem 1. Under Assumption 1, the vector field $F_T^{(i)}(\theta_i)$ is monotone modulus $m_g\lambda_1$, since

$$\begin{aligned}&\left(F_T^{(i)}(\theta) - F_T^{(i)}(\theta')\right)^{\mathrm{T}}(\theta - \theta')\\ &= \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1}^{\mathrm{T}}(\theta - \theta')\left(g\left(w_{t-\tau:t-1}^{\mathrm{T}}\theta\right) - g\left(w_{t-\tau:t-1}^{\mathrm{T}}\theta'\right)\right)\\ &\ge m_g\frac{1}{T}\sum_{t=1}^{T}\|w_{t-\tau:t-1}^{\mathrm{T}}(\theta - \theta')\|_2^2\\ &= m_g(\theta - \theta')^{\mathrm{T}}\frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1}w_{t-\tau:t-1}^{\mathrm{T}}(\theta - \theta')\\ &\ge m_g\lambda_1\|\theta - \theta'\|_2^2.\end{aligned}$$

In particular, we have:

$$\left(F_T^{(i)}(\hat{\theta}_i) - F_T^{(i)}(\theta_i^\star)\right)^{\mathrm{T}}(\hat{\theta}_i - \theta_i^\star) \ge m_g\lambda_1\|\hat{\theta}_i - \theta_i^\star\|_2^2.$$

Notice that our weak solution $\hat{\theta}_i$ is also a strong solution to the VI since the empirical vector field is continuous (cf. [Juditsky and Nemirovski, 2019]), which gives us

$$\left(F_T^{(i)}(\hat{\theta}_i)\right)^{\mathrm{T}}(\hat{\theta}_i - \theta_i^\star) \le 0.$$

By Cauchy Schwartz inequality, we also have

$$-\left(F_T^{(i)}(\theta_i^\star)\right)^{\mathrm{T}}(\hat{\theta}_i - \theta_i^\star) = -\Delta_i^{\mathrm{T}}(\hat{\theta}_i - \theta_i^\star) \le \|\Delta_i\|_2\|\hat{\theta}_i - \theta_i^\star\|_2.$$

Together with (10) in Proposition 1, we complete the proof. □

A.2 A Special Example

Decoupled Estimation. We consider a special case where $g(x) = x$. To ensure the model (2) can return a meaningful probability, we require the parameter θ_i to take value in $\Theta = \{\theta_i \in \mathbb{R}_+^d : 0 \le w_{t-\tau:t-1}^{\mathrm{T}}\theta_i \le 1,\ t = 1, \dots, T\}$. In this special case, the empirical vector field (3) becomes

$$\begin{aligned}F_T^{(i)}(\theta_i) &= \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1}w_{t-\tau:t-1}^{\mathrm{T}}\theta_i - \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1}y_t^{(i)}\\ &= \mathbb{W}_{1:T}\theta_i - \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1}y_t^{(i)},\end{aligned}$$

where

$$\begin{aligned}&\mathbf{w}_{1:T} = (w_{1-\tau:0}, \dots, w_{T-\tau:T-1}) \in \mathbb{R}^{d\times T},\\ &\mathbb{W}_{1:T} = \frac{1}{T}\mathbf{w}_{1:T}\mathbf{w}_{1:T}^{\mathrm{T}} = \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1}w_{t-\tau:t-1}^{\mathrm{T}} \in \mathbb{R}^{d\times d}.\end{aligned} \tag{11}$$

Most importantly, this vector field is indeed the gradient field of the least square objective, meaning that the weak solution to the corresponding VI is the following LS estimator [Juditsky *et al.*, 2020]:

$$\begin{aligned}\min_{\theta_i}\quad & \tfrac{1}{2T}\|\mathbf{w}_{1:T}^{\mathrm{T}}\theta_i - Y_{1:T}^{(i)}\|_2^2,\\ \text{subject to}\quad & \theta_i \ge \mathbf{0}_T,\ \mathbf{1}_T - \mathbf{w}_{1:T}^{\mathrm{T}}\theta_i \ge \mathbf{0}_T,\end{aligned} \tag{12}$$

where $Y_{1:T}^{(i)} = (y_1^{(i)}, \ldots, y_T^{(i)})^{\mathrm{T}}$, $\mathbf{0}_T$ and $\mathbf{1}_T$ are the column vectors of all zeros and ones in $\mathbb{R}^T$, respectively, and $\|\cdot\|_p$ denotes the vector ℓ_p norm.

Note that the equivalence between our proposed estimatorand LS estimator will only hold for linear link function, since the gradient field of LS objective with general link function is:

$$\frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1} g'\left(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i\right)\left(g\left(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i\right) - y_t^{(i)}\right).$$

One approach to solve (12) is to leverage the well-developed optimization tools, such as `Mosek` [ApS, 2019]. An alternative approach is through projected gradient descent, where the empirical vector field (3) is treated as the gradient. To be precise, we introduce dual variables $\eta_1 = (\eta_{1,1}, \ldots, \eta_{1,T})^{\mathrm{T}}$, $\eta_2 = (\eta_{2,1}, \ldots, \eta_{2,d})^{\mathrm{T}}$ and the Lagrangian is as follows:

$$\begin{aligned} L(\theta_i, \eta_1, \eta_2) = \frac{1}{2T}\|\mathbf{w}_{1:T}^{\mathrm{T}}\theta_i - Y_{1:T}^{(i)}\|_2^2 \\ + \eta_1^{\mathrm{T}}(\mathbf{w}_{1:T}^{\mathrm{T}}\theta_i - \mathbf{1}_T) - \eta_2^{\mathrm{T}}\theta_i. \end{aligned}$$

The Lagrangian dual function is $\min_{\theta_i} L(\theta_i, \eta_1, \eta_2)$. As we can see, the Lagrangian above is convex w.r.t. θ_i. By setting the derivative of $L(\theta_i, \eta_1, \eta_2)$ w.r.t. θ_i to zero, we have

$$\hat{\theta}_i = \frac{1}{T}\mathbb{W}_{1:T}^{-1}\left(\mathbf{w}_{1:T}Y_{1:T}^{(i)}/T - \eta_1\right) + \eta_2,$$

which minimizes the Lagrangian dual function. As pointed out in [Juditsky *et al.*, 2020], $\mathbb{W}_{1:T} \in \mathbb{R}^{d\times d}$ will be full rank with high probability when T is sufficiently large, and therefore $\mathbb{W}_{1:T}^{-1}$ exists. By plugging $\hat{\theta}_i$ into the Lagrangian dual function $\min_{\theta_i} L(\theta_i, \eta_1, \eta_2)$, we give the dual problem as follows:

$$\max_{\eta_1,\eta_2} L(\hat{\theta}_i, \eta_1, \eta_2), \text{ subject to } \eta_1, \eta_2 \geq \mathbf{0}_T.$$

As we can see, this dual problem can be easily solved by PGD. *Joint Estimation.* Now let us consider the joint estimation, where the vector field $F_T(\theta)$ (6) can be expressed as follows:

$$F_T(\theta) = \frac{1}{T}\mathbf{w}_{1:T}\mathbf{w}_{1:T}^{\mathrm{T}}\theta - \frac{1}{T}\mathbf{w}_{1:T}Y = \mathbb{W}_{1:T}\theta - \frac{1}{T}\mathbf{w}_{1:T}Y,$$

where $\mathbf{w}_{1:T} \in \mathbb{R}^{d\times T}$ is defined in 11. Similar to the example in decoupled estimation, the above vector field is the gradient field of the least square objective, and our proposed estimator boils down to LS estimator, which solves the following penalized optimization problem:

$$\begin{aligned} \min_{\theta\in\tilde{\Theta}} \quad & \frac{1}{2T}\|\mathbf{w}_{1:T}^{\mathrm{T}}\theta - Y\|_F^2 + \lambda\sum_{\ell=1}^{\tau}\left(\sum_{i\in I_{1,\ell}} \frac{e_{f_{i,\ell},d}^{\mathrm{T}}\theta e_{i,d_1}}{\hat{\alpha}_{ii\ell}}\right. \\ & + \sum_{i\notin I_{1,\ell}} \frac{e_{f_{i,\ell},d}^{\mathrm{T}}\theta e_{i,d_1}}{\Lambda} + \sum_{(i,j)\in I_{2,\ell}} \frac{e_{f_{j,\ell},d}^{\mathrm{T}}\theta e_{i,d_1} + e_{f_{i,\ell},d}^{\mathrm{T}}\theta e_{j,d_1}}{\delta_{2,\ell}(i,j)} \\ & \left. + \sum_{(i,j,k)\in I_{3,\ell}} \frac{e_{f_{j,\ell},d}^{\mathrm{T}}\theta e_{i,d_1} + e_{f_{k,\ell},d}^{\mathrm{T}}\theta e_{j,d_1} + e_{f_{i,\ell},d}^{\mathrm{T}}\theta e_{k,d_1}}{\delta_{3,\ell}(i,j,k)}\right), \end{aligned}$$

where $\|\cdot\|_F$ is the matrix F-norm. Therefore, the above optimization problem can be solved efficiently using PGD, where at each iteration the update rule is as follows:

$$\hat{\theta} \leftarrow \hat{\theta} - \eta F_T^{\mathrm{AL}}(\hat{\theta}),$$

where η is the step size/learning rate hyperparameter and $F_T^{\mathrm{AL}}(\cdot)$ is the penalized empirical field (5). Since the prediction of the i-th event at time t is determined by the estimated probability $w_{t-\tau:t-1}^{\mathrm{T}}\theta_i$ and a cut-off/threshold selected using the validation dataset, we can further relax the constraint $\mathbf{w}_{1:T}^{\mathrm{T}}\theta_i \leq \mathbf{1}_T$ and treat it as "score" instead of probability. Therefore, after the above update in each iteration, the projection onto the (relaxed) feasible region can be simply done by replacing all negative entries in $\hat{\theta}$ with zeros.

B Additional Numerical Experiments

We randomly generate ν and A using standard uniform distribution. Next, to ensure A stays in the feasible region Θ for the linear function case, we normalize each row to ensure it sums up to one. To be precise, we just update each entry in the row by dividing it with the row summation. To ensure A is DAG, we (i) first "sparse-ify" it by setting all entries smaller than the 95% percentile to zeros and (ii) next minimize the DAG characterization $h(A)$ using vanilla gradient descent (learning rate is 0.5 and we consider in total 5000 iterations). The reason of applying (i) is the highly non-convex optimization in (ii) — if we do not input a highly sparse graph, then we cannot shrink the DAG characterization to exactly zero with high probability. As for PGD approach to solve for VI estimator, we use 5×10^{-3} as the initial learning rate and decrease it by half every 2000 iterations (in total there are 6000 iterations). Here, we report additional results for completeness purpose. In particular, we report all four aforementioned metrics for Experiment 1 in Table 1, and plot the results for linear link function for Experiment 2 in Figure 5; please see the interpretation of those results in Section 5.

Table 1: Comparison of the mean (and standard deviation) of various performance metrics over 200 trials for different types of regularization. We report the matrix F-norm of the adjacency matrix estimation error (A err.), the ℓ_2 norm of the background intensity estimation error (ν err.), the "DAG-ness" measured by $h(A)$ and the Structural Hamming Distance (SHD).

Linear Link.

Dimension $d_1 = 10$, Time Horizon $T = 500$.

Penalty	None	Proposed	DAG	ℓ_1	Ada. ℓ_1
A err.	$0.3379_{(0.0988)}$	$0.2347_{(0.0698)}$	$0.3214_{(0.1009)}$	$0.2172_{(0.0674)}$	$0.2246_{(0.0872)}$
ν err.	$0.0970_{(0.0307)}$	$0.0661_{(0.0213)}$	$0.0822_{(0.0266)}$	$0.0636_{(0.0208)}$	$0.0584_{(0.0170)}$
$h(A)$	$0.0311_{(0.0163)}$	$0.0002_{(0.0011)}$	$0.0053_{(0.0041)}$	$0.0000_{(0.0000)}$	$0.0000_{(0.0000)}$
SHD	$44.3_{(5.24)}$	$12.74_{(4.73)}$	$32.7_{(6.34)}$	$31.77_{(5.48)}$	$26.22_{(6.32)}$

Dimension $d_1 = 20$, Time Horizon $T = 1000$.

Penalty	None	Proposed	DAG	ℓ_1	Ada. ℓ_1
A err.	$0.3764_{(0.0737)}$	$0.2035_{(0.0348)}$	$0.3382_{(0.0783)}$	$0.1820_{(0.0357)}$	$0.1819_{(0.0401)}$
ν err.	$0.1729_{(0.0329)}$	$0.0969_{(0.0253)}$	$0.1403_{(0.0259)}$	$0.0834_{(0.0225)}$	$0.0735_{(0.0170)}$
$h(A)$	$0.0573_{(0.0132)}$	$0.0000_{(0.0000)}$	$0.0086_{(0.0012)}$	$0.0000_{(0.0000)}$	$0.0000_{(0.0000)}$
SHD	$183.28_{(10.70)}$	$32.99_{(8.65)}$	$139.79_{(8.19)}$	$123.64_{(14.34)}$	$91.24_{(15.07)}$

Dimension $d_1 = 30$, Time Horizon $T = 1500$.

Penalty	None	Proposed	DAG	ℓ_1	Ada. ℓ_1
A err.	$0.4116_{(0.0448)}$	$0.1782_{(0.0239)}$	$0.3549_{(0.0491)}$	$0.1676_{(0.0213)}$	$0.1727_{(0.0226)}$
ν err.	$0.2486_{(0.0334)}$	$0.1104_{(0.0210)}$	$0.1995_{(0.0262)}$	$0.1013_{(0.0187)}$	$0.0987_{(0.0190)}$
$h(A)$	$0.0774_{(0.0113)}$	$0.0000_{(0.0000)}$	$0.0089_{(0.0011)}$	$0.0000_{(0.0000)}$	$0.0000_{(0.0000)}$
SHD	$411.81_{(12.18)}$	$73.43_{(11.03)}$	$306.52_{(11.69)}$	$277.25_{(19.99)}$	$199.15_{(26.89)}$

Exponential Link.

Dimension $d_1 = 10$, Time Horizon $T = 500$.

Penalty	None	Proposed	DAG	ℓ_1	Ada. ℓ_1
A err.	$0.4495_{(0.1457)}$	$0.2797_{(0.0914)}$	$0.4233_{(0.1452)}$	$0.2417_{(0.0620)}$	$0.2925_{(0.1167)}$
ν err.	$0.1061_{(0.0336)}$	$0.0720_{(0.0225)}$	$0.0889_{(0.0285)}$	$0.0666_{(0.0208)}$	$0.0644_{(0.0196)}$
$h(A)$	$0.0439_{(0.0243)}$	$0.0001_{(0.0006)}$	$0.0046_{(0.0039)}$	$0.0000_{(0.0000)}$	$0.0000_{(0.0000)}$
SHD	$43.75_{(5.00)}$	$12.96_{(5.11)}$	$30.77_{(5.71)}$	$31.66_{(5.28)}$	$24.7_{(6.44)}$

Dimension $d_1 = 20$, Time Horizon $T = 1000$.

Penalty	None	Proposed	DAG	ℓ_1	Ada. ℓ_1
A err.	$0.4731_{(0.0844)}$	$0.2118_{(0.0360)}$	$0.4206_{(0.0911)}$	$0.1898_{(0.0310)}$	$0.2136_{(0.0496)}$
ν err.	$0.1897_{(0.0385)}$	$0.0948_{(0.0201)}$	$0.1511_{(0.0298)}$	$0.0840_{(0.0187)}$	$0.0813_{(0.0174)}$
$h(A)$	$0.0799_{(0.0191)}$	$0.0002_{(0.0011)}$	$0.0087_{(0.001)}$	$0.0000_{(0.0000)}$	$0.0000_{(0.0000)}$
SHD	$183.83_{(10.18)}$	$34.59_{(10.79)}$	$134.53_{(7.80)}$	$125.7_{(12.75)}$	$90.8_{(17.29)}$

Dimension $d_1 = 30$, Time Horizon $T = 1500$.

Penalty	None	Proposed	DAG	ℓ_1	Ada. ℓ_1
A err.	$0.5048_{(0.0507)}$	$0.1841_{(0.0225)}$	$0.4277_{(0.0559)}$	$0.1738_{(0.0205)}$	$0.1888_{(0.0250)}$
ν err.	$0.2743_{(0.0361)}$	$0.1090_{(0.0170)}$	$0.2148_{(0.0274)}$	$0.1022_{(0.0168)}$	$0.1032_{(0.0162)}$
$h(A)$	$0.1090_{(0.0151)}$	$0.0000_{(0.0000)}$	$0.0089_{(0.0010)}$	$0.0000_{(0.0000)}$	$0.0000_{(0.0000)}$
SHD	$414.17_{(13.44)}$	$73.75_{(10.52)}$	$294.5_{(11.44)}$	$277.14_{(19.12)}$	$198.52_{(26.61)}$

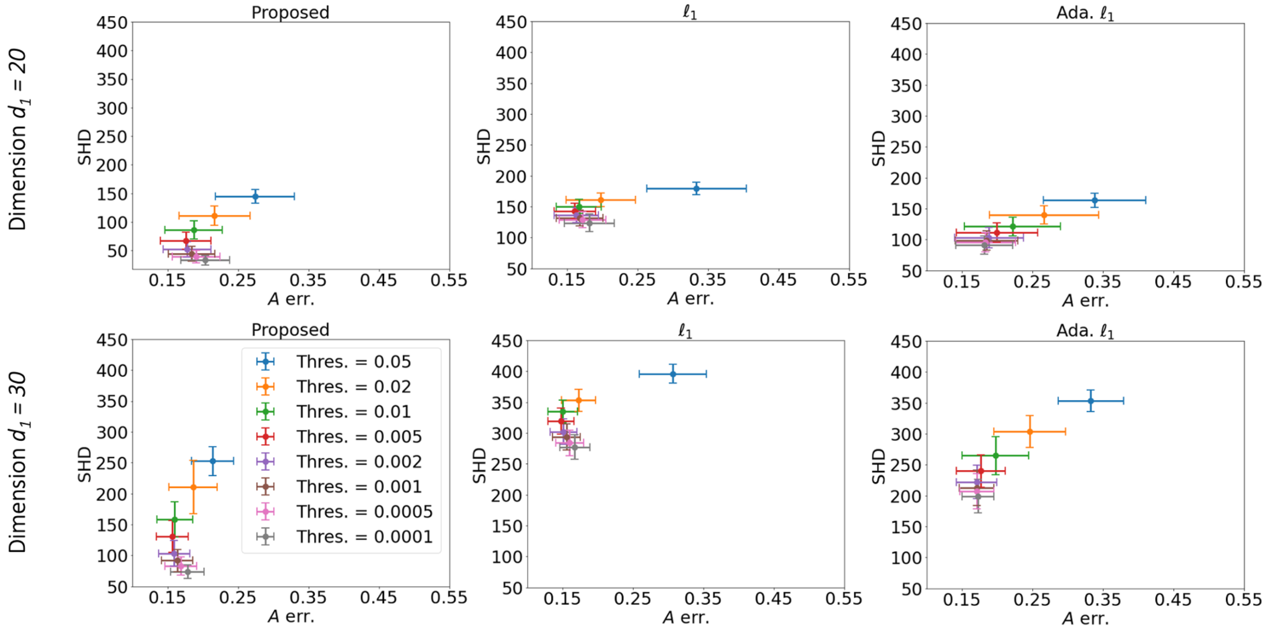

Figure 5: Effect of hyperparameter (continued). We consider the VI estimator with linear link function for completeness in this figure. We can observe similar patterns with Figure 4.