Causal Graph Identification Under Soft Intervention

Chen Peng and Urbashi Mitra
Department of Electrical & Computer Engineering
University of Southern California
Los Angeles, CA, US
Email: {cpeng732, ubli}@usc.edu

***Abstract*—In this paper, causal graph identification with natural observations as well as observations due to soft interventions is investigated. It is assumed that the graph is governed by linear structural equations; it is further assumed that both the causal topology and the distribution of interventions are unknown. The proposed causal graph learning approach is informed by prior work which proposed the decomposition of the problem into learning sub-graphs (all of the parents of a node) to learn the whole graph. A greedy algorithm that focuses on the reduction of the false negative rate (erroneously missing the presence of a causal relationship) is proposed. A sufficient condition is derived, under which the estimated graph is guaranteed to be free of false negatives, almost surely as the number of observations grows large. Numerical results indicate that the proposed scheme outperforms standard graph identification schemes by exploiting the sub-graph structure and by exploring a broader set of soft interventions. Compared to existing approaches, the proposed scheme achieves a 32% gain in false negative rate and a 62% gain in normalized Hamming distance.**

I. Introduction

Causal inference enables understanding of the underlying mechanisms in complex systems, with applications spanning social sciences [1], economics [2], biology [3], and machine learning [4]. Uncovering causal relationships facilitates the prediction of the effect of interventions and the design of effective policies, thus enhancing the understanding of system behavior. Causal structures are often represented by Bayesian networks in the form of directed acyclic graphs (DAGs).

Recent algorithms for causal discovery with purely observational data consider the exploitation of constraints or scores. Constraint-based methods, such as the inductive causation algorithm [5] and the PC algorithm [6], search for conditional independence among possible subsets of nodes. Alternatively, score-based methods (see *e.g.* [7]–[9]) evaluates different graph structures via their data-fitting quality. To improve scalability, some recent score-based approaches (see *e.g.* [10]–[12]) formulate graph identification as a continuous optimization problem by relaxing the DAG constraint. For example, DAGMA [12] employs an objective function consisting of the likelihood score and a log-determinant function to measure acyclicity.

A significant difficulty of pure observation-based causal discovery is the inherent ambiguity in distinguishing between competing causal structures [13]: multiple causal structures can exhibit similar statistical properties. Interventions (experiments) where certain variables are deliberately manipulated have shown promise in resolving these ambiguities [14]–[19]. Although interventions have gained recognition as a powerful tool in causal discovery, most existing approaches are based on hard (perfect) interventions that sever the causal links between the intervened node and its parents. While the hard intervention assumption simplifies causal modeling, in many applications, interventions do not completely eliminate causal effects (see *e.g.* [20]–[22]). Soft interventions, instead, are more aligned with real-world scenarios, where variables remain causally connected even under intervention.

Soft interventions present unique challenges and opportunities for causal discovery. Although they provide additional information for resolving ambiguities in causal structures, direct causal effects cannot be completely isolated. Causal discovery under standard soft intervention has been investigated recently [23]–[26]. In [24], we proposed a sub-graph learning scheme for the soft intervention setting that achieves high performance gain, for learning the entire graph, empirically. Moreover, it is observed that the false negative rate for link detection has a stronger impact on reward accumulation in causal bandits. Controlling for false negative or false positive rates results in new performance bounds and algorithms for causal discovery [27]. In contrast, the structure learning algorithm (GA-LCB-SL) with a graph error guarantee [26], identifies descendants of an intervened node by pairwise comparisons. This work is one of the few that consider soft interventions, similar to our framework, although it focuses on reward optimization rather than graph identification. We shall numerically compare our new method to GA-LCB-SL in the sequel.

While more complex graph identification methods offer strong performance, they are often challenging to analyze. In [8], a neighborhood selection scheme based on the minimum mean squared error (MMSE) and L_1 regularization is shown to be consistent for sparse graphs. A penalized maximum likelihood based estimator [28] has a convergence rate analysis. We underscore that these prior works did not consider the impact of false negatives or soft-interventions as we do herein.

Herein, we consider a more general type of soft intervention, which may alter both the topology and weights of a causal graph. Motivated by [24], we propose the *Causal Sub-graph Learning with Soft Intervention (CSL-SI)* scheme. By learning sub-graphs, the proposed scheme preserves low sample and computational complexities, while enabling asymptotic performance guarantees. The main contributions of this paper are:

1) We propose a sub-graph learning scheme with low

sample and computational complexities, tailored for the soft intervention setting.

2) The proposed algorithm is analyzed in the asymptotic regime and a sufficient condition is derived, under which the estimated graph is guaranteed to be free of false negatives, almost surely.
3) Based on the derived condition, the relationship between the design of soft interventions and structure identifiability is investigated and explained.
4) Numerical results indicate that the proposed scheme outperforms standard graph identification schemes by exploiting the sub-graph structure and by exploring a broader set of soft interventions.

II. Causal Graphical Model with Soft Interventions

To represent causal effects, consider a DAG with structure $(\mathcal{V}, \mathcal{B})$, where $\mathcal{V} = [N] \doteq \{1, \dots, N\}$ is the set of N nodes and $\mathcal{B}$ is the set of directed edges. The observational (without intervention) edge-weight matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ captures the strength of causal effects, where the (i, j)-th entry represents the weight of the edge $i \to j$.

To model causal effects under intervention, consider node-wise intervention, defined as

$$\boldsymbol{a} = (a_1, \dots, a_N)^\top \in \{0, 1\}^N, \tag{1}$$

where a_i represents whether node i is intervened (1) or not (0). Specifically, instead of hard interventions, we consider soft interventions, which do not necessarily cut off causal relationships between the intervened node and its parents, but change the incoming edges to the node. We denote the set of parents of node i by $\mathcal{P}_i(a_i)$, the estimated set of parents by $\hat{\mathcal{P}}_i(a_i)$. The set difference of the estimated and true parent sets is denoted by $\hat{\mathcal{P}}_i \backslash \mathcal{P}_i(a_i)$.

Further, we denote the interventional edge-weight matrix by $\boldsymbol{B}' \in \mathbb{R}^{N \times N}$, such that the post-intervention weight matrix $\boldsymbol{B}_{\boldsymbol{a}}$ can be constructed as

$$[\boldsymbol{B}_{\boldsymbol{a}}]_i = \mathbb{I}(a_i = 1)\boldsymbol{B}'_i + \mathbb{I}(a_i = 0)\boldsymbol{B}_i, \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $[\cdot]_i$ represents the i-th column of a matrix. The i-th column of the post-intervention weight matrix determines the set of parents of node i and how these parents causally influence node i.

As a result of the intervention, the vector of stochastic values associated with the nodes is represented by $\boldsymbol{x} \in \mathbb{R}^N$. The causal relationship among nodes is described by a linear structural equation model (LinSEM),

$$\boldsymbol{x} = (\boldsymbol{B}_{\boldsymbol{a}})^\top \boldsymbol{x} + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{e}$ is a vector of exogenous/noise variables. We assume that $\boldsymbol{e}$ contains independent elements, with known means and unknown variances represented by $\boldsymbol{\nu}$ and $\boldsymbol{\epsilon}$. The causal relationship described in (3) can be further manipulated, resulting in

$$\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{B}_{\boldsymbol{a}})^{-\top} \boldsymbol{e} \doteq (\boldsymbol{C}_{\boldsymbol{a}})^\top \boldsymbol{e}. \tag{4}$$

We define $\boldsymbol{C}_{\boldsymbol{a}}$ as the post-intervention **flow-weight matrix**, whose (i, j)-th entry represents the weight of the net flow from node i to j. In this way, each random variable x_i can be considered as a linear combination of exogenous variables in $\boldsymbol{e}$, weighted by the corresponding flow strength. Thus, under a specific intervention $\boldsymbol{a}$, $\boldsymbol{x}$ follows a multivariate distribution with mean and covariance defined as

$$\boldsymbol{\mu}(\boldsymbol{a}) \doteq \mathbb{E}[\boldsymbol{x}|\boldsymbol{a}] = (\boldsymbol{C}_{\boldsymbol{a}})^\top \boldsymbol{\nu}, \tag{5}$$

$$\boldsymbol{\Sigma}(\boldsymbol{a}) \doteq \mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{a}))(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{a}))^\top \middle| \boldsymbol{a}\right] \tag{6}$$

$$= (\boldsymbol{C}_{\boldsymbol{a}})^\top \mathbf{diag}(\boldsymbol{\epsilon})\ \boldsymbol{C}_{\boldsymbol{a}}. \tag{7}$$

Further, given an intervention selection strategy π, the flow weight matrix and the second moment matrix are defined as

$$\boldsymbol{C}^\pi \doteq \mathbb{E}_\pi[\boldsymbol{C}_{\boldsymbol{a}}] = \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}) \boldsymbol{C}_{\boldsymbol{a}}, \tag{8}$$

$$\boldsymbol{M}^\pi \doteq \mathbb{E}_\pi[\boldsymbol{x}\boldsymbol{x}^\top] = \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}) \big(\boldsymbol{\Sigma}(\boldsymbol{a}) + \boldsymbol{\mu}(\boldsymbol{a})\boldsymbol{\mu}(\boldsymbol{a})^\top\big). \tag{9}$$

III. The CSL-SI Algorithm

To estimate the causal structure, we employ the squared error as the score function, which is commonly used in the literature (see *e.g.* [10]–[12]). We shall see that our approach is tractable for analysis, while providing strong performance. With data collected under different soft interventions, the optimization problem can be formulated as

$$\min_{\hat{\boldsymbol{B}}, \hat{\boldsymbol{B}}'} \quad \sum_{\boldsymbol{a}} \left\| \boldsymbol{X}(\boldsymbol{a})\big(\boldsymbol{I} - \hat{\boldsymbol{B}}_{\boldsymbol{a}}\big) - \mathbf{1}\boldsymbol{\nu}^\top \right\|_F^2 \tag{10}$$

$$\text{s.t.} \quad \hat{\boldsymbol{B}}, \hat{\boldsymbol{B}}' \text{ represent DAGs}, \tag{11}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\boldsymbol{X}(\boldsymbol{a})$ represents the collection of samples under intervention $\boldsymbol{a}$. Further, we can rewrite the objective function as

$$\sum_{\boldsymbol{a}} \sum_{i \in [N]} \left\| \boldsymbol{X}(\boldsymbol{a}) \big[\boldsymbol{I} - \hat{\boldsymbol{B}}_{\boldsymbol{a}}\big]_i - \mathbf{1}\nu_i \right\|_2^2$$

$$= \sum_{i \in [N]} \sum_{a_i \in \{0,1\}} \left\| \boldsymbol{X}(\boldsymbol{a}) \big[\boldsymbol{I} - \hat{\boldsymbol{B}}_{\boldsymbol{a}}\big]_i - \mathbf{1}\nu_i \right\|_2^2 \tag{12}$$

$$= \sum_{i \in [N]} \Big(\left\| \boldsymbol{X}(a_i = 0) \big[\boldsymbol{I} - \hat{\boldsymbol{B}}\big]_i - \mathbf{1}\nu_i \right\|_2^2 + \left\| \boldsymbol{X}(a_i = 1) \big[\boldsymbol{I} - \hat{\boldsymbol{B}}'\big]_i - \mathbf{1}\nu_i \right\|_2^2 \Big). \tag{13}$$

The reformulation is enabled by the principle of independent mechanisms [13], which states that intervention on one mechanism does not affect the others. As in [24], the problem is decomposed into sub-problems of learning the local structures. Sub-graph learning is essentially the identification of the parents of a node. Moreover, the reformulation separates the estimation of the observational and interventional matrices into two independent tasks. In the sequel, we focus on learning the observational edge-weights and omit the intervention index a_i for brevity. The same strategy can be used to learn the interventional weight matrix, using corresponding samples.

Although the objective function can be decomposed, enforcing the DAG nature is a global constraint. Since the number of possible DAGs is super-exponential in the number of nodes [13], searching over this space is computationally infeasible. To avoid an exhaustive search, we propose a greedy approach. The graph is initialized with all possible edges, which minimizes the objective, but violates the DAG constraint. Rejection of an edge with the minimum increase to the objective is performed in every subsequent step, until the estimated graph becomes a DAG.

Specifically, to calculate the increase in the objective after rejection of an edge $i \to j$ with $i \in \hat{\mathcal{P}}_j$, we compare the squared residual norms with and without node i as a potential parent of node j. With the whole set of estimated parents, the estimated weights and residuals are given by MMSE estimation as

$$\hat{\boldsymbol{B}}_{j,\hat{\mathcal{P}}_j} = \left(\boldsymbol{X}_{\hat{\mathcal{P}}_j}^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j}\right)^{-1}\boldsymbol{X}_{\hat{\mathcal{P}}_j}^\top(\boldsymbol{X}_j - \mathbf{1}\nu_j), \tag{14}$$

$$\boldsymbol{r}_j(\hat{\mathcal{P}}_j) = \left[\boldsymbol{I} - \boldsymbol{X}_{\hat{\mathcal{P}}_j}\left(\boldsymbol{X}_{\hat{\mathcal{P}}_j}^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j}\right)^{-1}\boldsymbol{X}_{\hat{\mathcal{P}}_j}^\top\right](\boldsymbol{X}_j - \mathbf{1}\nu_j). \tag{15}$$

Note that $\boldsymbol{X}_{\mathcal{P}}$ represents the sub-matrix consisting of columns corresponding to the nodes in the set $\mathcal{P}$. Denote the projection matrices onto the column and left null space of $\boldsymbol{X}_{\mathcal{P}}$ by

$$\Phi(\boldsymbol{X}_{\mathcal{P}}) \doteq \boldsymbol{X}_{\mathcal{P}}\left(\boldsymbol{X}_{\mathcal{P}}^\top \boldsymbol{X}_{\mathcal{P}}\right)^{-1}\boldsymbol{X}_{\mathcal{P}}^\top, \tag{16}$$

$$\Phi^{\mathcal{C}}(\boldsymbol{X}_{\mathcal{P}}) \doteq \boldsymbol{I} - \boldsymbol{X}_{\mathcal{P}}\left(\boldsymbol{X}_{\mathcal{P}}^\top \boldsymbol{X}_{\mathcal{P}}\right)^{-1}\boldsymbol{X}_{\mathcal{P}}^\top. \tag{17}$$

which allow us to rewrite the residual vector as

$$\begin{aligned}\boldsymbol{r}_j(\hat{\mathcal{P}}_j) &= \left[\boldsymbol{I} - \Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j}\big)\right](\boldsymbol{X}_j - \mathbf{1}\nu_j) &(18)\\ &= \left[\boldsymbol{I} - \Phi\big(\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\setminus i}\big)\boldsymbol{X}_i\big)\right]\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\setminus i}\big)(\boldsymbol{X}_j - \mathbf{1}\nu_j) &(19)\\ &= \Phi^{\mathcal{C}}\big(\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\setminus i}\big)\boldsymbol{X}_i\big)\cdot \boldsymbol{r}_j(\hat{\mathcal{P}}_j\setminus i). &(20)\end{aligned}$$

Essentially, the residual vector is successively projected onto orthogonal subspaces [29]. Lastly, we define the *normalized difference in squared residual norms* as

$$\Delta_{ij} \doteq \left(\left\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j\setminus i)\right\|_2^2 - \left\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j)\right\|_2^2\right)/t_j, \tag{21}$$

where t_j denotes the number of time slots, or equivalently, number of samples of interest. Intuitively, a small Δ_{ij} indicates that keeping the edge (i,j) does not strongly improve the estimation quality of the value of node j. The edge with the smallest Δ_{ij} is removed from the edge set in each step to minimize the increase of the objective. The complete algorithm is provided in Algorithm 1.

Since the proposed algorithm stops once the estimated graph satisfies the DAG constraint, it does not depend on any predefined threshold. Nonetheless, there exists an implicit threshold on the objective induced by the stopping time. Since the threshold determines the number of edges to be rejected, it corresponds to a specific balance between the false negative rate (FNR) and false positive rate (FPR), defined as

$$\mathrm{FNR} \doteq \frac{\sum_{i,j}\mathbb{I}\left(B_{ij}\neq 0, \hat{B}_{ij} = 0\right)}{\sum_{i,j}\mathbb{I}\left(B_{ij}\neq 0\right)}, \tag{22}$$

$$\mathrm{FPR} \doteq \frac{\sum_{i,j}\mathbb{I}\left(B_{ij} = 0, \hat{B}_{ij} \neq 0\right)}{\sum_{i,j}\mathbb{I}\left(B_{ij} = 0\right)}. \tag{23}$$

Algorithm 1 The CSL-SI Algorithm

```
Require: The set of nodes $\mathcal{V}$ and node values $\boldsymbol{X}$.
Initialize the estimated edge set to include all possible directed edges: $\hat{\mathcal{B}} = \{(i,j) \mid i \neq j, \forall i,j \in \mathcal{V}\}$.
while $(\mathcal{V}, \hat{\mathcal{B}})$ is not a DAG do
    Compute differences in residual norms $\Delta_{ij}, \forall (i,j) \in \hat{\mathcal{B}}$, by MMSE estimation, with $\boldsymbol{X}$.
    Find the edge $(i,j) = \arg\min_{i,j}\Delta_{ij}$ and remove it from the estimated edge set $\hat{\mathcal{B}}$.
end while
Compute $[\hat{B}]_{ij}, \forall (i,j) \in \hat{\mathcal{B}}$, by linear MMSE estimation, with $\boldsymbol{X}$. For any $(i,j) \notin \hat{\mathcal{B}}$, set $\hat{B}_{ij} = 0$.
return Estimated weight matrix $\hat{\boldsymbol{B}}$.
```

Since the proposed algorithm does not reject edges once the estimate becomes a DAG, the corresponding implicit threshold cannot be reduced without violating the DAG constraint. Thus, one can interpret the proposed scheme as a greedy approach for minimizing the FNR over possible DAGs. Furthermore, Theorem 1 shows that if a sufficient condition is met, the proposed algorithm asymptotically achieves zero false negatives.

IV. Asymptotic Analysis

We next show that our proposed algorithm, asymptotically, achieves zero false negatives. To ensure that the estimated causal graph contains no false negative error, a sufficient condition is

$$\min_{j\in[N], i\in\mathcal{P}_j}\Delta_{ij} > \min_{j\in[N], i\in\hat{\mathcal{P}}_j\setminus\mathcal{P}_j}\Delta_{ij}, \tag{24}$$

which requires that every true parent has a larger normalized difference in residual norms than at least one non-parent. Thus, this condition guarantees that true parents will not be rejected if non-parents exist in the estimate. Note that the normalized differences in both cases depend on the intervention selection strategy π, which determines distribution of the samples. A good strategy results in large Δ_{ij} for true parents and small Δ_{ij} for non-parents.

To understand when Equation (24) is satisfied, we examine the range of Δ_{ij} by evaluating the asymptotic limits of the terms in (24).

Lemma 1. *If $\mathcal{P}_j \subseteq \hat{\mathcal{P}}_j$ and $i \in \hat{\mathcal{P}}_j\setminus\mathcal{P}_j$, Δ_{ij} converges almost surely to a limit for sufficiently large t_j,*

$$\begin{aligned}\Delta_{ij} &\xrightarrow{\text{a.s.}} \frac{\left[C_{ji}^{\pi}\epsilon_j^2 - \epsilon_j^2\,\boldsymbol{C}_{j,\hat{\mathcal{P}}_j\setminus i}^{\pi}\big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\setminus i}^{\pi}\big)^{-1}\boldsymbol{M}_{\hat{\mathcal{P}}_j\setminus i,i}^{\pi}\right]^2}{M_{ii}^{\pi} - \boldsymbol{M}_{i,\hat{\mathcal{P}}_j\setminus i}^{\pi}\big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\setminus i}^{\pi}\big)^{-1}\boldsymbol{M}_{\hat{\mathcal{P}}_j\setminus i,i}^{\pi}}\\ &\doteq \Delta_{ij}^{*}(\mathrm{F\ parent}).\end{aligned} \tag{25}$$

Proof Sketch: The proof exploits properties of the projection embedded in MMSE estimation and the fact that observations of the parents for node j reside in the subspace resulting from the projection. The strong law of large numbers is invoked as well as the continuous mapping theorem. The full proof is provided in [30].

There are several key remarks regarding Lemma 1.

Remark 1. *As node i is not a parent of node j, a small Δ_{ij} is desired in order to reject the edge $i \to j$. Since the limit is a function of the intervention selection strategy π, we can minimize Δ_{ij} by optimizing the strategy.*

Remark 2. *When node i is not a descendant of node j, the flow weight in the numerator is zero, $C_{ji}^{\pi} = 0$. In this case, the magnitude of Δ_{ij} is determined by the second term in the numerator, which essentially measures the correlation between node i and other potential parents.*

Remark 3. *When node i is a descendant of node j under certain interventions, a net flow weight C_{ji}^{π} exists and often it is larger than the second term. However, the causal confusion caused by descendants can be mitigated by applying a diverse set of soft interventions. Specifically, we have*

$$\left|\sum_{\boldsymbol{a}} \pi(\boldsymbol{a}) \left[C_{\boldsymbol{a}}\right]_{ji}\right| \leq \max_{\boldsymbol{a}} \left|\left[C_{\boldsymbol{a}}\right]_{ji}\right|, \tag{26}$$

which states that the averaged flow could be much smaller than the strongest causal flow under a particular intervention.

Lemma 2. *If $\mathcal{P}_j \subseteq \hat{\mathcal{P}}_j$ and $i \in \mathcal{P}_j$, Δ_{ij} converges almost surely to a limit for sufficiently large t_j, that is,*

$$\Delta_{ij} \xrightarrow{\text{a.s.}} \Delta_{ij}^{*}(\text{T parent}) \doteq \frac{\left[B_{ij} M_{ii}^{\pi} - \left(B_{ij} \boldsymbol{M}_{i,\hat{\mathcal{P}}_j \backslash i}^{\pi} + \epsilon_j^2 \boldsymbol{C}_{j,\hat{\mathcal{P}}_j \backslash i}^{\pi}\right)\left(\boldsymbol{M}_{\hat{\mathcal{P}}_j \backslash i}^{\pi}\right)^{-1} \boldsymbol{M}_{\hat{\mathcal{P}}_j \backslash i, i}^{\pi}\right]^2}{M_{ii}^{\pi} - \boldsymbol{M}_{i,\hat{\mathcal{P}}_j \backslash i}^{\pi}\left(\boldsymbol{M}_{\hat{\mathcal{P}}_j \backslash i}^{\pi}\right)^{-1} \boldsymbol{M}_{\hat{\mathcal{P}}_j \backslash i, i}^{\pi}} \tag{27}$$

Proof Sketch: As with the proof of Lemma 1, we use the fact that the observations at the parents for node j reside in the subspace resulting from the projection. We invoke the strong law of large numbers and the continuous mapping theorem for our asymptotic result. The full proof is provided in [30].

Remark 4. *As node i is a parent of node j, a large Δ_{ij} is desired in order to preserve the true edge $i \to j$. Maximization of Δ_{ij} can be achieved by optimization of the intervention strategy π as the limit is a function of π.*

Remark 5. *Equation* (27) *shows that the normalized difference Δ_{ij} is positively correlated with the strength of the causal link, B_{ij}. However, the strength of this true link can be weakened by the correlation between node i and other potential parents, as suggested by the second term in the numerator in* (27)*. To determine the true strength of the link, having a diverse set of interventions is helpful: even if strong correlation exists under certain interventions, the averaged correlation can be much weaker, due to the existence of weak or reversed correlation under other interventions.*

Finally, combination of the Lemmas enables us to develop the following sufficient condition for ensuring identification free of false negative errors.

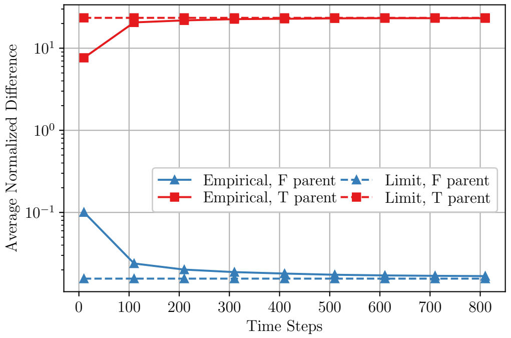

Fig. 1. Limits on Δ_{ij} from Lemmas 1, 2 and empirical values.

Theorem 1. *For sufficiently large t_j, $\forall j$, the proposed algorithm ensures no false negative error if the following condition is satisfied,*

$$\min_{j \in [N], i \in \mathcal{P}_j} \Delta_{ij}^{*}(\text{T parent}) > \min_{j \in [N], i \in \hat{\mathcal{P}}_j \backslash \mathcal{P}_j} \Delta_{ij}^{*}(\text{F parent}). \tag{28}$$

Proof: Since the algorithm starts with an estimated edge set that includes all possible edges, initially, no false negative error exists. In subsequent steps, if $\forall j$, $\mathcal{P}_j \subseteq \hat{\mathcal{P}}_j$ is satisfied before edge rejection, Lemma 1, Lemma 2 and the condition (28) guarantee that

$$\min_{j \in [N], i \in \mathcal{P}_j} \Delta_{ij} > \min_{j \in [N], i \in \hat{\mathcal{P}}_j \backslash \mathcal{P}_j} \Delta_{ij}, \tag{29}$$

which suggests that the minimum Δ_{ij} corresponds to a nonexistent causal edge. Thus, after rejecting this edge, $\forall j$, $\mathcal{P}_j \subseteq \hat{\mathcal{P}}_j$ remains satisfied. Applying mathematical induction completes the proof. □

V. Numerical Results

In this section, we numerically evaluate the performance of the proposed CSL-SI algorithm, for a linear structural equation model and soft interventions. For each Monte Carlo run, the causal structure is randomly generated, with the edge weights randomly sampled from the uniform distribution $\mathcal{U}(-2, 2)$. The exogenous variables are independently sampled in each time step from the Gaussian distribution $\mathcal{N}(1, 1)$. For each set of parameters, we repeat the Monte Carlo run $M = 100$ times.

To better understand the behavior of the normalized difference, we compute both the empirical values and the limits provided in Lemmas 1 and 2. We take $N = 10$ and average over 100 Monte Carlo runs, with results presented in Fig. 1. In each time step, an intervention is conducted and a vector of node values is collected. As expected, empirical values converge to the corresponding limits, and true parents have larger normalized difference Δ_{ij} compared with false parents.

To evaluate graph identification performance, we consider the DAGMA algorithm [12] and the GA-LCB-SL algorithm [26] for comparison. The DAGMA algorithm estimates the weight matrix by minimizing the following objective function,

$$-\log \mathcal{L}(\boldsymbol{X}; \hat{\boldsymbol{B}}) + \beta_1 \left\|\hat{\boldsymbol{B}}\right\|_1 - \log\det(\beta_2 \boldsymbol{I} - \hat{\boldsymbol{B}} \circ \hat{\boldsymbol{B}}) + N \log \beta_2,$$

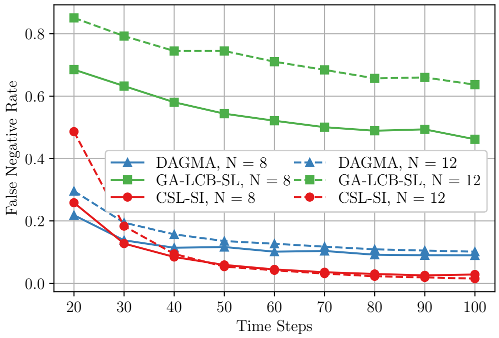

Fig. 2. False negative rate as a function of time steps.

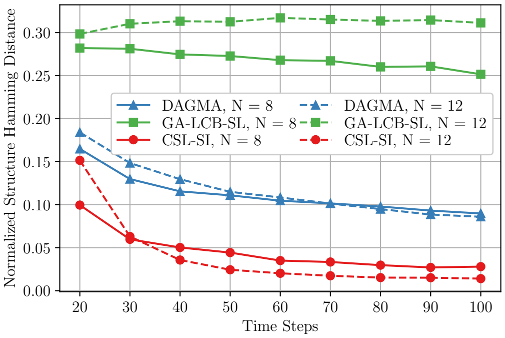

Fig. 3. Normalized structural Hamming distance as a function of time steps.

where $\mathcal{L}(\cdot)$ denotes the likelihood function, β_1, β_2 are weights for the penalty and log-determinant terms and the operator $\circ$ denotes the Hadamard product. Since DAGMA identifies entire graphs, half of the interventions are set as $\mathbf{0}$ to learn $\boldsymbol{B}$ while the other half are set as $\mathbf{1}$ for $\boldsymbol{B}'$. The GA-LCB-SL algorithm first estimates the descendant set of each node as

$$\left\{j \in [N] : \left|\hat{\mu}_j(\mathbf{0}) - \hat{\mu}_j(a_i = 1, a_k = 0, \forall k)\right| > \frac{\eta}{2}\right\}, \quad (30)$$

where η is the threshold given by the regularity assumption. The empirical mean for node j for the cases of with and without atomic interventions on node i are compared to determine whether node j is a descendant of node i. Then, the parent set of each node is determined by the Lasso regression [31] on the ancestors of that node.

Figure 2 plots the FNR (defined in (22)) as a function of time steps, for two different graph sizes. As expected, FNR of all algorithms decrease as more samples are collected. For $N = 8$, the average FNR achieved by the proposed CSL-SI scheme is lower than the DAGMA algorithm and the GA-LCB-SL algorithm by 34.8% and 85.9% respectively. For $N = 12$, the gain in FNR becomes 29.4% and 85.4%, compared with DAGMA and GA-LCB-SI. We also observe that, initially, the FNR of CSL-SI is higher compared to DAGMA, but decreases rapidly and becomes almost zero at the end. This phenomenon confirms that the proposed scheme has low sample complexity and vanishing FNR.

In Fig. 3, the normalized Hamming distance is plotted as a function of time steps, which is defined as

$$d(\boldsymbol{B}, \hat{\boldsymbol{B}}) \doteq \sum_{i,j} \big[\mathbb{I}(B_{ij} = 0, \hat{B}_{ij} \neq 0) + \mathbb{I}(B_{ij} \neq 0, \hat{B}_{ij} = 0) - \mathbb{I}(B_{ij} \neq 0, \hat{B}_{ij} = 0, \hat{B}_{ji} \neq 0)\big]/N^2. \quad (31)$$

Compared to DAGMA and GA-LCB-SI, the average performance gains provided by the CSL-SI scheme are 59.6% and 83.2% for $N = 8$, 66.2% and 87.3% for $N = 12$. An interesting observation is that, both DAGMA and GA-LCB-SI perform better on a smaller graph ($N = 8$), while the proposed scheme performs better on a larger graph. This result suggests that as N increases, the number of mistaken edges increases more slowly than the normalization factor, which is quadratic in N.

The observed gain in both FNR and normalized structural Hamming distance for CSL-SI has a two-fold explanation. First, CSL-SI leverages the power of sub-graph learning to achieve lower sample complexity, which partially explains the improvement over DAGMA. Second, CSL-SI fully exploits soft interventions to reduce causal ambiguities. As discussed in the Remarks, enforcing a diverse set of soft interventions can improve sub-structures identification. Specifically, the number of utilized soft interventions is 2 for DAGMA, $N+1$ for GA-LCB-SI, and 2^N for CSL-SI. For GA-LCB-SI, we emphasize that although pair-wise comparison in means (see (30)) enables a finite sample performance analysis, it also imposes restrictions on intervention selection, degrading performance in the limited data region. Moreover, since GA-LCB-SI is designed for reward optimization, which may not require an accurate causal graph, the comparison is not completely fair. However, as noted previously, GA-LCB-SI does employ soft interventions in contrast to most prior work.

VI. Conclusions

In this paper, we investigated causal graph identification under soft intervention. A sub-graph learning based greedy algorithm is proposed, focusing on the reduction of the false negative rate. Further, we derived a sufficient condition, under which the proposed algorithm ensures no false negative error, almost surely as the number of observations grows large. Numerical results show that the proposed algorithm outperforms existing approaches by exploiting the sub-graph structure and exploring a broader set of soft interventions.

Acknowledgments

This work has been funded by one or all of the following grants: ONR N00014-22-1-2363, ARO W911NF2410094, ARO W911NF1910269, NSF CCF-2311653, NSF DBI-2412522, NSF A25-0139-S009, UCSD/NSF CCF-2148313.

References

[1] M. Gangl, “Causal inference in sociological research,” *Annual review of sociology*, vol. 36, no. 1, pp. 21–47, 2010.

[2] H. R. Varian, “Causal inference in economics and marketing,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7310–7315, 2016.

[3] B. Shipley, *Cause and correlation in biology: A user’s guide to path analysis, structural equations and causal inference with R*. Cambridge University Press, 2016.

[4] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, “Causal inference and counterfactual prediction in machine learning for actionable healthcare,” *Nature Machine Intelligence*, vol. 2, no. 7, pp. 369–375, 2020.

[5] J. Pearl, *Causality*. Cambridge University Press, 2009.

[6] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT Press, 2001.

[7] D. M. Chickering, “Optimal structure identification with greedy search,” *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[8] N. Meinshausen and P. Bühlmann, “High-dimensional graphs and variable selection with the lasso,” *The Annals of Statistics*, pp. 1436–1462, 2006.

[9] M. Koivisto, “Advances in exact bayesian structure discovery in bayesian networks,” in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 2006, pp. 241–248.

[10] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, “Dags with no tears: Continuous optimization for structure learning,” *Advances in neural information processing systems*, vol. 31, 2018.

[11] I. Ng, A. Ghassami, and K. Zhang, “On the role of sparsity and dag constraints for learning linear dags,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 943–17 954, 2020.

[12] K. Bello, B. Aragam, and P. Ravikumar, “Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 8226–8239, 2022.

[13] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[14] F. Eberhardt, “Almost optimal intervention sets for causal discovery,” *arXiv preprint arXiv:1206.3250*, 2012.

[15] J. M. Mooij, S. Magliacane, and T. Claassen, “Joint causal inference from multiple contexts,” *Journal of machine learning research*, vol. 21, no. 99, pp. 1–108, 2020.

[16] P. Lippe, T. Cohen, and E. Gavves, “Efficient neural causal discovery without acyclicity constraints,” in *International Conference on Learning Representations*, 2021.

[17] L. Lorch, S. Sussex, J. Rothfuss, A. Krause, and B. Schölkopf, “Amortized inference for causal structure learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 104–13 118, 2022.

[18] C. Qiu and K. Yang, “Interventional causal structure discovery over graphical models with convergence and optimality guarantees,” *IEEE Transactions on Network Science and Engineering*, 2024.

[19] B. Varıcı, D. Katz, D. Wei, P. Sattigeri, and A. Tajer, “Interventional causal discovery in a mixture of dags,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 86 574–86 601, 2024.

[20] R. Bonneau, M. T. Facciotti, D. J. Reiss, A. K. Schmid, M. Pan, A. Kaur, V. Thorsson, P. Shannon, M. H. Johnson, J. C. Bare *et al.*, “A predictive model for transcriptional control of physiology in a free living cell,” *Cell*, vol. 131, no. 7, pp. 1354–1365, 2007.

[21] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, “Counterfactual reasoning and learning systems: The example of computational advertising.” *Journal of Machine Learning Research*, vol. 14, no. 11, 2013.

[22] N. Meinshausen, A. Hauser, J. M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann, “Methods for causal inference from gene perturbation experiments and validation,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7361–7368, 2016.

[23] M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim, “Characterization and learning of causal graphs with latent variables from soft interventions,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] C. Peng, D. Zhang, and U. Mitra, “Asymmetric graph error control with low complexity in causal bandits,” *IEEE Transactions on Signal Processing*, pp. 1–15, 2025.

[25] J. Zhang, K. Greenewald, C. Squires, A. Srivastava, K. Shanmugam, and C. Uhler, “Identifiability guarantees for causal disentanglement from soft interventions,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[26] Z. Yan and A. Tajer, “Linear causal bandits: Unknown graph and soft interventions,” *arXiv preprint arXiv:2411.02383*, 2024.

[27] J. Shaska and U. Mitra, “Neyman-pearson causal inference,” in *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024, pp. 1269–1274.

[28] J. Peters and P. Bühlmann, “Identifiability of gaussian structural equation models with equal error variances,” *Biometrika*, vol. 101, no. 1, pp. 219–228, 2014.

[29] G. Strang, *Introduction to linear algebra*. SIAM, 2022.

[30] C. Peng and U. Mitra, “Supplementary file for causal graph identification under soft intervention,” https://github.com/CalixPeng/Causal_Disc_SI, 2025.

[31] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.