Intervention Efficient Algorithms for Approximate Learning of Causal Graphs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of learning the causal relationships between a set of observed variables in the presence of latents, while minimizing the cost of interventions on the observed variables. We assume access to an undirected graph G on the observed variables whose edges represent either all direct causal relationships or, less restrictively, a superset of causal relationships (identified e.g., via conditional independence tests or a domain expert). Our goal is to recover the directions of all causal or ancestral relations in G, via a minimum cost set of interventions.

It is known that constructing an exact minimum cost intervention set for an arbitrary graph G is NP-hard. We further argue that, conditioned on the hardness of approximate graph coloring, no polynomial time algorithm can achieve an approximation factor better than the trivial $\Theta(\log n)$, where n is the number of observed variables in G. To overcome this limitation, we introduce a bi-criteria approximation goal that lets us recover the directions of all but ϵn^2 edges in G, for some specified error parameter $\epsilon > 0$. Under this relaxed goal, we give polynomial time algorithms that achieve intervention cost within a small constant factor of the optimal. Our algorithms combine work on efficient intervention design and the design of low-cost *separating set systems*, with ideas from the literature on graph property testing.

1 Introduction

Discovering causal relationships is one of the fundamental problems of causality [24]. In this paper, we study the problem of *learning a causal graph* where we seek to identify all the causal relations between variables in our system (nodes of the graph). It has been shown that, under certain assumptions, observational data alone lets us recover the existence of a causal relationship between, but not the direction of all relationships. To recover the direction, we use the notion of an intervention (or an experiment) described in Pearl's Structural Causal Models (SCM) framework [24].

An intervention requires us to fix a subset of variables to each value in their domain, inducing a new distribution on the free variables. For example, we may intervene to require that some patients in a study follow a certain diet and others do not. As performing interventions is costly, a widely studied goal is to find a minimum set of interventions for learning the causal graph [26]. This goal however does not address the fact that interventions may have different costs. For example, interventions that fix a higher number of variables will be more costly. Additionally, there may be different intervention costs associated with different variables. For example, in a medical study, intervening on certain variables might be impractical or unethical. Hyttinen et al. [16] address the need for such cost models and give results for the special case of learning complete graphs when the cost of an intervention is equal to the number of variables it intervenes on. Generalizing this notion, we study a *linear cost model* where the cost of an intervention on a set of variables is the sum of (possibly non-uniform) costs for each variable in the set. This model was first introduced in [18] and has received recent attention [2, 21].

Significant prior work on efficient intervention design assumes *causal sufficiency*, i.e., there are no unobserved (latent) variables in the system. In this setting, there is an exact characterization of the interventions required to learn the causal graph, using the notion of *separating set systems* [8, 26].

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Recently, the problem of learning the causal graph with latents using a minimum number of interventions has received considerable attention with many known algorithms that depend on various properties of the underlying causal graph [2, 19, 20]. However, the intervention sets used by these algorithms contain a large number of variables, often as large as $\Omega(n)$, where n is the number of observable variables. Thus, they are generally not efficient in the linear cost model. Some work has considered efficient intervention design in the linear cost model for recovering the ancestral graph containing all indirect causal relations [2]. Other algorithms such as IC* and FCI aim to learn the causal graph in the presence of latents using only observational data; however, they can only learn some causal relations not the full graph [30, 31].

Our Results. We address these shortcomings by considering two settings: in the first we assume that we are given an undirected graph that contains all causal relations between observable variables, but must identify their directions. This undirected graph may be obtained e.g., by running algorithms that identify conditional dependencies and consulting a domain expert to identify causal links. In the second setting, we study a relaxation where we are given a supergraph H of G containing all causal edges and other additional edges which need not be causal. The second setting is less restrictive, modeling the case where we can ask a domain expert or use observational data to identify a superset of possible causal relations. From H we seek to recover edges of the ancestral graph of G, a directed graph containing all ancestral causal relations between the observable variables.

Depending on the method by which H is obtained, it may have special properties that can be leveraged for efficient intervention design. E.g., if we use FCI/IC* [30] to recover a partial ancestral graph from observational data, the remaining undirected edges form a chordal graph [33]. Past work has also considered the worst case when H is the complete graph [2]. In this work we do not assume anything about how H is obtained and thus give results holding for general graphs.

In both settings, we show a connection to separating set systems – to solve the recovery problems it is necessary and sufficient to use an intervention set that corresponds to a (strongly) separating set system on the given graph (either the undirected causal graph G or the supergraph H). A separating set system in one in which each pair of nodes connected by an edge is separated by at least one intervention – one variable is intervened on and the other is free. A strongly separating set system requires that every connected pair is separated by two interventions – where one of the nodes is intervened on, that the other is free in.

Unfortunately, finding a minimum cost (strongly) separating set system for an arbitrary graph G is known to be NP-Hard [16, 21]. We give trivial algorithms that achieve $O(\log n)$ approximation and further argue that, conditioned on the hardness of approximate graph coloring, no polynomial time algorithm can achieve an approximation factor of $o(\log n)$, where n is the number of observed variables. To overcome this limitation, we introduce a *bi-criteria approximation* goal that lets us recover all but ϵn^2 edges in the causal or ancestral graph. For this goal, it suffices to use a relaxed notion of a set system, which we show can be found efficiently using ideas from graph property testing literature [11].

In the setting where we are given the causal edges in G and must recover their directions, we give a polynomial time algorithm that finds a set of interventions from which we can recover all but ϵn^2 edges and with cost at most $2 + \gamma$ times the optimal cost for learning the full graph, where $\epsilon, \gamma > 0$ are specified error parameters. In the setting of ancestral graph recovery we similarly show how to recover all but ϵn^2 edges with intervention cost at most $4 + \gamma$ times the optimal cost for recovering all edges. Our result improves upon [2] which gives a 2-approximation to the minimum cost strongly separating set system assuming the worst case when the supergraph H is a complete graph. Their algorithm does not translate to an approximation guarantee better than $\Omega(\log n)$ for general graphs.

Other Related Work. Assuming causal sufficiency (no latents), most work focuses on recovering causal relationships based on just observational data. Examples include algorithms like *IC* [24] and *PC* [30], which have been wide studied [13–15, 22, 27]. It is well-known that to disambiguate a causal graph from an equivalence class of possible causal structures, interventional, rather than just observational data is required [8, 9, 12]. There is a growing body of recent work devoted to minimizing the number of interventions [19, 20, 26] and costs of intervention [18, 21]. Since causal sufficiency is often too strong an assumption [4], many algorithms avoiding the causal sufficiency assumption, such as IC* [31] and FCI [30], and using just observational data have been developed. There is a growing interest in optimal intervention design in this setting [16, 19, 20, 23, 29].

2 Preliminaries

Causal Graph Model. Following the SCM framework [24], we represent a set of random variables by $V \cup L$ where V contains the endogenous (observed) variables that can be measured and L contains the exogenous (latent) variables that cannot be measured. We define a directed causal graph $\mathcal{G} = \mathcal{G}(V \cup L, \mathcal{E})$ on these variables where an edge corresponds to a causal relation between the corresponding variables: a directed edge (v_i, v_j) indicates that v_i causes v_j.

We assume that all causal relations belong to one of two categories : (i) $E \subseteq V \times V$ containing direct causal relations between the observed variables and (ii) $E_L \subseteq L \times V$ containing relations from latents to observables. Thus, the full edge set of our causal graph is $\mathcal{E} = E \cup E_L$. We also assume that every latent $l \in L$ influences exactly two observed variables, i.e., $(l, u), (l, v) \in E_L$ and no other edges are incident on l following. This *semi-Markovian* assumption is widely used in prior work [19, 28] (see Appendix A for a more detailed discussion). Let $G(V, E)$ denote the subgraph of $\mathcal{G}$ restricted to observable variables, referred to as the observable graph. Let $\mathrm{Anc}(G)$ denote the *ancestral graph* of G [25]. Here, we consider a restricted version of $\mathrm{Anc}(G)$ which contains only directed edges (v_i, v_j) if there is a directed path from v_i to v_j in G (equivalently in $\mathcal{G}$ due to the semi-Markovian assumption). Throughout we denote $n = |V|$.

Intervention Sets. Our primary goal is to recover either G or $\mathrm{Anc}(G)$ via interventions on the observables. We assume the ability to perform an *intervention* on a set of variables $S \subseteq V$ which fixes $S = s$ for each s in the domain of S. We then perform a conditional independence test answering for all v_i, v_j *"Is v_i independent of v_j in the interventional distribution* $\mathrm{do}(S = s)$*?* Here $\mathrm{do}(S = s)$ uses Pearl's do-notation to denote the interventional distribution when the variables in S are fixed to s. An intervention set is a collection of subsets $\mathcal{S} = \{S_1, \ldots, S_m\}$ that we intervene on in order to recover edges of the observable or ancestral graph. It will also be useful to associate a matrix $L \in \{0, 1\}^{n \times m}$ with the collection where the ith column is the characteristic vector of set S_i. We can also think of L as a collection of $n = |V|$ length-m binary vectors that indicate which of the m intervention sets $S_1, \ldots, S_m$ each variable v_i belongs to.

As is standard, we assume that $\mathcal{G}$ satisfies the *causal Markov condition* and assume *faithfulness* [30], both in the observational and interventional distributions following [13]. This ensures that conditional independence tests lead to the discovery of true causal relations rather than spurious associations.

Cost Model and Approximate Learning. In our cost model, each node $u \in V$ has a cost $C(u) \in [1, W]$ for some $W \geq 1$ and the cost of intervention on a set $S \subseteq V$ has the linear form $C(S) = \sum_{u \in S} C(u)$. That is, interventions that involve a larger number of, or more costly nodes, are more expensive. Our goal is to find an intervention set $\mathcal{S}$ minimizing $C(\mathcal{S}) = \sum_{S \in \mathcal{S}} \sum_{u \in S} C(u)$, subject to a constraint m on the number of interventions used. This *min cost intervention design* problem was first introduced in [18]. Letting $L \in \{0, 1\}^{n \times m}$ be the matrix associated with an intervention set $\mathcal{S}$, the cost $C(\mathcal{S})$ can be written as $C(L) = \sum_{j=1}^{n} C(v_j) \cdot \|L(j)\|_1$, where $\|L(j)\|_1$ is the *weight* of L's j^{th} row, – the number of 1's in that row and number of interventions v_j is involved in.

We study two variants of causal graph recovery, in which we seek to recover the observable graph G or the ancestral graph $\mathrm{Anc}(G)$. We say that an intervention set $\mathcal{S}$ is α-optimal for a given recovery task if $C(\mathcal{S}) \leq \alpha \cdot C(\mathcal{S}^*)$, where $\mathcal{S}^*$ is the minimum cost intervention set needed for that task.

For both recovery tasks we consider a natural approximate learning guarantee:

Definition 2.1 (ϵ-Approximate Learning)**.** *An algorithm ϵ-approximately learns $G(V, E)$ (analogously, $\mathrm{Anc}(G)$) if it identifies the directions of a subset $\widetilde{E} \subseteq E$ of edges with $|E \setminus \widetilde{E}| \leq \epsilon n^2$.*

Generally, we will seek an intervention set $\mathcal{S}$ that lets us ϵ-approximately learn G or $\mathrm{Anc}(G)$, and which has cost bounded in terms of $\mathcal{S}^*$, the minimum cost intervention set needed to *fully* learn the graph. In this sense, our algorithms are bicriteria approximations.

Independent Sets. Our intervention set algorithms will be based on finding large independent sets of variables, that can be included in the same intervention sets, following the approach of [21]. Given $G(V, E)$, a subset of vertices $Z \subseteq V$ forms an independent set if there are no edges between any vertices in Z, i.e., $E[Z] = \emptyset$ where $E[Z]$ is set of edges in the sub-graph induced by Z. Given a cost function $C : V \to \mathbb{R}^+$, an independent set Z is a maximum cost independent set (MIS) if $C(Z) = \sum_{u \in Z} C(u)$ is maximized over all independent sets in G. Since finding MIS is hard [6], we will use the following two notions of a near-MIS in our approximate learning algorithms:

Definition 2.2 ((γ,ϵ)-near-MIS)**.** *A set of nodes $S \subseteq V$ is a (γ,ϵ)-near-MIS in $G = (V,E)$ if $C(S) \geq (1-\gamma)C(T)$ and $|E[S]| \leq \epsilon n^2$ where T is a maximum cost independent set (MIS) in G.*

Definition 2.3 ((ρ,γ,ϵ)-Independent-Set)**.** *A set of nodes $S \subseteq V$ is a (ρ,γ,ϵ)-independent-set in $G = (V,E)$ if $C(S) \geq \rho(1-\gamma) \cdot C(V)$ and $|E[Z]| \leq \epsilon n^2$.*

3 Separating Set Systems

Our work focuses on two important classes of intervention sets which we show in Sections 4 and 5 are necessary and sufficient for recovering G and $\mathrm{Anc}(G)$ in our setting. Missing details from this section are collected in Appendix B.

Definition 3.1 ((Strongly) Separating Set System)**.** *For any undirected graph $G(V,E)$, a collection of subsets $\mathcal{S} = \{S_1, \cdots, S_m\}$ of V is a separating set system if every edge $(u,v) \in E$ is separated – i.e., there exists a subset S_i with $u \in S_i$ and $v \notin S_i$ or with $v \in S_i$ and $u \notin S_i$. The set system is* strongly separating *if there exist two subsets S_i and S_j such that $u \in S_i \setminus S_j$ and $v \in S_j \setminus S_i$.*

For a separating set system, each pair of nodes connected in G must simply have different assigned row vectors in the matrix $L \in \{0,1\}^{n \times m}$ corresponding to $\mathcal{S}$ (i.e., the rows of L form a valid coloring of G). For a strongly separating set system, the rows must not only be distinct, but one cannot have support which is a subset of the other's. We say that such rows are *non-dominating*: there are distinct $i,j \in [m]$ such that $L(u,i) = L(v,j) = 0$ and $L(u,j) = L(v,i) = 1$. When $\mathcal{S}$ is a (strongly) separating set system for G we call its associated matrix L a *(strongly) separating matrix* for G.

Finding an exact minimum cost (strongly) separating set system is NP-Hard [16, 21] and thus we focus on approximation algorithms. We say the $\mathcal{S}$ is an α-optimal (strongly) separating set system if $C(\mathcal{S}) \leq \alpha \cdot C(\mathcal{S}^*)$, where $\mathcal{S}^*$ is the minimum cost (strongly) separating set system. Equivalently, for matrices L, L^* corresponding to $\mathcal{S}, \mathcal{S}^*$, $C(L) \leq \alpha \cdot C(L^*)$.

Unfortunately, even when approximation is allowed, finding a low-cost set system for an arbitrary graph G is still hard. In particular, we prove the conditional lower bound:

Theorem 3.2. *Assuming 3-colorable graphs cannot be colored with sub-polynomial colors in polynomial time, there is no polynomial time algorithm for finding an $o(\log n)$-optimal (strongly) separating set system for an arbitrary graph G with n nodes when $m = \beta \log n$ for some constant $\beta > 2$.*

Achieving sub-polynomial coloring for 3-colorable graphs in polynomial time is a longstanding open problem [5, 17, 32], with the current best known algorithm [3] achieving an approximation factor $O(n^{0.2111})$. Thus Theorem 3.2 shows the hardness of finding near optimal separating set systems, barring a major breakthrough on this classical problem.

It is easy to check that for a strongly separating set system, every node must appear in at least one intervention, and so the set system has cost as least $\sum_{v \in V} C(v)$. At the same time, with $m \geq 2\log n$, we can always find a strongly separating set system where each node appears in $\log n$ interventions. In particular, we assign each node to a unique vector with weight $\log n$. Such an assignment is non-dominating and since $\binom{2\log n}{\log n} \geq n$, is feasible. It achieves cost $C(\mathcal{S}) = \log n \cdot \sum_{v \in V} C(v)$, giving a simple $\log n$-approximation for the minimum cost strongly separating set system problem. For a separating set system, a simple $O(\log n)$-approximation is also achievable by first computing an approximate minimum weight vertex cover and assigning all nodes in its complementary independent set the weight 0 vector i.e., assigning them to no interventions (See Appendix B for details). By Theorem 3.2, it is hard to improve significantly on either of these bounds.

Given the hardness result of Theorem 3.2 we focus on finding relaxed separating set systems in which some variables are not separated. We will see that these set systems still suffice for approximately learning G and $\mathrm{Anc}(G)$ under the notion of Definition 2.1.

Definition 3.3 (ϵ-(Strongly) Separating Set System)**.** *For any undirected graph $G(V,E)$, a collection of subsets $\mathcal{S} = \{S_1, \cdots, S_m\}$ of V is an ϵ-separating set system if, letting $L \in \{0,1\}^{n \times m}$ be the matrix corresponding to $\mathcal{S}$, $|\{(v_i,v_j) \in E : L(i) = L(j)\}| < \epsilon n^2$. It is strongly separating if $|\{(v_i,v_j) \in E : L(i), L(j)$ are not non-dominating$\}| < \epsilon n^2$.*

For ϵ-strongly separating set systems, when the number of interventions is large, specifically $m \geq 1/\epsilon$, a simple approach is to partition the nodes into $1/\epsilon$ groups of size $\epsilon \cdot n$. We then assign the same weight 1 vector to nodes in the same group and different weight 1 vectors to nodes in different groups.

For ϵ-separating set system, we first find an approximate minimum vertex cover, and then apply the above partitioning. In Appendix B, we show that we get within a 2 factor of the optimal (strongly) separating set system. Therefore, for the remainder of this paper we assume $m < 1/\epsilon$. While m is an input parameter, smaller m corresponds to fewer interventions and this is the more interesting regime.

4 Observable Graph Recovery

We start by considering the setting where we are given all edges in the observable graph G (i.e., all direct causal relations between observable variables) e.g., by a domain expert, and wish to identify the direction of these edges. It is known that, assuming causal sufficiency (no latents), a separating set system is necessary and sufficient to learn G [8]. In Appendix C we show that this is also the case in the presence of latents when we are given the edges in G but not their directions. We also show that an ϵ-separating set system is sufficient to approximately learn G in this setting:

Claim 4.1. *Under the assumptions of Section 2, if $\mathcal{S} = \{S_1, S_2, \cdots S_m\}$ is an ϵ-separating set system for G , $\mathcal{S}$ suffices to ϵ-approximately learn G.*

In particular, if $\mathcal{S}$ is an ϵ-separating set system, we can learn all edges in G that are separated by $\mathcal{S}$ up to ϵn^2 edges which are not separated. Given Claim 4.1, our goal becomes to find an ϵ-separating matrix L_ϵ for G satisfying for some small approximation factor α, $C(L_\epsilon) \leq \alpha \cdot C(L^*)$ where L^* is the minimum cost separating matrix for G. Missing technical details of this section are collected in Appendix C.

We follow the approach of Lindgren et al. [21], observing that every node in an independent set of G can be assigned the same vector in a valid separating matrix. They show that if we greedily peel off maximum independent sets from G and assign them the lowest remaining weight vector in $\{0,1\}^m$ not already assigned as a row in L, we will find a 2-approximate separating matrix. Their work focuses on chordal graphs where an MIS can be found efficiently in each step. However for general graphs G, finding an MIS (even approximately) is hard (see Appendix A). Thus, in Algorithm 1, we modify the greedy approach and in each iteration we find a near independent set with cost nearly as large as the true MIS in G (Def. 2.2). Each such set has few internal edges, this leads to few non-separating assignments between edges of G in L_ϵ. Let γ be error in our approximation which is scaled appropriately (See Appendix C for more details) along with ϵ, δ when we call NEAR-MIS.

Algorithm 1 ϵ-SEPARATING MATRIX$(G, m, \gamma, \epsilon, \delta)$

Input : Graph $G = (V, E)$, cost function $C : V \to \mathbb{R}^+$, m, error ϵ, γ, and failure probability δ.
Output : ϵ-Separating Matrix $L_\epsilon \in \{0,1\}^{n \times m}$.
Mark all vectors in $\{0,1\}^m$ as available.
while $|V| > 0$ **do**
    $S \leftarrow$ NEAR-MIS $(G, \gamma/8m, \epsilon^2, \epsilon\delta)$
    $\forall\, v_j \in S$, Set $L_\epsilon(j)$ to smallest weight vector available from $\{0,1\}^m$ and mark it unavailable.
    Update G by $E \leftarrow E \setminus E[S]$ and $V \leftarrow V \setminus S$.
return L_ϵ

Observe that any subset of fewer than ϵn nodes has at most $\epsilon^2 n^2$ internal edges and so the NEAR-MIS $(G, \gamma/8m, \epsilon^2, \epsilon\delta)$ routine employed in Algorithm 1 always returns at least ϵn nodes. Thus the algorithm terminates in $1/\epsilon$ iterations. Across all $1/\epsilon$ near MIS's there are at most $\epsilon^2 n^2 \cdot 1/\epsilon = \epsilon n^2$ edges with endpoints assigned the same vector in L_ϵ, ensuring that L_ϵ is indeed ϵ-separating for G.

In Algorithm 2, we implement the NEAR-MIS routine by using the notion of a (ρ, γ, ϵ)-independent-set (Definition 2.3). We find a value of ρ that achieves close to the true MIS cost via a search over decreasing powers of $(1+\gamma)$. In Algorithm 3 we show how to obtain a (ρ, γ, ϵ)-independent-set whenever the cost of true MIS in G is at least $\rho \cdot C(V)$.

We extend ideas from property testing for graphs with unit vertex costs [11]. Suppose S is a fixed MIS in G with $|S| \geq \rho n$ and $U \subset S$. Let $\Gamma(u)$ represent the set of nodes that are neighbors of node u in G and $\Gamma(U) = \bigcup_{u \in U} \Gamma(u)$. Let $\overline{\Gamma}(U) = V \setminus \Gamma(U)$ be the set of nodes with no edges to any node of U. Observe that $S \subseteq \overline{\Gamma}(U)$ since $U \subset S$. Further, [11] proves that, if U is sampled randomly from S, taking the lowest degree nodes in the induced subgraph on $\overline{\Gamma}(U)$ will with high probability yield a near MIS for G. Intuitively, the nodes in $\overline{\Gamma}(U)$ have no connections to U and thus are unlikely to have

many connections to S. To find such a U, since we assume $|S| \geq \rho n$ (in the unit cost case), we can simply sample a small set of nodes in G, which will contain with good probability a representative proportion of nodes in S. We can then brute force search over all subsets of this sampled set until we hit U which is entirely contained in S and for which our procedure on $\overline{\Gamma}(U)$ returns a near-MIS.

Algorithm 2 NEAR-MIS

Input : Graph $G(V, E)$, cost function $C : V \to \mathbb{R}^+$, error ϵ, γ, and failure probability δ.
Output : Set of nodes S that is a $(4\gamma, \epsilon)$-near-MIS in G.
Initialize $\rho = 1$, and let T be the set of $\sqrt{\epsilon}n$ nodes in G with the highest cost.
while $\rho \geq \sqrt{\epsilon}$ **do**
    $S \leftarrow \rho$-INDSET$(\rho, \gamma, \epsilon, \delta')$ where $\delta' = 2\gamma\delta/\log(1/\epsilon)$
    if $C(S) \geq C(T)$ **and** $|E[S]| \leq \epsilon n^2$ **then**
        break
    $\rho = \rho/(1+\gamma)$
return $\arg\max_{X \in \{S,T\}} C(X)$

In the weighted case, when S is a high cost MIS, may not contain a large number of nodes, making it more difficult to identify via sampling. To handle this, we partition the nodes based on their costs in powers of $(1+\gamma)$ into $k = O(\gamma^{-1} \log W)$ partitions $V_1, \ldots, V_k$. A *good partition* is one that contains a large fraction of nodes in S: at least $\gamma\rho|V_i|$. Focusing on these partitions suffices to recover an approximation to S. Intuitively, all bad partitions have few nodes in S and thus ignoring nodes in them will not significantly affect the MIS cost.

Definition 4.2 ((γ, ρ)-good partition)**.** *Let S be an independent set in G with cost $\geq \rho C(V)$. Then $F_{(\gamma,\rho)} = \{i \mid |V_i \cap S| \geq \gamma\rho|V_i|\}$ is the set of* good partitions *of V with respect to S.*

Claim 4.3. *Suppose S is an independent set in G with cost $C(S) \geq \rho C(V)$, then, there exists an independent set $S' \subseteq S$ such that $C(S') \geq \rho(1-2\gamma)C(V)$ and $S' \cap V_i = S \cap V_i$ for all $i \in F_{(\gamma,\rho)}$.*

While we do not a priori know the set of good partitions, if we sample a small number t of nodes uniformly from each partition, with good probability, for each good partition we will sample $\gamma\rho t/2$ nodes in S. We will brute force search over all possible $\mathcal{U} = U_1 \cup U_2 \cdots \cup U_k$ where $|U_i| = \gamma\rho t/2$ and in at least one instance will have all U_i in good partitions fully contained in S.

Let $\overline{\Gamma}_i(U_i) = V_i \setminus \Gamma(U_i)$ be the nodes in V_i with no connections to U_i and let $Z(\mathcal{U}) := \bigcup_i \overline{\Gamma}_i(U_i)$. Analogously to unit cost case, we sort the nodes in each $\overline{\Gamma}_i(U_i)$ by their degree in the induced subgraph on $Z(\mathcal{U})$. We select low degree nodes from each partition until the sum of the total degrees of the nodes selected is $\epsilon n^2/k$. We output union of all such nodes iff it is a $(\rho, 3\gamma, \epsilon)$-independent set.

Algorithm 3 ρ-INDSET

Input : Graph $G = (V, E)$, cost function $C : V \to \mathbb{R}^+$, parameters ρ, γ, ϵ and δ
Output : $(\rho, 3\gamma, \epsilon)$ independent set in G if one exists.
For $i = 1, \ldots, k$, define $V_i = \{v \in V \mid (1+\gamma)^{i-1} \leq C(v) < (1+\gamma)^i\}$ where $k = \gamma^{-1} \log W$
Sample $t = O(\frac{\log(2k/(\epsilon\delta))}{\epsilon\gamma\rho})$ nodes $\widetilde{V}_i$ in each partition V_i.
for every partition $\mathcal{U} = U_1 \cup U_2 \cup \cdots \cup U_k$ such that $U_i \subseteq \widetilde{V}_i$ with size $\gamma\rho t/2$ for all i **do**
    Let $Z(\mathcal{U}) := \bigcup_{i=1}^k V_i \setminus \Gamma(U_i)$.
    for $i = 1 \ldots k$ **do**
        Sort nodes in $Z(\mathcal{U}) \cap V_i$ in increasing order of degree in the induced graph on $Z(\mathcal{U})$.
        Let $\widehat{Z}_i(\mathcal{U}) \subseteq Z(\mathcal{U}) \cap V_i$ be set of nodes obtained by considering the nodes in the sorted order until the total degree is $\epsilon n^2/k$.
    Let $\widehat{Z}(\mathcal{U}) = \bigcup_{i=1}^k \widehat{Z}_i(\mathcal{U})$.
    return $\widehat{Z}(\mathcal{U})$ if $C(\widehat{Z}(\mathcal{U})) \geq \rho(1-3\gamma)C(V)$.

By construction, our output, denoted by $\widehat{Z}(\mathcal{U})$ will have at most ϵn^2 internal edges. Thus, the challenge lies in analyzing its cost. We argue that in at least one iteration, all chosen U_i for good partitions will not only lie within the MIS S, but their union will accurately represent connectivity to S. Specifically, any vertex $v \in \overline{\Gamma}_i(U_i)$, i.e., with no edges to U_i for some $i \in F_{(\gamma,\rho)}$, should have few edges to S.

Definition 4.4. *(ϵ_2-IS representative subset) $R \subseteq \bigcup_{i \in F_{(\gamma,\rho)}} (S \cap V_i)$ is an ϵ_2-IS representative subset of S if for all but $\epsilon_2 n$ nodes of V it satisfies the following property: suppose $v \in V_i$ and $i \in F_{(\gamma,\rho)}$ if $\Gamma(v) \cap (R \cap V_i) = \emptyset$ then $|\Gamma(v) \cap S| \leq \epsilon_2 n$.*

We show that there is a *IS representative subset* containing at least $\gamma\rho t/2$ nodes from each good partition among our sampled nodes $\bigcup_{i=1}^{k} \widetilde{V}_i$. Setting $\epsilon_2 = \epsilon/k$ we have:

Claim 4.5. *If t nodes are uniformly sampled from each partition V_i to give $\widetilde{V}_i$, with probability $1-\delta$, there exists an ϵ/k-IS representative subset R such that, for every $i \in F_{(\gamma,\rho)}$, $|\widetilde{V}_i \cap R| = \gamma\rho t/2$.*

Claim 4.5 implies that in at least one iteration, our guess $\mathcal{U}$ restricted to the good partitions is in fact an ϵ/k-IS representative subset. Thus, nearly all nodes in $Z(\mathcal{U})$ lying in good partitions have at most $\epsilon n/k$ connections to S. Letting $F = \bigcup_{i \in F(\gamma,\rho)} V_i$ contain all nodes in good partitions, since we know that $\mathcal{U} \cap F \subseteq S \cap F$, we have $S \cap F \subseteq \overline{\Gamma}(S) \cap F \subseteq Z(\mathcal{U}) \cap F$.

In the graph induced by nodes of $Z(\mathcal{U})$, with edge set $E[Z(\mathcal{U})]$, consider the degree incident on nodes of $S \cap V_i$ for each partition V_i. As there are at most n nodes in V_i, from Def. 4.4, we have the total degree incident on $S \cap V_i$ is $\epsilon n^2/k$. Thus, including the nodes with lowest degrees in $\widehat{Z}_i(\mathcal{U})$ until the total degree is $\epsilon n^2/k$ will yield a larger set of nodes than $S \cap V_i$. Since all nodes in V_i have cost within a $1 \pm \gamma$ factor of each other, we will have $C(\widehat{Z}_i(\mathcal{U})) \geq (1-\gamma) \cdot C(S \cap V_i)$. As the cost of S in the bad partitions is small, using Claim 4.3, we have $\widehat{Z}(\mathcal{U}) = \bigcup_{i=1}^{k} \widehat{Z}_i(\mathcal{U})$ is a $(\rho, O(\gamma), \epsilon)$-independent set.

Overall, Algorithm 3 implements ρ-IndSet as required by Algorithm 2 to compute a near-MIS at each iteration of Algorithm 1. It just remains to show that, by greedily peeling off near-MISs from G, Algorithm 1 achieves a good approximation guarantee for the ϵ-Approximate Learning G. To do this, we use the analysis of a previous work from [21]. In their work, an exact MIS is computed at each step, since their graph is chordal so the MIS problem is polynomial time solvable. However, the analysis extends to when the MIS has near optimal cost and some violated constraints, allowing us to achieve near 2-factor approximation, as is achieved in [21]. Our final result is:

Theorem 4.6. *For any $m \geq \eta \log 1/\epsilon$ for some constant η, with probability $\geq 1-\delta$, Algorithm 1 returns L_ϵ with $C(L_\epsilon) \leq (2 + \gamma + \exp(-\Omega(m))) \cdot C(L^*)$, where L^* is the min-cost separating matrix for G. Moreover L_ϵ ϵ-separates G. Algorithm 1 has a running time $O(n^2 f(W, \gamma, \epsilon, \delta))$ where $f(W, \gamma, \epsilon, \delta) = \frac{1}{\epsilon^2 \gamma} \log \frac{1}{\epsilon} \cdot \exp(O(\frac{\log W}{\gamma \epsilon^3} \cdot \log \frac{1}{\gamma\epsilon} \cdot \log \frac{\log 1/\epsilon \log W}{\gamma \epsilon \delta}))$*

By Theorem 4.6 with $m = O(\log(1/\epsilon))$ interventions we can ϵ-approximately learn any causal graph G. For full learning, $m \geq \log \chi$ interventions are necessary, where χ is the chromatic number of G, since the rows of $L \in \{0,1\}^{n \times m}$ must be a valid coloring of G.

5 Ancestral Graph Recovery

In Section 4, we assumed knowledge of the edges in the observable graph G and sought to identify their directions. In this section, we relax the assumption, assuming we are given any undirected supergraph H of G i.e., it includes all edges of G and may also include edges which do not represent causal edges. When given such a graph H, we cannot recover G itself. However we still seek to recover all directed edges of the ancestral graph $\mathrm{Anc}(G)$ appearing in H, which we denote $\mathrm{Anc}(G) \cap H$. This problems strictly generalizes that of Section 4, in which $H = G$ since $\mathrm{Anc}(G) \cap H = G$. Missing details of this section are collected in Appendix D.

We show in Appendix D that to recover $\mathrm{Anc}(G) \cap H$, a strongly separating system (Def 3.1) for H is both necessary and sufficient. An ϵ-strongly separating system suffices for approximate learning:

Lemma 5.1. *Under the assumptions of Section 2, if $\mathcal{S} = \{S_1, S_2, \cdots S_m\}$ is an ϵ-strongly separating set system for H , $\mathcal{S}$ suffices to ϵ-approximately learn $\mathrm{Anc}(G) \cap H$.*

Given Claim 5.1, our goal becomes to find an ϵ-strongly separating matrix for H, L_ϵ with cost within an α factor of the optimal strongly separating matrix for H, for some small α. To do so, our algorithm builds on the separating set system algorithm of Section 4. We first run Algorithm 1 to obtain an ϵ-separating matrix L_ϵ^S and construct $S_1, S_2, \cdots S_{1/\epsilon}$ where each set S_i contains all nodes assigned the same vector in L_ϵ^S – i.e., S_i corresponds to the near-MIS computed at step i of Algorithm 1. We form a new graph on $1/\epsilon$ vertices V_S by contracting all nodes in each S_i into a single *super node*.

In [2], the authors give a 2-approximation algorithm for finding a strongly separating matrix on a set of nodes, provided the graph on these nodes is complete. As H is an arbitrary super graph of G, the contracted graph on V_S is also arbitrary. However we simply assume the worst case, and run the Algorithm of [2] on it to produce L_ϵ, which strongly separates the complete graph on V_S. It is easy to show that as a consequence, L_ϵ ϵ-strongly separates G.

Algorithm 4 ANCESTRAL GRAPH$(V, m, \gamma, \epsilon, \delta)$

1: $L_\epsilon^S :=$ ϵ-SEPARATING MATRIX$(G, m, \gamma/2, \epsilon, \delta)$
2: Construct $S_1, S_2, \cdots S_{1/\epsilon}$ where each set S_i contains nodes assigned the same vectors in L_ϵ^S
3: Construct a set of nodes V_S by representing S_i as a single node w_i and $C(w_i) = \sum_{u \in S_i} C(u)$
4: $L_\epsilon^{SS} :=$ SSMATRIX(V_S, m) from [2]
5: **return** L_ϵ^{SS}

To prove the approximation bound, we extend the result of [2], showing that their algorithm actually achieves a cost at most 2 times the optimal cost of a *separating matrix* for the complete graph on V_S (not strongly separating) which satisfies two additional restrictions: (1) it does not assign the all zeros vector to any node and (2) it assigns the same number of weight one vectors as the optimal strongly separating matrix. Further, we show via a similar analysis to Theorem 4.6 that this cost on V_S is bounded by $2 + 0.5\gamma$ times the cost of the optimal strongly separating matrix on the contracted graph over V_S. Combining these bounds yields the final $4 + \gamma$ approximation guarantee of Theorem 5.2.

Theorem 5.2. *Let $m \geq \eta \log 1/\epsilon$ for some constant η and L_ϵ^{SS} be matrix returned by Algorithm 4. Then with probability $\geq 1 - \delta$, L_ϵ^{SS} is an ϵ-strongly separating matrix for G and $C(L_\epsilon^{SS}) \leq (4 + \gamma + \exp(-\Omega(m))) \cdot C(L^*)$ where L^* is the min-cost strongly separating matrix for G. Algorithm 4 runs in time $O(n^2 f(W, \gamma, \epsilon, \delta))$ where $f(W, \gamma, \epsilon, \delta) = \frac{1}{\epsilon^2 \gamma} \log \frac{1}{\epsilon} \cdot \exp(O(\frac{\log W}{\gamma \epsilon^3} \cdot \log \frac{1}{\gamma \epsilon} \cdot \log \frac{\log 1/\epsilon \log W}{\gamma \epsilon \delta}))$*

6 Experiments

In this section, we empirically demonstrate the performance of our algorithms for ϵ-approximately learning a causal (ancestral) graph G (See Appendix E for more details). We consider random graphs generated using Erdös-Rényi random graphs $G(n, p)$ and include our results in Figure 1 with parameters $p = 0.3, \epsilon = 0.1, \delta = 0.01$.

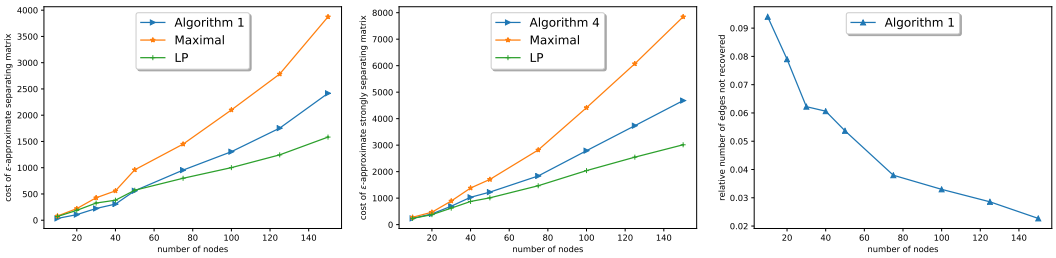

Figure 1: cost of ϵ-approximate learning of causal (or ancestral) graphs

Baselines. For learning causal graphs, we compare the cost of ϵ-separating matrix from Algorithm 1 with two baselines (i) a greedy procedure, similar to Algorithm 1 where we identify a maximal independent set by sorting the nodes and including the high cost nodes until it remains an independent set (ii) LP relaxation [18, 21] that is a lower bound on the true cost obtained by viewing the problem as graph coloring. For learning $\mathrm{Anc}(G)$, we compare cost of ϵ- strongly separating matrix from Algorithm 4 with two baselines (i) we construct maximal independent sets as described above and run Algorithm SSMATRIX from [2] on these sets. (ii) we consider a modified LP relaxation of separating set system, where we require nodes to have vectors of weight at least 1. Again, this relaxation gives a lower bound on the true optimum.

Results. For ϵ-approximate learning causal graph G, we observe that the cost of interventions obtained using Algorithm 1 are better than the maximal independent set baseline and closer to the LP optimum value for low values of n. For ϵ-approximate learning of $\mathrm{Anc}(G)$, we observe a similar pattern for Algorithm 4 with respect to our baselines. We also note that the cost for learning $\mathrm{Anc}(G)$ is higher than the cost for learning G. We also observe that the number of edges that are not separating (parameterized by ϵ) returned by Algorithm 1 decays very quickly with n, with a similar behaviour observed for learning $\mathrm{Anc}(G)$ using Algorithm 4.

Broader Impact

Our work studies the fundamental problem of intervention (experiment) design for causal inference. Causality plays a pivotal role in understanding how various components of a system interact and so causal inference is one of the key goals in scientific discovery. While our work is mostly theoretical, we hope that better general understanding of cost aware intervention design will prove useful to practitioners in a wide range of fields. As just one example of the potential impact of work based on causal inference: the work of the recent recipients of Nobel Prize in Economics, Abhijit Banerjee, Esther Duflo and Michael Kremer [1] used randomized control trials [7] (an analog of interventions) to fight global poverty. As a direct result of one of their studies, more than five million Indian children have benefitted from effective programs of remedial tutoring in schools [1].

As more machine learning based systems are integrated into the real world, there is a need to understand the decisions made by these systems and their causal interactions with societal outcomes. Improved understanding of causal inference can help in doing this.

We do not foresee any direct negative outcomes of our work. As with all theoretical work, our results are based on simplified models of the real world, and this is important to keep in mind. In designing interventions for, e.g., a medical study, blindly attempting to minimize intervention cost under one simple metric while ignoring other complex factors would be irresponsible.

References

[1] Press release. URL https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/.

[2] Raghavendra Addanki, Shiva Prasad Kasiviswanathan, Andrew McGregor, and Cameron Musco. Efficient intervention design for causal discovery with latents. *arXiv 2005.11736. International Conference on Machine Learning*, 2020.

[3] Sanjeev Arora and Eden Chlamtac. New approximation guarantee for chromatic number. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 215–224, 2006.

[4] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[5] Avrim Blum and David Karger. An O$(n^{3/14})$-coloring algorithm for 3-colorable graphs. *Information processing letters*, 61(1):49–53, 1997.

[6] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT press, 2009.

[7] Esther Duflo, Rachel Glennerster, and Michael Kremer. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962, 2007.

[8] Frederick Eberhardt. Causation and intervention. *PhD Thesis, Carnegie Mellon University*, 2007.

[9] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

[10] Uriel Feige, Shafi Goldwasser, Laszlo Lovász, Shmuel Safra, and Mario Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM (JACM)*, 43(2):268–292, 1996.

[11] Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

[12] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

[13] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

[14] Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

[15] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[16] Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Experiment selection for causal discovery. *The Journal of Machine Learning Research*, 14(1):3041–3071, 2013.

[17] David Karger, Rajeev Motwani, and Madhu Sudan. Approximate graph coloring by semidefinite programming. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 2–13. IEEE, 1994.

[18] Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1875–1884. JMLR. org, 2017.

[19] Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems*, pages 7018–7028, 2017.

[20] Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems*, pages 14346–14356, 2019.

[21] Erik Lindgren, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Experimental design for cost-aware learning of causal graphs. In *Advances in Neural Information Processing Systems*, pages 5279–5289, 2018.

[22] 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.

[23] Pekka Parviainen and Mikko Koivisto. Ancestor relations in the presence of unobserved variables. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 581–596. Springer, 2011.

[24] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge university press, 2009.

[25] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *Ann. Statist.*, 30(4): 962–1030, 08 2002. doi: 10.1214/aos/1031689015.

[26] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, pages 3195–3203, 2015.

[27] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[28] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 1219–1226, 2006.

[29] Ricardo Silva, Richard Scheine, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246, 2006.

[30] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[31] Thomas Verma and Judea Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in artificial intelligence*, pages 323–330. Elsevier, 1992.

[32] Avi Wigderson. Improving the performance guarantee for approximate graph coloring. *Journal of the ACM (JACM)*, 30(4):729–735, 1983.

[33] Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9 (Jul):1437–1474, 2008.

[34] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.