Greedy Relaxations of the Sparsest Permutation Algorithm

Wai-Yin Lam 1  
Bryan Andrews 1  
Joseph Ramsey 1

1 Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Abstract

There has been an increasing interest in methods that exploit permutation reasoning to search for directed acyclic causal models, including the “Ordering Search” of Teyssier and Kohler and GSP of Solus, Wang and Uhler. We extend the methods of the latter by a permutation-based operation tuck, and develop a class of algorithms, namely GRaSP, that are efficient and pointwise consistent under increasingly weaker assumptions than faithfulness. The most relaxed form of GRaSP outperforms many state-of-the-art causal search algorithms in simulation, allowing efficient and accurate search even for dense graphs and graphs with more than 100 variables.

1 INTRODUCTION

Searching for causal models by identifying patterns of conditional independence in observational data has become a well-established activity, though it is not without detractors. For one thing, it is commonly believed that the only correct method for establishing causal relationships is through experimental manipulation, as is done in a randomized controlled trial. Accordingly, causal inference from observational data alone can be seen as second-rate. This is not completely unreasonable; many causal search algorithms, even in seemingly ideal conditions, demonstrate poor performance, calling into question whether their inferences can be relied upon. Furthermore, the theoretical assumptions made by these algorithms are often criticized for being too strong. More specifically, these algorithms assume that the true model belongs to a model class with no latent variables or no cycles, and that the patterns of conditional independence in the data generating distribution can be represented by the assumed model class exactly. The latter assumption is called the causal faithfulness condition and can be violated (or almost violated) by unexpected patterns of conditional independence that arise from subtleties in the distribution, such as (near) determinism or (almost) path cancellation.

The most common model class assumed by causal search algorithms is characterized by directed acyclic graphs (DAGs). Many algorithms for causal inference search the space of causal DAGs, such as the PC (“Peter and Clark”, [Spirtes et al., 2000]) and GES (“Greedy equivalence Search”, [Chickering, 2002]) algorithms, and provably return a set of DAGs that contains the true model under the causal faithfulness condition. However, the performance of these algorithms appear to fall short of their theoretical claims, especially when the true model is densely connected. One hypothesis for this phenomenon is that almost-violations of the causal faithfulness condition frequently occur and impede search procedures [Uhler et al., 2013]. Accordingly, the performance of these algorithms might be improved by relaxing the causal faithfulness condition, as is done by the SP (“Sparsest Permutation”) algorithm of Raskutti and Uhler [2018]. SP considers the space of variable orderings and builds a DAG using a procedure inspired by Verma and Pearl [1988], where the parents of each variable are selected from the preceding variables in the permutation. Ultimately, the permutations that induce DAGs with the minimal edge count are selected.

Raskutti and Uhler [2018] proved that if the data generating distribution is a graphoid, then the set of DAGs returned by SP contains the true model under an assumption strictly weaker than the causal faithfulness condition. While SP recovers the set of all frugal models, it is super exponential in the number of variables, that is, if there are $n$ variables, then there are $n!$ permutations that must be visited. In practice, it is limited to a maximum of about nine variables due to its computational complexity. This naturally raises the question: is there an algorithm that is equally accurate in most cases for such data, but that can scale to larger problems?
Teyssier and Koller [2005] give a clever search and score procedure, “Ordering Search”, over variable permutations, pointing out that when two adjacent variables in a permutation are swapped, only local scores for the swapped variables need to be recalculated, the rest of the score calculation remains unchanged—this swapping operation is called an adjacency transposition (AT). The Ordering Search algorithm greedily traverses the space of permutations with adjacency transpositions using a hill-climbing approach, random restarts, and a tabu list. However, they do not give any consistency guarantees.

The ESP (“Edge Sparsest Permutation”) algorithm of Solus et al. [2021] iterates upon the Ordering Search algorithm by greedily traversing the space of permutations by sequences of ATs where each AT leads to an equal or less edge count, found by depth first search (DFS), to achieve asymptotic correctness. In addition, their TSP (“Triangle Sparsest Permutation”) algorithm uses the theory of Chickering [2002] to navigate the space of DAGs, more efficiently than ESP, under a stronger assumption. A simulation study using a Python implementation of TSP [Solus et al., 2021] suggests that this procedure is fast, but has difficulty scaling accurately to moderate or large sized graphs [Lu et al., 2021].

To address the scaling problem for both accuracy and timing, in this paper we explore different ways of traversing the space of permutations that get closer to the performance and assumption relaxation of Raskutti and Uhler while maintaining scalability. As part of this effort, we also use the “Grow-Shrink” algorithm from Margaritis and Thrun [1999] to learn the DAG.

In what follows, we give an elaboration of the theoretical background of our set of permutation-based procedures, GRaSP (“Greedy Relaxations of Sparsest Permutation”). GRaSP has three tiers, GRaSP$_0$ (basically equivalent to TSP), GRaSP$_1$ (basically equivalent to ESP), and GRaSP$_2$ (a novel relaxation); we show how moving from a lower tier to a higher tier results in a gradual theoretical relaxation of accuracy. W e then follow this with a study of oracle behavior for GRaSP$_0$, GRaSP$_1$, and GRaSP$_2$ on exhaustive lists of independence models with violations of faithfulness for all 4-variable regular Gaussian and positive discrete distributions and all 5-variable unfaithful DAGs with added marginal independencies between a pair of variables. We also give a detailed simulation study for the linear, Gaussian case for larger possibly dense models of up to 100 variables, with consistently accurate results using GRaSP$_2$. Further, we study an empirical example to test GRaSP$_2$. We then give a conclusion and discussion where we point out areas of immediate future work.

2 CONTRIBUTIONS

The most salient contribution is that GRaSP$_2$ can scale to at least 100 variables with average degree at least 10 on a laptop with high adjacency and arrowhead precision and recall for the linear, Gaussian case, addressing the longstanding practical problem of dense graph causal search in a meaningful way.

Second, theoretical development over assumptions on causal discovery from previous works has been simplified, in places corrected, and reworked as a structured study of casual razors. Accordingly, the proof that GRaSP$_0$, TSP, and by implication GSP, require faithfulness is a logical discovery. Also, the proof that faithfulness is equivalent to unique Pearl-minimality is a novel contribution.

Third, we extended the discussion of unit tests initiated in [Solus et al., 2021] considerably, using the criterion that a wide variety of unit tests should systematically pass on all initial permutations using a d-separation oracle. More specifically, we run GRaSP on models detailed in Šimecek [2006a,b] and those listed in Appendix [G].

Finally, the tuck operation is a novel transformation that has not been considered in the literature before. We show that traversing any edge of the DAG-associalhedron can be equivalently done via a tuck. Reframing TSP in terms of the tuck operation allows TSP and ESP to be neatly placed into a hierarchy. Moreover, it admits the natural generalization to GRaSP$_2$ (by not restricting which edges can be tucked).

3 BACKGROUND

Throughout this paper, italicized letters are used to denote variables (e.g., $X_1, Y$) and boldfaced letters for sets of variables (e.g., $X$). Graphical definitions and notations related to directed acyclic graphs (DAGs) are provided in Appendix [A]. A DAG $\mathcal{G}$ over a set of measured variables $V = \{X_1, \ldots, X_m\}$ consists of $m$ vertices $v = \{1, \ldots, m\}$ where each vertex $i$ associates to the variable $X_i$, and each directed edge between two distinct vertices $j \rightarrow k$ represents the direct causal influence from $X_j$ to $X_k$. We write $i \perp_{\mathcal{G}} j | k$ to denote the d-separation relation between $i$ and $j$ given $k$ in $\mathcal{G}$ for any pairwise disjoint subsets of vertices $i, j, k \subseteq v$. Similarly, given a joint probability distribution $\mathcal{P}$ over $V$, denote $X \perp_{\mathcal{P}} Y | Z$ as the conditional independence (CI) relation between $X$ and $Y$ given $Z$ for any pairwise disjoint subsets of variables $X, Y, Z \subseteq V$.

A model is a pair $(\mathcal{G}, \mathcal{P})$ where $\mathcal{G}$ is a DAG and $\mathcal{P}$ is a joint probability distribution over the same set of measured variables $V$. We use $\mathcal{G}^*$ to refer to the true data-generating DAG such that $(\mathcal{G}^*, \mathcal{P})$ is the true model assumed to always exist. Certain standard properties of a model can be defined in terms of the d-separation relations in $\mathcal{G}$ and the CI relations in $\mathcal{P}$. Denote $I(\mathcal{G}) = \{(X_j, X_k \mid X_l) : j \perp_{\mathcal{G}} k \mid l\}$
A causal search algorithm is a procedure of recovering the causal information of the true DAG from its underlying joint probability distribution. Let \( \text{MEC}(G) \) be the Markov equivalence class (MEC) of \( G \) such that \( \text{I}(G) = \text{I}(G') \) for each \( G' \in \text{MEC}(G) \). One crucial goal of causal search is the identification of \( \text{MEC}(G^*) \) from \( P \). With regard to this goal, a causal search algorithm is correct if its output DAG (or the DAG induced by its output) is in \( \text{MEC}(G^*) \). All known causal search algorithms assume the Markov assumption, and some well-known algorithms in the relevant literature (e.g., GES) assume faithfulness as well. Nevertheless, as pointed out by Uhler et al. [2013], learning CI relations from data by hypothesis testing is error-prone, and almost-violations of faithfulness are common. This motivates the exploration of causal search algorithms which rely on assumptions strictly weaker than faithfulness. These assumptions, faithfulness included, are what we refer to as causal razors.

One recent approach proposed by Raskutti and Uhler [2018] is the SP algorithm, which identifies the set of sparsest permutations defined over \( v \) under the causal razors. Let \( E(G) \) be the set of directed edges in a DAG \( G \).

Definition 3.3 (U-frugality) For any joint probability distribution \( P \), define \( \text{uFr}(P) = \{ G \in \text{CMC}(P) : \neg \exists G' \in \text{CMC}(P) \text{ s.t. } |E(G')| < |E(G)| \} \) and \( \text{uFr}(P) = \{ G \in \text{Fr}(P) : \neg \exists G' \in \text{Fr}(P) \text{ s.t. } G' \notin \text{MEC}(G) \} \) as the sets of frugal DAGs and uniquely frugal, or u-frugal, DAGs respectively. \( (G^*, P) \) satisfies the u-frugality assumption if \( G^* \in \text{uFr}(P) \).

In words, u-frugality requires that \( G^* \) is not only the sparsest Markovian DAG, but also that all sparsest Markovian DAGs belong to the same MEC as \( G^* \). Raskutti and Uhler [2018] showed that SP is correct under u-frugality which is strictly weaker than faithfulness. Below we introduce some necessary notations of permutation-based algorithms. To begin with, we refer the readers to Appendix A.2 for the graphoid axioms. Generally speaking, every joint probability distribution is a semigraphoid, strictly positive distributions are graphoids, and regular Gaussian distributions are compositional graphoids.

Given \( V = \{ X_1, ..., X_n \} \), let \( \Pi(v) \) be the set of all permutations over \( v = \{ 1, ..., n \} \). For each \( \pi \in \Pi(v) \), let \( \pi_i \) be the \( i \)-th vertex in \( \pi \), and \( \pi[j] \) be the index of vertex \( j \) in \( \pi \) (s.t. \( \pi[1] = j \)), and \( \text{Pre}(j, \pi) = \{ \pi_i : 1 \leq i < \pi[j] \} \) be the set of vertices that precede \( j \)’s index in \( \pi \). We say that \( \pi \in \Pi(v) \) is a causal order of \( G \in \text{DAG}(V) \) if \( i \in \text{Pre}(j, \pi) \) for each \( j \in v \) and each \( i \in \text{An}(j, \pi) \) (i.e., the set of \( j \)'s ancestors in \( \pi \)). Given a graphoid \( P \) over \( V \), each \( \pi \in \Pi(v) \) induces a DAG \( G_\pi \) satisfying the following condition:

\[
\pi \in \text{Pre}(k, \pi) \quad \text{and} \quad X_j \in \text{Pre}(k, \pi) \quad \Rightarrow \quad (j \to k) \in E(G_\pi).
\]

(RU) is the method of constructing a unique DAG from \( P \) and \( \Pi(v) \) discussed in Raskutti and Uhler [2018]. It is derived from a more general method in Verma and Pearl [1988]. The two methods will be compared in Appendix A.3. But we refer to \( G_\pi \) as the DAG induced from \( \pi \) and the graphoid \( P \) using (RU) unless specified otherwise. Obviously, \( \pi \) is a causal order of \( G_\pi \). Below is an important feature of \( G_\pi \).

Definition 3.4 (SGS-minimality) For any joint probability distribution \( P \), define \( \text{SGS}(P) = \{ G \in \text{CMC}(P) : \neg \exists G' \in \text{CMC}(P) \text{ s.t. } E(G') \subset E(G) \} \) as the set of SGS-minimal DAGs.

Theorem 3.5 [Verma and Pearl 1988, Raskutti and Uhler 2018] Given a graphoid \( P \) over \( V \), \( G_\pi \) induced by \( \pi \) using (RU) is Markovian and SGS-minimal for every \( \pi \in \Pi(v) \).

The theorem above states that, for every permutation \( \pi \), the induced DAG \( G_\pi \) is Markovian and no subgraph of \( G_\pi \) is Markovian. By identifying the sparsest permutation \( \hat{\pi} = \text{argmin}_{\pi \in \Pi(v)} |E(G_\pi)| \), \( G_\pi \) returned by SP is guaranteed to be in \( \text{MEC}(G^*) \) with u-frugality satisfied. Nevertheless, SP needs to examine all \( ||v||! \) permutations in \( \Pi(v) \) to identify the sparsest one and hence lacks scalability. Solus et al. [2021] introduce a greedy version of SP, namely Triangle SP (TSP), which is proven to be correct under faithfulness. Below, we provide a quick and simple sketch of this result.

\[ ^3 \text{This assumption is named as sparsest Markov representation (SMR) in [Raskutti and Uhler 2018].} \]
TSP borrows the *Chickering algorithm* in [Chickering 2002] to perform their *depth-first search (DFS)* procedure. For each vertex $i \in V$, let $Pa(i, G)$ be the set of parents in $G$. A directed edge $j \rightarrow k$ is covered in $G$ if $Pa(j, G) = Pa(k, G) \setminus \{j\}$.

**Theorem 3.6 (Chickering sequences) [Chickering 2002]**

Given a set of variables $V$, for every pair of DAGs $G, H \in \text{DAG}(V)$, if $I(H) \subseteq I(G)$, there exists a sequence of DAGs, call it a Chickering sequence ($H = G_1, G_2, \ldots, G_k = G$) from $H$ to $G$ s.t. $I(G^t) \subseteq I(G^{t+1})$ and $G^{t+1}$ is obtained from $G^t$ by either reversing a covered edge or deleting a directed edge for each $1 \leq i < k$.

A sequence of DAGs $(G^1, \ldots, G^k)$ is said to be weakly decreasing if $|E(G^t)| \geq |E(G^{t+1})|$ for each $1 \leq i < k$. Obviously, every Chickering sequence is weakly decreasing. Given an arbitrary initial permutation $\pi \in \Pi(V)$, TSP uses DFS to search for a Chickering sequence from $G_\pi$ to some SGS-minimal DAG $G_\pi$, where $|E(G_\pi)| > |E(G_\pi^\pi)|$, and update $G_\pi$ as $G_\pi$ until no such $G_\pi$ is found. Now we demonstrate TSP’s correctness under faithfulness.

**Definition 3.7 (U-P-minimality)** For any joint probability distribution $P$, define $P_m(P) = \{G \in \text{CMC}(P) : \neg \exists G' \in \text{CMC}(P) \text{ s.t. } I(G) \subset I(G') \}$ as the sets of $P$-minimal DAGs and uniquely $P$-minimal DAGs respectively. $(G^*, P)$ satisfies the $u$-$P$-minimality assumption if $G^* \in P_m(P)$.

**Theorem 3.8 [Zhang 2013]** For any joint probability distribution $P$, $\text{CFC}(P) = P_m(P) = \text{MEC}(G^*)$ if faithfulness holds.

A DAG being P-minimal, as in **Definition 3.7**, states that there exists no Markovian DAG which can entail a proper superset of CI relations, and its unique variant further requires that all P-minimal DAGs belong to the same MEC as $G^*$. We elaborate the importance of u-P-minimality in the next section. By **Theorem 3.6**, TSP guarantees that its output $G_\pi$ is P-minimal. When faithfulness holds, **Theorem 3.8** ensures that $G_\pi \in \text{MEC}(G^*)$, and hence TSP is correct.

Notice that the identification of a Chickering sequence from $G_\pi$ to a P-minimal $G_\pi$ is essentially a DAG-based operation. In the next section, we introduce our permutation-based operation to converge to a P-minimal DAG, and propose a class of greedy permutation-based algorithms which employs weaker causal razors than TSP does.

In addition to TSP, Solus et al. [2021] introduced another greedy algorithm, namely *Edge SP* (ESP), which is defined by weakly decreasing traversals over the *DAG associahedron* (i.e., the permutohedron contracted by $I(P)$). These technical terms are defined in the Appendix. ESP is shown to be assuming a weaker causal razor than TSP. In the next section, we will draw a logical discovery on how ESP is connected to our novel permutation-based operation.

### 4 METHODS

In this section, we introduce a class of permutation-based algorithms with a generic name *Greedy Relaxations of Sparsest Permutation* (GRaSP). Three tiers of relaxation will be studied: GRaSP0 is our basic algorithm, GRaSP1 relaxes the search criterion of GRaSP0 while GRaSP2 further relaxes that of GRaSP1. This hierarchy allows the identification of MEC($G^*$) under progressively weaker causal razors. In addition, we show that GRaSP0 is logically equivalent to TSP, and GRaSP1 to ESP. All proofs are left in Appendix. First, we introduce our characteristic permutation-based operation $\text{tuck}$ and how it operates under different types of directed edges.

**Definition 4.1 (Tuck)** Consider any graphoid $P$ over $V$, any $\pi \in \Pi(V)$, and any $j, k \in V$ where $\pi[j] \neq \pi[k]$. Rewrite $\pi$ as $(\delta_1, j, \delta_2, k, \delta_3)$ where each $\delta_i$ is a (possibly empty) sub-sequence of $\pi$. Let $\gamma$ and $\gamma'$ be the subsequences $(i \in \delta_2 : i \in \text{Pa}(k, G_\pi))$ and $(i \in \delta_2 : i \notin \text{Pa}(k, G_\pi))$ respectively. Define $\text{tuck}(\pi, j, k) = \begin{cases} (\delta_1, \gamma, k, j, \gamma', \delta_3) & \text{if } (j \rightarrow k) \in E(G_\pi) \\ \pi & \text{otherwise.} \end{cases}$

**Definition 4.2** Given a DAG $G$, a directed edge $(j \rightarrow k) \in E(G)$ is said to be singular if there exists no unidirectional directed path from $j$ to $k$ in $G$ except $j \rightarrow k$. Define $E^t(G) = \begin{cases} \text{covered edges in } E(G) & \text{if } t = 0 \\ \text{singular edges in } E(G) & \text{if } t = 1 \\ E(G) & \text{if } t = 2. \end{cases}$

Readers can verify that $E^0(G) \subseteq E^1(G) \subseteq E^2(G)$ holds for any DAG $G$. The introduction of singular edges is crucial to our logical discovery that every move ESP takes in the DAG associahedron (as defined in Appendix) corresponds to tucking a unique singular edge. Figure provides an example on how $\text{tuck}$ works for each defined type of edges. As seen in the example, $\text{tuck}$ is an operation that aims to change a permutation minimally to obtain a differently induced DAG, while a broader class of directed

---

4The original theorem in [Chickering 2002] is expressed in terms of addition of directed edges. This modification helps by indicating that every Chickering sequence is a weakly decreasing sequence. In addition, one can easily observe that there does not exist any Chickering sequence from $H$ to $G$ if $I(H) \not\subseteq I(G)$.

5P-minimality refers to the minimality condition discussed in [Pearl 2000].

---

To be precise, $\delta_1 = \langle \pi_1 : 1 \leq i < \pi[j] \rangle$, $\delta_2 = \langle \pi_1 : \pi[j] > i < \pi[k] \rangle$, and $\delta_3 = \langle \pi_1 : \pi[k] < i \leq \pi[n] \rangle$. 
edges generally leads to more possible re-orderings of the vertices.

covered: 1 2 4 3 4 5 6 7
singular: 1 2 3 5 4 6 7
general: 3 4 1 2 3 4 5 6 7

Figure 1: Consider \( \pi = \langle 1, 2, 3, 4, 5, 6, 7 \rangle \) and its induced \( G_\pi \) shown on the right. Each of the three orderings on the left illustrates how a directed edge between two darkly shaded vertices is tucked to obtain a new permutation. For example, consider 1 \( \rightarrow \) 4 which is not singular due to the unidirectional path 1 \( \rightarrow \) 3 \( \rightarrow \) 4. Performing \( \text{tuck}(\pi, 1, 4) \) requires the identification of the intermediate vertices between 1 and 4 in \( \pi \) which are ancestors of 4 in \( G_\pi \) (i.e., the lightly shaded 3). Then, while the positions of other vertices remain intact, 3 and 4 are moved to the front of 1.

After clarifying how tuck works, we can define a sequence of tuck operations, particularly when applied to covered edges, and how a sequence of covered tucks (ct) is connected to a Chickering sequence.

Definition 4.3 (ct-sequence) Given a graphoid \( \mathcal{P} \) over \( V \), for any \( \pi, \tau \in \Pi(\mathcal{V}) \), \( \tau \) is said to be a ct-mutation of \( \pi \) if there exist \( j, k \in V \) s.t. \( (j \rightarrow k) \in E(G_\pi) \) is covered and \( \tau = \text{tuck}(\pi, j, k) \). Also, \( (\pi^1, ..., \pi^m) \) is said to be a ct-sequence if \( \pi^{i+1} \) is a ct-mutation of \( \pi^i \) for each \( 1 \leq i < m \), and \( (G_{\pi^1}, G_{\pi^i}) \) are pairwise distinct for any \( 1 \leq i < 1 \leq m \).

Lemma 4.4 [Appendix B] Given a graphoid \( \mathcal{P} \), for any \( \pi \in \Pi(\mathcal{V}) \) and any Chickering sequence from \( G_\pi \) to some \( \mathcal{H} \in \text{SGS}(\mathcal{P}) \) considered by TSP, there exists a ct-sequence \( \langle \pi, ..., \tau \rangle \) s.t. \( G_\tau = \mathcal{H} \).

Similar to the DAG-based DFS over Chickering sequences employed by TSP, the lemma above motivates our permutation-based DFS over ct-sequences as shown in Algorithm 1.

Algorithm 1: DFS: dfs(\( \mathcal{P}, \pi, d, d_{cur}, t \))

Input: (a) \( \mathcal{P} \): a graphoid over \( V \); (b) \( \pi \in \Pi(\mathcal{V}) \); (c) \( d \): depth bound; (d) \( d_{cur} \): recorder of the recursive call; (e) \( t \): type of directed edges

Output: \( \tau \in \Pi(\mathcal{V}) \) where \( \text{score}(\tau) \geq \text{score}(\pi) \)

1 foreach \( (j \rightarrow k) \in E(G_\pi) \) do
2 \( \tau \leftarrow \text{tuck}(\pi, j, k) \)
3 if \( \text{score}(\tau) = \text{score}(\pi) \) and \( d_{cur} < d \) then
4 \( \tau \leftarrow \text{dfs}(\mathcal{P}, \tau, d, d_{cur} + 1, t) \)
5 if \( \text{score}(\tau) > \text{score}(\pi) \) then
6 return \( \tau \)
7 return \( \pi \)

Algorithm 2: GRaSP\( _t \): grasp(\( \mathcal{P}, \pi, d, t \))

Input: (a) \( \mathcal{P} \): a graphoid over \( V \); (b) \( \pi \in \Pi(\mathcal{V}) \); (c) \( d \): depth bound; (d) \( t \): tier of GRaSP

Output: \( \tau \in \Pi(\mathcal{V}) \) where \( \text{score}(\tau) \geq \text{score}(\pi) \)

2 if \( t \neq 0 \) then
3 \( \pi \leftarrow \text{grasp}(\mathcal{P}, \pi, d, t - 1) \)
4 do
5 \( \pi \leftarrow \tau \)
6 \( \tau \leftarrow \text{dfs}(\mathcal{P}, \pi, d, 1, t) \)
7 while \( \text{score}(\tau) > \text{score}(\pi) \)
8 return \( \tau \)

Algorithm 2 [Consider \( t = 0 \) in particular. Given an arbitrary initial permutation \( \pi \), Algorithm 1 performs a greedy procedure to identify a ct-sequence from \( \pi \). Figure 2 shows a simple example. Then Algorithm 2 iterates the DFS in Algorithm 1 until no sparser permutation can be found. The theorem below ensures that \( \hat{\tau} \) returned by Algorithm 2 induces \( G_{\hat{\tau}} \in \text{Pm}(\mathcal{P}) \).]

Theorem 4.5 [Appendix B] Given a graphoid \( \mathcal{P} \) over \( V \) and any \( \pi \in \Pi(\mathcal{V}) \), if \( G_{\pi} \notin \text{Pm}(\mathcal{P}) \), then there exists a ct-sequence \( \xi = \langle \pi, ..., \tau \rangle \) s.t. \( G_{\xi} \in \text{Pm}(\mathcal{P}) \).

By Theorem 4.5 the correctness of unbounded GRaSP\( _0 \) under faithfulness follows immediately from Theorem 5.8

As shown by [Forster et al. 2020], FCF(\( \mathcal{P} \)) = Fr(\( \mathcal{P} \)) holds under faithfulness. Since Algorithm 2 requires that the permutation returned by a higher tier of GRaSP cannot be denser than that returned by a lower tier, the correctness of unbounded GRaSP\( _1 \) and unbounded GRaSP\( _2 \) under faithfulness immediately follows. The sample version of GRaSP can be obtained by substituting the graphoid \( \mathcal{P} \) with an i.i.d. observational dataset \( D \), and \( \text{score}(\pi) \) with the BIC score of \( G_{\pi} \) from \( D \) (defined in Appendix B). Pointwise consistency under faithfulness directly follows from the local consis-
and Appendix G

Figure 2: Example of a ct-sequence $\langle \pi^1, \pi^2, \pi^3 \rangle$ where $I(\mathcal{P}) = \{(X_1, X_2) \mid \emptyset \}$. The blue (covered) edges indicate how a subsequent permutation is obtained by tuck. For example, $3 \rightarrow 1$ in (a) specifies that $\pi^2$ is obtained from $tuck(\pi^1 ; 3, 1)$. Also, Algorithm 1 returns $\pi^3 = \{1, 2, 3\}$ since the DAG in (c) is sparser than those in (a) and (b).

tency of BIC

Corollary 4.6 Unbounded GRaSP$_0$, GRaSP$_1$, and GRaSP$_2$ are correct and pointwise consistent under faithfulness.

Next, we want to highlight two logical discoveries with respect to the discussion of TSP and GRaSP$_0$.

Theorem 4.7 [Appendix B] Given a graphoid $\mathcal{P}$ and an initial permutation, the DAG returned by TSP is the same as the DAG induced by the output of unbounded GRaSP$_0$.

The theorem above suggests that TSP and GRaSP$_0$ are logically equivalent. Additionally, contrary to what Solus et al. [2021] argued, faithfulness is a necessary condition for TSP.

Theorem 4.8 [Appendix B] Given a graphoid $\mathcal{P}$, faithfulness is necessary for the correctness of TSP.

This theorem is entailed by a novel logical result that $\text{CFC}(\mathcal{P}) = u\text{Pa}(\mathcal{P})$ as proven in Appendix B. Thus, the two theorems together prompt the usage of GRaSP with a higher tier. Extending $E^0(\cdot)$ to $E^3(\cdot)$ and $E^3(\cdot)$ licenses a higher tier of GRaSP to attain a strictly sparser permutation under unfaithfulness. Examples of this sort will be studied in Section 5.1 and Appendix D.

Corollary 4.9 Given a graphoid $\mathcal{P}$, unbounded GRaSP$_2$ is correct under a strictly weaker causal razor than unbounded GRaSP$_1$, which is correct under a strictly weaker causal razor than unbounded GRaSP$_0$.

Further, in Appendix C we show the logical equivalence between unbounded GRaSP$_1$ and ESP. As a consequence, unbounded GRaSP$_2$ is a relaxation beyond the two causal razors discussed in [Solus et al., 2021]. That said, we are aware of cases where unbounded GRaSP$_2$ is incorrect under u-frugality. Such a counterexample will be studied in Section 5.1 and Appendix D.

We conclude this section by discussing how to use the DAG-inducing method in [Verma and Pearl, 1988] based on BIC scores. This facilitates our simulations done in Section 5.2. Given a semigraphoid $\mathcal{P}$ over $\mathcal{V}$, each $\pi \in \Pi(\mathcal{V})$ induces a DAG $\mathcal{G}_\pi$ satisfying the following condition:

$$X_j \in \mathcal{M} \iff (j \rightarrow k) \in E(\mathcal{G}_\pi) \quad \text{(VP)}$$

where $\mathcal{M}$ is a Markov boundary of $X_k$ relative to $X_{\text{Pre}(k, \pi)}$ (defined in Appendix A.3). Lemma A.4 highlights that the DAGs induced by (VP) and (RU) are equivalent when $\mathcal{P}$ is a graphoid. But (VP) is preferred since we can estimate the unique Markov boundary by the Grow-Shrink (GS) algorithm from [Margaritis and Thrun, 1999] using BIC scores and avoid hypothesis testing needed in (RU). We leave the discussion of the GS algorithm in Appendix E. In Section 5.2, we are going to evaluate the performance of GRaSP through (VP) and GS in light of finite samples.

5 SIMULATIONS

In this section, we review empirical results of unfaithful u-frugal models with respect to DAGs and algorithmic performance on Gaussian distributed data generated under a variety of situations. References to the code and instantiated models with replicability instructions are included on a GitHub site for the project. Also referenced will be a running version of GRaSP in the Tetrad project [Ramsey et al., 2018]) as well as tabular data for all simulations. A scalable Python translation of GRaSP$_2$ using (VP) with a linear, Gaussian BIC score is included in the causal-learn Python package.

5.1 U-FRUGAL FAITHFULNESS VIOLATIONS

In what follows, we consider three sets of u-frugal models that violate faithfulness. The sets of models correspond to: regular Gaussian distributions over four variables [Šimeček, 2006a], discrete distributions over four variables satisfying the intersection graphoid axiom and the Spohn condition (this includes all positive discrete distributions) [Šimeček, 2006b] (see Appendix A.2), and unfaithful DAGs (uDAGs) over five variables where a path cancellation in-
We evaluate the capabilities of GRaSP₀, GRaSP₁, and GRaSP₂ to recover u-frugal DAGs using an independence oracle on models from each set. We say that a GRaSP variant recovers the u-frugal model if it can do so from every permutation; if the algorithm can reach the u-frugal model from every permutation, then the correctness of the variant will be independent of the DFS implementation.

Table 1: The number of u-frugal models recovered by GRaSP₀, GRaSP₁, and GRaSP₂ from three sets of u-frugal models that violate faithfulness. A model is considered to be recovered if it is recovered from every permutation.

|          | GRaSP₀ | GRaSP₁ | GRaSP₂ | Total |
|----------|--------|--------|--------|-------|
| Gaussian | 0      | 7      | 10     | 10    |
| Discrete | 0      | 79     | 84     | 84    |
| uDAGs    | 0      | 19     | 49     | 61    |

Table 1 provides a computational proof that there are GRaSP₁ models not found by GRaSP₀, and GRaSP₂ models not found by GRaSP₁. These results support the claims in Corollary 4.9.

5.2 LINEAR GAUSSIAN SIMULATIONS

We studied GRaSP’s performance in the linear Gaussian case by varying simulations parameters around a configuration with 60 variables, an average degree of 6, and a sample size of 1,000 against two standard algorithms: fGES [Chickering, 2002, Ramsey et al., 2017] and PC [Spirtes et al., 2000]. In Figure 4 we vary the number of measured variables from 20 to 100 with values 20, 30, 40, 50, 60, 70, 80, 90, and 100. In Figure 5, we vary the average degree from 2 to 10 with values 2, 3, 4, 5, 6, 7, 8, 9, and 10. For Figure 5, we vary the sample size from 200 to 100,000, with values 200, 500, 1,000, 2,000, 5,000, 10,000, 20,000, 50,000, and 100,000. In all cases, we draw coefficient values uniformly from $U(-1, 1)$ and incorporate independent additive exogenous noise distributions set to $N(0, 1)$. All statistics are averaged over 20 independent runs. Finally, in Figure 6 we give the running times for our Java implementation of the algorithms. All of the algorithms except PC used BIC with a parameter penalty multiplier of 2 as a score; PC used partial correlation with a significance threshold of 0.001 as a conditional independence test. For the GRaSP variants, we allow tucks of covered edges up to depth 3, and tucks of non-covered edges at depth 1 when applicable.

In Figure 3, we show the result of varying the number of measured variables while holding the average degree constant in the estimated graphical model for each algorithm, respectively. For adjacency statistics, a true adjacency is a directed edge in the CPDAG of the generative graphical model and a positive adjacency is a directed edge in the CPDAG of the estimated DAG, with negative and false adjacencies are pairs of vertices that are (not) adjacent in the estimated graphical model for each algorithm, respectively. For arrowhead statistics, a true arrowhead is a directed edge in the CPDAG of the generative graphical model and a positive arrowhead is a directed edge in the CPDAG of the estimated DAG, with negative and false arrowheads indicating the absence of these directed edges in their respective CPDAGs.

In these figures, precision = $TP / (TP + FP)$ and recall = $TP / (TP + FN)$, where $TP$ is the number of true positives, $FP$ is the number of false positives, and $FN$ is the number of false negatives. We give precision and recall statistics for adjacencies and arrowheads separately. For adjacencies, true (false) adjacencies are pairs of vertices that are (not) adjacent in the generative graphical model, and positive (negative) adjacencies are pairs of vertices that are (not) adjacent in the estimated graphical model for each algorithm, respectively. For arrowhead statistics, a true arrowhead is a directed edge in the CPDAG of the generative graphical model and a positive arrowhead is a directed edge in the CPDAG of the estimated DAG, with negative and false arrowheads indicating the absence of these directed edges in their respective CPDAGs.

Figure 3 shows that algorithmic performance is strongly dependent on the average degree. While the compared algorithms generally perform well on sparse models, their performance drops off as the density increases. The exception is GRaSP₂, which dominates this group of algorithms, with a strong performance for both adjacencies and arrowheads as average degree is increased.

Figure 4 shows the result of varying the number of measured variables. Notably, increasing the number of measured variables while holding the average degree constant decreases graph density. We see upward trends for some arrowhead statistics corresponding to this decrease in density.

---

11 In the Java implementation of the algorithm, we include parameters for uncovered depth and non-singular depth to provide the user with more control over this heuristic.

12 A CPDAG (a.k.a. “pattern”) is a graphical representation of the Markov equivalence class for a DAG. See [Spirtes et al., 2000] for details.
Again, GRaSP$_2$ dominates this group of algorithms, with strong precision and recall for both adjacencies and arrowheads.

All the algorithms compared in these simulations claim pointwise consistency, however, only GRaSP$_2$ seems to back up these claims in Figure 5. This might suggest that GRaSP$_2$ is better equipped to handle almost-violations of faithfulness in linear Gaussian models. As with previous figures, GRaSP$_2$ dominates this group of algorithms for precision and recall for both adjacencies and arrowheads for all sample sizes studied.

Figure 6 shows that all the algorithms on average return in under two minutes for the studied scenarios. However, given the log scale, it should be noted that the computation time for GRaSP$_2$ increases exponentially with respect to the average degree of the graph and with respect to the number of measured variables. Other algorithms see similar slow-downs, but, other than GRaSP$_1$, none of the other algorithms experience as significant of a slow-down.

In this paper, we focused on algorithms that can run on a 100 variable problem in a reasonable amount of time on a laptop. However, we would be remiss if we did not mention a recent algorithm by Lu et al. [2021] called Triplet A$^*$ that performs in terms of accuracy as well as, if not better than, GRaSP$_2$. We declined to directly compare the Triple A$^*$ algorithm in our Figures because it was unable to finish our simulations in reasonable time; for instance, the point they give in their Figure 6 for the 60-variable, average degree 5 case was already as slow as could be managed (personal communication); we took our simulations out to an average degree of 10. In lieu of this, we include in Appendix F.1 results of running GRaSP$_2$ on their published simulation data.

6 EMPIRICAL EXAMPLE

We give a simple empirical example, the 6-variable Airfoil example from the Irvine Machine Learning Repository.

---

[13] All simulations in this paper were run on a MacBook Pro laptop computer, M1, 2020, with 16G of RAM, using the Corretto 18 Java SDK. Memory is the main resource constraint on the procedure, which is needed for caching scores. Thanks to the comment of an anonymous reviewer, a machine with 256GB of RAM may be useful for analyses significantly larger than the ones studied.
The experiment measures sound pressure elicited by an airfoil in a wind tunnel. The variables in the example are as follows: (1) Velocity of the wind in the tunnel, (2) chord length of the airfoil, (3) angle of attack of the airfoil, (4) displacement of the wind away from the airfoil, (5) frequency of the elicited sound, and (6) measured pressure of the elicited sound. (1), (2), and (3) are experimental variables and thus exogenous; (6) in the experiment is endogenous. The GRaSP₂, PC and GES graphs are given in Appendix F.2. The GRaSP₂ model (which is the same as the SP model) is uniquely frugal; background knowledge is satisfied, except possibly for (3), which looks to be not exogenous in the model; here, it helps to remember that latent variables might exist. This raises the question as to whether a causally insufficient algorithm might find a model consistent with (3) being exogenous. We will explore how GRaSP₂ may be used to do latent variable reasoning to see whether (3) remains non-exogenous in general.

This example has a number of advantages: (a) It is an experiment so readily interpretable as a causal system; (b) because it is an experiment, partial ground truth for the system can easily be ascribed to the data, and since this produces a single model, we can simply compare the output of GRaSP₂ to the output of SP to show that GRaSP₂ finds the optimal BIC model.

Further empirical examples with SP (where possible), GRaSP₂, fGES, and PC are given on our GitHub site.

7 DISCUSSION

Permutation-based reasoning in designing causal search algorithms is increasingly influential in the literature, including the methods from Teysier and Koller [2005] and Raskutti and Uhler [2018]. We propose a class of algorithms under the generic name GRaSP characterized by an efficient permutation-based operation, tuck. All tiers of GRaSP are shown to be correct and pointwise consistent under the assumption of faithfulness. Also, we show that the two lower tiers of GRaSP are logically equivalent to the algorithms TSP and ESP discussed in Solus et al. [2021]. We further prove that the final tier of GRaSP makes a strictly weaker assumption than its lower-tier counterparts and demonstrate that it outperforms the lower-tier algorithms and two standard causal search algorithms, PC and fGES, in simulations.

Discussion of GRaSP can be extended in several directions. First, we have already begun to explore even higher tiers of GRaSP which relax the search criterion even further. Figure 3 suggests that GRaSP may provide tools helpful for the discussion of dense graph search. Given the hierarchy of GRaSP, higher tiers will hopefully improve the performance statistics and employ weaker assumption than the existing tiers. Ultimately, we hope to develop a tier of GRaSP that is correct under u-frugality alone.

Second, many advances have been made in the area of more or completely general modeling of discrete distributions, with corresponding improvements in accuracy of causal search for algorithms taking general modeling assumptions into account. It would be helpful to consider how such ideas can be incorporated into GRaSP. For example, Huang et al. [2018] show how a consistent general score can be incorporated into GES; it will be interesting to see whether GRaSP is able to show similar improvement in applicability when using such a score.

Third, we have analyzed Gaussian simulations in Section 5 but some simulation work needs to be done to show that GRaSP works well for discrete distributions (where the theory is already applicable) and also for mixed Gaussian/discrete distributions studied in Andrews et al. [2019].

Fourth, the discussion of this paper is built upon the assumptions of causal sufficiency, that is, no latent common causes, and no selection bias. Causal search without these assumptions was pioneered by the FCI algorithm from Spirtes et al. [2000] and Zhang [2008]. To improve empirical performance of FCI, Ogarrio et al. [2014] initiated a hybrid algorithm GFCI which combines GES with FCI. To follow suit, we plan to explore an algorithm that incorporates GRaSP into GFCI (in place of GES), further improving this empirical performance.

Fifth, more direct comparisons to other algorithms need ideally to be done. As a step in this direction, we include figures on our GitHub site using the simulation parameters in Lu et al. [2021], corresponding to their Figures 6, so there is oblique comparison to the algorithms in those figures, including GES and PC in the PCALG package [Kalisch et al., 2012], Triplet A∗ [Lu et al., 2021], NOTEARS [Zheng et al. 2018], the GSP implementation in the Python causaldag package, LiNGAM [Shimizu et al., 2006], and MMHC [Tsamardinos et al., 2006]. The reader is invited to explore those comparisons.

Finally, we have taken up just one real data example in this paper, but it is useful to point out in a forward-looking way that improvements in the ability to handle latent and mixed continuous/discrete variables in a scalable and accurate causal search algorithm would put one in a good position to analyze a number of otherwise difficult real data examples. Accurate preliminary results consistent with ground truth using the suggested modification of GFCI for a number of mixed datasets from the Irvine Machine Learning Repository [Dua and Graff, 2017], for instance, suggest that this would be a good direction to look for new practical methods (cf. Raghu et al. 2018).
Author Contributions

WL contributed theoretical results, with input from BA, while BA and JR worked on the algorithm implementations and contributed empirical results. All authors contributed to algorithmic development.

Acknowledgements

We thank Greg Cooper, Clark Glymour, Ignavier Ng, Peter Spirtes, Jiji Zhang, and Kun Zhang for discussion and feedback, and the anonymous reviewers for detailed and insightful comments.

References

Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. Learning high-dimensional directed acyclic graphs with mixed data-types. In The 2019 ACM SIGKDD Workshop on Causal Discovery, pages 4–21. PMLR, 2019.

David Maxwell Chickering. A transformational characterization of equivalent Bayesian network structures. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, page 87–98, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

David Maxwell Chickering. Optimal structure identification with greedy search. Journal of machine learning research, 3:507–554, 2002.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml

Malcolm Forster, Garvesh Raskutti, Reuben Stern, and Natali Weinberger. The frugal inference of causal relations. The British Journal for the Philosophy of Science, 2020.

Dominique M. A. Haughton. On the choice of a model to fit data from an exponential family. The Annals of Statistics, 16:342 – 355, 1988.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1551–1560, 2018.

Markus Kalisch, Martin Mächler, Diego Colombo, Marco H Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. Journal of statistical software, 47:1–26, 2012.

Robert E Kass and Paul W Vos. Geometrical Foundations of Asymptotic Inference. John Wiley & Sons, Hoboken, NJ, 2011.

Ni Y Lu, Kun Zhang, and Changhe Yuan. Improving causal discovery by optimal Bayesian network learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 8741–8748, 2021.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. Advances in neural information processing systems, 12, 1999.

FatemehMohammadi, Caroline Uhler, Charles Wang, and Josephine Yu. Generalized permutohedra from probabilistic graphical models. SIAM Journal on Discrete Mathematics, 32:64–93, 2018.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In Conference on probabilistic graphical models, pages 368–379. PMLR, 2016.

Judea Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., 1988.

Judea Pearl. Causality: Models, Reasoning and Inference. Cambridge University Press, USA, 2nd edition, 2009.

Vineet K Raghu, Joseph D Ramsey, Alison Morris, Dimitrios V Manatakis, Peter Spirtes, Panos K Chrysanthis, Clark Glymour, and Panayiotis V Benos. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. International journal of data science and analytics, 6:33–45, 2018.

Joseph Ramsey, Jiji Zhang, and Peter Spirtes. Adjacency-faithfulness and conservative causal inference. In Proceedings of the conference on Uncertainty in artificial intelligence, pages 401–408, 2006.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. International journal of data science and analytics, 3:121–129, 2017.

Joseph D Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez-Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A Barnes, and Clark Glymour. Tetrad—a toolbox for causal discovery. In 8th International Workshop on Climate Informatics, 2018.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. Stat, 7, 2018.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7, 2006.
Petr Šimecek. Gaussian representation of independence models over four random variables. In Proceedings of the International Conference on Computational Statistics. Citeseer, 2006a.

Petr Šimecek. A short note on discrete representability of independence models. In Proceedings of the Workshop on Probabilistic Graphical Models, pages 287–292. Citeseer, 2006b.

Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108:795–814, 2021.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Wolfgang Spohn. On the properties of conditional independence. In Paul Humphreys, editor, *Patrick Suppes: Scientific Philosopher: Volume I. Probability and Probabilistic Causality*, pages 173–196. Springer Netherlands, 1994.

Milan Studený. *Probabilistic Conditional Independence Structures*. Information Science and Statistics. Springer, 2005.

Marc Teyssier and Daphne Koller. Ordering-based search: a simple and effective algorithm for learning Bayesian networks. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, pages 584–590, 2005.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Thomas Verma and Judea Pearl. Causal networks: semantics and expressiveness. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, pages 69–78, 1988.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172:1873–1896, 2008.

Jiji Zhang. A comparison of three Occam’s razors for Markovian causal models. *The British journal for the philosophy of science*, 64:423–448, 2013.

Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2): 239–271, 2008.

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.
A BACKGROUND MATERIALS

A.1 GRAPHICAL DEFINITIONS

A directed graph \( G \) over a set of measured variables \( V = \{X_1, ..., X_m\} \) consists of \( m \) vertices \( v = \{1, ..., m\} \) where each vertex \( i \in v \) associates to a variable \( X_i \in V \), and each edge in \( G \) is directed with the form \( j \rightarrow k \) and no vertex has a directed edge to itself. A directed path \( p \) is a sequence of vertices \( (i_1, i_2, ..., i_k) \) for some \( k \geq 2 \) where \( i_j \in v \) for each \( 1 \leq j \leq k \), and \( i_j \) and \( i_{j+1} \) are connected by a directed edge (i.e., \( i_j \rightarrow i_{j+1} \) or \( i_{j+1} \rightarrow i_j \)). Such a path \( p \) is unidirectional if \( i_1 \rightarrow i_{j+1} \) for each \( 1 \leq j < k \). A directed acyclic graph (DAG) is a directed graph where no vertex can have a unidirectional directed path to itself.

Denote \( \text{E}(G) \) as the set of directed edges in \( G \). A pair of DAGs \( G_1, G_2 \) over the same set of variables \( V \) are equivalent if and only if \( \text{E}(G_1) = \text{E}(G_2) \). Let \( \text{Pa}(j, G) = \{k \in v : (k \rightarrow j) \in \text{E}(G)\} \) be the set of parents of \( j \) in \( G \), and \( \text{Ch}(j, G) = \{k \in v : (j \rightarrow k) \in \text{E}(G)\} \) be the set of children of \( j \) in \( G \). \( \text{An}(j, G) \), the ancestors of \( j \) in \( G \), is defined by the transitive closure of \( \text{Pa}(j, G) \). Similarly, \( \text{De}(j, G) \), the descendants of \( j \) in \( G \), is defined by the transitive closure of \( \text{Ch}(j, G) \) and union with \( \{j\} \) itself (i.e., \( j \) is its own descendant). Further let \( \text{Nd}(j, G) = v \setminus \text{De}(j, G) \) be the set of \( j \)'s non-descendants.

A pair of vertices \( j, k \in v \) are said to be adjacent in \( G \) if \( (j \rightarrow k) \in \text{E}(G) \) or \( (k \rightarrow j) \in \text{E}(G) \). For any triple of pairwise distinct vertices \( i, j, k \in v \), we say that \( (i, j, k) \) is unshielded if \( (i, j) \) and \( (j, k) \) are adjacent pairs in \( G \), but not \( (i, k) \). \( (i, j, k) \) forms a triangle if they are pairwise adjacent. If \( (i, j, k) \) is an unshielded triple or is a triangle, \( j \) is a collider (on the directed path \( (i, j, k) \)) if \( (i \rightarrow j), (k \rightarrow j) \in \text{E}(G) \), and a non-collider otherwise. A directed path \( p \) is a trek if it contains no collider.

For any \( j, k \in v \) and any \( i \in v \setminus \{j, k\} \), \( j \) and \( k \) are \( d \)-connected given \( i \) in \( G \) if there exists a directed path \( p \) between \( j \) and \( k \) in \( G \) such that no non-collider on \( p \) is in \( i \), and each collider \( l \) on \( p \) or \( l \)'s descendant is in \( i \). \( j \) and \( k \) are \( d \)-separated given \( i \) in \( G \) if \( j \) and \( k \) are not \( d \)-connected given \( i \). For any disjoint subsets of vertices \( j, k, i \subseteq v \), \( j \) and \( k \) are \( d \)-separated given \( i \) in \( G \) if \( j \) and \( k \) are \( d \)-separated by \( i \) in \( G \) for every \( j \in j \) and every \( k \in k \).

Given a model \( (G, P) \) over \( V \), \( G \) is said to be local Markov to \( P \) if \( X_j \perp_{G, P} X_{\text{Nd}(j, G)} \setminus X_{\text{Pa}(j, G)} \mid X_{\text{Pa}(j, G)} \) for every \( j \in v \). It is a well-known fact that \( G \) is local Markov to \( P \) if and only if \( I(G) \subseteq I(P) \) (i.e., global Markov as defined by d-separation).

A.2 GRAPHOID AXIOMS

For any pairwise disjoint sets of variables \( W, X, Y, \) and \( Z, \)

\[
\begin{align*}
X \perp Y \mid Z & \Rightarrow Y \perp X \mid Z \quad \text{(symmetry)} \\
X \perp Y \cup W \mid Z & \Rightarrow (X \perp Y \mid Z) \land (X \perp W \mid Z) \quad \text{(decomposition)} \\
X \perp Y \cup W \mid Z & \Rightarrow X \perp Y \mid Z \cup W \quad \text{(weak union)} \\
(X \perp Y \mid Z) \land (X \perp W \mid Z \cup Y) & \Rightarrow X \perp Y \cup W \mid Z \quad \text{(contraction)} \\
(X \perp Y \mid Z \cup W) \land (X \perp W \mid Z \cup Y) & \Rightarrow X \perp Y \cup W \mid Z \quad \text{(intersection)} \\
(X \perp Y \mid Z) \land (X \perp W \mid Z) & \Rightarrow X \perp Y \cup W \mid Z \quad \text{(composition)}
\end{align*}
\]

A distribution \( P \) is a semi-graphoid if \( I(P) \) is closed under symmetry, decomposition, weak union, and contraction. A semi-graphoid \( P \) is a graphoid if \( I(P) \) is closed under intersection. A graphoid \( P \) is compositional if \( I(P) \) is closed under composition. See Chapter 2 of [Student, 2005] for a more comprehensive study of graphoid axioms. In addition, applications of symmetry in our upcoming proofs will be done implicitly for the sake of simplicity.

Additionally, Spohn [1994] notes that the following property necessarily holds in the independence models induced by positive discrete probability distributions. For any pairwise disjoint sets of variables \( W, X, Y, \) and \( Z, \)

\[
(X \perp Y \mid W \cup Z) \land (W \perp Z \mid X \cup Y) \land (W \perp Z \mid X) \Rightarrow \left[ (W \perp Z \mid Y) \Leftrightarrow (W \perp Z \mid \emptyset) \right]
\]

Spohn condition

A.3 DAG INDUCED FROM A PERMUTATION

Definition A.1 Given a semi-graphoid \( P \) over \( V \), for every \( X \in V \), we say that \( M \subseteq V \) is a Markov blanket of \( X \) relative to \( Z \subseteq V \setminus \{X\} \) if
We want to show that the two DAG-inducing methods are equivalent when the underlying distribution is a graphoid.

Immediate from Proof.

We divide the proof into two directions: (VP) and (RU), and (VP) is the construction of a boundary DAG in [Verma and Pearl, 1988]. On the other hand, given a graphoid $P$ over $V$, each $\pi \in \Pi(v)$ induces a DAG satisfying the following condition:

$$j \in \text{Pre}(k, \pi) \text{ and } X_j \perp_{P} X_k | X_{\text{Pre}(k, \pi) \setminus \{j\}} \iff (j \rightarrow k) \in E(G_{\pi}).$$  \hfill (RU)

We want to show that the two DAG-inducing methods are equivalent when the underlying distribution is a graphoid.

Lemma A.4 Given a graphoid $P$ over $V$, consider any $\pi \in \Pi(v)$. Let $G_{\pi}$ be the DAG induced from $\pi$ by (VP), and $H_{\pi}$ be the DAG induced from $\pi$ by (RU). Then $G_{\pi} = H_{\pi}$.

Proof. We divide the proof into two directions: (VP) $\Rightarrow$ (RU), and (VP) $\Leftarrow$ (RU). Consider any $j, k \in v$ where $\pi[j] < \pi[k]$ such that $j \in \text{Pre}(k, \pi)$. Let $M$ be the unique Markov boundary $\text{MB}(X_k, X_{\text{Pre}(k, \pi)})$.

$[\Rightarrow]$ Suppose that $(j \rightarrow k) \notin E(G_{\pi})$. We have $X_j \notin M$. By Definition A.1 (ii), we then have,

$$(i) \ M \subseteq Z;$$

$$(ii) \ X \perp_{P} (Z \setminus M) | M.$$

Such a Markov blanket $M$ is said to be a Markov boundary if it further satisfies the following condition:

$$(iii) \text{ there does not exist } M' \subset M \text{ s.t. } X \perp_{P} (Z \setminus M') | M'.$$

Lemma A.2 [Verma and Pearl 1988] Given a graphoid $P$ over $V$, for every $X \in V$ and every $Z \subseteq V \setminus \{X\}$, there is a unique Markov boundary of $X$ relative to $Z$.

In the following, we use $\text{MB}(X, Z)$ to refer to the unique Markov boundary of $X$ relative to $Z$. The subscript $P$ will be suppressed if the underlying graphoid is clear from context.

Lemma A.3 Given a graphoid $P$ over $V$, for every $X \in V$ and every $Z \subseteq V \setminus \{X\}$, if $M$ is a Markov blanket of $X$ relative to $Z$, then $\text{MB}(X, Z) \subseteq M$.

Proof. Immediate from Definition A.1 and Lemma A.2

Next, we revisit the two methods of inducing a DAG from a permutation. Given a semigraphoid $P$ over $V$, each $\pi \in \Pi(v)$ induces a DAG satisfying the following condition:

$$X_j \in M \iff (j \rightarrow k) \in E(G_{\pi})$$  \hfill (VP)

where $M$ is a Markov boundary of $X_k$ relative to $X_{\text{Pre}(k, \pi)}$. (VP) is the construction of a boundary DAG in [Verma and Pearl, 1988].

Theorem A.5 [Pearl 1988] Given a semigraphoid $P$ over $V$, $G_{\pi}$ induced by $\pi$ using (VP) is Markovian and SGS-minimal for any $\pi \in \Pi(v)$.

Theorem A.5 Given a graphoid $P$ over $V$, $G_{\pi}$ induced by $\pi$ using (RU) is Markovian and SGS-minimal for any $\pi \in \Pi(v)$.

Proof. Immediate from Lemma A.4 and Theorem A.5\(^{14}\)

\(^{14}\)The two DAG-inducing methods were not differentiated in [Raskutti and Uhler, 2018]. Thus, we provide a proof of Theorem A.5.
B Correctness Results

First, we introduce some permutation-based notations to facilitate our coming proofs. In this section, we use $G_π$ to denote the DAG induced by $π$ from a graphoid $P$ using (RU) unless specified otherwise.

Given a set of variables $V$, consider any $π ∈ Π(V)$ and any pair $j, k ∈ V$ where $π[j] < π[k]$. $π$ can be written as $(δ_{<j}, j, δ_{j<k}, k, δ_{>k})$ such that $δ_{<j} = \{π_i : 1 ≤ i < π[j]\}$, $δ_{j<k} = \{π_i : π[j] < i < π[k]\}$, and $δ_{>k} = \{π_i : π[k] < i ≤ |Π|\}$. When $δ_{j<k} = ∅$, we say that $j$ and $k$ are π-adjacent. In that case, $π$ can be written as $(δ_{<j}, j, k, δ_{>k})$ instead.

Definition B.1 Given a set of variables $V$, for any $π, τ ∈ Π(V)$,

(a) $τ$ is said to be $(j, k)$-different from $π$ for some $j, k ∈ V$ if $j$ and $k$ are π-adjacent (i.e., $π = (δ_{<j}, j, k, δ_{>k})$) and $τ = (δ_{<j}, j, k, δ_{>k})$;

(b) $π$ and $τ$ are said to be in adjacent transposition (AT) if they are $(j, k)$-different for some $j, k ∈ V$.

Lemma B.2 Given a graphoid $P$ over $V$, consider any $H ∈ ΠMC(P)$. If $π ∈ Π(V)$ is a causal order of $G_π$, then $E(G_π) ⊆ E(H)$. Also, $G_π = H$ if $H ∈ SGS(P)$.

Proof. Consider any $k ∈ V$ and $H ∈ ΠMC(P)$ (i.e., the set of $k$’s non-descendants in $H$). Since $H ∈ ΠMC(P)$, it follows that $X_k ⊥ π X_{Nd(k, H)} \setminus X_{Pa(k, H)} \setminus X_{Pa(k, H)}$. Also, we have $Pa(k, H) ⊆ Pre(k, π) ⊆ Nd(k, H)$ from π’s being a causal order of $H$. By decomposition, we have $X_k ⊥ π X_{Pa(k, π)} \setminus X_{Pa(k, H)} | X_{Pa(k, H)}$ such that $X_{Pa(k, H)}$ is a Markov blanket of $X_k$ relative to $X_{Pa(k, π)}$. By Lemma A.3, we have $MB(X_k, X_{Pre(k, π)}) ⊆ Pa(k, H)$. Consider $G_π$ induced by (VP). The above entails that $E(G_π) ⊆ E(H)$ since $Pa(k, G_π) ⊆ MB(X_k, X_{Pre(k, π)}) ⊆ Pa(k, H)$ for each $k ∈ V$. Due to Lemma A.4, $E(G_π) ⊆ E(H)$ still holds even if $G_π$ is induced by (RU). Lastly, $G_π = H$ follows from Definition B.4 if $H ∈ SGS(P)$. □

Lemma B.3 (Solus et al. [2021]) Given a graphoid $P$ over $V$, consider any $π, τ ∈ Π(V)$ where $τ$ is $(j, k)$-different from $π$ for some $j, k ∈ V$. Then $G_π = G_τ$ if and only if $X_j ⊥ π X_k | X_{Pre(j, π)}$.

Proof. Suppose that $X_j ⊥ π X_k | X_{Pre(j, π)}$. By (RU), we have $(j → k) ∈ E(G_π)$. Note that $(j → k) ∉ E(G_τ)$ since $τ[k] < τ[j]$ and $τ$ is a causal order of $G_τ$ by construction. Hence, $G_π ≠ G_τ$.

On the other hand, suppose that $X_j ⊥ π X_k | X_{Pre(j, π)}$. Since $τ$ is $(j, k)$-different from $π$, we have $π = (δ_{<j}, j, k, δ_{>k})$ and $τ = (δ_{<j}, j, k, δ_{>k})$ according to Definition B.1 (a). By (RU), we know that $(k → j) ∉ E(G_τ)$. Hence, $π$ is a causal order of $G_τ$. By Theorem 3.5, $G_τ ⊆ SGS(P)$. Therefore, it follows from Lemma B.2 that $G_τ = G_π$. □

Lemma B.4 Given a graphoid $P$ over $V$, consider any $π ∈ Π(V)$. Suppose that $G_π$ contains a covered edge $j → k$ where $π = (δ_{<j}, j, δ_{j<k}, k, δ_{>k})$. If $τ = (δ_{<j}, j, k, δ_{j<k}, δ_{>k})$, then $G_π = G_τ$.

Proof. Since $j → k$ is a covered edge in $G_π$, it follows that $(i → k) ∉ E(G_π)$ for each $i ∈ δ_{j<k}$, and thus $X_i ⊥ π X_k | X_{Pre(i, π)}$. By (RU), $G_π$ can be obtained after |δ_{j<k}| applications of Lemma B.3 □

Theorem B.5 [Zhang [2013]] Given a set of variables $V$, for any $G, H ∈ ΠMC(V)$, if $E(G) ⊆ E(H)$, then $I(H) ⊆ I(G)$.

Lemma B.6 [Chickering [1995]] Consider any DAG $G$. Let $H$ be the result of reversing $(i → j) ∈ E(G)$. Then $H ∈ ΠMEC(G)$ if and only if $i → j$ is a covered edge.

Theorem B.7 [Chickering [1995]] Consider any pair of DAGs $G$ and $H$ over the same set of variables s.t. $H ∈ ΠMEC(G)$, and for which there are $k$ edges in $G$ that have opposite orientation in $H$. Then there exists a sequence of $k$ distinct covered edge reversals in $G$ s.t. $G$ becomes $H$ after all reversals.

Lemma B.8 Given a graphoid $P$ over $V$, consider any $π ∈ Π(V)$. Suppose that $(j → k) ∈ E(G_π)$ is a covered edge, and let $H$ be the DAG resulted from reversing $(j → k)$ in $G_π$. If $τ = tuck(π, j, k)$, then

(a) $τ$ is a causal order of $H$;
(b) \( E(G_\pi) \subseteq E(H) \);
(c) \( |E(G_\pi)| \leq |E(G_\tau)| \);
(d) \( I(G_\pi) \subseteq I(G_\tau) \).

Proof. A similar lemma has been shown in [Solus et al., 2021]. First, we write \( \pi = (\delta_{<j}, j, \delta_{j \sim k}, k, \delta_{>k}) \) as usual. Consider \( \pi' = (\delta_{<j}, j, k, \delta_{j \sim k}, \delta_{>k}) \). By Lemma [B.4] we have \( G_\pi = G_{\pi'} \). Note that \( \tau = \text{tuck}(\pi, j, k) = (\delta_{<j}, k, j, \delta_{j \sim k}, \delta_{>k}) \) because \( \text{tuck}(\pi, j, k) \) is a covered tuck. Thus, \( \tau \) is \((j, k)\)-different from \( \pi' \). Also, since \( \pi' \) is a causal order of \( G_\pi \), it follows that \( \tau \) is a causal order of \( H \) and thus (a) is proven.

Next, observe that \( I(G_\pi) = I(H) \) from Lemma [B.6]. From \( G_\pi \in \text{CMC}(P) \) by Theorem [B.3] we know that \( H \in \text{CMC}(P) \). Thus, (b) immediately follows from (a) and Lemma [B.2]. Also, (c) is entailed by \( |E(G_\tau)| \leq |E(H)| = |E(G_\pi)| \). Finally, by Theorem [B.3] we have \( I(G_\pi) = I(H) \subseteq I(G_\tau) \) as desired in (d). 

Before we compare TSP and unbounded GRaSP\(_0\), we want to make an assumption related to how the set of covered edges in any particular DAG is ordered. To see the importance of such an assumption, observe that different orderings of \( E^0(G_\pi) \) (i.e., the set of covered edges in an induced DAG \( G_\pi \)) can alter the output of TSP and also GRaSP\(_0\). For example, suppose that \((j \rightarrow k), (j' \rightarrow k') \in E^0(G_\pi)\). Say the DFS of GRaSP\(_0\) starts with performing \( \text{tuck}(\pi, j, k) \) and leads to some permutation \( \tau \). However, choosing to perform \( \text{tuck}(\pi, j', k') \) instead at the beginning of the DFS procedure can lead to some \( \tau' \) where \( G_{\pi'} \neq G_{\tau'} \). Hence, we enforce the assumption that the ordering of \( E^0(G) \) for any DAG \( G \) is fixed arbitrarily. For instance, \((j \rightarrow k) \) precedes \((j' \rightarrow k') \) in \( E^0(G) \) if \( j < j' \), or if \( j = j' \) and \( k < k' \). Consequently, the issue of order-dependence can be avoided even when comparing a Chickering sequence found by TSP and a ct-sequence found by unbounded GRaSP\(_0\). In the following, this assumption will be made implicitly.

Now we revisit how TSP works. Given a graphoid \( P \) over \( V \) and an initial permutation \( \pi \in \Pi(v) \), TSP begins with setting \( G \) as the induced \( G_\pi \). Starting with the root \( G \), TSP performs DFS to identify a SGS-minimal DAG \( H \) connected by a Chickering sequence from \( G \) such that \( |E(G)| > |E(H)| \). TSP returns \( G \) if no such \( H \) is found. Otherwise, it updates \( G \) as \( H \) and repeat the procedure.

The DFS procedure of TSP aims to traverse from one SGS-minimal DAG to another SGS-minimal DAG by the construction of a Chickering sequence. Though we know that a Chickering sequence is obtained by the reversals of covered edges and deletions of directed edges, Solus et al. [2021] did not specify any ordering of these operations. Below we provide a more precise definition of the Chickering sequences considered by TSP.

**Definition B.9** Given a graphoid \( P \) over \( V \), a TSP-Chickering sequence \( \zeta = \langle G^1, \ldots, G^m \rangle \) is a Chickering sequence satisfying the following condition:

(a) \( G^1, G^m \in \text{SGS}(P) \);
(b) \( G^i \) and \( G^{i'} \) are pairwise distinct for \( 1 \leq i < i' \leq m \);
(c) if \( |E(G^1)| = |E(G^m)| \), then \( G^1, \ldots, G^m \in \text{SGS}(P) \) where they differ by the reversals of some covered edges;
(d) otherwise, there exists a turning index \( 1 < l < m \) such that (i) \( G^1, \ldots, G^{l-1} \in \text{SGS}(P) \), (ii) \( G^1, \ldots, G^l \) differ by the reversals of some covered edges, and (iii) \( G^{l+1} \) is obtained from deleting a directed edge in \( G^l \notin \text{SGS}(P) \) for each \( l \leq i < m \).

Readers are suggested to find the original pseudocode of TSP in [Solus et al., 2021] to verify that our Definition [B.9] is a fair description of the Chickering sequences considered by TSP. Conditions (a) and (b) are straightforward. (c) refers to the case where TSP cannot find a sparser SGS-minimal DAG. So if any \( G^i \) in \( \zeta \) were non-SGS-minimal, then TSP would have obtained a proper subgraph of \( G^i \) which is SGS-minimal by a series of edge-deletion. (d) refers to the case where TSP manages to find a sparser SGS-minimal DAG. Notice that \( G^2 \) must be obtained by a covered edge reversal from \( G^1 \) since \( G^1 \in \text{SGS}(P) \). If \( G^2 \notin \text{SGS}(P) \), then TSP can obtain the desired SGS-minimal DAG by a series of edge-deletion from \( G^2 \). But if \( G^2 \in \text{SGS}(P) \), the procedure above repeats until finding the turning index \( l \) such that \( G^l \notin \text{SGS}(P) \) and then the sparser \( G^m \in \text{SGS}(P) \) can be obtained by a series of edge-deletion from \( G^l \).

Now we compare TSP and unbounded GRaSP\(_0\) by considering their respective sequences traversed in the DFS procedure.

**Lemma B.10** Given a graphoid \( P \) over \( V \), consider any \( \pi \in \Pi(v) \) and \( \tau = \text{tuck}(\pi, j, k) \) where \( (j \rightarrow k) \in E^0(G_\pi) \). Given that \( \Sigma = (\pi, \tau) \) is a ct-sequence,
(a) if $|E(G_j)| = |E(G_\pi)|$, then $C = \langle G_\pi, G_\tau \rangle$ is a TSP-Chickering sequence where $G_\tau$ is obtained from reversing $(j \to k) \in E^0(G_j)$.

(b) otherwise, there exists a TSP-Chickering sequence $C = \langle G_\pi = G^1, \ldots, G^m = G_\tau \rangle$ s.t. $G^2$ is obtained from reversing $(j \to k) \in E^0(G_\pi)$, and $G^{i+1}$ is obtained from deleting a directed edge in $G^i$ for each $2 \leq i < m$.

Proof. First, consider the DAG $H$ obtained from reversing $(j \to k) \in E^0(G_\pi)$. We start with the case in (a) where $|E(G_\pi)| = |E(G_\tau)| = |E(H)|$. We want to show that $G_\pi = H$. By Lemma B.8(b), we have $E(G_\pi) \subseteq E(H)$. If $E(G_\tau) \subseteq E(H)$ holds, then $|E(G_\pi)| = |E(H)|$ will be violated. Hence, we have $G_\tau = H$ and thus $C = \langle G_\pi, H = G_\tau \rangle$ is our desired TSP-Chickering sequence.

For (b), it follows from Lemma B.8(c) that $|E(G_\pi)| < |E(H)| = |E(G_\tau)|$. Let $G_\pi$ and $H$ be $G^1$ and $G^2$ respectively. By Lemma B.8(b) again, we have $E(G_\tau) \subseteq E(G^2)$ such that we can remove a directed edge from $G^2$ once at a time until obtaining $G_\tau$. Therefore, we have the desired TSP-Chickering sequence in (b).

Lemma B.11 Given a graphoid $P$ over $V$, consider any TSP-Chickering sequence $C = \langle G^1, \ldots, G^m \rangle$. Let $\pi^1$ be a causal order of $G^1$. Then

(a) if $|E(G^1)| = |E(G^m)|$, then $G^{i+1} = G_{\pi^i+1} = G_{\text{tuck}(\pi^i,j,k)}$ where $j \to k$ is the covered edge reversed to obtain $G^{i+1}$ from $G^i$ for each $1 \leq i < m$ s.t. $\pi = (\pi^1, \ldots, \pi^m)$ is a ct-sequence;

(b) otherwise, then $G^{i+1} = G_{\pi^i+1} = G_{\text{tuck}(\pi^i,j,k)}$ where $j \to k$ is the covered edge reversed to obtain $G^{i+1}$ from $G^i$ for each $1 \leq i < l$ where $l$ is the turning index of $C$ and $G_{\pi^l} = G^m$ s.t. $\pi = (\pi^1, \ldots, \pi^l)$ is a ct-sequence.

Proof. (a) can be easily shown by Lemma B.8(a) and Lemma B.2. For (b), the proof of $G^i = G_{\pi^i}$ for each $1 \leq i < l$ is similar to that in (a). So we consider $l$ where $G^l \notin \text{SGS}(P)$ according to Definition B.9(d). However, it follows from Lemma B.8(a) that $\pi^l$ is a causal order of $G^l$. Since $E(G^m) \subseteq E(G^l)$, we know that $\pi^l$ is also a causal order of $G^m$. Lastly, given that $G^m \in \text{SGS}(P)$, it follows from Lemma B.2 that $G_{\pi^l} = G^m$.

Lemma 4.4 Given a graphoid $P$, for any $\pi \in \Pi(v)$ and any Chickering sequence from $G_\pi$ to some $H \in \text{SGS}(P)$ considered by TSP, there exists a ct-sequence $\langle \pi, \ldots, \tau \rangle$ s.t. $G_\tau = H$.

Proof. Given that a Chickering sequence considered by TSP is simply a TSP-Chickering sequence defined in Definition B.9 the lemma follows immediately from Lemma B.11.

Theorem 4.7 Given a graphoid $P$ over $V$ and any initial permutation $\pi \in \Pi(v)$, the DAG induced by the output of unbounded GRaSP$_0$ is equivalent to the DAG returned by TSP.

Proof. Immediate from Lemma B.10 and Lemma B.11.

Now we turn to the discussion on the correctness of GRaSP$_0$ under faithfulness.

Lemma B.12 Given a graphoid $P$ over $V$ and any $\pi \in \Pi(v)$, if $G_\pi \notin \text{Pm}(P)$, then there exists a ct-sequence $\tau = \langle \pi, \ldots, \tau \rangle$ s.t. $I(G_\pi) \subset I(G_\tau)$.

Proof. Suppose that $G_\pi \notin \text{Pm}(P)$. By Definition 3.7 it follows that there exists $H \in \text{CMC}(P)$ s.t. $I(G_\pi) \subset I(H) \subset I(P)$. By Theorem 3.6 we know that there exists a Chickering sequence $C_0 = \langle G_\pi = G^1, \ldots, G^l = H \rangle$. Without loss of generality, suppose that $C_0$ is the shortest Chickering sequence where each $G^{i+1}$ differs from $G^i$ in SGS($P$) by the reversal of a covered edge in $E^0(G^i)$ for each $1 \leq i < l - 1$, and $G^l$ is obtained from deleting a directed edge in $G^{l-1}$. Notice that $|E(G^1)| < |E(G^i)|$ due to the edge deletion. If $G^l \in \text{SGS}(P)$, then $C_0$ is a TSP-Chickering sequence. Otherwise, we can easily construct a TSP-Chickering sequence $C = \langle G^1, \ldots, G^m \rangle$ with $l - 1$ as the turning index and $G^m \in \text{SGS}(P)$ obtained by repeated edge-deletion from $G^l$ such that $I(G^m) \subset I(G^l) \subset I(G^m)$. By Lemma B.11(b), we have the desired ct-sequence.

Theorem 4.5 Given a graphoid $P$ over $V$ and any $\pi \in \Pi(v)$, if $G_\pi \notin \text{Pm}(P)$, then there exists a ct-sequence
\[
\Xi = (\pi, ..., \tau) \text{ s.t. } G_{\tau} \in \operatorname{PM}(P).
\]

**Proof.** Immediate from **Lemma B.12** \(\square\)

**Theorem B.13** Unbounded GRaSP\(_0\) is correct and pointwise consistent under faithfulness.

**Proof.** We review the argument for the correctness of unbounded GRaSP\(_0\) under faithfulness given in the main paper. Given a graphoid \(P\) over \(V\), consider any initial permutation \(\pi \in \Pi(v)\). Given that unbounded GRaSP\(_0\) greedily search for a ct-sequence from \(\pi\), it is guaranteed by \textbf{Theorem 4.5} that \(\tau\) returned by unbounded GRaSP\(_0\) in \textbf{Algorithm 2} induces a P-minimal DAG. Under faithfulness, we have \(G_{\tau} \in \operatorname{MEC}(G^*)\) due to \textbf{Theorem 3.8} and hence unbounded GRaSP\(_0\) is correct.

Alternatively, the correctness and pointwise consistency of unbounded GRaSP\(_0\) can also be proven directly from **Theorem 4.7** and the corresponding results of TSP in \cite{Solus2021}.

**Corollary 4.6** Unbounded GRaSP\(_0\), GRaSP\(_1\), and GRaSP\(_2\) are correct and pointwise consistent under faithfulness.

In the following, we want to prove that faithfulness is not only sufficient, but also necessary for the correctness of TSP and unbounded GRaSP\(_0\). We first want to prove an interesting and novel equivalence between two causal razors: faithfulness and u-P-minimality.

**Lemma B.14** Given a joint probability distribution \(P\) over \(V\), for any \(<X_i, X_j \mid X_k> \in I(P)\), there exists \(G \in \operatorname{DAG}(V)\) s.t. \(I(G) = \{(X_i, X_j \mid X_k)\}\).

**Proof.** Consider \(V = \{X_1, ..., X_m\}\). An empty DAG suffices when \(m = 2\). So assume that \(m \geq 3\). Without loss of generality, consider \(<X_1, X_{k+2} \mid X_k> \in I(P)\) where \(k = (2, ..., k + 1)\), and the remaining vertices are \((k + 3, ..., m)\). We propose a procedure which guarantees the existence of the desired DAG \(G\).

1. \(G \leftarrow \) a complete undirected graph over \(v\)
2. remove the adjacency \(1 \rightarrow k + 2\) in \(G\)
3. foreach \((j, k)\) that are adjacent in \(G\) do
   4. \(\text{if } j < k\) then
      5. orient \(j \rightarrow k\) in \(G\)
6. return \(G\)

Line 3 to 5 guarantee that \(G\) is a DAG since all edges are directed and pointing from lower indices to higher indices such that no directed cycle can occur. Finally, \(1 \perp_{G} k + 2 \mid k\) holds because all directed paths from 1 to \(k + 2\) either contain a non-collider \(i \in k\) or contain a collider \(i \notin k\). Therefore, \(I(G) = \{(X_1, X_{k+2} \mid X_k)\}\) because no other d-separation relations hold in \(G\). \(\square\)

**Theorem B.15** For any joint probability distribution \(P\), CFC\((P) = \operatorname{uPM}(P)\).

**Proof.** [\(\subseteq\)] Suppose that \(G \in \operatorname{CFC}(P)\). It follows that \(G \in \operatorname{PM}(P)\) by **Definition 3.7**. For any \(G' \in \operatorname{CMC}(P)\), if \(I(G') \subset I(G)\), then \(G' \notin \operatorname{PM}(P)\). Hence, if \(G' \in \operatorname{PM}(P)\), then \(I(G') = I(G)\). Hence, \(G \in \operatorname{uPM}(P)\).

[\(\supseteq\)] Suppose that \(G \notin \operatorname{CFC}(P)\). Since \(\operatorname{uPM}(P) \subseteq \operatorname{PM}(P)\) by **Definition 3.7**, if \(G \notin \operatorname{PM}(P)\), we have \(G \notin \operatorname{uPM}(P)\) immediately. So consider the case where \(G \in \operatorname{PM}(P)\). It follows from \(G \notin \operatorname{CFC}(P)\) that there exists a CI relation \(\psi \in I(P) \setminus I(G)\). By **Lemma B.14**, we can construct a DAG \(G^0\) such that \(I(G^0) = \{\psi\}\). Consequently, there exists \(G^1 \in \operatorname{PM}(P)\) such that \(I(G^1) \subseteq I(G^0) \subseteq I(P)\). Since \(\psi \in I(G^1)\), we know that \(G^1 \notin \operatorname{MEC}(G)\). Given that both \(G, G^1 \in \operatorname{PM}(P)\), we have \(G \notin \operatorname{uPM}(P)\). \(\square\)
Theorem 4.8  Given a graphoid $\mathcal{P}$, faithfulness is necessary for the correctness of TSP.

Proof. Suppose that $(\mathcal{G}^*, \mathcal{P})$ is unfaithful. We consider the two kinds of unfaithfulness in [Zhang and Spirtes, 2008]: detectable (i.e., CFC$(\mathcal{P}) = \emptyset$) versus undetectable (i.e., $\mathcal{G}^* \in$ CFC$(\mathcal{P})$ where $\mathcal{G}^* \notin$ MEC$(\mathcal{G}^*)$). For the latter, TSP can identify $\mathcal{G}_\tau \in$ Pm$(\mathcal{P})$ = CFC$(\mathcal{P})$ = MEC$(\mathcal{G}^*)$. However, TSP is incorrect because $\mathcal{G}_\tau \notin$ MEC$(\mathcal{G}^*)$.

On the other hand, consider the case that CFC$(\mathcal{P}) = \emptyset$. By Theorem B.15 there exists $\mathcal{G} \in$ Pm$(\mathcal{P})$ such that $\mathcal{G} \notin$ MEC$(\mathcal{G}^*)$ even if $\mathcal{G}^* \in$ Pm$(\mathcal{P})$. Recall that Chickering algorithm can only allow us to traverse to a DAG $\mathcal{H}$ from $\mathcal{G}$ satisfying $I(\mathcal{G}) \subseteq I(\mathcal{H})$. It entails that Chickering algorithm can only obtain DAGs that are in MEC$(\mathcal{G})$ since $\mathcal{G} \in$ Pm$(\mathcal{P})$ and hence never be able to reach $\mathcal{G}^*$ where $I(\mathcal{G}_\tau) \nsubseteq I(\mathcal{G}^*)$. Therefore, by setting $\pi$ as the initial permutation to TSP where $\mathcal{G}_\pi = \mathcal{G}$, TSP will return $\mathcal{G}_\pi$ incorrectly. □

Notice that Theorem 4.8 is contrary to what Solus et al. [2021] suggested. They proposed an example arguing that TSP can be correct even under (detectable) unfaithfulness. However, the distribution used in the example is not a semigraphoid. This renders their example illegitimate because every joint probability distribution is a semigraphoid.

C ESP AND GRASP-1

As shown in Theorem 4.8 in the last section, TSP cannot be correct under unfaithfulness by choosing an arbitrary initial permutation. Consequently, one important question is how to relax the search space of TSP to identify a sparser permutation under unfaithfulness. Solus et al. [2021] proposed the Edge SP (ESP) algorithm based on an assumption strictly weaker than that assumed by TSP. However, unlike TSP, they did not provide an operational version of ESP in their work. In this section, we are going to show a theorem similar to Theorem 4.7 but with respect to ESP and unbounded GRSp$_1$. In other words, unbounded GRSp$_1$ is an operational version of ESP. In the following, we first examine some technical notations used in [Mohammadi et al., 2018] and [Solus et al., 2021]. Readers are strongly suggested to visit [Solus et al., 2021] for the full discussion of ESP and relevant notations.

Given a set of measured variables $X$, a permutohedron on $v$, denoted $\mathcal{A}_v$, is the convex hull in $\mathbb{R}^{|v|}$ of all permutations in $\Pi(\mathcal{V})$. In simpler terms, $\mathcal{A}_v$ is the state space with each state being a permutation $\pi \in \Pi(\mathcal{V})$. The neighborhood of states in $\mathcal{A}_v$ is defined by adjacent transpositions (ATs) as in Definition B.1 (b).

Notice that different states in $\mathcal{A}_v$ can induce the same DAG given a graphoid $\mathcal{P}$. Thus, a natural way to narrow down the search space is to identify permutations inducing the same DAG. Lemma B.3 provides such a characterization. Construct $\mathcal{A}_v(\mathcal{P})$ by contracting neighborhood in $\mathcal{A}_v$ to ATs that correspond to the CI relations in $\Pi(\mathcal{P})$ specified in Lemma B.3. To be more specified, the contracted permutohedron $\mathcal{A}_v(\mathcal{P})$, also known as the DAG associahedron, is the state space with each state being an induced DAG. Two states $\mathcal{G}_1, \mathcal{G}_2$ in $\mathcal{A}_v(\mathcal{P})$ are neighbors if and only if there exist $\pi^1, \pi^2 \in \Pi(\mathcal{V})$ s.t. $\mathcal{G}_{\pi^1} = \mathcal{G}^1, \mathcal{G}_{\pi^2} = \mathcal{G}^2$, and $\pi^1$ and $\pi^2$ are neighbors in the permutohedron $\mathcal{A}_v$. As shown by Mohammadi et al. [2018], the DAG associahedron is a convex polytope where each vertex of $\mathcal{A}_v(\mathcal{P})$ corresponds to a different DAG.

To draw a clearer picture, consider any $\pi, \tau \in \Pi(\mathcal{V})$ where $\tau$ is $(j, k)$-different from $\pi$ for some $j, k \in \mathcal{V}$. They are neighbors in $\mathcal{A}_v$ but they do not necessarily induce the same DAG. If $X_j \perp_{\mathcal{P}} X_k | X_{\text{Pre}(j, \pi)}$ holds, they induce the same DAG and thus correspond to the same state $\mathcal{G}_\pi$ in the DAG associahedron $\mathcal{A}_v(\mathcal{P})$. But if the CI relation does not hold, then $\mathcal{G}_\pi$ and $\mathcal{G}_\tau$ are neighbors in $\mathcal{A}_v(\mathcal{P})$. See Figure 7 for an example from [Solus et al., 2021].

Observe that each state in the DAG associahedron $\mathcal{A}_v(\mathcal{P})$ corresponds to a SGS-minimal DAG according to Theorem 3.5. ESP performs a greedy DFS in $\mathcal{A}_v(\mathcal{P})$. Given an initial permutation $\pi \in \Pi(\mathcal{V})$, set $\mathcal{G}$ as the induced $\mathcal{G}_\pi$ and traverse through $\mathcal{A}_v(\mathcal{P})$ by a weakly decreasing walk to obtain $\mathcal{H}$ where $|E(\mathcal{H})| < |E(\mathcal{G}_\tau)|$. If no such $\mathcal{H}$ exists, ESP returns $\mathcal{G} = \mathcal{G}_\pi$; else $\mathcal{G}$ is reset as $\mathcal{H}$ and repeat.

As noted by Solus et al. [2021], the construction of $\mathcal{A}_v(\mathcal{P})$ is inefficient since one is only required to know the neighboring states instead of the entire $\mathcal{A}_v(\mathcal{P})$ to perform the traversal. Below we show that unbounded GRSp$_1$ can efficiently learn the

---

15See Figure 2 in the supplementary materials of [Solus et al., 2021].
16One can equivalently express each state in the DAG associahedron as the set of permutations which induce the same DAG. This is the original representation in [Mohammadi et al., 2018]. However, we prefer the representation given in [Solus et al., 2021] in the sense that one can easily compare DAGs that are in neighborhood.
17In [Solus et al., 2021], their pseudocode does not indicate that such a walk needs to be weakly decreasing but such a requirement is imposed in the description of the algorithm.
For the forward direction, given that π^m and π^5 are (j, k)-different but induce different DAGs, it follows from the definition of A_v(\mathcal{P}) that G_{π^m} \neq G_{π^5} and G_{π^5} are neighbors in A_v(\mathcal{P}). Also, we know that (j \rightarrow k) \in E(G_{π^m-1}) and (k \rightarrow j) \in E(G_{π^m}) by Lemma B.3 and (RU). The fact that G_{π^1} = G_{π^m-1} and G_{π^m} are (j, k)-reverse follows immediately from (RU) and the assumption that π^m−1 is (j, k)-different from π^m.

For the backward direction, suppose that G_{τ} and G_{τ'} are neighbors in A_v(\mathcal{P}) that are (j, k)-reverse. It entails from (RU) that there exist π', τ' \in Π(v) such that π' and τ' are (j, k)-different where G_{τ} = G_{τ'}. Hence, (π', τ') is our desired DAG-changing walk relative to (j, k) in A_v.

Lemma C.4 Given a graphoid \mathcal{P} over V, consider any pair π^1, τ^1 \in Π(v) such that π^1 = \langle δ_1, j, k, δ_2 \rangle for some sub-sequences δ_1, δ_2 of π^1, and τ^1 = \langle ζ_1, j, k, ζ_2 \rangle for some sub-sequences ζ_1, ζ_2 of τ^1. Further consider τ^2 = \langle ζ_1, k, j, ζ_2 \rangle and τ^2 = \langle ζ_1, k, j, ζ_2 \rangle. If G_{π^1} = G_{τ^1}, then G_{π^2} = G_{τ^2}.
Proof. Notice that $G_{\tau^2} \in SGS(\mathcal{P})$ by Theorem [3.5]. If we can show that $\tau^2$ is a causal order of $G_{\tau^2}$, it follows from Lemma [B.2] that $G_{\tau^2} = G_{\tau^2}$. To do so, it suffices to show the following. For any $i \in v \setminus \{j,k\}$,

(i) if $(i \rightarrow j) \in E(G_{\tau^2})$, then $i \in \zeta_1$;
(ii) if $(i \rightarrow k) \in E(G_{\tau^2})$, then $i \in \zeta_1$;
(iii) if $(j \rightarrow i) \in E(G_{\tau^2})$, then $i \in \zeta_2$;
(iv) if $(k \rightarrow i) \in E(G_{\tau^2})$, then $i \in \zeta_2$.

For (i), suppose that $(i \rightarrow j) \in E(G_{\tau^2})$. If $(i \rightarrow j) \in E(G_{\tau^1})$ as well, then $(i \rightarrow j) \in E(G_{\tau^1})$ since $G_{\tau^1} = G_{\tau^1}$. This entails that $i \in \zeta_1$. On the other hand, consider the case that $(i \rightarrow j) \notin E(G_{\tau^1})$. Then

\[ X_i \not\perp\!\!\!\!\perp X_j \mid X_{\delta_1 \setminus \{i\}} \cup \{X_k\} \quad \because (i \rightarrow j) \in E(G_{\tau^2}) \quad (4) \]
\[ X_i \not\perp\!\!\!\!\perp X_j, X_k \mid X_{\delta_1 \setminus \{i\}} \quad \because (4), \text{weak union} \quad (5) \]
\[ X_i \not\perp\!\!\!\!\perp X_j \mid X_{\delta_1 \setminus \{i\}} \quad \because (i \rightarrow j) \notin E(G_{\tau^1}) \quad (6) \]
\[ X_i \not\perp\!\!\!\!\perp X_j \mid X_{\delta_1 \setminus \{i\}} \cup \{X_k\} \quad \because (5), (6), \text{contraction} \quad (7) \]

By (RU), (8) entails that $(i \rightarrow k) \in E(G_{\tau^1}) = E(G_{\tau^1})$. Since $\tau^1$ is a causal order of $G_{\tau^1}$, we have $i \in \zeta_1$.

For (ii), suppose that $(i \rightarrow k) \in E(G_{\tau^2})$. Similar to (i), the case for $(i \rightarrow k) \in E(G_{\tau^1})$ is simple. So consider the case where $(i \rightarrow k) \notin E(G_{\tau^1})$.

\[ X_i \not\perp\!\!\!\!\perp X_k \mid X_{\delta_1 \setminus \{i\}} \quad \because (i \rightarrow k) \in E(G_{\tau^2}) \quad (8) \]
\[ X_i \not\perp\!\!\!\!\perp X_j, X_k \mid X_{\delta_1 \setminus \{i\}} \quad \because (8), \text{decomposition} \quad (9) \]
\[ X_i \not\perp\!\!\!\!\perp X_k \mid X_{\delta_1 \setminus \{i\}} \cup \{X_j\} \quad \because (i \rightarrow k) \notin E(G_{\tau^1}) \quad (10) \]
\[ X_i \not\perp\!\!\!\!\perp X_j \mid X_{\delta_1 \setminus \{i\}} \quad \because (9), (10), \text{contraction} \quad (11) \]

By (RU), (13) entails that $(i \rightarrow j) \in E(G_{\tau^1}) = E(G_{\tau^1})$ and hence $i \in \zeta_1$.

For (iii), suppose that $(j \rightarrow i) \in E(G_{\tau^2})$. Then we have $(j \rightarrow i) \in E(G_{\tau^1})$ by (RU) because Pre$(i, \pi^1) = Pre(i, \pi^2)$. Hence $(j \rightarrow i) \in E(G_{\tau^1})$ since $G_{\tau^1} = G_{\tau^1}$. Given that $\tau^1$ is a causal order of $G_{\tau^1}$, we have $i \in \zeta_2$. (iv) is analogous to (iii).

\[ \square \]

Lemma C.5 Given a graphoid $\mathcal{P}$ over $V$, consider any two DAG-changing walks $\mathcal{W} = \langle \pi^1, ..., \pi^m \rangle$ and $\mathcal{W}' = \langle \tau^1, ..., \tau^n \rangle$ in $A_\pi$ where $\pi^1 = \tau^1$. If $\mathcal{W}$ and $\mathcal{W}'$ are both relative to the same $(j, k)$ for some $j, k \in V$, then $G_{\pi^m} = G_{\tau^n}$.

Proof. Immediate from Definition C.1 and Lemma C.4

\[ \square \]

Lemma C.6 Given a graphoid $\mathcal{P}$ over $V$, consider any DAG-changing walk $\mathcal{W} = \langle \pi^1, ..., \pi^m \rangle$ in $A_\pi$ which is relative to $(j,k)$ for some $j,k \in V$. Then $j \rightarrow k$ is a singular edge in $G_{\pi^m}$.

Proof. Let $\mathcal{W}_0$ denotes the DAG-preserving walk $\langle \pi^1, ..., \pi^{m-1} \rangle$. Given that $\pi^m$ is $(j,k)$-different from $\pi^{m-1}$, it follows from Lemma [B.3] and (RU) that $(j \rightarrow k) \in E(G_{\pi^{m-1}})$. Since $\mathcal{W}_0$ is a DAG-preserving walk in $A_\pi$, we have $(j \rightarrow k) \in E(G_{\pi^m}) = E(G_{\pi^{m-1}})$.

Next, suppose by reductio that $j \rightarrow k$ is not a singular edge in $G_{\pi^m}$. Then there is a unidirectional path from $j$ to $k$ other than $j \rightarrow k$ in $G_{\pi^m}$. So there exists $l \in V$ such that $l \in \text{De}(j, G_{\pi^m}) \cap \text{An}(k, G_{\pi^m})$. In order to ensure that $j$ and $k$ are $\pi^{m-1}$-adjacent, either $\pi^{m-1}[l] < \pi^m[j]$ or $\pi^{m-1}[l] > \pi^m[k]$ holds. However, either case will violate that $\pi^{m-1}$ is a causal order of $G_{\pi^{m-1}} = G_{\pi^m}$.

\[ \square \]

Lemma C.7 Given a graphoid $\mathcal{P}$ over $V$, consider $\pi \in \Pi(v)$ where $(j \rightarrow k) \in E(G_{\pi})$ is a singular edge for some $j,k \in V$. Then there exists a DAG-changing walk $\mathcal{W} = \langle \pi, ..., \tau \rangle$ in $A_\pi$ relative to $(j,k)$ where $\tau = \text{tuck}(\pi, j, k)$.

Proof. First, we rewrite $\pi = \langle \delta_{j,k}, \delta_{j \rightarrow k}, \delta_{\pi} \rangle$ as usual. Then we partition $\delta_{j \rightarrow k}$ as follows: $\zeta_a = \{ i \in \delta_{j \rightarrow k} : i \in \text{An}(k, G_{\pi}) \}$, and $\zeta_b = \{ i \in \delta_{j \rightarrow k} : i \notin \text{An}(k, G_{\pi}) \}$. Given that $(j \rightarrow k)$ is a singular edge, we know that $\text{De}(j, G_{\pi}) \cap$
Proof. For (a), we know that (i) each vertex in $\zeta_a$ has no ancestor in $\delta_{j\sim k}\setminus\zeta_b$ in $G_x$ and (ii) each vertex in $\zeta_b$ has no descendant in $\delta_{j\sim k}\setminus\zeta_b$ in $G_x$.

Now consider the permutation $\tau' = (\delta_{xj}, \zeta_a, j, k, \zeta_b, \delta_{j\sim k})$ in particular. We want to show that there exists a DAG-preserving walk from $\tau$ to $\tau'$. Such a walk is easy to construct. First, perform repeated ATs by moving each $i \in \zeta_a$ prior to $j$ from left to right, and then repeated ATs by moving each $i \in \zeta_b$ behind $k$ from right to left. The two sets of ATs are licensed by (i) and (ii) respectively. Hence, we have $G_{\tau'} = G_\tau$. Finally, consider $\tau = \text{tuck}(\pi, j, k) = (\delta_{xj}, \zeta_a, j, k, \zeta_b, \delta_{j\sim k})$ which is $(j, k)$-different from $\tau'$. By (RU) and Lemma B.3, we know that $G_{\tau} \neq G_{\tau'}$, and thus $\{\pi, \ldots, \tau', \tau\}$ is a DAG-changing walk in $A_v$ relative to $(j, k)$.

\[\Box\]

Theorem C.8 Given a graphoid $P$ over $V$, consider any DAG-changing walk $\Omega = \langle \pi^1, \ldots, \pi^m \rangle$ in $A_v$ which is relative to $(j, k)$ for some $j, k \in V$. Then $G_{\pi^m} = G_\tau$ where $\tau = \text{tuck}(\pi^1, j, k)$.

Proof. We obtain a DAG-changing walk $\Omega' = \langle \pi^1, \ldots, \tau \rangle$ in $A_v$ relative to $(j, k)$ by Lemma C.7. Since both $\Omega$ and $\Omega'$ are relative to the same $(j, k)$, it follows from Lemma C.5 that $G_{\pi^m} = G_\tau$. \[\Box\]

Similar to the discussion in Appendix B, we want to fix the ordering of the set of singular edges in any DAG. This ensures that ESP and unbounded GRaSP$_1$ will not yield different DAGs simply due to the issue of order-dependence. Below we prove that ESP and unbounded GRaSP$_1$ are equivalent algorithms.

Theorem C.9 Given a graphoid $P$ and any initial permutation $\pi \in \Pi(v)$, the DAG induced by the output of unbounded GRaSP$_1$ is equivalent to the DAG returned by ESP.

Proof. Consider any $j, k \in V$. By Lemma C.6 and Lemma C.7, every DAG-changing walk $\Omega = \langle \pi^1, \ldots, \pi^m \rangle$ in $A_v$ relative to $(j, k)$ corresponds to a tuck operation of the singular edge $j \rightarrow k$ in $E(G_x)$. Hence, by Lemma C.3, we know that $\text{tuck}(\pi^1, j, k)$ corresponds to the neighboring relation between $G_{\pi^1}$ and $G_{\pi^m}$ in $A_v(P)$ that are $(j, k)$-reverse. Therefore, every step taken by ESP to move to a neighboring state in $A_v(P)$ (relative to a unique pair of vertices) is equivalent to the tuck operation taken by GRaSP$_1$ over the same pair of vertices. \[\Box\]

D CAUSAL RAZORS AND GRASP

In this section, we first provide a logical analysis of the causal razors discussed in the main text.\[\footnote{There are other causal razors discussed in the literature, including, but not limited to, adjacency-faithfulness and orientation-faithfulness in \[\cite{Ramsay2006}\] and triangle-faithfulness in \[\cite{Zhang2013}\]. But they do not have a strong connection with our discussion of GRaSP and so will not analyzed in this work.}

Then we construct new causal razors with respect to each tier of GRaSP, and show how a higher tier of GRaSP requires a strictly weaker causal razor.

Theorem D.1 The following statements are true:

(a) For any joint probability distribution $P$, $uP_{\pi^m}(P) = \text{CFC}(P) \subseteq uF_{\pi^m}(P) \subseteq Fr(P) \subseteq Pm(P) \subseteq \text{SGS}(P)$.

(b) For any joint probability distribution $P$, if faithfulness is satisfied, $\text{CFC}(P) = uF_{\pi^m}(P) = Fr(P) = Pm(P)$.

(c) There exists a joint probability distribution $s.t. \text{CFC}(P) \subset uF_{\pi^m}(P)$.

(d) There exists a joint probability distribution $s.t. uF_{\pi^m}(P) \subset Fr(P)$.

(e) There exists a joint probability distribution $s.t. Fr(P) \subset Pm(P)$.

(f) There exists a joint probability distribution $s.t. Pm(P) \subset \text{SGS}(P)$.

Proof. For (a), $uP_{\pi^m}(P) = \text{CFC}(P)$ is our result in Theorem B.15. $\text{CFC}(P) \subseteq uF_{\pi^m}(P)$ is proven in \[\cite{Raskutti2018}\]. $uF_{\pi^m}(P) \subseteq Fr(P)$ is true by Definition 3.3. $Fr(P) \subseteq Pm(P)$ in \[\cite{Forster2020}\]. $\text{SGS}(P) \subseteq Pm(P)$ in \[\cite{Zhang2013}\]. (b) is a direct consequence of (a) and Theorem C.3.

For (c), see \[\cite{Raskutti2018}\] Theorem 2.4.1. For (d), see \[\cite{Forster2020}\] Figure 6. For (e), see \[\cite{Raskutti2018}\] Theorem 2.5. For (f), see \[\cite{Zhang2013}\] Figure 2. Additionally, the example in Theorem D.6 and its corresponding Figure 8 verifies (c) and (e); $G^* \in uF_{\pi^m}(P) \setminus \text{CFC}(P)$ and $G_{\pi^m} \in Pm(P) \setminus Fr(P)$. On the other hand, each of $G_{\pi^1}, G_{\pi^2}, G_{\pi^3},$ and $G_{\pi^4}$ in the DAG-associahedron in Figure 7 is in $\text{SGS}(P) \setminus Pm(P)$ verifying (f). \[\Box\]
Definition D.2 (TSP-razor and ESP-razor) Given a graphoid \( \mathcal{P} \) over \( \mathcal{V} \), let \( tsp(\mathcal{P}, \pi) \) be the DAG returned by TSP on \( \mathcal{P} \) by setting \( \pi \) as the initial permutation. Define \( TSP(\mathcal{P}) = \{ G \in DAG(\mathcal{V}) : \pi \in \Pi(\mathcal{V}) \text{ and } G = tsp(\mathcal{P}, \pi) \} \) as the set of DAGs returned by TSP on \( \mathcal{P} \) over each initial permutation in \( \Pi(\mathcal{V}) \). Further define

\[
TSP(\mathcal{P}) = \{ G \in TSP(\mathcal{P}) : -\exists G' \in TSP(\mathcal{P}) \text{ s.t. } G' \notin MEC(G) \}.
\]

\((G^*, \mathcal{P})\) satisfies the TSP-razor if \( G^* \in TSP(\mathcal{P}) \). Similarly for ESP, esp, ESP, ESPx, and ESP-razor.

One can observe that \( TSP(\mathcal{P}) = tsp(\mathcal{P}) \) if every DAG in \( TSP(\mathcal{P}) \) belongs to the same MEC, and \( TSP(\mathcal{P}) = \emptyset \) otherwise. The same is also true for ESP(\mathcal{P}) and ESP(\mathcal{P}). These definitions will be proven useful when we compare them with the classes of DAGs discussed in Theorem D.1. Below we provide a similar definition for each tier of GRaSP.

Definition D.3 (GRaSPt-razor) Given a graphoid \( \mathcal{P} \) over \( \mathcal{V} \), for \( t \in \{0, 1, 2\} \), define \( GRaSP_t(\mathcal{P}) = \{ G' \in DAG(\mathcal{V}) : \pi \in \Pi(\mathcal{V}) \text{ and } \tau = grasp(\mathcal{P}, \pi, |\mathcal{V}|, t) \} \) as the set of DAGs returned by unbounded GRaSPt on \( \mathcal{P} \) over each initial permutation in \( \Pi(\mathcal{V}) \). Further define

\[
GRaSP_t(\mathcal{P}) = \{ G \in GRaSP_t(\mathcal{P}) : -\exists G' \in GRaSP_t(\mathcal{P}) \text{ s.t. } G' \notin MEC(G) \}.
\]

\((G^*, \mathcal{P})\) satisfies the GRaSPt-razor if \( G^* \in GRaSP_t(\mathcal{P}) \).

**Theorem D.4** Given a graphoid \( \mathcal{P} \), the following statement is true:

\[
CFC(\mathcal{P}) = TSP(\mathcal{P}) = GRaSP_0(\mathcal{P}) \subseteq ESP(\mathcal{P}) = GRaSP_1(\mathcal{P}) \subseteq GRaSP_2(\mathcal{P}) \subseteq uFr(\mathcal{P}).
\]

**Proof.** \( CFC(\mathcal{P}) = TSP(\mathcal{P}) = GRaSP_0(\mathcal{P}) \) is directly entailed by Theorem 4.7 and Theorem 4.8 from Solus et al. [2021] showed that \( TSP(\mathcal{P}) \subseteq ESP(\mathcal{P}) \). \( ESP(\mathcal{P}) = GRaSP_1(\mathcal{P}) \) is entailed by Theorem C.9

Next, to show that \( GRaSP_1(\mathcal{P}) \subseteq GRaSP_2(\mathcal{P}) \), notice that \( GRaSP_1(\mathcal{P}) = GRaSP_1(\mathcal{P}) \) when all DAGs in \( GRaSP_1(\mathcal{P}) \) belong to the same MEC, and \( GRaSP_1(\mathcal{P}) = \emptyset \) otherwise. The latter case validates \( GRaSP_1(\mathcal{P}) \subseteq GRaSP_2(\mathcal{P}) \) trivially. Now consider the former case where all DAGs in the non-empty \( GRaSP_1(\mathcal{P}) \) belong to the same MEC and so they have the same number of edges. Now consider any \( \pi \in \Pi(\mathcal{V}) \) satisfying \( G_\pi \in Fr(\mathcal{P}) \) (where \( Fr(\mathcal{P}) \) is necessarily non-empty). We know that \( G_\pi \in GRaSP_1(\mathcal{P}) \). This is because every initial permutation in \( \Pi(\mathcal{V}) \) is considered and unbounded \( GRaSP_1 \) will never return a denser permutation than its initial permutation. Hence, every DAG in \( GRaSP_1(\mathcal{P}) \) is the sparsest Markovian DAG. (The same also holds when \( GRaSP_2(\mathcal{P}) \neq \emptyset \).) Then the construction of Algorithm 2 entails that \( GRaSP_2 \) will return the same permutation as \( GRaSP_1 \). Hence, \( GRaSP_1(\mathcal{P}) = GRaSP_2(\mathcal{P}) \) when all DAGs in \( GRaSP_1(\mathcal{P}) \) belong to the same MEC.

Lastly, to show that \( GRaSP_2(\mathcal{P}) \subseteq uFr(\mathcal{P}) \), we use a proof similar to the above. First, the case where \( GRaSP_2(\mathcal{P}) = \emptyset \) is trivial. Consider the case where \( GRaSP_2(\mathcal{P}) = GRaSP_2(\mathcal{P}) \) s.t. all DAGs in \( GRaSP_2(\mathcal{P}) \) are in the same MEC. Using a similar inference used in the last paragraph, we know that every DAG in \( GRaSP_2(\mathcal{P}) \) is the sparsest Markovian DAG. Therefore, \( GRaSP_2(\mathcal{P}) = uFr(\mathcal{P}) \) when all DAGs in \( GRaSP_2(\mathcal{P}) \) belong to the same MEC.

**Theorem D.5** There exists a graphoid \( \mathcal{P} \) s.t. \( GRaSP_0(\mathcal{P}) \subset GRaSP_1(\mathcal{P}) \).

**Proof.** Given the equivalence between TSP and unbounded \( GRaSP_0 \) shown in Theorem 4.7, and that between ESP and unbounded \( GRaSP_1 \) in Theorem C.9, we can borrow the example from Solus et al. [2021] on how ESP requires a strictly weaker causal razor than TSP. We refer the readers to Figure 3 in the supplementary materials of Solus et al. [2021].

In the remainder of this section, we discuss two examples: how unbounded \( GRaSP_2 \) requires a strictly weaker causal razor than unbounded \( GRaSP_1 \), and how unbounded \( GRaSP_2 \) requires a strictly stronger causal razor than u-frugality. The joint distribution of each example below is a compositional graphoid. For the sake of simplicity, we only include CI relations that hold between two singleton sets of variables such that all other CI relations entailed by each of the graphoid axioms discussed in Appendix A.2 are understood.

**Theorem D.6** There exists a graphoid \( \mathcal{P} \) s.t. \( GRaSP_1(\mathcal{P}) \subset GRaSP_2(\mathcal{P}) \).
Proof. Given $V = \{X_1, ..., X_4\}$, consider the unfaithful model $(G^*, P)$ where the true DAG $G^*$ is shown on the left in Figure 8 and $I(P) = \Phi \cup \Psi$ where $\Phi$ is the set of faithful CI relations and $\Psi$ is the set of unfaithful CI relations as listed below:

$\Phi = \{\phi_1 = \langle X_1, X_3 \mid \varnothing \rangle, \phi_2 = \langle X_2, X_4 \mid \{X_1, X_3\} \rangle, \phi_3 = \langle X_2, X_3 \mid \varnothing \rangle, \phi_4 = \langle X_2, X_3 \mid \{X_1\} \rangle, \phi_5 = \langle X_2, X_5 \mid \{X_1, X_3, X_4\} \rangle\};$

$\Psi = \{\psi_1 = \langle X_2, X_4 \mid \varnothing \rangle\}.$

For every $G \in \mathcal{CMC}(P)$ where $\psi_1 \in I(G)$, we have $|E(G)| > |E(G^*)| = 4$. Also, all 4-edge Markovian DAGs are in the same MEC. Hence, u-frugality is satisfied. Consider feeding the initial permutation $\pi = (2, 4, 1, 3)$ to unbounded GRaSP_1. It will return the same $\pi$ after the DFS procedure and the induced $G_\pi$, as shown on the right in Figure 8, contains 5 edges. Therefore, unbounded GRaSP_1 fails to return the sparsest permutation under some initial permutation and $\text{GRaSP}_1 r(P) = \varnothing$.

On the contrary, $|\pi| = 24$ initial permutations have been tested on unbounded GRaSP_2 and it returns $\hat{\pi}$ where $G_{\hat{\pi}} \in \mathcal{MEC}(G^*)$ for each initial permutation. Hence, $\text{GRaSP}_2 r(P) \neq \varnothing$. □

![Figure 8: An unfaithful model satisfying u-frugality. The true DAG $G^*$ is shown on the left where the two shaded vertices indicate the unfaithful marginal independence $X_2 \perp\!\!\!\perp X_4 \mid \varnothing$. Unbounded GRaSP_1 returns its initial permutation $\pi = (2, 4, 1, 3)$. The induced DAG $G_\pi$ is shown on the right with 5 edges. However, unbounded GRaSP_2 manages to return one of the sparsest permutations under every initial permutation.

**Theorem D.7** There exists a graphoid $P$ s.t. $\text{GRaSP}_2 r(P) \subset \text{uFr}(P)$.

Proof. The example below is one of the uDAGs studied in Table 1 in Section 5.1 where GRaSP_2 fails to return one of the sparsest permutations under u-frugality. Given $V = \{X_1, ..., X_5\}$, consider the unfaithful model $(G^*, P)$ where the true DAG $G^*$ is shown on the left in Figure 8 and $I(P) = \Phi \cup \Psi$ where $\Phi$ is the set of faithful CI relations and $\Psi$ is the set of unfaithful CI relations as listed below:

$\Phi = \{\phi_1 = \langle X_1, X_2 \mid \varnothing \rangle, \phi_2 = \langle X_1, X_2 \mid \{X_3\} \rangle, \phi_3 = \langle X_2, X_3 \mid \varnothing \rangle, \phi_4 = \langle X_2, X_3 \mid \{X_1\} \rangle, \phi_5 = \langle X_2, X_5 \mid \{X_1, X_3, X_4\} \rangle\};$

$\Psi = \{\psi_1 = \langle X_1, X_5 \mid \varnothing \rangle\}.$

For every $G \in \mathcal{CMC}(P)$ where $\psi_1 \in I(G)$, we have $|E(G)| > |E(G^*)| = 7$. Also, all 7-edge Markovian DAGs are in the same MEC and there exists no sparser Markovian DAG. Hence, u-frugality is satisfied s.t. $\text{uFr}(P) \neq \varnothing$.

Next, consider feeding the initial permutation $\pi = (5, 1, 3, 4, 2)$ to unbounded GRaSP_2. It will return the same $\pi$ after the DFS procedure and the induced $G_{\pi}$, as shown on the right in Figure 8, contains 8 edges. Therefore, unbounded GRaSP_2 fails to return one of the sparsest permutations under some initial permutation and $\text{GRaSP}_2 r(P) = \varnothing$. □
Figure 9: An unfaithful model satisfying u-frugality. The true DAG $G^*$ is shown on the left where the two shaded vertices indicate the unfaithful marginal independence $X_1 \perp \perp X_5 \mid \emptyset$. Unbounded GRaSP returns its initial permutation $\pi = (5, 1, 3, 4, 2)$. The induced DAG $G_\pi$ is shown on the right with 8 edges. Hence, GRaSP is not correct under u-frugality alone.

Corollary 4.9 Given a graphoid $\mathcal{P}$, unbounded GRaSP is correct under a strictly weaker causal razor than unbounded GRaSP, which is correct under a strictly weaker causal razor than unbounded GRaSP.

### E GROW-SHRINK ALGORITHM AND ITS PROPERTIES

**Definition E.1** Given an observational dataset $D$ with $n$ i.i.d. observations from a joint probability distribution $\mathcal{P}$ over $V$ that belongs to a curved exponential family\(^{19}\) for every $X \in V$ and every $\mathcal{M} \subseteq V \setminus X$,

$$BIC_D(X, \mathcal{M}) = \ell_{X|\mathcal{M}}(\hat{\theta}_{\text{MLE}} \mid D) + c \frac{|\hat{\theta}_{\text{MLE}}|}{2} \log(n)$$

where $\ell_{X|\mathcal{M}}$ is the conditional log likelihood function, $|\hat{\theta}_{\text{MLE}}|$ is the absolute value of the maximum likelihood estimate, and $c$ is a multiplier for the parameter penalty.

BIC score is a decomposable scoring function in the sense that the BIC score of any DAG $G$ (over the same set of variables $V$ as the observational dataset $D$), denoted as $BIC_D(G)$, satisfies the following:

$$BIC_D(G) = \sum_{i \in V} BIC_D(X_i, X_{\text{Pa}(i, G)})$$

In addition, since we will be using BIC throughout this appendix, we assume that every joint probability distribution $\mathcal{P}$

---

\(^{19}\)See [Kass and Vos, 2011] for an in-depth analysis of curved exponential families.
Algorithm 3: GROW: $\text{grow}(\mathcal{D}, X, Z)$

**Input:** (a) $\mathcal{D}$: an observational dataset over $V$; (b) $X \in V$; (c) $Z \subseteq V \setminus \{X\}$

**Output:** $M_{gr} \subseteq Z$

1. $s \leftarrow \text{BIC}_D(X, \emptyset)$
2. $s' \leftarrow s$
3. $M_{gr} \leftarrow \emptyset$
4. **do**
5.   
6.     $s' \leftarrow \max_{Y \in Z \setminus M_{gr}} \text{BIC}_D(X, M_{gr} \cup \{Y\})$
7.     $Y' \leftarrow \arg\max_{Y \in Z \setminus M_{gr}} \text{BIC}_D(X, M_{gr} \cup \{Y\})$
8.     **if** $s' > s$ **then**
9.         $M_{gr} \leftarrow M_{gr} \cup \{Y'\}$
10. **while** $s' > s$
11. return $M_{gr}$

Algorithm 4: SHRINK: $\text{shrink}(\mathcal{D}, X, Z)$

**Input:** (a) $\mathcal{D}$: an observational dataset over $V$; (b) $X \in V$; (c) $Z \subseteq V \setminus \{X\}$

**Output:** (i) $M_{sh} \subseteq Z$; (ii) $s = \text{BIC}_D(X, M_{sh})$

1. $s \leftarrow \text{BIC}_D(X, Z)$
2. $s' \leftarrow s$
3. $M_{sh} \leftarrow Z$
4. **do**
5.   
6.     $s' \leftarrow \max_{Y \in M_{sh}} \text{BIC}_D(X, M_{sh} \setminus \{Y\})$
7.     $Y' \leftarrow \arg\max_{Y \in M_{sh}} \text{BIC}_D(X, M_{sh} \setminus \{Y\})$
8.     **if** $s' > s$ **then**
9.         $M_{sh} \leftarrow M_{sh} \setminus \{Y'\}$
10. **while** $s' > s$
11. return $M_{sh}, s$

**Theorem E.2** [Chickering, 2002] Given an observational dataset $\mathcal{D}$ with $n$ i.i.d. observations from a joint probability distribution $\mathcal{P}$ over $V$, consider $\mathcal{G}, \mathcal{G}' \in \text{DAG}(V)$ where $\mathcal{G}'$ is resulted from adding the edge $j \rightarrow k$ in $\mathcal{G}$. In the large sample limit of $n$,

(a) if $X_j \perp_{\mathcal{P}} X_k \mid X_{Pa(k,g)}$, then $\text{BIC}_D(\mathcal{G}') > \text{BIC}_D(\mathcal{G})$;

(b) if $X_j \perp_{\mathcal{P}} X_k \mid X_{Pa(k,g)}$, then $\text{BIC}_D(\mathcal{G}') < \text{BIC}_D(\mathcal{G})$.

The theorem above is known as the *local consistency* of BIC score over DAGs. We can easily derive a lemma which concerns the BIC score of a variable (relative to a set of variables).

**Lemma E.3** Given an observational dataset $\mathcal{D}$ with $n$ i.i.d. observations from a joint probability distribution $\mathcal{P}$ over $V$, consider any distinct $j, k \in v$ and $i \subseteq v \setminus \{j,k\}$. In the large sample limit of $n$,

(a) if $X_j \perp_{\mathcal{P}} X_k \mid X_i$, then $\text{BIC}_D(X_k, X_i) > \text{BIC}_D(X_k, X_j)$;

(b) if $X_j \perp_{\mathcal{P}} X_k \mid X_i$, then $\text{BIC}_D(X_k, X_i \cup \{X_j\}) < \text{BIC}_D(X_k, X_i)$. 

**Proof.** Construct a DAG $\mathcal{G} \in \text{DAG}(V)$ by drawing all and only directed edges from each vertex in $i$ to $k$, and another DAG $\mathcal{G}' \in \text{DAG}(V)$ by adding $j \rightarrow k$ in $\mathcal{G}$. Then the lemma immediately follows from Theorem E.2 and the decomposable feature of BIC scores. □
Lemma E.4 Consider an observational dataset $D$ with $n$ i.i.d. observations from a compositional graphoid $\mathcal{P}$ over $V$. In the large sample limit of $n$, for any $X \in V$ and any $Z \subseteq V \setminus \{X\}$, $MB(X, Z) \subseteq M_{gr}$ where $M_{gr} = grow(D, X, Z) \subseteq Z$.

Proof. First, Algorithm requires that $M_{gr} \subseteq Z$, and $BIC_{D}(X, M_{gr}) \setminus \{Y\} < BIC_{D}(X, M_{gr})$ for every $Y \in Z \setminus M_{gr}$. By Lemma E.3 we have $X \perp_{p} Y | M_{gr}$ for each $Y \in Z \setminus M_{gr}$. By composition, we have $X \perp_{p} (Z \setminus M_{gr}) | M_{gr}$. Therefore, by Definition A.1 and Lemma A.2 we have $MB(X, Z) \subseteq M_{gr}$. □

Lemma E.5 Consider an observational dataset $D$ with $n$ i.i.d. observations from a graphoid $\mathcal{P}$ over $V$. In the large sample limit of $n$, for any $X \in V$ and any $Z \subseteq V \setminus \{X\}$, $MB(X, Z) = M_{sh}$ where $M_{sh} = shrink(D, X, Z) \subseteq Z$.

Proof. We show the lemma by $M_{sh} \subseteq MB(X, Z)$ and $M_{sh} \supseteq MB(X, Z)$.

By reductio, suppose that there exists $Y \in M_{sh} \subseteq Z$ but $Y \notin MB(X, Z)$. Let $S$ be $M_{sh} \setminus \{Y\}$. Algorithm requires that $BIC_{D}(X, M_{sh} \setminus \{Y\}) < BIC_{D}(X, M_{sh})$. In other words, we have $BIC_{D}(X, S) < BIC_{D}(X, S \cup \{Y\})$. By Lemma E.3 we have $X \perp_{p} Y | S$.

Let $W = S \setminus MB(X, Z)$. From $Y \notin MB(X, Z)$ and $Y \notin S$, we have $\{Y\} \cup W \subseteq Z \setminus MB(X, Z)$. Recall Definition A.1 that $X \perp_{p} Z \setminus MB(X, Z) | MB(X, Z)$. Thus,

$$X \perp_{p} \{Y\} \cup W | MB(X, Z) \quad \therefore X \perp_{p} Z \setminus MB(X, Z) | MB(X, Z), \text{ decomposition (12)}$$

$$X \perp_{p} Y | MB(X, Z) \cup W \quad \therefore (12), \text{ weak union (13)}$$

$$X \perp_{p} Y | S \quad \therefore (13), W = S \setminus MB(X, Z) \quad (14)$$

Contradiction arises with $X \not\perp_{p} Y | S$.

Notice that $M^{-1} = M^{i} \cup \{W_{i}\}$. Algorithm requires that $BIC_{D}(X, M^{i}) > BIC_{D}(X, M^{i-1}) = BIC_{D}(X, M^{i} \cup \{W_{i}\})$. We then have

$$X \perp_{p} W_{1} | M^{1} \quad \therefore BIC_{D}(X, M^{1}) > BIC_{D}(X, M^{0}) \text{, Lemma E.3} \quad (15)$$

$$X \perp_{p} W_{2} | M^{2} \quad \therefore BIC_{D}(X, M^{2}) > BIC_{D}(X, M^{1}) \text{, Lemma E.3} \quad (16)$$

$$X \perp_{p} W_{1} | M^{2} \cup \{W_{2}\} \quad \therefore (15), M^{1} = M^{2} \cup \{W_{2}\} \quad (17)$$

$$X \perp_{p} \{W_{1}, W_{2}\} | M^{2} \quad \therefore (16), (17), \text{ contraction (18)}$$

$$\vdots$$

$$X \perp_{p} \{W_{1}, ..., W_{k}\} | M^{k} \quad \therefore ..., \text{ contraction (19)}$$

$$X \perp_{p} W | M_{sh} \quad \therefore (19), W = \{W_{1}, ..., W_{k}\} \text{ and } M^{k} = M_{sh} \quad (20)$$

$$X \perp_{p} Z \setminus M_{sh} | M_{sh} \quad \therefore (20), W = Z \setminus M_{sh} \quad (21)$$

Hence, it follows from Definition A.1 that $M_{sh} \supseteq MB(X, Z)$. □

Theorem E.6 Consider an observational dataset $D$ with $n$ i.i.d. observations from a compositional graphoid $\mathcal{P}$ over $V$. In the large sample limit of $n$, for any $X \in V$ and any $Z \subseteq V \setminus \{X\}$, $MB(X, Z) = M_{gs}$ where $M_{gs} = shrink(D, X, grow(D, X, Z))$.

Proof. Immediate from Lemma E.4 and Lemma E.5 □

Theorem E.7 Consider an observational dataset $D$ with $n$ i.i.d. observations from a (compositional) graphoid $\mathcal{P}$ over $V = \{X_{1}, ..., X_{m}\}$, and any $\pi \in \Pi(v)$. Let $s_{\pi}$ and $M_{\pi}$ be the score and the set of variables returned by $shrink(D, X_{1}, X_{\mathcal{P}(\pi)})$ (or $shrink(D, X_{1}, grow(D, X_{1}, X_{\mathcal{P}(\pi)})$) if $\mathcal{P}$ is a compositional graphoid) respectively. Denote $s_{\pi}$ as $\sum_{i \in V}s_{i}$. In the large sample limit of $n$, $BIC_{D}(\mathcal{G}_{\pi}) = s_{\pi}$ where $\mathcal{G}_{\pi}$ is induced from $\pi$ by (VP).
Proof. Immediate from the decomposable feature of BIC scores, Lemma E.5 and Theorem E.6. □

Lastly, though a compositional graphoid is a sufficient condition for the correct identification of the unique Markov boundary using the grow-shrink algorithm, we are aware of an assumption weaker than compositional graphoid to validate such an identification. Nevertheless, this discussion will be beyond the scope of this paper and we will leave the formal proof to future work.

F ADDITIONAL EXAMPLES

F.1 LU ET AL. COMPARISON

Reported below are average statistics obtained by running GRaSP₂ on the published datasets used to generate Figure 6 in [Lu et al., 2021]. We cannot compare these results to Lu et al. precisely, since their statistics are given in figures and not exactly in tables, though judging from their figures it appears that GRaSP₂ is dominating for adjacency precision and recall, arrowhead recall, and most results for arrowhead precision. Timing results are not reported by Lu et al.; we include these to show that GRaSP₂ returns quickly for all of these examples, where we know (personal communication) that some of the results for Triple A* take much longer. Adjacencies in these graphs are sampled with uniform probability, “Edge-prob”.

| Edge-prob | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 |
|-----------|------|------|------|------|------|------|
| Precision | 0.964 | 0.976 | 0.979 | 0.980 | 0.982 | 0.976 |
| Recall    | 0.985 | 0.982 | 0.986 | 0.986 | 0.985 | 0.985 |
| F1        | 0.974 | 0.979 | 0.983 | 0.983 | 0.983 | 0.980 |

Table 2: GRaSP₂ Adjacency Statistics

| Edge-prob | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 |
|-----------|------|------|------|------|------|------|
| Precision | 0.907 | 0.914 | 0.933 | 0.949 | 0.946 | 0.945 |
| Recall    | 0.897 | 0.916 | 0.933 | 0.952 | 0.952 | 0.955 |
| F1        | 0.898 | 0.913 | 0.932 | 0.950 | 0.948 | 0.950 |

Table 3: GRaSP₂ Arrowhead Statistics

| Edge-prob | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 |
|-----------|------|------|------|------|------|------|
| Seconds   | 0.405 | 0.755 | 1.403 | 2.703 | 4.795 | 7.161 |

Table 4: GRaSP₂ Timing Statistics

F.2 AIRFOIL EXAMPLE

Figure 10 gives the results of running GRaSP₂, PC, and fGES on the Airfoil empirical example described in Section 6. GRaSP₂ gets the same uniquely frugal result as SP. To improve readability, we use the names of the variables (instead of numerals) to label the vertices.

Note that both the GRaSP₂ and FGES results use the linear, Gaussian BIC score with a penalty multiplier of 2. For the GRaSP₂ result in (a), Attack is not exogenous, which is counter-intuitive, since it is experimentally controlled. Allowing for latent variables could resolve this issue. However, we leave the development of such an algorithm to future work. On the other hand, the FGES result in (b) is notably not the same as the SP result and so is not frugal. Also, the orientation between Attack and Displacement is reversed.

The PC result in (c), which uses the zero partial correlation test with a significance level of 0.01, in fact has fewer edges than the frugal result and makes Chord, another experimental variable, endogenous. Causally, PC is giving incorrect and incomplete information.

https://github.com/ninalu/urlearning-cpp/tree/master/triplet_data
We consider path cancellations in DAGs between pairs of vertices, one of which is exogenous, connected by two or more unique treks. Furthermore, the path cancellations we consider elicit a marginal independence between the two vertices in question. Below, we enumerate all possible path cancellations of this type (up to vertex relabeling). Each graph illustrates a case where an unfaithful marginal independence is elicited between the two gray vertices due to path cancellation. A complete list of all unfaithful CI relations (symmetry assumed) where the independent sets are singletons is also provided for each graph.

(a) GRaSP result

(b) fGES result

(c) PC result

Figure 10: Results of algorithms on NASA airfoil experiment.

G UNIT TESTS

We consider path cancellations in DAGs between pairs of vertices, one of which is exogenous, connected by two or more unique treks. Furthermore, the path cancellations we consider elicit a marginal independence between the two vertices in question. Below, we enumerate all possible path cancellations of this type (up to vertex relabeling). Each graph illustrates a case where an unfaithful marginal independence is elicited between the two gray vertices due to path cancellation. A complete list of all unfaithful CI relations (symmetry assumed) where the independent sets are singletons is also provided for each graph.
$X_1 \perp X_4$

$X_1 \perp X_4$

$X_1 \perp X_5$

$X_1 \perp X_5$
Figure 1: Lu et al. Fig. 3
Figure 2: Lu et al. Fig. 3
Figure 3: Lu et al. Fig. 6
Figure 4: Lu et al. Fig. 6
Figure 5: GRa

Figure 6: Lu et al.