A Divide and Conquer Approach for Solving Structural Causal Models

Anna Rodum Bjøru ANNA.R.BJORU@NTNU.NO
Department of Computer Science
Norwegian University of Science and Technology, Trondheim (Norway)
Rafael Cabañas RCABANAS@UAL.ES
Department of Mathematics and
Centre for the Development and Transfer of Mathematical Research to Industry (CDTIME)
University of Almería (Spain)
Helge Langseth HELGE.LANGSETH@NTNU.NO
Department of Computer Science
Norwegian University of Science and Technology, Trondheim (Norway)
Antonio Salmerón ANTONIO.SALMERON@UAL.ES
Department of Mathematics and
Centre for the Development and Transfer of Mathematical Research to Industry (CDTIME)
University of Almería (Spain)

Editors: J.H.P. Kwisthout & S. Renooij

Abstract

Structural causal models permit causal and counterfactual reasoning, and can be regarded as an extension of Bayesian networks. The model consists of endogenous and exogenous variables, with exogenous variables often being of unknown semantic interpretation. Consequently, they are typically non-observable, with the result that counterfactual queries may be unidentifiable. In this setting, standard inference algorithms for Bayesian networks are insufficient. Recent methods attempt to bound unidentifiable queries through imprecise estimation of exogenous probabilities. However, these approaches become unfeasible with growing cardinality of the exogenous variables. This paper proposes a divide and conquer method that transforms a general causal model into a set of models with low-cardinality exogenous variables, for which any query can be calculated exactly. Bounds for a query in the original model are then efficiently approximated by aggregating the results for the set of smaller models. Experimental results demonstrate that these bounds can be computed with lower error levels and less resource consumption compared to existing methods.

Keywords: Structural causal models; causality; counterfactual reasoning; Satisfiability; Heuristic search.

1. Introduction

Structural causal models (SCMs) with discrete variables (Pearl, 2009; Bareinboim et al., 2022) are a type of probabilistic graphical model (PGM) for causal and counterfactual reasoning. SCMs enable reasoning about hypothetical scenarios, such as estimating the probability of recovery for a deceased patient in a medical trial if they had received a different treatment. SCMs consist of endogenous (observable) and exogenous (usually latent)

variables, with endogenous values determined from exogenous ones through structural equations. Often, the exogenous probabilities are unavailable due to the lack of data for these variables. Consequently, many queries are non-identifiable and cannot be calculated.

One of the first approaches for addressing this problem was proposed by Kang and Tian (2006), who presented a systematic technique to derive constraints on a causal query, albeit with exponential growth. Sachs et al. (2023) introduced a method for deriving bounds on causal effects. Zhang et al. (2022) proposed approximating credible intervals based on sampling algorithms. More related to our work, Zaffalon et al. (2020) proposed transforming SCMs into credal networks (Cozman, 2000), requiring the solution of various linear programming problems. However, this approach may be infeasible due to the large cardinality of exogenous variables. More recently, Zaffalon et al. (2024) introduced EMCC for approximating the bounds of any non-identifiable query. This involves repeatedly running the expectation-maximization (EM) algorithm (Koller and Friedman, 2009) to obtain precise specifications of exogenous distributions. Queries can then be separately calculated and aggregated to approximate the bounds. The problem with EMCC is that each individual EM run necessitates an exceptionally large number of iterations to achieve low error.

This paper introduces the *Divide and Conquer for Causal Computation* (DCCC) method, which integrates elements of the two previously mentioned approaches. It aims to obtain precise specifications of exogenous distributions, from which any query can be calculated. DCCC reduces SCMs by removing certain exogenous states, transforming the SCMs into collections of less complex models. Then, various linear programming problems with unique solutions are solved in the reduced models. Experimental results show that DCCC achieves these bounds with lower error levels and in less time compared to EMCC.

2. Background

2.1. Basic notation

With respect to the general notation, upper-case letters are used to denote random variables and lower-case for their possible values (or states). That is, given a variable V, v is an element of its domain, denoted by Ω_V. We assume that all the variables are discrete. Similarly, $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$ denotes a set of variables and $\mathbf{v}$ a joint state of its domain $\Omega_{\mathbf{V}} = \times_{V \in \mathbf{V}} \Omega_V$. For the sake of simplicity, variables are omitted from assignments when their context is clear. For instance, $P(V = v)$ will be denoted simply as $P(v)$. In a directed graph, Pa_V are the parents (i.e., the immediate predecessors) of V.

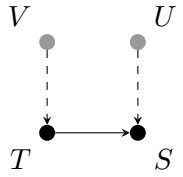
2.2. Structural causal models

Structural Causal Models (SCMs) (Pearl, 2009) are a class of probabilistic graphical models (PGMs) used for causal and counterfactual reasoning, consisting of two types of nodes: *endogenous* nodes, which represent the internal variables of the modeled problem, and *exogenous* nodes, which represent factors outside the model. SCMs can be more formally defined as follows (Bareinboim et al., 2022).

Definition 1 *A structural causal model (SCM) is defined as a 5-tuple* $\langle \mathbf{U}, \mathbf{V}, \mathcal{G}, \mathcal{F}, \mathcal{P} \rangle$, *where* $\mathbf{U}$ *and* $\mathbf{V}$ *are respectively the sets of exogenous and endogenous variables;* $\mathcal{G}$ *is a directed acyclic graph (DAG) representing the causal relationships among variables in* $\mathbf{U} \cup \mathbf{V}$*;*

$\mathcal{F}$ is a set of structural equations (SEs) $\{f_{V_1}, f_{V_2}, ..., f_{V_n}\}$, such that each of them is a map $f_X : \Omega_{\mathrm{Pa}_V} \to \Omega_V$; $\mathcal{P}$ is a set containing a probability distribution $P(U)$ for each $U \in \mathbf{U}$.

When the distributions for the exogenous variable are unknown, we say that the model is a partially specified SCM. Such models are denoted by calligraphic letters, like $\mathcal{M}$. By contrast, if all such distributions are provided, we say instead that it is fully specified and denoted as, e.g., M. If we do not explicitly state whether a model is fully specified or not, we use the most general representation, namely $\mathcal{M}$. As an example, let us consider the SCM shown in Figure 1 modeling a drug study involving 700 patients (Mueller et al., 2021). The causal graph depicted on the left includes the endogenous variables $\mathbf{V} = \{T, S\}$ representing the *treatment* and the *survival* respectively. The aim is to analyze whether being treated ($T = 1$) helps in survival ($S = 1$). On the other hand, $\mathbf{U} = \{V, U\}$ is the set of exogenous variables, which are assumed to be root and having as children endogenous variables. In this paper, only *Markovian* models will be considered, meaning that each endogenous variable has a single exogenous parent and each exogenous variable has a single endogenous child (as the model in the example). In other words, models that do not have hidden confounders.



f_T

T	v_0	v_1
0	1	0
1	0	1

f_S

T	S	u_0	u_1	u_2	u_3
0	0	1	1	0	0
1	0	1	0	1	0

$\tilde{P}(T)$	$T=0$	$T=1$
	0.337	0.662

$\tilde{P}(S\|T)$	$T=0$	$T=1$
$S=0$	0.462	0.323
$S=1$	0.538	0.677

Figure 1: Elements of an SCM: (left) causal graph, (center) structural equations and (right) empirical distribution computed from the data.

Figure 1 (center) shows the SEs as deterministic CPTs of the form $P(T|V)$ and $P(S|T, U)$. If not provided from expert knowledge, SEs can be automatically inferred from the causal graph, without any loss of generality, via a *canonical specification*. This is the case of the SEs shown in the example, where the states of an exogenous variable will then represent all possible deterministic mechanisms between its children and their respective endogenous parents. Conversely, under the non-canonical specification, some of the exogenous states are assumed to be impossible and directly omitted from the CPTs. In a Markovian model, the cardinality of each exogenous variable U is at most $|\Omega_Y|^{|\Omega_{\mathbf{X}}|}$, where Y is its only child and $\mathbf{X}$ the set of endogenous parents of Y.

When doing inference with an SCM $\mathcal{M}$ and a dataset $\mathcal{D}$, typically only observations for the endogenous variables are available. For example, in the model under consideration, it is possible to calculate from $\mathcal{D}$ the empirical distributions $\tilde{P}(T)$ and $\tilde{P}(S|T)$ with the values shown in Figure 1 (right). The task of doing inference in a partial SCM given a dataset involves estimating a set of fully-specified SCMs that are compatible with the data. This idea, which is fundamental for our method, is extended in the following section.

3. Imprecise characterisation

As initially proposed by Zaffalon et al. (2020) and later expanded upon by Zaffalon et al. (2024), it has been demonstrated that an SCM can be accurately mapped into a *credal*

network (Cozman, 2000), which is essentially a generalization of a Bayesian network with an imprecise specification of its parameters. In this context, each node is associated with a set of probability mass functions known as a *credal set*. Specifically, when mapping SCMs into credal networks, the endogenous observations impose linear constraints on the probabilities of the exogenous variables. Consequently, we can derive a separate credal set for each exogenous variable, which here will be called the *solution set*, and formally defined as follows.

Definition 2 (Solution set for an exogenous variable) *A solution for an exogenous variable U and a dataset $\mathcal{D}$ is the credal set defined as*

$$\mathcal{K}(U) := \left\{ P(U) : \sum_{u \in \Omega_U} P(u) \cdot P(Y|\mathbf{X}, u) = \tilde{P}(Y|\mathbf{X}) \right\} \tag{1}$$

where Y is the only child of U, $\mathbf{X}$ the set of endogenous parents of Y and $\tilde{P}(Y|\mathbf{X})$ is the empirical distribution computed from $\mathcal{D}$.

In other words, $\mathcal{K}(U)$ is the convex set of all the distributions over U leading to the same distribution over the endogenous children after marginalizing out the exogenous parents. In our problem, we are interested in finding all these distributions. Each element $P(U) \in \mathcal{K}(U)$ will be called a solution for U. This idea can be extended to an SCM as follows:

Definition 3 (Solution for an SCM) *A solution for a partially-defined SCM with exogenous variables $\mathbf{U}$ given dataset $\mathcal{D}$ is a set of distributions $\{P(U)\}_{U \in \mathbf{U}}$ where each $P(U)$ is in $\mathcal{K}(U)$, i.e., it is a solution for the given exogenous variable.*

Example 1 *For the model $\mathcal{M}$ and the empirical distribution (from a dataset $\mathcal{D}$) shown in Figure 1, there is a unique $\mathcal{S}_{\mathcal{M},\mathcal{D}}$ for variable V, namely $\mathcal{K}(V) = \{P(V) = [0.337, 0.662]\}$. On the other hand, solutions for U are all the distributions that are a convex combination of $P_1(U)$ and $P_2(U)$, defined as*

$$P_1(U) = \begin{matrix} u_0 & u_1 & u_2 & u_3 \\ [0.323 & 0.139 & 0 & 0.538] \end{matrix}, \quad P_2(U) = \begin{matrix} u_0 & u_1 & u_2 & u_3 \\ [\,0 & 0.462 & 0.323 & 0.215] \end{matrix}.$$

Intuitively, each fully-specified SCM that is a solution is also a model that can have produced the available endogenous data. Thus, we can introduce the following solvability condition:

Definition 4 (Solvable SCM) *A partially-defined SCM is* solvable *for a given dataset $\mathcal{D}$ iff there exists a non-empty solution set $\mathcal{K}(U)$ for each exogenous variable $U \in \mathbf{U}$.*

In other words, a partially-defined SCM is solvable (for the available endogenous data) if there exists at least one distribution for each exogenous variable satisfying the linear constraints. When this happens, it is said that the dataset is *M-compatible* (Zaffalon et al., 2024). Frequently, the solution for an SCM will not be unique, and hence it is required to define a solution set for an SCM as follows.

Definition 5 (Solution set for an SCM) *A solution set for a partially-defined SCM $\mathcal{M}$ given dataset $\mathcal{D}$, denoted $\mathcal{S}_{\mathcal{M},\mathcal{D}}$, is the set of all the fully-specified SCMs such that $P(U) \in \mathcal{K}(U)$ for each $U \in \mathbf{U}$.*

That is, each SCM contained in $\mathcal{S}_{\mathcal{M},\mathcal{D}}$ is a solution for the given dataset and partially-defined SCM. This set can be represented as a credal network where each exogenous variable has associated a credal set as defined by Equation (1) whereas the endogenous variables, T and S, instead of a credal set, have associated a single conditional distribution corresponding to the SEs, i.e. f_S. Intuitively, this precise model represents all the fully-specified SCMs that might have produced the available endogenous data. Typical counterfactual queries in an SCM are *the probability of sufficiency* (PS) or *probability of necessity* (PN). For further details see (Pearl, 2009) or (Cabañas et al., 2024). Any counterfactual query $q_{\mathcal{M},\mathcal{D}}$ can be transformed into a query in the credal network defined as

$$q_{\mathcal{M},\mathcal{D}} := \{q_M \mid M \in \mathcal{S}_{\mathcal{M},\mathcal{D}}\} \tag{2}$$

where q_M is the same query but computed in each fully-specified SCM M which is member of the solution set. In practice, previous set of queries will be summarized by the lower and upper bounds:

$$\left[\min_{M\in\mathcal{S}_{\mathcal{M},\mathcal{D}}} q_M, \max_{M\in\mathcal{S}_{\mathcal{M},\mathcal{D}}} q_M\right]. \tag{3}$$

Example 2 *In the context of the running example, the set of models contained in $\mathcal{S}_{\mathcal{M},\mathcal{D}}$ is any fully-specified SCM where $P(w_0) = 0.337, P(w_1) = 0.662$ and $P(U)$ is a convex combination of $P_1(U)$ and $P_2(U)$. Then PS is bounded to the interval $[0.301, 1.0]$.*

4. Divide and conquer algorithm

4.1. Model reduction

The underlying idea of our method consists of transforming a complex SCM into a simpler one with exogenous variables of smaller cardinality. The transformation proposed is essentially the removal of some states from exogenous domains, namely a *reduction*, which can be defined as follows.

Definition 6 *Let $\mathcal{M}$ be a partial SCM whose set of exogenous variables is $\mathbf{U}$ and let $u \in \Omega_U$ with $U \in \mathbf{U}$. Then the reduction operation, denoted $R(\mathcal{M}, u)$, produces a new partial SCM $\mathcal{M}'$ by removing assignments from $\mathcal{F}$ and $\mathcal{P}$ in $\mathcal{M}$ that are consistent with u.*

This reduction operation will be applied to various states of the different exogenous variables, i.e., to a set defined as $\mathcal{A}_{\mathbf{U}} := \{u^{(i)}\}_{i=1}^{m}$ s.t. $u^{(i)} \in \Omega_U$ and $U \in \mathbf{U}$. For simplicity we can recursively define this as $R(\mathcal{M}, \mathcal{A}_{\mathbf{U}}) = R\left(R(\mathcal{M}, u^{(1)}), \mathcal{A}_{\mathbf{U}} \setminus \{u^{(1)}\}\right)$. The reduction operation $R(\mathcal{M}, u)$ is equivalent to imposing the additional constraint that $P(u) = 0$ on $\mathcal{M}$. Thus, when looking for the solution set of a reduced SCM, we can equivalently look for all the solutions from the original solution set that are consistent with that constraint. The following theorem simplifies calculating the set of queries given in Equation (2).

Theorem 1 *Let $\mathcal{M}$ be a partially-defined SCM, and $\mathcal{D}$ an M-compatible dataset. If $\mathcal{M}' = R(\mathcal{M}, \mathcal{A}_{\mathbf{U}})$ is solvable for $\mathcal{D}$, then it holds that $\mathcal{S}_{\mathcal{M}',\mathcal{D}} \subseteq \mathcal{S}_{\mathcal{M},\mathcal{D}}$.*

Proof Any fully-specified model member of $\mathcal{S}_{\mathcal{M},\mathcal{D}}$ by definition satisfies the linear constraints specified in Equation (1) for each $U \in \mathbf{U}$. The same constraints are satisfied by any member of $\mathcal{S}_{\mathcal{M}',\mathcal{D}}$. Additionally, any fully specified SCM in the latter set also respects the additional constraints imposed by $R(\mathcal{M}, \mathcal{A}_{\mathbf{U}})$ stating that $P\left(u^{(i)}\right) = 0$ on $\mathcal{M}$ for all $\mathcal{A}_{\mathbf{U}} = \{u^{(i)}\}_{i=1}^{m}$ s.t $u^{(i)} \in \Omega_U$ and $U \in \mathbf{U}$. Given that the set of constraints associated to $\mathcal{S}_{\mathcal{M},\mathcal{D}}$ is a subset of those associated to $\mathcal{S}_{\mathcal{M}',\mathcal{D}}$, it holds that if $M \in \mathcal{S}_{\mathcal{M}',\mathcal{D}}$ then $M \in \mathcal{S}_{\mathcal{M},\mathcal{D}}$, and the result follows. ■

Corollary 1 *Let $\mathcal{M}$ be a partially-defined SCM, and $\mathcal{D}$ an M-compatible dataset. If $\mathcal{M}' = R(\mathcal{M}, \mathcal{A}_{\mathbf{U}})$ is solvable for $\mathcal{D}$, then it holds that $q_{\mathcal{M}',\mathcal{D}} \subseteq q_{\mathcal{M},\mathcal{D}}$.*

Proof According to Equation (2), and taking into account that, as stated by Theorem 1, $\mathcal{S}_{\mathcal{M}',\mathcal{D}} \subseteq \mathcal{S}_{\mathcal{M},\mathcal{D}}$,

$$q_{\mathcal{M}',\mathcal{D}} = \{q_M \mid M \in \mathcal{S}_{\mathcal{M}',\mathcal{D}}\} \subseteq \{q_M \mid M \in \mathcal{S}_{\mathcal{M},\mathcal{D}}\} = q_{\mathcal{M},\mathcal{D}}.$$

■

Example 3 *If the reduction $R(\mathcal{M}, u_2)$ is applied to the running example, the solution set for U is $\mathcal{K}(U) = \{[P(u_0) = 0.323, P(u_1) = 0.139, P(u_3) = 0.538]\}$ and PS is 0.301, which is the lower bound in Example 2.*

4.2. Scope of the method

The most general specification of a Markovian SCM $\mathcal{M}$ lets each endogenous variable Y have a single exogenous parent U with states that index all the possible deterministic structural equations from the set of Y's endogenous parents in $\mathcal{M}$ to Y. With this general specification, for a dataset $\mathcal{D}$, any model M that is compatible with $\mathcal{D}$ is contained in $\mathcal{S}_{\mathcal{M},\mathcal{D}}$. For such a model, the size of the state space of an exogenous variable U is determined by the size of the state spaces of its endogenous child variable Y and of Y's endogenous parents $\mathbf{X} = \text{Pa}_Y \setminus \{U\}$, with $|\Omega_U| = |\Omega_Y|^{|\Omega_{\mathbf{X}}|}$. Letting $m = |\Omega_Y|^{|\Omega_{\mathbf{X}}|}$, define $\Omega_U = \{u_i\}_{i=0}^{m-1}$. A mapping between the possible functions f_{u_i} from $\mathbf{X}$ to Y and states in Ω_U then completes the definition of the structural function $f_Y(\mathbf{X}, U) = f_U(\mathbf{X})$ in $\mathcal{M}$.

Assuming all endogenous variables in $\mathcal{M}$ are binary, a structured approach to defining a consistent mapping from u_i to f_{u_i} for n endogenous parents of Y is defined as follows: For a specific ordering $(\mathbf{x}_i)_{i=1}^{2^n}$ of the 2^n states of the endogenous parents $\mathbf{X}$, the function indexed by u_i is given by the binary encoding of i to 2^n digits, such that the digit at position j correspond to the output y of $f_{u_i}(\mathbf{x}_j)$. If n=2, with the order of states for $\mathbf{X} = (X_1, X_2)$ as $((0,0),(0,1),(1,0),(1,1))$, u_0 is defined by binary string $0_{10} = 0000_2$ such that $f_Y(X_1, X_2, u_0) = 0$ for all states, while for u_3, $3_{10} = 0011_2$ defines $f_Y(X_1, X_2, u_3) = X_1$, i.e. for the first two states $(0,0),(0,1)$ of X_1, X_2, the output is 0 and for the next two states $(1,0),(1,1)$ the output is 1. Tables 1 and 2 details the complete mapping for the case where Y has one endogenous parent X and where Y has two endogenous parents X_1, X_2,

X	Y	U				$P(Y\|X)$
		u_0	u_1	u_2	u_3	
0	0	1	1	0	0	$P(Y=0\|X=0)=P(u_0)+P(u_1)$
0	1	0	0	1	1	$P(Y=1\|X=0)=P(u_2)+P(u_3)$
1	0	1	0	1	0	$P(Y=0\|X=1)=P(u_0)+P(u_2)$
1	1	0	1	0	1	$P(Y=1\|X=1)=P(u_1)+P(u_3)$
$f_U(X)=$		0	X	$\overline{X}$	1	

Table 1: A canonical specification of Ω_U when endogenous child Y of U has one endogenous parent X. For columns with header u_i, a 1 indicates that the function $f_{u_i}(X) = Y$ as defined in the respective column is consistent with the pair of values for X, Y as listed in the respective row, i.e. that Y takes on its value with probability 1 given the value of X and U. Conversely, a 0 indicates 0 probability of the value combination in question. While the complete table shows probabilities of all state combinations, this can be translated into the corresponding binary representation column-wise by selecting the Y-value of probability 1 for each pair of consecutive rows, which is equivalent to reading only the probabilities for the $P(Y = 1|X)$-rows per column. The bottom row summarises the functions defined. The rightmost column shows the resulting linear equations defining the credal set $\mathcal{K}(U)$.

X_1	X_2	Y	U																$P(Y\|X_1,X_2)$
			u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{13}	u_{14}	u_{15}	
0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	$P(Y=0\|X_1=0,X_2=0)$
0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	$P(Y=1\|X_1=0,X_2=0)$
0	1	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	$P(Y=0\|X_1=0,X_2=1)$
0	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	$P(Y=1\|X_1=0,X_2=1)$
1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	$P(Y=0\|X_1=1,X_2=0)$
1	0	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	$P(Y=1\|X_1=1,X_2=0)$
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	$P(Y=0\|X_1=1,X_2=1)$
1	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	$P(Y=1\|X_1=1,X_2=1)$
$f_U(X_1,X_2)=$			0	$X_1 \wedge X_2$	$X_1 \wedge \overline{X}_2$	X_1	$\overline{X}_1 \wedge X_2$	X_2	$X_1 \oplus X_2$	$X_1 \vee X_2$	$\overline{X}_1 \wedge \overline{X}_2$	$X_1 \oplus \overline{X}_2$	$\overline{X}_2$	$X_1 \vee \overline{X}_2$	$\overline{X}_1$	$\overline{X}_1 \vee X_2$	$\overline{X}_1 \vee \overline{X}_2$	1	

Table 2: A canonical specification of Ω_U when endogenous child Y of U has two endogenous parents X_1 and X_2. The table is read analogously to Table 1. The probabilities in the rightmost column are again the sum of the probabilities of all u_i for which the respective column evaluates to 1.

respectively. Given this definition of $f_Y(\mathbf{X}, U)$, the credal set $\mathcal{K}(U)$ is now defined by the following linear system:

$$\sum_{u \in \Omega_U^{j,k}} P(u) = P(Y = y_k|\mathbf{X} = \mathbf{x}_j), \quad y_k \in \Omega_Y, \mathbf{x}_j \in \Omega_{\mathbf{X}} \tag{4}$$

with $\Omega_U^{j,k} = \{u_i \in \Omega_U, i = 0, \ldots, |U|-1$ s.t. position j in the binary encoding of i equals $y_k\}$ (See the rightmost column in Table 1 for an example). This linear system is underdetermined with 2^{2^n} unknowns, such that there are infinitely many solution models for $\mathcal{M}$, for which $\mathcal{D}$ is M-compatible. Now, consider reductions R for which the resulting model $\mathcal{M}' = R(\mathcal{M}, \mathcal{A}_U)$ has a single solution M. The following theorem bounds the size of the domain of an exogenous variable U' of a reduced model $\mathcal{M}'$ for which the linear system has a unique solution:

Theorem 2 *Let $\mathcal{M}$ be a partially-defined Markovian SCM with binary endogenous variables, and $\mathcal{D}$ an M-compatible dataset. For an exogenous variable U with endogenous child variable Y in $\mathcal{M}$, let n denote the number of endogenous parents of Y. If a reduction R is applied to $\mathcal{M}$ returning $\mathcal{M}'$ such that the resulting linear system defining $\mathcal{K}(U')$ has exactly one solution M for $\mathcal{D}$, then $|\Omega_{U'}| \leq 2^n + 1$, where U' is the image of U under R.*

Proof The equation $\sum_i P(u_i) = 1$ along with $\{P(Y = 0|\mathbf{X} = \mathbf{x}_j)\}_{j=1}^{2^n}$ ($\{P(Y = 1|\mathbf{X} = \mathbf{x}_j)\}_{j=1}^{2^n}$ being dependent given $\sum_i P(u_i) = 1$) together form a system of 2^n+1 independent linear equations. For a number of unknowns after reduction $|\Omega_{U'}| > 2^n + 1$, the system solution space is infinite.

■

Such a model reduction R, for which $|\mathcal{K}(U')| = 1$, effectively reduces the number of unknowns in $\{P(u_i)\}_{u_i \in \Omega_U}$ by fixing $2^{2^n} - (2^n+1)$ of these probabilities to be 0. If the unique solution of the resulting linear system given data $\mathcal{D}$ respects $P(u_i) \geq 0$ and $\sum_i P(u_i) = 1$, then this solution is a model $M \in \mathcal{S}_{\mathcal{M},\mathcal{D}}$, for which $q_M \in q_{\mathcal{M},\mathcal{D}}$.

4.3. Satisfiability

Algorithm 1 outlines the steps of the process of finding fully-specified SCMs given a partially-specified model $\mathcal{M}$ and a dataset $\mathcal{D}$, where Line 4 corresponds to finding a reduction such that the resulting model has a unique solution. Now, Theorem 2 states that such a reduction R may reduce the complexity of the model to be solved significantly in replacing variable U, for which $|\Omega_U| = 2^{2^n}$, by U' with $|\Omega_{U'}| \leq 2^n + 1$. However, the space of possible reductions consists of a total of $\binom{2^{2^n}}{2^n+1}$ distinct R's, for which U' has the required domain size $2^n + 1$. Moreover, out of this set of reduced models, only a potentially small subset will be consistent with a given dataset $\mathcal{D}$. Thus, this search is not straight forward when n grows.

A connection to the satisfiability problem is introduced next. Defining a new set of variables $\{z_i\}_{i=0}^{m-1}$ such that $z_i = \begin{cases} \text{True}, & \text{if } P(u_i)>0 \\ \text{False}, & \text{if } P(u_i)=0 \end{cases}$, a Conjunctive Normal Form (CNF) formula may be formed from the left hand side expressions of the equation set given by Equation (4). For each sum $\sum_{u \in \Omega_U^{j,k}} P(u)$, a disjunction clause over the variables in the corresponding set $\{z_i : i \text{ s.t. } u_i \in \Omega_U^{j,k}\}$ is added to the CNF formula, such that the complete formula is the conjunction of all $2 \cdot 2^n$ disjunction clauses. For $n = 1$, the complete equation set as shown in the rightmost column of Table 1 corresponds to the CNF formula $g(\mathbf{z}) = (z_0 \vee z_1) \wedge (z_2 \vee z_3) \wedge (z_0 \vee z_2) \wedge (z_1 \vee z_3)$. Now, in order for any subset $\{P(u_i)\}_{i \in \mathcal{I}}, |\mathcal{I}| = 2^n+1$, to solve the linear system given by Equation (4), the variables in $\{z_i\}_{i \in \mathcal{I}}$ set to True must be a solution to the corresponding CNF formula $g(\mathbf{z})$. If not, $P(u) = 0 \; \forall u \in \Omega_U^{j,k}$ for some pair of values $\mathbf{x}_j, y_k$, which in general will violate Equation (4). Thus it is only necessary to consider $\mathcal{I}$ such that $g(\mathbf{z})$ is satisfied, to be possible model solutions of the linear system.

Thus, Line 4 of Algorithm 1 may be accomplished by first finding a set $\{z_i\}_{i \in \mathcal{I}}$ that solves $g(\mathbf{z})$, and only then solving the linear system, by setting all $\{P(u_j)\}_{j=0}^{m-1} \setminus \{P(u_i)\}_{i \in \mathcal{I}}$ to 0 and then inverting the resulting $(2^n+1) \times (2^n+1)$ square matrix. If this unique solution satisfies $\sum_{i \in \mathcal{I}} P(u_i) = 1$ and $P(u_i) \geq 0, \forall i \in \mathcal{I}$, the solution corresponds to a fully-specified

Algorithm 1 Learning

input: $\mathcal{M}$ (partially-specified SCMs), $\mathcal{D}$ (endogenous dataset), N (number of runs)
output: $\mathcal{S} = \{M_1, M_2, \ldots, M_N\}$ (set of fully-specified SCMs)

$\mathcal{S} \leftarrow \emptyset$
for $i \in \{1, \ldots, N\}$ **do**
  for $U \in \mathbf{U}$ **do**
    Find a $\mathcal{A}_U \subset \Omega_U$ s.t. $\mathcal{M}' = R(\mathcal{M}, \mathcal{A}_U)$ has a single solution for U.
    $P'(U) \leftarrow$ solve U in $\mathcal{M}'$ given the data $\mathcal{D}$.
  end for
  $M_i \leftarrow$ build an SCM with $\{P'(U)\}_{U \in \mathbf{U}}$.
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{M_i\}$
end for
return $\mathcal{S}$

Algorithm 2 Inference

input: $\mathcal{S}$ (set of fully-specified SCMs), q (causal or counterfactual query)
output: bounds of q

$\mathbf{q} \leftarrow \emptyset$
for $M \in \mathcal{S}$ **do**
  $\mathbf{q} \leftarrow \mathbf{q} \cup \{q_M\}$
end for
return $(\min(\mathbf{q}), \max(\mathbf{q}))$

model $M \in \mathcal{S}_{\mathcal{M},\mathcal{D}}$, of which q_M lies in $q_{\mathcal{M},\mathcal{D}}$ to be approximated. Algorithm 2 details the approximation of $q_{\mathcal{M},\mathcal{D}}$ given a set of fully-specified models found by Algorithm 1.

For $n = 1$, all subsets of $\{P(u_i)\}_{i=0}^{3}$ of size $2^1 + 1 = 3$ correspond to subsets of $\{z_i\}_{i=0}^{3}$ that satisfy $g(\mathbf{z})$, and the 3 equations given by Equation (4) may be solved for each of the $\binom{2^{2^1}}{2^1+1} = \binom{4}{3} = 4$ possible subsets, such that a solution model $M \in \mathcal{S}_{\mathcal{M},\mathcal{D}}$ is found if the solution corresponds to a probability distribution.

Example 4 *There are four possible reductions $R(\mathcal{M}, u_i)$ for $i \in \{0, 1, 2, 3\}$, for variable U of model $\mathcal{M}$ of Figure 1. With data $\mathcal{D}$ of Figure 1, the solution set $\mathcal{K}(U)$ for $R(\mathcal{M}, u_2)$ is as shown in Example 3. The reduction $R(\mathcal{M}, u_0)$ returns a model with the solution set $\mathcal{K}(U) = \{[P(u_1) = 0.462, P(u_2) = 0.323, P(u_3) = 0.215]\}$, while both $R(\mathcal{M}, u_1)$ and $R(\mathcal{M}, u_3)$ are models with empty solution sets for $\mathcal{D}$.*

For $n = 2$, it is no longer the case that all of the size $2^2 + 1 = 5$ subsets of $\{P(u_i)\}_{i=0}^{15}$ satisfies $g(\mathbf{z})$. The total solution space of $\binom{16}{5} = 4368$ possible solutions may still be searched exhaustively, tested for satisfiability against $g(\mathbf{z})$, then tested by solving the equation system for possible solution models. For $n = 3$ however, $\binom{2^{2^n}}{2^n+1} = \binom{256}{9} \approx 10^{16}$, and a complete search is no longer feasible. Thus, a heuristics based search approach is described here to allow faster retrieval of models in $\mathcal{S}_{\mathcal{M},\mathcal{D}}$, based on ensuring satisfiability of the CNF formula.

4.4. Heuristic-based solution search

Now, instead of testing every possible subset among the $\binom{2^{2^n}}{2^n+1}$ possibilities, the approach presented here will generate sets $\{z_i\}_{i \in \mathcal{I}}$ of size $|\mathcal{I}| = 2^n + 1$ in such a way that 1) the CNF formula is guaranteed to be satisfied and 2) the solution for $\{P(u_i)\}_{i \in \mathcal{I}}$ is more likely to be a probability distribution than for randomly sampled subsets. Per conditional distribution $P(Y|\mathbf{X} = \mathbf{x}_j)$ over binary Y, each $P(u_i)$ contributes to exactly one of $P(Y = 0|\mathbf{X} = \mathbf{x}_j)$ or

$P(Y = 1|\mathbf{X} = \mathbf{x}_j)$. Thus for a pair of distribution clauses, any z_i not part of one clause is part of the other. This can be seen in Figure 2, which shows the CNF disjunction clauses when Y has two endogenous parents X_1, X_2.

CNF disjunction clauses	$P(Y\|X_1, X_2)$
$z_0 \vee z_1 \vee z_2 \vee z_3 \vee z_4 \vee z_5 \vee z_6 \vee z_7$	$P(Y = 0\|X_1 = 0, X_2 = 0)$
$z_8 \vee z_9 \vee z_{10} \vee z_{11} \vee z_{12} \vee z_{13} \vee z_{14} \vee z_{15}$	$P(Y = 1\|X_1 = 0, X_2 = 0)$
$z_0 \vee z_1 \vee z_2 \vee z_3 \vee z_8 \vee z_9 \vee z_{10} \vee z_{11}$	$P(Y = 0\|X_1 = 0, X_2 = 1)$
$z_4 \vee z_5 \vee z_6 \vee z_7 \vee z_{12} \vee z_{13} \vee z_{14} \vee z_{15}$	$P(Y = 1\|X_1 = 0, X_2 = 1)$
$z_0 \vee z_1 \vee z_4 \vee z_5 \vee z_8 \vee z_9 \vee z_{12} \vee z_{13}$	$P(Y = 0\|X_1 = 1, X_2 = 0)$
$z_2 \vee z_3 \vee z_6 \vee z_7 \vee z_{10} \vee z_{11} \vee z_{14} \vee z_{15}$	$P(Y = 1\|X_1 = 1, X_2 = 0)$
$z_0 \vee z_2 \vee z_4 \vee z_6 \vee z_8 \vee z_{10} \vee z_{12} \vee z_{14}$	$P(Y = 0\|X_1 = 1, X_2 = 1)$
$z_1 \vee z_3 \vee z_5 \vee z_7 \vee z_9 \vee z_{11} \vee z_{13} \vee z_{15}$	$P(Y = 1\|X_1 = 1, X_2 = 1)$

Figure 2: The CNF formula disjunction clauses over variables $\{z_i\}_{i=0}^{15}$ when Y has two endogenous parents X_1, X_2. Each pair of consecutive rows make up a single conditional distribution. Conditional probabilities are shown in the rightmost column, with left hand clauses corresponding to conditional probabilities for $Y = 0$ given different values for parents X_1, X_2, and right hand clauses correspond to conditional probabilities for $Y = 1$. Over the set of left hand clauses, two examples of variable sets are circled that both satisfy the full formula. Both sets $\{z_7, z_{11}, z_{13}, z_{14}\}$ (in red) and $\{z_3, z_{13}, z_{14}\}$ (in blue) are such that at most one of the variables is present in each clause, such that the remaining variables must be present in the corresponding right hand clause.

Now, this relationship between the clauses within a distribution may be exploited in order to generate subsets of variables that satisfy the CNF formula. Specifically, for any collection of exactly one clause per configuration of endogenous parents $\mathbf{x}$, selecting a set of two or more variables such that no two variables of the set appear in the same clause will satisfy the formula, due to the inverse symmetry across distributions. See Figure 2 for an example, where both sets $\{z_7, z_{11}, z_{13}, z_{14}\}$ and $\{z_3, z_{13}, z_{14}\}$ are identified as CNF-solutions for $n = 2$ over the clauses for $P(Y = 0|X_1 = x_1, X_2 = x_2), \forall x_1 \in \Omega_{X_1}, x_2 \in \Omega_{X_2}$.

Furthermore, this can be done systematically such that distinct partial solutions are generated across all 2^n possible sets of clauses. The approach generates partial CNF-solutions of size m, where $2 \leq m \leq 2^n$, but any choice of additional variables may be included not affecting satisfiability, in order to find complete $2^n + 1$-size solutions. While the approach so far guarantees that all solutions $\mathcal{I}$ searched are such that $\{z_i\}_{i\in\mathcal{I}}$ solves the CNF-formula, the corresponding unique solution to (4) will most of the time not correspond to a probability distribution. Thus, in order to focus the search towards the probability simplex, the approach may consider only some of the 2^n subsets of clauses to build the partial solutions across. Specifically, clauses may be selected according to their corresponding probability: For each distribution, choose the clause of lowest probability, and for this set of lowest probability clauses, build partial solutions. An example is shown in Figure 3.

This approach biases the search towards solutions that will have more non-zero components in equations that sum to probabilities > 0.5, and fewer non-zero components in equations that sum to < 0.5. Expanding the partial solutions to $2^n + 1$ size could similarly be approached by favouring variables that appear most often in higher probability clauses.

CNF disjunction clauses	$P(Y\|X_1, X_2)$
$z_0 \vee z_1 \vee z_2 \vee z_3 \vee z_4 \vee z_5 \vee z_6 \vee z_7$ $z_8 \vee z_9 \vee z_{10} \vee z_{11} \vee z_{12} \vee z_{13} \vee z_{14} \vee z_{15}$	$P(Y=0\|X_1=0, X_2=0)=0.95$ $P(Y=1\|X_1=0, X_2=0)=0.05$
$z_0 \vee z_1 \vee z_2 \vee z_3 \vee z_8 \vee z_9 \vee z_{10} \vee z_{11}$ $z_4 \vee z_5 \vee z_6 \vee z_7 \vee z_{12} \vee z_{13} \vee z_{14} \vee z_{15}$	$P(Y=0\|X_1=0, X_2=1)=0.78$ $P(Y=1\|X_1=0, X_2=1)=0.22$
$z_0 \vee z_1 \vee z_4 \vee z_5 \vee z_8 \vee z_9 \vee z_{12} \vee z_{13}$ $z_2 \vee z_3 \vee z_6 \vee z_7 \vee z_{10} \vee z_{11} \vee z_{14} \vee z_{15}$	$P(Y=0\|X_1=1, X_2=0)=0.07$ $P(Y=1\|X_1=1, X_2=0)=0.93$
$z_0 \vee z_2 \vee z_4 \vee z_6 \vee z_8 \vee z_{10} \vee z_{12} \vee z_{14}$ $z_1 \vee z_3 \vee z_5 \vee z_7 \vee z_9 \vee z_{11} \vee z_{13} \vee z_{15}$	$P(Y=0\|X_1=1, X_2=1)=0.39$ $P(Y=1\|X_1=1, X_2=1)=0.61$

Figure 3: The CNF formula disjunction clauses sorted by probability distributions. Clauses to the left correspond to lowest probabilities. Choosing partial subsets over these clauses bias the selected variable set such that variables appear more often in high-probability clauses. Circled in red is the size 4 partial CNF-solution $\mathcal{Z}_1 = \{z_1, z_2, z_7, z_{11}\}$ and circled in blue is the size 3 partial CNF-solution $\mathcal{Z}_2 = \{z_1, z_2, z_{15}\}$. Indeed, for each clause on the right hand side corresponding to higher probabilities, $|\mathcal{Z}_1| - 1 = 3$ of the variables in set $\mathcal{Z}_1$ appear, as does $|\mathcal{Z}_2| - 1 = 2$ of the variables in set $\mathcal{Z}_2$.

Finally, note the special case in which $P(Y = y_k|\mathbf{X} = \mathbf{x}_j) = 0$ for one or more pairs of values $(y_k, \mathbf{x}_j) \in \Omega_Y \times \Omega_{\mathbf{X}}$. If a conditional probability equals 0, it forces every component of the corresponding sum $\sum_{u \in \Omega_U^{k,j}} P(u)$ to be 0, and by such reduces the number of equations in the system by 1. Thus solution sets $\mathcal{I}$ are now to be of size 2^n. This also affects the CNF formula in the following way: The disjunction clause corresponding to the equation in question becomes negated, ensuring no variable part of that clause can evaluate to True while solving the formula. This similarly extends if more than one conditional equals 0, reducing both the number of equations and the size of the set of variables to chose from.

In order for the heuristics-based search to deal with such special cases most efficiently, for any equation $\sum_{u \in \Omega_U^{k,j}} P(u) = 0$ part of the system, the solutions searched are restricted to contain exactly one variable z_i for which $u_i \in \Omega_U^{k,j}$. Then, $P(u_i)$ is solved to be 0 as part of a now still unique solution to the $2^n + 1$ equations of the linear system.

5. Experiments

For validation, we consider a benchmark of 380 randomly generated Markovian SCMs with all the endogenous variables being binary. All the models have the inverted tree topology discussed in Section 4.2, where a variable Y has a set of independent parents $\mathbf{X}$. Among all the models, 41 have three parents, whereas the rest have two. The SEs are randomly selected to be either canonical or non-canonical. After randomly initializing the exogenous distributions, an M-compatible dataset of 1000 instances is sampled. For each model, different queries are considered, specifically PS and PN with Y as the effect and each of the parents in $\mathbf{X}$ as the cause. The benchmark and Python code for replicating the experiments is available in a dedicated GitHub repository[1].

1. https://github.com/PGM-Lab/2024-PGM-DCCC

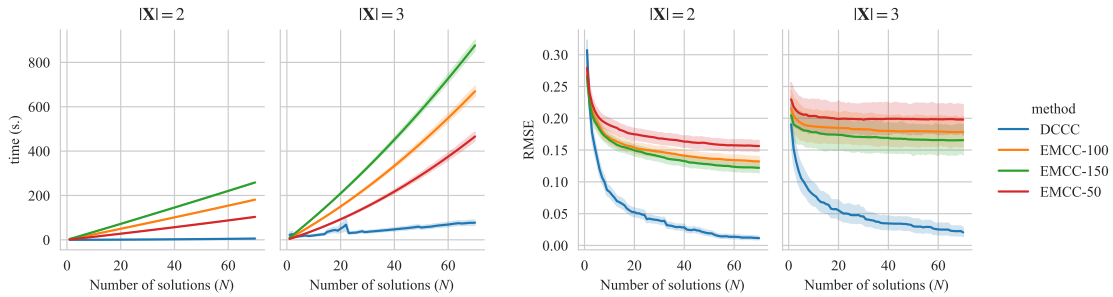


Figure 4: Average computation time (left) and error with respect the exact bounds (right).

The results of the experimentation are summarized in Figure 4, where our method (DCCC) is compared against EMCC with a fixed number of iterations of 50, 100, and 150. The two subfigures on the left show the average computation time (i.e., learning and inference times) for an increasing number of generated solutions (i.e, parameter N from Algorithm 1). For the two-parent case, DCCC is applied exhaustively, whereas a heuristic search is used for models with three parents. In all cases, DCCC is the most efficient approach. The two subfigures on the right depict the RMSE (*root mean squared error*) with respect to the exact bounds computed using the exact method proposed by Zaffalon et al. (2020). DCCC achieves the lowest error levels for a given number of visited solutions. Although EMCC could potentially reach a similar error with a large number of iterations, this would be extremely time-consuming.

6. Conclusions and future work

In this paper we propose a method for bounding unidentifiable queries in SCMs using a divide and conquer strategy to transform a general causal model into a set of models with low-cardinality exogenous variables, in which we can calculate any query using standard Bayesian network inference. Bounds for the query in the original model are then efficiently approximated by aggregating the results from these smaller models. The experiments show that the proposed method is more efficient than current state of the art. We envision at least three directions for future research: Firstly, the theoretical properties of the algorithm need further investigation. Secondly, we want to extend the applicability of the approach to more general model classes, like non-Markovian SCMs. Finally, we want to investigate other ideas for generating efficient heuristics for which reductions to prioritize (for instance, make the selection of reductions informed by the query being calculated).

Acknowledgements

Grant PID2022-139293NB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe. R.C. acknowledges the support by Spanish Ministry of Science, Innovation and Universities through the "María Zambrano" grant (RR_C_2021_01) funded with NextGenerationEU funds. AJB and HL acknowledge the support by the Norwegian Research Council through grant 304843.

References

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 507–556. ACM, 2022.

R. Cabañas, A. D. Maldonado, M. Morales, P. A. Aguilera, and A. Salmerón. Counterfactual reasoning with probabilistic graphical models for analyzing socioecological systems. *arXiv preprint arXiv:2401.10101*, 2024.

F. G. Cozman. Credal networks. *Artificial intelligence*, 120(2):199–233, 2000.

C. Kang and J. Tian. Inequality constraints in causal models with hidden variables. In R. Dechter and T. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, page 233–240. AUAI Press, 2006.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

S. Mueller, A. Li, and J. Pearl. Causes of effects: Learning individual responses from population data. *arXiv preprint arXiv:2104.13730*, 2021.

J. Pearl. *Causality. Models, inference and reasoning. Second edition*. Cambridge University Press, New York, 2009.

M. C. Sachs, E. E. Gabriel, A. Sölander, and E. E. Gabriel. Symbolic computation of tight causal bounds. *Journal of Computational and Graphical Statistics*, 32(2):567–576, 2023.

M. Zaffalon, A. Antonucci, and R. Cabañas. Structural causal models are (solvable by) credal networks. In *International Conference on Probabilistic Graphical Models*, pages 581–592. PMLR, 2020.

M. Zaffalon, A. Antonucci, R. Cabañas, D. Huber, and D. Azzimonti. Efficient computation of counterfactual bounds. *International Journal of Approximate Reasoning*, page 109111, 2024.

J. Zhang, J. Tian, and E. Bareinboim. Partial counterfactual identification from observational and experimental data. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *Proceedings of the Thirty-Ninth International Conference on Machine Learning*, volume 162 of *ICML'22*, pages 26548–26558. JMLR.org, 2022.