Tractable Bounding of Counterfactual Queries by Knowledge Compilation

David Huber[1] **Yizuo Chen**[2] **Alessandro Antonucci**[1] **Adnan Darwiche**[2] **Marco Zaffalon**[1]

[1]IDSIA, Lugano, Switzerland
[2]UCLA, Los Angeles, US

Abstract

We discuss the problem of bounding partially identifiable queries, such as counterfactuals, in Pearlian structural causal models. A recently proposed iterated EM scheme yields an inner approximation of those bounds by sampling the initialisation parameters. Such a method requires multiple (Bayesian network) queries over models sharing the same structural equations and topology, but different exogenous probabilities. This setup makes a compilation of the underlying model to an *arithmetic circuit* advantageous, thus inducing a sizeable inferential speed-up. We show how a single *symbolic* knowledge compilation allows us to obtain the circuit structure with symbolic parameters to be replaced by their actual values when computing the different queries. We also discuss parallelisation techniques to further speed up the bound computation. Experiments against standard Bayesian network inference show clear computational advantages with up to an order of magnitude of speed-up.

1 INTRODUCTION

Causal inference is an important direction for modern AI. Following Pearl's *ladder of causation* [Bareinboim et al., 2022], observational data are sufficient to compute correlational queries, while answering interventional queries requires an additional structure such as the causal graph and dedicated computational schemes such as the popular *do calculus* [Pearl, 2009]. Moving further into counterfactual inference requires the full specification of the underlying causal model, including the structural equations and the exogenous parameters. While the equations might be available (or sampled), the exogenous parameters are typically latent and unavailable. Most counterfactuals are therefore *partially identifiable* and only bounds are obtained for the corresponding queries [Shpitser and Pearl, 2007].

Despite the hardness of the task [Zaffalon et al., 2021], approximate bounding schemes exist. These include polynomial programming [Duarte et al., 2021], credal networks inference [Zaffalon et al., 2020], sampling [Zhang et al., 2022], and EM [Zaffalon et al., 2021]. The latter, in particular, allows us to derive credible intervals while reducing the bounds' computation to iterated (Bayesian network) inferences in a fully specified structural causal model. Such a method requires multiple queries over models sharing the same structural equations but different exogenous probabilities.

Tractable *arithmetic circuits* (e.g., Darwiche [2022b]) offer a graphical formalism to represent generative probabilistic models and compute standard inferential tasks in linear time by a circuit traversal. The ACE library[1] allows Bayesian network compilation to arithmetic circuits with state-of-the-art performances [Agrawal et al., 2021].

The goal of this paper is to adopt the above compilation strategy to achieve a sizeable inferential speed-up in the computation of bounds for counterfactual queries. In particular, we consider a *symbolic* knowledge compilation as in Darwiche [2022a] to obtain the circuit structure with the symbolic parameters to be replaced by their actual values when computing the different queries (Sect. 3). We also present parallelisation techniques to further speed up the bound computation. Experiments based on ACE against standard Bayesian network algorithms report computational speed-ups up to an order of magnitude (Sect. 4). This contribution appears to be the first application of knowledge compilation to counterfactual inference. A discussion on the outlooks of these strategies is in Sect. 5.

[1]`http://reasoning.cs.ucla.edu/ace`.

Accepted for the 6th Workshop on Tractable Probabilistic Modeling at UAI (TPM 2023).

2 NOTATION AND BASICS

Variable X takes values from a finite set Ω_X, θ_X is a probability mass function (PMF) over X, θ_x denote the probability of $X = x$, and λ_x the indicator function of that event.

Bayesian Networks (BNs) Given variables Y and X, a conditional probability table (CPT) $\theta_{Y|X}$ is a collection of PMFs over Y indexed by the values of X. Given a joint variable $\boldsymbol{X} := (X_1, \ldots, X_n)$ and a directed acyclic graph $\mathcal{G}$ with nodes in a one-to-one correspondence with the variables in $\boldsymbol{X}$, a BN is a collection of CPTs $\boldsymbol{\theta} := \{\theta_{X_i|\mathrm{Pa}_{X_i}}\}_{i=1}^n$, where Pa_{X_i} denotes the *parents* of X_i according to $\mathcal{G}$ (see, e.g., Fig. 1). A BN induces a PMF $\theta_{\boldsymbol{X}}$ s.t. $\theta_{\boldsymbol{x}} = \prod_{i=1}^n \theta_{x_i|\mathrm{pa}_{X_i}}$, for each $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$.

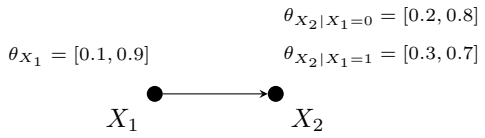

Figure 1: A BN over two Boolean variables.

Arithmetic Circuits (ACs) We can express the joint PMF of a BN as a multi-linear function of the CPT parameters, i.e., $\theta_{\boldsymbol{x}} = \sum_{x'_1,\ldots,x'_n} \prod_i \theta_{x'_i|\mathrm{pa}_{X_i}} \lambda_{x_i}$. Such an exponential-size representation becomes more compact by exploiting the BN conditional independence relations induced by $\mathcal{G}$ and consequently moving the sums inside the products. The representation might be even more compact if different CPT parameters take the same value. Common examples are context-specific independence relations and CPTs implement deterministic relations through *degenerate* (i.e., 0/1) values only. Such functions are graphically depicted as ACs composed by leaves, annotated by CPT probabilities and indicator functions, and inner nodes containing sums and multiplications (e.g., Fig. 2). Those ACs are called *tractable*, as they allow to answer some queries in linear-time, through feed-forward passes on the circuit structure. A number of *compilation* algorithms have been proposed to build compact AC representations of BNs.

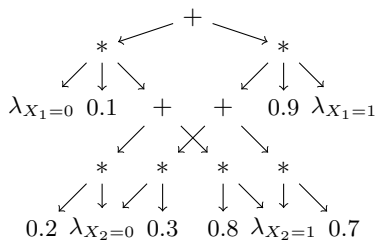

Figure 2: An AC implementing the BN in Fig. 1.

Structural Causal Models (SCMs) A *structural equation* (SE) f associated with variable Y and based on the input variable(s) X, is a surjective function $f : \Omega_X \to \Omega_Y$ that determines the value of Y from that of X. Given two joint variables $\boldsymbol{U}$ and $\boldsymbol{V}$, called respectively *exogenous* and *endogenous*, a collection of SEs $\{f_V\}_{V \in \boldsymbol{V}}$ such that, for each $V \in \boldsymbol{V}$ the input variables of f_V are in $(\boldsymbol{U}, \boldsymbol{V})$, is called a *partially specified* SCM (PSCM). A PSCM induces a directed graph $\mathcal{G}$ with nodes in correspondence with the variables in $(\boldsymbol{U}, \boldsymbol{V})$ and such that there is an arc between two variables if and only if the first variable is an input variable for the SE of the second (e.g., Fig. 3). We focus on *semi-Markovian* PSCMs, i.e., those PSCMs that lead to acyclic graphs. A *fully specified* SCM (FSCM) is just a PSCM M paired with a collection of marginal PMFs, one for each exogenous variable. As SEs induce (degenerate) CPTs, an FSCM defines a BN over $(\boldsymbol{U}, \boldsymbol{V})$ based on $\mathcal{G}$.

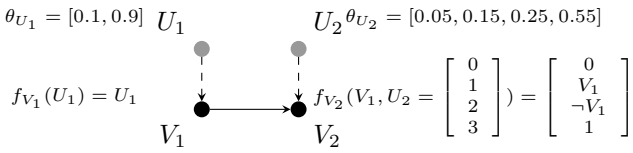

Figure 3: A FSCM over two endogenous (black) and two exogenous variables (grey nodes).

Causal Queries in FSCMs BN algorithms allow to compute inferences in FSCMs. This is trivially the case for observational queries involving joint or conditional states of the endogenous variables. For interventional queries, this can be also done provided that the SEs of the intervened variables are replaced by constant maps pointing to the selected state. For counterfactual queries, where the same variable may be observed as well as subject to intervention, albeit in distinct *worlds*, we use auxiliary structures where different copies of the endogenous variables and their SEs are considered in each world. Han et al. [2023] provides a precise characterisation of the computational complexity of those inferences in terms of treewidth.

Partially Identifiable Causal Queries in PSCMs FSCMs are rarely available. Considering a PSCM specification with a dataset $\mathcal{D}$ of endogenous observations represents a more common setup. This is not critical for observational queries: a BN over the endogenous variables can be obtained by deriving its graph from that of the PSCM and the CPTs from $\mathcal{D}$ [Tian, 2002]. Interventional queries can be possibly reduced to observational queries by the *do calculus* [Pearl, 2009]. If this is not possible, we say that the query is only *partially identifiable*. In those cases, a characterisation is still provided by the bounds spanned by the values of the query computed for all the FSCMs consistent with the PSCM and the endogenous BN (e.g., Zaffalon et al. [2020]). Counterfactual queries are very often only partially identifiable.

3 TRACTABLE BOUNDING OF COUNTERFACTUALS

Bounding partially identifiable queries PSCMs is an NP-hard problem even on polytrees [Zaffalon et al., 2021, Theorem 2].

Zhang et al. [2022] have proposed a Bayesian sampling procedure that eventually approximates the bounds via credible intervals. The sampling is query-driven; new queries will require new sampling. The accuracy of the approximation is unclear in general as a systematic experimental analysis is missing.

EM Approach The algorithm proposed by Zaffalon et al. [2021] samples the initialisation of the exogenous chances, which are used to start an EM scheme returning a compatible FSCM specification. Alg. 1 depicts a single EM run. The interval spanned by the values of the query computed on the FSCMs returned by the EM for each run provides an (inner) approximation of the expectation bounds. This approach is 'agnostic' w.r.t. the query. It aims at reconstructing the uncertainty related to exogenous variables (via sets of probabilities). Once this is done, different (counterfactual) queries will use the same sets of probabilities to compute the wanted bounds—no more sampling is needed.

Algorithm 1 In a PSCM paired with an endogenous dataset $\mathcal{D}$, given a random initialisation $\{\theta_U^{(0)}\}_{U\in\boldsymbol{U}}$ in input, the algorithm returns the exogenous chances $\{\theta_U\}_{U\in\boldsymbol{U}}$ obtained after likelihood convergence.

1: $t \leftarrow 0$
2: **while** $P(\mathcal{D}|\{\theta_U^{t+1}\}_{U\in\boldsymbol{U}}) \geq P(\mathcal{D}|\{\theta_U^{t}\}_{U\in\boldsymbol{U}})$ **do**
3: **for** $U \in \boldsymbol{U}$ **do**
4: $\theta_U^{t+1} \leftarrow |\mathcal{D}|^{-1} \sum_{\boldsymbol{v}\in\mathcal{D}} \theta_{U|\boldsymbol{v}}^{t}$
5: $t \leftarrow t+1$
6: **end for**
7: **end while**

An approximate bounding scheme based on Alg. 1 may suffer from two potential bottlenecks: (i) an insufficient number of runs leading to a poor inner bound approximation; (ii) the time needed by the FSCM inferences required by the exogenous queries (line 4) and the likelihood evaluation (line 2).

Regarding (i), Zaffalon et al. [2022] derived a characterisation of the accuracy of the bounds in terms of credible intervals, and the EM scheme has been proven to yield accurate bounds with relatively few runs.

Here we address instead (ii) by first noticing that the queries needed by Alg. 1 are computed on different FSCMs based on the same PSCM, thus having possibly different exogenous chances, but always the same endogenous CPTs implementing the SEs of the PSCM. This is true for the models corresponding to different time steps t, but also when different exogenous initialisations are considered in input. In practice the algorithm requires the computation of inferences in different BNs having the same CPTs for the non-root nodes, but different marginal PMFs on the root nodes. This simple remark suggests the use of AC compilation to achieve faster inferences.

Symbolic Knowledge Compilation Consider the AC compilation of two BNs over the same variables and with the same graph but different CPT parameters. Suppose these parameters, separately for each BN, have no repeated values. In that case, the compiler minimises the size of the ACs by only exploiting the independence relations induced by the BN graph. As these are the same for the two BNs, the two ACs returned by the compiler should share the same inner nodes and the same indicators on the leaves while differing only on the chances in the leaves.

This fact allows for a *symbolic* compilation achieved by regarding the chances in the leaves as symbolic parameters to be replaced by their actual values during an inferential computation. Compilers can quickly implement symbolic compilation by replacing the BN parameters with unique numerical identifiers to be eventually retrieved in the AC returned by the compiler.

Returning to the queries of interest for the EM scheme, we might intend PSCM compilation as a symbolic compilation achieved by treating the exogenous PMFs as parameters. In contrast, the endogenous CPTs, implementing the SEs and remaining the same for all the models, are treated as constant numerical values. The degenerate nature of the CPTs can be exploited by the compiler to achieve smaller ACs and hence faster inferences (e.g., with the FSCM in Fig. 3 as input, ACE returns an AC with 96 arcs if the determinism of the CPTs is not exploited and 23 arcs otherwise). After the PSCM symbolic compilation, the AC of each FSCM required by Alg. 1 is obtained in linear time (w.r.t. the AC size) by replacing the parameters of the *symbolic* AC with the actual values in the particular FSCM.

The queries required by Alg. 1 are, for each $\boldsymbol{v} \in \mathcal{D}$, the computation of endogenous marginal $\theta_{\boldsymbol{v}}$ and the exogenous posterior $\theta_{u|\boldsymbol{v}}$, to be computed for each $U \in \boldsymbol{U}$ and $u \in \Omega_U$. We therefore focus on the computation of the joint query $\theta_{u,\boldsymbol{v}}$ for each $U \in \boldsymbol{U}$, $u \in \Omega_U$ and $\boldsymbol{v} \in \mathcal{D}$. This is performed in linear time by a bottom-up traversal of the AC after instantiating the indicators of the variables in $\boldsymbol{V}$ and U.

Parallelisation Alg. 1 allows for a straightforward parallelisation at the run level. A more sophisticated parallelisation can be based on *c-components* [Tian, 2002]. In a PSCM, a c-component is a set of variables connected through undirected paths consisting solely of exogenous-to-endogenous arcs. For each c-component, we define a sub-graph consisting of the nodes in the c-component and its direct parents,

with all other variables and edges removed. The corresponding sub-model might yield the chances of the exogenous variables in the c-component through Alg. 1. The procedure can be executed in parallel, separately for each c-component.

4 EXPERIMENTS

To evaluate the benefits of the proposed AC approach when running the EM scheme in Alg. 1, we compare the AC execution times against those based on standard BN inference for a synthetic benchmark of 335 PSCMs. The PSCM graphs have a random topology (Erdös-Rényi sampling), the number of nodes ranges between 5 and 21 (avg. 9.9), and the number of root nodes (i.e., exogenous variables) between 2 and 10 (avg. 4.5). All the endogenous variables are binary, while the cardinalities of the exogenous ones range between 3 and 256 (avg. 29.6). Each PSCM comes with a dataset of endogenous observations of size between $1,000$ and $5,000$ records obtained by sampling a compatible FSCM. The benchmark and the code used for the simulations are available in a dedicated repository.[2]

The code is built on the top of CREDICI[3] [Cabañas et al., 2020], a Java library implementing the EM scheme and embedding a BN inference engine. Here we consider inferences based on variable elimination with the min-fill heuristics. The symbolic compilation is instead developed within the Java/C++ ACE compiler (see Footnote 1). The experiments are run on a dual 2.20GHz Intel(R) Xeon(R) Silver 4214 CPU Dell PowerEdge R540 server running Ubuntu 20.04.6 LTS. All the experiments are performed using a fixed seed for the random initialisation and, as expected, resulted in the exact same set of PSCMs.

For each PSCM, we perform 200 runs of 500 iterations. We set a timeout to 15 minutes for each experiment. The BN approach based on the whole model often reaches this limit. Thus, as a baseline for the BN approach, we consider the faster BNC approach based on queries in the sub-BNs associated with the model c-components. The number of c-components for the benchmark models ranges between 1 and 10 (avg. 4.2). The parallelisation of BNC over the different components is denoted instead as BNP. We similarly denote as ACC the method based on the (symbolic) compilation of the sub-BNs and as ACP its parallelisation. The overall execution times (in hours) on the whole benchmark for the four methods are $T_{\mathrm{BNC}} = 17.0$, $T_{\mathrm{BNP}} = 7.3$, $T_{\mathrm{ACC}} = 2.4$, and $T_{\mathrm{ACP}} = 1.3$. This clearly shows the advantage of the (symbolic) knowledge compilation.

A deeper analysis is provided by computing, separately for each PSCM, the ratio between the EM execution time of a particular approach and that of BNC. Fig. 4 shows the boxplots of the different approaches. In practice, using ACs makes the bounding of the counterfactual queries one order of magnitude faster. Note also that we considered PSCM of bounded size (≤ 21 nodes) just to permit a comparison against the BN approaches, which cannot handle bigger networks in reasonable time limits.

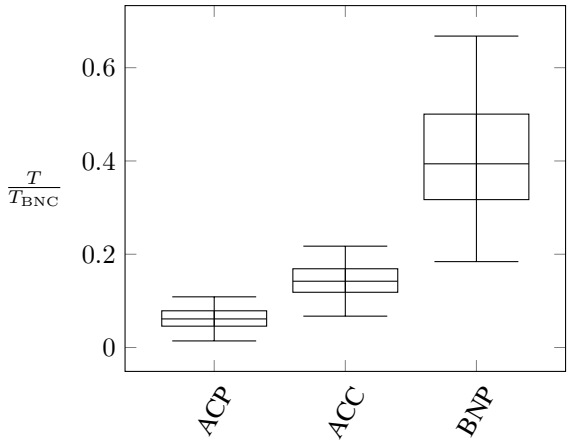

Figure 4: Runtime savings w.r.t. BNC.

5 CONCLUSIONS

In this study, we have investigated the potential of knowledge compilation within the framework of partially identifiable queries, such as counterfactuals, in structural causal models. We have assumed that structural equations are given together with a dataset of endogenous observations. From these we reconstruct the uncertainty about the exogenous variables with sets of probabilities.

The advantages of using knowledge compilation appear clear: the new approach leads to one order of magnitude speed-up compared to pre-existing models based on Bayesian nets.

As future work we intend to use the knowledge compilation approach to execute the EM scheme in very large models along two dimensions: the size of the network as well as the cardinality of exogenous variables. The latter is in particular an important factor to represent general 'canonical' specifications of PSCMs. These specifications enable one to be dispensed of the requisite to provide structural equations in input: a causal graph with endogenous data would suffice to compute counterfactual inference.

We also intend to explore more in-depth problems with network structures with large treewidth that may thus be intractable by variable elimination.

References

D. Agrawal, Y. Pote, and K. S. Meel. Partition function estimation: A quantitative study. *arXiv preprint*

[2] `anonymous.4open.science/r/uai-E5D7`.

[3] `github.com/idsia/credici`.

arXiv:2105.11132, 2021.

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, page 507–556. ACM, New York, US, 2022.

R. Cabañas, A. Antonucci, D. Huber, and M. Zaffalon. CREDICI: a Java library for causal inference by credal networks. In *Proceedings of PGM*, 2020.

A. Darwiche. Causal inference using tractable circuits. *arXiv preprint arXiv:2202.02891*, 2022a.

A. Darwiche. Tractable Boolean and arithmetic circuits. *arXiv preprint arXiv:2202.02942*, 2022b.

G. Duarte, Finkelsteinm N., D. Knox, J. Mummolo, and I. Shpitser. An automated approach to causal inference in discrete settings. *arXiv preprint arXiv:2109.13471*, 2021.

Y. Han, Y. Chen, and A. Darwiche. On the complexity of counterfactual reasoning. *arXiv preprint arXiv:2211.13447*, 2023.

J. Pearl. *Causality*. Cambridge University Press, 2009.

I. Shpitser and J. Pearl. What counterfactuals can be tested. In *Proceedings of UAI*, 2007.

J. Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, UCLA, 2002.

M. Zaffalon, A. Antonucci, and R. Cabañas. Structural causal models are (solvable by) credal networks. In *Proceedings of PGM*, 2020.

M. Zaffalon, A. Antonucci, and R. Cabañas. Causal Expectation-Maximisation. In *WHY-21 @ NeurIPS*, 2021.

M. Zaffalon, A. Antonucci, R. Cabañas, D. Huber, and D. Azzimonti. Bounding counterfactuals under selection bias. In *Proceedings of PGM*, 2022.

J. Zhang, J. Tian, and E. Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of ICML*, 2022.