Stability of causal inference

Leonard J. Schulman Piyush Srivastava

Abstract

We consider the sensitivity of causal identification to small perturbations in the input. A long line of work culminating in papers by Shpitser and Pearl (2006) and Huang and Valtorta (2008) led to a complete procedure for the causal identification problem. In our main result in this paper, we show that the identification function computed by these procedures is in some cases extremely unstable numerically. Specifically, the "condition number" of causal identification can be of the order of $\Omega(\exp(n^{0.49}))$ on an identifiable semi-Markovian model with n visible nodes. That is, in order to give an output accurate to d bits, the empirical probabilities of the observable events need to be obtained to accuracy $d + \Omega(n^{0.49})$ bits.

1 Introduction

The gold standard for estimating the causal effect of one part of a system on another is the *controlled experiment*: the experimenter controls, or *intervenes* with, the *stimulus* variables in a way such that they are not affected by any non-measurable confounding factors, and then observes the distribution of the *response* variables as the stimuli are varied. Unfortunately, in a variety of important applications, the controlled experiment is not available as a method for reasons of ethics or practicality: a popular example of such a scenario is the question of whether a lifestyle choice such as smoking causes lung cancer. It can be argued (and in this particular case, has been argued (Ohlemeyer, 1999)!) that the strong observed correlations may be due to hidden confounding factors (here environmental or genetic).

In a series of seminal papers starting with (Pearl, 1995), Judea Pearl and others proposed and analyzed a framework for computing causal effects of hypothetical interventions solely from passively observed (i.e., non-experimental) data. The starting point of this framework is a model of the system as a directed graphical model with hidden nodes representing the non-measurable confounding variables. The goal is to take as input the joint distribution of the observed nodes in the model, and to deduce from them the *intervention distributions* that would result if an hypothetical controlled experiment were to be performed. A long line of work (Pearl, 1995; Pearl and Robins, 1995; Kuroki and Miyakawa, 1999; Halpern, 2000; Tian, 2002) on this framework culminated in papers by Shpitser and Pearl (2006) and Huang and Valtorta (2008) which gave a complete characterization of models in which this is achievable: in particular, they provided an algorithm which on input a directed graphical model and the set of stimulus and response variables outputs *either* a procedure that will compute the intervention distribution given the joint distributions of the observed nodes, or a certificate that the intervention distribution is not determined uniquely by the observed joint distribution. In the former case, the causal effect of the stimuli upon the response variables is said to be *identifiable* in the model.

This paper is concerned with the numerical properties of the identification problem. Note that an inference process such as causal identification as described above will always run on empirical inputs. We therefore ask: when the causal effect is identifiable, how sensitive is it to small inaccuracies either in the knowledge of the model, or of the observed distribution? Our main result (Theorem 1.2) in fact shows that causal inference can in fact be extremely sensitive to small errors: we give example of models on n nodes where *any* numerical algorithm for computing the intervention distribution from the observed distribution will lose roughly $\Theta(\sqrt{n})$ bits of precision. This sets an extraordinary demand on the accuracy of the input data of such a system.

As we discuss in more detail in Section 1.2, there are several natural sources of errors in the input to a causal identification problem: these include errors incurred in measurements of the observed distribution; round-off; and as we observe in Section 1.2.1, inexact descriptions of models. Our results therefore point to a new line of investigation

concerning the classification of graphical models based on the sensitivity of causal identification to such perturbations in the input.

We begin in the next subsection by formalizing first the identification question, then the appropriate notion of stability for causal identification.

1.1 Pearl's notion of causal identifiability

In Pearl's framework, the system being studied is modeled as a *semi-Markovian* graphical model. A semi-Markovian graphical model is a directed acyclic graph $G = (V, E, U, H)$, which has observed nodes and edges V and E, and hidden nodes and edges U and H, and is constrained so that the observed edges E lie between the observed vertices V, while the hidden edges in H go from a hidden vertex in U to an observed vertex in V.

The observed nodes V of the model are identified with the measurable components of the system, while the hidden nodes in G represent confounding variables that are not accessible to measurement. The edges model dependencies between these random variables: every variable is independent of its ancestors in G conditioned on its immediate predecessors. A probability distribution that satisfies this constraint is said to *respect* G. Equivalently, a probability distribution $\mathbb{P}$ respects G if it factorizes as

$$\mathbb{P}\left(V_1 = v_1, \ldots, V_n = v_n, U = u\right) = \mathbb{P}\left(U = u\right) \prod_{V_i \in V} \mathbb{P}\left(V_i = v_i \mid \mathrm{pa}(V_i) = v_{\mathrm{pa}(V_i)}\right),$$

where $\mathrm{pa}(S)$ is the set of parents of the node S.

However, since the hidden nodes in U are not measurable, any measurement can only estimate the *observed marginal*

$$\mathbb{P}\left(V = v\right) = \sum_{u \in \Omega(U)} \mathbb{P}\left(U = u\right) \prod_{V_i \in V} \mathbb{P}\left(V_i = v_i \mid \mathrm{pa}(V_i) = v_{\mathrm{pa}(V_i)}\right),$$

where $\Omega(U)$ denotes the range of the set U of hidden variables.

Pearl (1995) proposed that a natural representation of an experimental intervention on some subset X of observed variables is to remove from them any effect of their ancestors. Formally, the *intervention distribution*, denoted $\mathbb{P}\left(V - X \mid \mathrm{do}(X = x)\right)$, of the nodes in $V - X$ under the intervention $X = x$ can therefore be defined as follows:

$$\mathbb{P}\left(V - X = v_{V-X} | \, \mathrm{do}(X = x)\right) = \sum_{u \in \Omega(U)} \mathbb{P}\left(U = u\right) \prod_{V_i \in V - X} \mathbb{P}\left(V_i = v_i \mid \mathrm{pa}(V_i) = v_{\mathrm{pa}(V_i)}\right). \tag{1}$$

The marginals of the above distribution define the distribution $\mathbb{P}\left(\cdot \mid \mathrm{do}(X = x)\right)$ on all subsets of $V - X$.

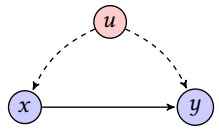
Directly computing the intervention distribution using eq. (1) requires knowledge of the distributions of the hidden variables, as well as their effect on the observed nodes. These are, of course, not measurable in practice. This leads to the question: when is the intervention distribution in eq. (1) efficiently computable (or *identifiable*) from a knowledge of only the observed statistics? More formally, given a semi-Markovian graph $G = (V, E, U, H)$ and disjoint subsets X, Y of V, $\mathbb{P}\left(Y \mid \mathrm{do}\, X\right)$ is said to be *identifiable* in G if and only if there exists a function

$$\mathbf{ID}(G, X, Y) : P(V) \mapsto P(Y \mid \mathrm{do}\, X)$$

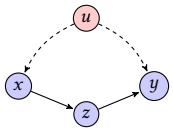
which maps observed distributions $P(V)$ to intervention distributions $P(Y \mid \mathrm{do}\, X)$. The question then is to decide, given G, X, and Y, whether such a map exists, and if yes, to compute it.

As expected, the answer to the question is not always positive. For example, in the graph in fig. 1a (where u is a hidden node), it is not possible to express $\mathbb{P}\left(y | \, \mathrm{do}(x)\right)$ in terms of the marginal distribution $\mathbb{P}\left(x, y\right)$. This is intuitive, since any observed correlation between X and Y is equally well attributable as being due to the hidden variable U as due to a causal effect of X of Y. However, in the similar model in fig. 1b, which has just one more observed node, the distribution $\mathbb{P}\left(y | \, \mathrm{do}(x)\right)$ *is* identifiable.

The algorithms of Shpitser and Pearl (2006) and Huang and Valtorta (2008), on input G, X, and Y as in the above discussion, either output a description of the map $\mathbf{ID}(G, X, Y)$, or give a certificate that the causal effect of X on Y is not identifiable in G. Here we study the stability of the map $\mathbf{ID}(G, X, Y)$ when it exists; we also show how identification can be applied when the map "almost" exists.



(a) A Simple Unidentifiable Case



(b) A Simple Identifiable Case

Figure 1: Simple Graphical Models

1.2 Results

Before embarking on our study of the sensitivity of the map $\mathbf{ID}(G, X, Y)$ to errors in the input, we make a few comments about the sources of such errors. Conventionally, two such sources are considered: errors introduced due to limitations in measuring the input, and errors introduced due to rounding off the input to a fixed finite floating point precision. These sources of errors are fairly generic and apply to almost any function, hence we defer their discussion in the context of the **ID** map to Remark 1.1. Here, we discuss another kind of error in the input specific to the problem of causal inference: error arising from inaccuracies in the knowledge of the graphical model of the system under study. We start by analyzing such errors in Section 1.2.1. Finally, in Section 1.2.2, we formalize the notion of the *condition number* which captures all the three forms of errors described above and then state our results in terms of this notion.

1.2.1 Errors in the Model Description

We now consider the effect of ignoring some edges in the input graphical model. Let $G = (V, E, U, H)$ be a semi-Markovian graph with observed nodes and edges V and E, and hidden nodes and edges U and H respectively. Let X, Y be disjoint subsets of V, and suppose that $\mathbb{P}(Y \mid \mathrm{do}\, X)$ is not identifiable in G, but identifiable in the subgraph $G' = (V, E, U, H - \{e\})$ in which a certain edge e has been removed. Another way to frame the situation is that we start with the model G' in which the requisite intervention is identifiable, and then consider the effect of adding the edge e to the model which destroys identifiability.

In particular, we wish to quantify the non-identifiability induced by the addition of the edge e to G', as a function of some measure of the "strength" of the edge e. A natural measure of the strength of an edge (A, B) is the amount by which it can affect the conditional probability at its child vertex B, when all the other parents of B are held fixed. More formally, we propose the following measure of strength:

Definition 1.1 (ϵ**-weak edge**)**.** Let $e = (A, B)$ be any edge in a semi-Markovian graph $G = (V, E, U, H)$ and let P be a model respecting G. Let $\Xi(B)$ denote the set of parents of B in $(V \cup U) \setminus \{A\}$. We say that e is *ϵ-weak* with respect to G and P if for every setting b of B and ξ of $\Xi(B)$, and any two values a and a' in the range of A, we have

$$-\epsilon \leq \log \frac{P(B = b \mid \Xi(B) = \xi, A = a)}{P(B = b \mid \Xi(B) = \xi, A = a')} \leq \epsilon.$$

Now suppose that $e = (A, B) \in E \cup H$ is an ϵ-weak edge in $G = (V, E, U, H)$. We ask: given an observed distribution P, what is the error incurred if we perform causal inference in G' instead of G? We answer this in:

Proposition 1.1. *Consider a semi-Markovian graph $G = (V, E, U, H)$ and a distribution $P(V, U)$ respecting it. Let $R = \{e_i = (A_i, V_i) \mid 1 \leq i \leq q\}$ be a set of k edges in $E \cup H$ such that e_i is ϵ_i-weak, with $\epsilon := \sum_{i=1}^{k} \epsilon_i$. Suppose that X, Y are disjoint subsets of V for which $\mathbb{P}(Y \mid \mathrm{do}\, X)$ is not identifiable in G, but identifiable in $G' = (V, E \setminus R, U, H \setminus R)$. Then there exists a distribution $\tilde{P}(V, U)$ respecting G' such that*

$$-\epsilon \leq \log \frac{\tilde{P}(V)}{P(V)} \leq \epsilon, \quad \textit{and} \quad -\epsilon \leq \log \frac{\tilde{P}(Y \mid \mathrm{do}\, X)}{P(Y \mid \mathrm{do}\, X)} \leq \epsilon.$$

Note that $\tilde{P}(Y \mid \text{do}\, X)$ is computable (by the algorithms of Shpitser and Pearl (2006) and Huang and Valtorta (2008)) given $\tilde{P}(V)$, but $P(Y \mid \text{do}\, X)$ is not even uniquely determined given only the observed marginal $P(V)$.

The proof of this proposition can be found in Appendix A.

1.2.2 Uncertainty in the Input Distribution

Proposition A.1 shows that there exists a distribution $\tilde{P}$ on the subgraph G' for which the intervention distribution is both close to that of P, and also computable only from the projection of $\tilde{P}$ to the observed nodes. On the other hand, the proposition does not provide a method to produce such a $\tilde{P}$ given the projection of P to the observed variables in G (or G'). However, it does guarantee that the observed marginals of P and $\tilde{P}$ are ϵ-close in the following sense:

Definition 1.2 (ϵ-close distributions). Two probability distributions P and Q on the same domain Ω are said to be ϵ-close, denoted $P \overset{\epsilon}{\sim} Q$, to each other if for every $\omega \in \Omega$,

$$-\epsilon \leq \log \frac{P(\omega)}{Q(\omega)} \leq \epsilon.$$

We therefore want to study the effect of doing causal inference with observed marginals that are only ϵ-close to the actual observed marginal. To make this precise, we first recall that if $G = (V, E, U, H)$ is a semi-Markovian graph and X, Y are disjoint subsets of V, then $\mathbb{P}(Y \mid \text{do}\, X)$ is said to be *identifiable* in G if and only if there exists a map

$$\mathbf{ID}(G, X, Y) : P(V) \mapsto P(Y \mid \text{do}\, X)$$

which maps observed distributions $P(V)$ to intervention distributions $P(Y \mid \text{do}\, X)$. Our *statistical stability* question then is the following:

Question 1. *Suppose G, X, Y are such that the map $\mathbf{ID}(G, X, Y)$ exists. How sensitive is the map $\mathbf{ID}(G, X, Y)$ to uncertainties in the input P?*

The standard solution concept for studying such a question is the notion of the *condition number* of the map (see, e.g., (Bürgisser and Cucker, 2013, *Overture*)). We specialize here to the so-called "componentwise condition number".

Definition 1.3 (Condition number). Let $f : \mathbb{R}^k \to \mathbb{R}^\ell$ be an arbitrary vector valued function. The condition number of f at $a \in \mathbb{R}^k$, denoted $\kappa_f(a)$, is defined as

$$\kappa_f(a) := \lim_{\delta \downarrow 0} \sup_{\substack{a' \in \mathbb{R}^k \\ \mathsf{Rel}(a, a') \leq \delta}} \frac{\mathsf{Rel}(f(a), f(a'))}{\mathsf{Rel}(a, a')},$$

where for real vectors a, a' in the Euclidean space $\mathbb{R}^t$, the *relative error* $\mathsf{Rel}(a, a')$ is defined as $\max_{1 \leq i \leq t} \frac{|a_i - a'_i|}{|a_i|}$. The condition number of f over a domain $\mathcal{D}$, denoted as κ_f when $\mathcal{D}$ is clear from the context, is $\sup_{a \in \mathcal{D}} \kappa_f(a)$.

Note that the condition number is a property of the function f, and not of a particular algorithm for numerically computing f. In particular, $\kappa_f(a)$ can informally be construed as a derivative of the coordinate-wise logarithm of f as a function of the coordinate-wise logarithm of a.

Armed with the definition of the condition number, we can now further refine our earlier Question 1.

Question 2. *What is the condition number of the map $\mathbf{ID}(G, X, Y)$ for a given G and subsets X, Y, provided in the first place that this map exists?*

Remark 1.1. A few remarks about the sources of errors are in order. First, note that Proposition A.1 already provides a natural setting in which the "error model" used in the definition of the condition number is the right one: in the notation of that proposition, if we computed the intervention distributions in G' by applying the map $\mathbf{ID}(G', X, Y)$ to P instead of $\tilde{P}$, the worst case relative error in the output will be lower bounded by roughly $\epsilon \cdot \kappa_{\mathbf{ID}(G', X, Y)}(\tilde{P})$, independent of the algorithm used for computation. However, we point out that there is another—arguably even

more natural—source of errors that is best captured in terms of relative errors: errors introduced due to rounding in fixed precision floating point systems. We refer the reader to, e.g., the textbook by Bürgisser and Cucker (2013, Section O.3) for a formalization of such systems.

The final type of errors that we discuss here are sampling errors arising due to the finiteness of the sampling procedures used to estimate the observed marginals that are fed as input to the map **ID**. These sampling errors are more likely to be additive (as opposed to relative) in nature. When the input coordinates are elements of the interval $[0, 1]$ (as is the case in our application) an additive error of a given magnitude ϵ always corresponds to a relative error that is at least as large as ϵ in magnitude. Hence, upper bounds imposed on the relative error in the output using an upper bound on the condition number can only worsen if the error guarantees on the input are only additive. In particular, if we show that the condition number defined with respect to relative errors is large, the instability of the problem with respect to additive errors in the input also follows.

Our first result regarding statistical stability demonstrates that the condition number can in fact be sub-exponentially large in the size of the model.

Theorem 1.2. *For every $0 < \alpha < 1/2$, there exists an infinite sequence of semi-Markovian graphs $G_N = (V_N, E_N, U_N, H_N)$ with $|V_N| = N$, and disjoint subsets S_N and T_N of V_N such that*

$$\kappa_{\mathbf{ID}(G_N, T_N, S_N)} = \Omega(\exp(N^{\alpha})).$$

The proof of this theorem appears in Section 2. We now isolate one important class of special cases where the condition number is not so bad.

Proposition 1.3. *Let $G = (V, E, U, H)$ be a semi-Markovian graph, and let X be a node in V such that it is not possible to reach a child of X from X using only the edges in H (with their directions ignored). Then, for any subset S of V not containing X.*

$$\kappa_{\mathbf{ID}(G, X, S)} = O(|V|).$$

The hypothesis of the above proposition has appeared earlier as a sufficient condition for the identifiability of $P(S \mid \mathrm{do}\, X)$ in the early work of Tian and Pearl (2002). While this condition is not necessary for identifiability, we show that it carries a distinct advantage: when it holds, the condition number of the identification function is relatively small. The proof of Proposition 1.3 appears in Section 3.

2 Ill-conditioned examples

In this section, we prove Theorem 1.2. We begin with a brief outline of our general strategy.

Our main object of study will be semi-Markovian models $\mathcal{G}_n^k$ indexed by positive integers n and k, such that $\mathcal{G}_n^k$ has $\Theta(nk)$ visible nodes, and $\Theta(n + k)$ hidden nodes. The maximum degree of $\mathcal{G}_n^k$ will be $\Theta(k)$ for the observed nodes, and $\Theta(n)$ for the hidden nodes.

Let U and V denote the hidden nodes and the observed nodes, respectively, of $\mathcal{G}_n^k$. In our construction, the variables in both U and V will be binary valued. The crux of our proof is a construction of two probability distributions: the first of these, Q, will be a distribution on the states of the nodes in $U \cup V$ which respects $\mathcal{G}_n^k$. The second, $\tilde{Q}$, will be a distribution only on the states of V, such that it is ϵ-close to the marginal of Q on V. Q and $\tilde{Q}$ will be designed to ensure that when k is chosen to be an appropriate function of n, the values of a certain intervention distribution on $\mathcal{G}_n^k$ computed according to $\tilde{Q}$ differ from the correct answer (i.e., the one computed according to Q) by a factor of $1 \pm \epsilon'$ where ϵ' is larger then ϵ by a factor $\Omega(\exp(N^{\alpha}))$ (for any $\alpha < 1/2$), for $N = nk + n - 1$, the size of $\mathcal{G}_n^k$.

2.1 The gadget

We now define the semi-Markovian graph $\mathcal{G}_n^k$ formally. The visible nodes V of the graph partition into three classes: the "X" nodes, of which there are $k - 1$, the S nodes, of which there are n, and the "Y" nodes, of which there are $(n - 1)k$, arranged in $n - 1$ "towers" of k each (see fig. 2). Formally, we have

$$V := \{X_i \mid 2 \le i \le k\} \cup \{S_i \mid 1 \le i \le n\} \cup \left\{Y_{i,j} \mid 1 \le i \le n - 1, 1 \le j \le k\right\}.$$

We now describe the visible edges. First, each S node is a child of each of the Y nodes in the tower immediately to its left. Each Y node is a child of the Y node immediately below it in its tower, of the S node immediately to the left of its tower, and, if it is in the leftmost tower, of the X node at the same "level" as itself (see fig. 2). Formally,

$$\begin{aligned} E := &\{(X_i, Y_{1,i}) \mid 2 \le i \le k\} \\ &\cup \left\{(S_i, Y_{i,j}) \mid 1 \le i \le n-1, 1 \le j \le k\right\} \\ &\cup \left\{(Y_{i,j}, S_{i+1}) \mid 1 \le i \le n-1, 1 \le j \le k\right\} \\ &\cup \left\{(Y_{i,j}, Y_{i,j+1}) \mid 1 \le i \le n-1, 1 \le j \le k-1\right\}. \end{aligned}$$

Our final task is to describe the hidden nodes and variables: the structure of these defines the C-components of the model, and hence they will dictate the sequence of operations in the Shpitser-Pearl algorithm for causal identification applied to $\mathcal{G}_n^k$.[1] In order to make sure that the S nodes are always in the same C-component, we stipulate an unnamed hidden variable for each adjacent pair of the S_i, which has both of the elements of the pair in question as its children. In addition, we have further (named) hidden variables $\{U_i \mid 2 \le i \le k\}$, such that U_i has as children all the S nodes, the nodes X_i, and all the X and Y nodes at "levels"strictly below i. Formally, the hidden edges incident on these named hidden nodes are:

$$\begin{aligned} H := &\left\{(U_i, X_j) \mid 2 \le i \le k, 2 \le j \le i\right\} \\ &\cup \left\{(U_i, S_j) \mid 2 \le i \le k, 1 \le j \le n\right\} \\ &\cup \{(U_i, Y_{s,t}) \mid 2 \le i \le k, 1 \le s \le n-1, 1 \le t < i\}. \end{aligned}$$

See fig. 2 for a depiction of the gadget $\mathcal{G}_6^4$. In the figure, the named hidden variables U_i and their edges are not included for clarity. Instead, the hyperedges depicting the hidden variables U_i are depicted by the different shaded regions: the lowest region includes all the visible nodes that are children of U_2, the next higher region includes all the visible nodes that are children of U_3 and the topmost region includes all the visible nodes that are children of U_4.

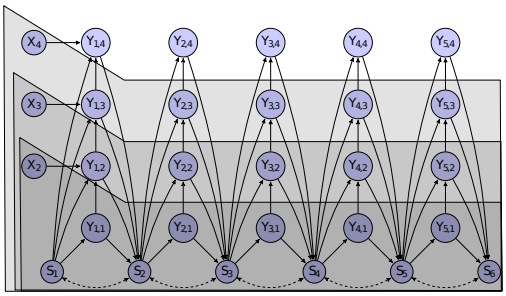


Figure 2: The Graph $\mathcal{G}_6^4$.

2.2 Identification on $\mathcal{G}_n^k$: the peel-off operator

Notation. For any set S of indices, and a symbol A, we denote by A_S the set $\{A_i \mid i \in S\}$ of indexed symbols. Similarly, for sets S and T of indices, we denote by $A_{S,T}$ the set $\left\{A_{i,j} \mid i \in S, j \in T\right\}$. For integers $a \le b$, $[a, b]$ denotes the set of integers between a and b, inclusive. For a positive integer a, we use $[a]$ as a shorthand for the set $[1, a]$. We also denote vectors of values by boldface fonts; in particular, $\mathbf{1}_{[l]}$ denotes a vector of length l all whose entries are 1.

Consider now the computation of the following intervention distribution in $\mathcal{G}_n^k$:

$$P\left(S_{[n]} = \mathbf{1}_{[n]} \mid \text{do}\left(X_{[2,k]} = \mathbf{1}_{[k-1]}, Y_{[n-1],[k]} = \mathbf{1}_{[(n-1)k]}\right)\right). \tag{2}$$

[1] A *C-component* is a maximal set of visible vertices which are reachable from each other through paths consisting only of hidden edges (the directions of the hidden edges are ignored). We refer the reader to, e.g., Shpitser and Pearl (2006) for a discussion of the importance of C-components to causal identification.

The gadget is defined so as to make the Shpitser-Pearl algorithm iterate a sequence of "multiplication" and "marginalization" steps alternately in the computation of this distribution: our goal ultimately is to amplify errors in the multiplication step, and to attempt to preserve the amplification in the marginalization step. However, before seeing how this can be done, we first abstract the operation of the Shpitser-Pearl algorithm on $\mathcal{G}_n^k$ in terms of a *peel-off* operator which clubs the alternating "multiplication" and "marginalization" steps.

We begin by noting that the gadget $\mathcal{G}_n^{k-1}$ can be viewed as a subgraph of the gadget $\mathcal{G}_n^k$ in a canonical manner by identifying the vertices present in the both the gadgets. These "identified" vertices include all the hidden and visible vertices of $\mathcal{G}_n^k$ except X_k, U_k and $\{Y_{i,k} \mid 1 \le i \le n-1\}$. Let $\mathcal{P}_n^k$ denote the set of probability distributions on states of the observed variables of $\mathcal{G}_n^k$. We define an operator π that acts on a distribution in $\mathcal{P}_n^k$ by "peeling off" the top layer of variables and produces an object in $\mathcal{P}_n^{k-1}$ as the output. Although the action of the operator depends upon the values of n and k, we drop its dependence upon these parameters for ease of notation.

Definition 2.1 (Operator π). Given a probability distribution $P \in \mathcal{P}_n^k$ the probability distribution $\pi(P) \in \mathcal{P}_n^{k-1}$ is defined as

$$\begin{aligned}\pi(P)\left(X_{[2,k-1]}, S_{[n]}, Y_{[n-1],[k-1]}\right) := \sum_x & P\left(X_k = x, X_{[2,k-1]}\right) \\ & \cdot \prod_{i=1}^{n-1} P(S_i, Y_{i,[k-1]} \mid X_k = x, Y_{[i-1],k} = \mathbf{1}_{i-1}, X_{[2,k-1]}, S_{[i-1]}, Y_{[i-1],[k-1]}) \\ & \cdot P(S_n \mid X_k = x, Y_{[n-1],k} = \mathbf{1}_{n-1}, X_{[2,k-1]}, S_{[n-1]}, Y_{[n-1],[k-1]}),\end{aligned}$$

where x ranges over all possible values of the X_k, and 1 is assumed to be in the range of $Y_{i,k}$ for all $i \in [n-1]$.

Remark 2.1. Note that the above definition is valid only for $k \ge 2$. However, we can extend it to the case $k = 1$ by "ignoring" the summation over x when $k = 1$ (or equivalently, by assuming that there exists a variable X_1 with no incident edges).

The algorithm of Shpitser and Pearl (2006) for computing the intervention distribution in eq. (2) then amounts to iterating the operator π k times on the observed distribution P:

$$P\left(S_{[n]} = \mathbf{1}_n \mid \text{do}\left(X_{[2,k]} = \mathbf{1}_{k-1}, Y_{[n-1],[k]} = \mathbf{1}_{k(n-1)}\right)\right) = \pi^k(P)\left(S_{[n]} = \mathbf{1}_n\right).$$

Our high level strategy is to take advantage of the multiplication in the definition of π to amplify errors in each step. Intuitively, if each factor in the product in the definition of $\pi(P)$ has an error factor of $(1+\epsilon)$, then we might expect the error in $\pi(P)$ to be of the order of $(1 + \Theta(n)\epsilon)$. We might then expect such an effect to propagate through the k levels so that the final error is of the order of $(1 + \Theta(n^k)\epsilon)$. However, the marginalization over x can destroy this propagation effect, and we will need to be careful to get around this. This will be done by biasing the distribution Q away from being a uniform distribution, using the bias function defined below (see also Remark 2.2).

2.3 The adversarial model

Our goal now is to define a model Q on $\mathcal{G}_n^k$ and a "perturbed" version $\tilde{Q}$ of the observed marginal Q, such that for every observation $\boldsymbol{\xi}$, Q and $\tilde{Q}$ are ϵ-close to each other.To show lower bounds on the condition number of causality, we will need to show that even when Q and $\tilde{Q}$ are ϵ-close, it is possible to arrange matters so that (in the limit $\epsilon \downarrow 0$)

$$\log\left(\frac{\pi^k(\tilde{Q})\left(S_{[n]} = \mathbf{1}_n\right)}{\pi^k(Q)\left(S_{[n]} = \mathbf{1}_n\right)}\right) = \log(1+\epsilon) \cdot \Omega\left(\left(\frac{n}{ck}\right)^k\right), \tag{3}$$

for some positive constant c independent of n and k.

For ease of exposition, we work directly with some marginals of the distributions Q and $\tilde{Q}$. We defer the construction of Q and $\tilde{Q}$ achieving these marginals to a later section. For further ease of notation, we denote the set of vertices $S_{[i]} \cup Y_{[i],[j]}$ as A_{ij}: in terms of fig. 2, this is the set of vertices of height up to j in the first i "towers" (if we

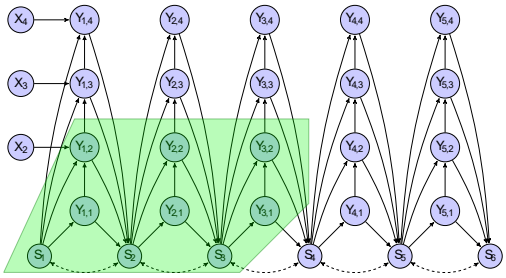


Figure 3: The Set $A_{3,2}$

also consider each of the S vertices to be part of the "tower" of Y vertices immediately to their right; see fig. 3 for an illustration of the set $A_{3,2}$). We further use $A_{ij} = \mathbf{1}$ as a shorthand for the conditioning $S_i = \mathbf{1}_i, Y_{[i-1],[k]} = \mathbf{1}_{(i-1)k}$. We now describe the marginals of Q required in our proof, in terms of a *bias function* as defined below. Only conditional expectations of the form described below will be required in the recursive computation of $\pi^k(Q)\,(S_{[n]} = \mathbf{1}_n)$.

Definition 2.2 (Bias function). The bias function $b : \{0,1\} \to [0,1]$ is defined as

$$b(x) = \begin{cases} \frac{3}{2} & x = 1, \\ 1 & x = 0. \end{cases}$$

$$\begin{aligned}
Q(X_{[2,k]} = \cdot) &= \frac{1}{2^{k-1}}, \\
Q(S_i = 1 \mid A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2} \cdot b(X_2), \text{ for } i \in [n], \\
Q(S_i = 1, Y_{i,[j]} = \mathbf{1}_j \mid A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2^{j+1}} \cdot b\left(X_{j+2}\right), \text{ for } i \in [n-1], j \in [k-2], \\
Q(S_i = 1, Y_{i,[k-1]} = \mathbf{1}_{k-1} \mid A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2^k}, \text{ for } i \in [n-1].
\end{aligned}$$

After one application of π, we have the following expressions for conditional expectations of the above form:

$$\begin{aligned}
\pi(Q)(X_{[2,k-1]} = \cdot) &= \frac{1}{2^{k-2}}, \\
\pi(Q)(S_i = 1 \mid A_{i-1,k-1} = \mathbf{1}, X_{[2,k-1]}) &= \frac{1}{2} \cdot b(X_2), \text{ for } i \in [n], \\
\pi(Q)(S_i = 1, Y_{i,[j]} = \mathbf{1}_j \mid A_{i-1,k-1} = \mathbf{1}, X_{[2,k-1]}) &= \frac{1}{2^{j+1}} \cdot b\left(X_{j+2}\right) \text{ for } i \in [n-1], j \in [k-3], \\
\pi(Q)(S_i = 1, Y_{i,[k-2]} = \mathbf{1}_{k-2} \mid A_{i-1,k-1} = \mathbf{1}, X_{[2,k-1]}) &= \frac{1}{2^{k-1}} \cdot \frac{b(0) + b(1)}{2}, \text{ for } i \in [n-1].
\end{aligned}$$

This process can be continued, and we finally obtain

$$\pi^{k-1}(Q)(S_i = 1 \mid S_{[i-1]} = \mathbf{1}_{i-1}, Y_{[i-1],1} = \mathbf{1}_{i-1}) = \frac{1}{2} \cdot \frac{b(0) + b(1)}{2} = \frac{5}{8}, \text{ for } i \in [n],$$

so that

$$\pi^k(Q)(S_{[n]} = \mathbf{1}_n) = \left(\frac{5}{8}\right)^n.$$

We will now list our requirements for the conditionals of the perturbed distribution $\tilde{Q}$. The construction of a $\tilde{Q}$ that achieves these marginals and that is ϵ-close to Q can be found in Section 2.3.1. Here, we only note that the

construction will be such that $\tilde{Q}$ will differ from Q only in the probabilities of those observations in which all the X_i are simultaneously equal to 1.

$$\begin{aligned}
\tilde{Q}(X_{[2,k]} = \cdot) &= \frac{1}{2^{k-1}}, \\
\tilde{Q}(S_i = 1 \mid A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2} \cdot b\left(X_2\right), \text{ for } i \in [n], \\
\tilde{Q}(S_i = 1, Y_{i,[j]} = \mathbf{1}_j \mid A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2^{j+1}} \cdot b\left(X_{j+2}\right), \text{ for } i \in [n-1], j \in [k-2], \\
\tilde{Q}(S_i = 1, Y_{i,[k-1]} = \mathbf{1}_{k-1} \mid A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2^k} \cdot \begin{cases} (1+\epsilon/2) & X_{[2,k]} = \mathbf{1}_{k-1} \\ 1 & X_{[2,k]} \neq \mathbf{1}_{k-1} \end{cases} \\
&\qquad \text{for } i \in [n-1].
\end{aligned}$$

We now compute the relevant iterated applications of π on $\tilde{Q}$ using the above marginals.To simplify notation, define:

$$\nu := (1 + \epsilon/2).$$

We now compute the corresponding expressions for $\pi(\tilde{Q})$. It will also be convenient to use the following "error-propagator" function.

Definition 2.3 (Error propagator). Given a bias function b as defined above, the error propagator function η_b (abbreviated to η when b is clear from the context) is $\eta\left(x\right) = \eta_b(x) := \frac{b(0)+xb(1)}{1+x}$.

We are now ready to describe $\pi(\tilde{Q})$:

$$\begin{aligned}
\pi(\tilde{Q})(X_{[2,k-1]} = \cdot) &= \frac{1}{2^{k-2}}, \\
\pi(\tilde{Q})(S_i = 1 \mid A_{i-1,k-1} = \mathbf{1}, X_{[2,k-1]}) &= \frac{1}{2} \cdot b\left(X_2\right), \text{ for } i \in [n], \\
\pi(\tilde{Q})(S_i = 1, Y_{i,[j]} = \mathbf{1}_j \mid A_{i-1,k-1} = \mathbf{1}, X_{[2,k-1]}) &= \frac{1}{2^{j+1}} \cdot b\left(X_{j+2}\right) \text{ for } i \in [n-1], j \in [k-3], \\
\pi(\tilde{Q})(S_i = 1, Y_{i,[k-2]} = \mathbf{1}_{k-2} \mid A_{i-1,k-1} = \mathbf{1}, X_{[2,k-1]}) &= \frac{1}{2^{k-1}} \cdot \begin{cases} \eta\left(\nu^{i-1}\right) & X_{[2,k-1]} = \mathbf{1}_{k-2} \\ \eta\left(1\right) & X_{[2,k-1]} \neq \mathbf{1}_{k-2} \end{cases} \\
&\qquad \text{for } i \in [n-1].
\end{aligned}$$

Note that only the last of these expressions differs from the case of Q, and it differs only for those observations in which all the remaining X_i nodes are set to 1. This pattern persists for further iterates of π.

Remark 2.2. We can now make precise how the bias function allows propagation of errors through π in spite of the marginalization step. Note that if we had no bias, i.e., $b\left(x\right) \equiv 1$, then the discrepancy from $\pi(Q)$ in the last line in the definition of $\pi(\tilde{Q})$ will not arise at all.

In order to describe the evolution of this discrepancy, we define the quantities $\nu_{l,i}$ as ratios between the intermediate marginals computed from $\tilde{Q}$ and Q respectively after l applications of ν:

$$\nu_{l,i} := \frac{\pi^l(\tilde{Q})(S_i = 1, Y_{i,[k-l-1]} = \mathbf{1}_{k-l-1} \mid A_{i-1,k-l} = \mathbf{1}, X_{[2,k-l]} = \mathbf{1}_{k-l-1})}{\pi^l(Q)(S_i = 1, Y_{i,[k-l-1]} = \mathbf{1}_{k-l-1} \mid A_{i-1,k-l} = \mathbf{1}, X_{[2,k-l]} = \mathbf{1}_{k-l-1})}.$$

From the base case of $\tilde{Q}$ and the definition of operator π, we then have

$$\begin{aligned}
\nu_{0,i} &= \nu = (1+\epsilon/2), && \text{for } i \in [n-1] \\
\nu_{l,n} &= 1 && \text{for } 0 \leq l \leq k-1 \\
\nu_{l,i} &= \frac{\eta\left(\prod_{j=1}^{i-1} \nu_{l-1,j}\right)}{\eta\left(1\right)}, && \text{for } 1 \leq l \leq k-1, i \in [n-1].
\end{aligned}$$

Note that we have

$$\frac{\pi^k(\tilde{Q})(S_{[n]} = \mathbf{1}_n)}{\pi^k(Q)(S_{[n]} = \mathbf{1}_n)} = \prod_{i=1}^{n-1} \nu_{k-1,i}, \tag{4}$$

so that we only need to upper bound the $\nu_{l,i}$ appropriately. In order to do this, we will use the following lemma:

Lemma 2.1. *There exists a $\delta > 0$ such that for $x \in [1, 1+\delta)$, $\frac{\eta(x)}{\eta(1)} \geq x^{1/11}$. The parameter δ can be chosen to be at least 1.*

Proof. The claim of the lemma is equivalent to the existence of a positive δ such that

$$f(x) := 5x^{12/11} + 5x^{1/11} - 6x - 4 \leq 0 \text{ for } x \in [1, 1+\delta).$$

To prove the latter fact, we observe that $f(1) = 0$, and that $f'(1) = \left[\frac{60}{11}x^{1/11} + \frac{5}{11}x^{-10/11} - 6\right]_{x=1} = -\frac{1}{11} < 0$. Indeed, a direct computation shows that δ can be chosen to be 1, since $f(2) < 0$ and f is convex in $[1, 2)$. □

We will now use the above lemma to prove by induction the following lower bound on the $\nu_{l,i}$.

Lemma 2.2. *Suppose that for $1 \leq i \leq n-1$ and $0 \leq l \leq k-1$, we have $\nu_{l,i} \geq 1$ and $\frac{1}{11^{l-1}}\binom{i-1}{l} \log \nu < \log 2$. Then, for such l and i, $\log \nu_{l,i} \geq \frac{1}{11^l}\binom{i-1}{l} \log \nu$.*

Proof. The base case $l = 0$ is true by the definition of the $\nu_{l,i}$. For the induction, $l \geq 1$, and we start with the recursive definition of $\nu_{l,i}$ (for $i \leq n-1$) in terms of $\nu_{l-1,j}$:

$$\begin{aligned}
\nu_{l,i} &= \frac{1}{\eta(1)} \cdot \eta\left(\prod_{j=1}^{i-1} \nu_{l-1,j}\right) \\
&\geq \frac{1}{\eta(1)} \cdot \eta\left(\prod_{j=1}^{i-1} \exp\left(\frac{1}{11^{l-1}}\binom{j-1}{l-1}\log \nu\right)\right) &(5)\\
&= \frac{1}{\eta(1)} \cdot \eta\left(\exp\left(\frac{1}{11^{l-1}}\binom{i-1}{l}\log \nu\right)\right) &(6)\\
&\geq \exp\left(\frac{1}{11^l}\binom{i-1}{l}\log \nu\right), &(7)
\end{aligned}$$

where eq. (5) uses the induction hypothesis and the fact that η is an increasing function, eq. (6) employs the elementary combinatorial identity $\sum_{j=1}^{i-1}\binom{j-1}{l-1} = \binom{i-1}{l}$, and eq. (7) uses the hypotheses of the lemma to apply Lemma 2.1. □

We are now ready to complete the proof of Theorem 1.2.

Proof of Theorem 1.2. Substituting the bounds on the $\nu_{l,i}$ from Lemma 2.2 in eq. (4), we get

$$\begin{aligned}
\frac{\pi^k(\tilde{Q})(S_{[n]} = \mathbf{1}_n)}{\pi^k(Q)(S_{[n]} = \mathbf{1}_n)} &= \prod_{i=1}^{n-1} \nu_{k-1,i} &(8)\\
&\geq \exp\left(\frac{\log \nu}{11^{k-1}} \sum_{i=1}^{n-1}\binom{i-1}{k-1}\right) &(9)\\
&= \exp\left(\frac{\log \nu}{11^{k-1}}\binom{n-1}{k}\right) &(10)\\
&\geq \exp\left(11 \log \nu \left(\frac{n-1}{11k}\right)^k\right), &(11)
\end{aligned}$$

where eq. (9) uses Lemma 2.2, eq. (10) is again based on the elementary identity used in eq. (6), and eq. (11) is an application of the standard inequality $\binom{a}{b} \geq \left(\frac{a}{b}\right)^b$. Now, let $k = (n-1)^{\alpha'}/11$ for $\alpha' \in [0,1)$, and set $M := 11(n-1)k$. Note that $M = O(N)$ where N is the number of visible nodes in $\mathcal{G}_n^k$. We then have

$$\frac{\pi^k(\tilde{Q})(S_{[n]} = \mathbf{1}_n)}{\pi^k(Q)(S_{[n]} = \mathbf{1}_n)} \geq \exp\left(11 \log \nu \exp\left(\frac{(1-\alpha')M^{\alpha'/(\alpha'+1)} \log(n-1)}{11}\right)\right).$$

Choosing $\alpha = \alpha'/(\alpha'+1)$ completes the proof. □

2.3.1 Definitions of Q and $\tilde{Q}$

We now supply the promised definitions of Q and $\tilde{Q}$. We first define Q by providing the appropriate conditional distributions of the bits at different nodes in the graph. For the sake of brevity, we only specify the conditional distributions which are *not* uniform: all the unspecified distributions are assumed to be uniform over $\{0,1\}$. We recall the definition of the bias function used earlier:

Definition 2.4 (Bias function). The bias function $b : \{0,1\} \to [0,1]$ is defined as

$$b(x) = \begin{cases} \frac{3}{2} & x = 1, \\ 1 & x = 0. \end{cases}$$

Q is now defined as follows:

$$\begin{aligned}
Q(X_i = U_i \mid U_i) &= 1, \text{ for } 2 \leq i \leq k, \\
Q(S_i = 1 \mid Y_{i-1,[k]} = \mathbf{1}_k, U_{[2,k]}) &= \frac{1}{2} \cdot b(U_2) \text{ for } i \in [n], \\
Q(Y_{i,j} = 1 \mid S_i = 1, Y_{i,j-1} = 1, U_{[j+1,k]}) &= \frac{1}{2} \cdot \begin{cases} b\left(U_{j+2}\right) & U_{j+1} = 0, \\ \frac{1}{b\left(1-U_{j+2}\right)} & U_{j+1} = 1 \end{cases} \\
&\qquad \text{for } i \in [n-1], j \in [k-2]. \\
Q(Y_{i,k-1} = 1 \mid S_i = 1, Y_{i,k-2} = 1, U_k) &= \frac{1}{2} \cdot \frac{1}{b(U_k)}, \text{ for } i \in [n-1].
\end{aligned}$$

The above construction of Q leads, in particular, to the following observed conditional distributions for Q:

$$\begin{aligned}
Q(X_{[2,k]} = \cdot) &= \frac{1}{2^{k-1}} \\
Q(S_i = 1 \mid A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2} \cdot b(X_2), \text{ for } i \in [n], \\
Q(Y_{i,j} = 1 \mid S_i = 1, Y_{i,[j-1]} = \mathbf{1}_{j-1}, A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2} \cdot \begin{cases} b\left(X_{j+2}\right) & X_{j+1} = 0, \\ \frac{1}{b\left(1-X_{j+2}\right)} & X_{j+1} = 1, \end{cases} \\
&\qquad \text{for } i \in [n-1], j \in [k-2]. \\
Q(Y_{i,k-1} = 1 \mid S_i = 1, Y_{i,[k-2]} = \mathbf{1}_{k-2}, A_{i-1,k} = \mathbf{1}, X_{[2,k]}) &= \frac{1}{2} \cdot \frac{1}{b(X_k)}, \text{ for } i \in [n-1].
\end{aligned}$$

The Perturbed Distribution We now define the perturbation $\tilde{Q}$ of Q used in the proof. We only specify those entries of $\tilde{Q}$ which are different from those of Q.

$$\begin{aligned}
\frac{\tilde{Q}(X_{[2,k]} = \mathbf{1}_{k-1}, S_i = 1, A_{i-1,k} = \mathbf{1}, Y_{i,[k-2]} = \mathbf{1}_{k-2}, Y_{i,[k-1,k]} = 00, \star)}{Q(X_{[2,k]} = \mathbf{1}_{k-1}, S_i = 1, A_{i-1,k} = \mathbf{1}, Y_{i,[k-2]} = \mathbf{1}_{k-2}, Y_{i,[k-1,k]} = 00, \star)} &= (1-\epsilon/2), \text{ for } i \in [n-1], \\
\frac{\tilde{Q}(X_{[2,k]} = \mathbf{1}_{k-1}, S_i = 1, A_{i-1,k} = \mathbf{1}, Y_{i,[k-2]} = \mathbf{1}_{k-2}, Y_{i,[k-1,k]} = 10, \star)}{Q(X_{[2,k]} = \mathbf{1}_{k-1}, S_i = 1, A_{i-1,k} = \mathbf{1}, Y_{i,[k-2]} = \mathbf{1}_{k-2}, Y_{i,[k-1,k]} = 10, \star)} &= (1+\epsilon), \text{ for } i \in [n-1].
\end{aligned}$$

3 A well-conditioned class

In this section we provide a counterpoint to our main result: we exhibit a useful class of well-conditioned causal identification problems, by proving Proposition 1.3. The proof follows almost immediately from an earlier causal identification result of Tian and Pearl (2002).

Proof of Proposition 1.3. We restrict our attention to the condition number of $\mathbf{ID}(G, V - \{X\}, X)$: since $P(S \mid \mathrm{do}\, X)$ can be obtained from $P(V - \{X\} \mid \mathrm{do}\, X)$ by a marginalization operation, an upper bound on the condition number for $P(V - \{X\} \mid \mathrm{do}\, X)$ is also an upper bound on the condition number of $P(S \mid \mathrm{do}\, X)$.

For a sufficiently small ϵ, let $\tilde{P}$ be a probability distribution that is ϵ-close to the actual empirical distribution P. This implies, in particular, that all *conditional* probabilities computed according to $\tilde{P}$ are 2ϵ-close to the true conditional probabilities computed according to P. Formally, for any disjoint subsets S and T of V, $P(S = s \mid T = t)$ and $\tilde{P}(S = s \mid T = t)$ are 2ϵ-close: this follows from the fact that $P(S = s \mid T = t) = P(S = s, T = t)/P(T = t)$.

Let Z be the set of nodes (except X) in the same C-component as X. Using the identifiability result of Tian and Pearl (2002), we have

$$P(V - \{X\} = v_{V-\{X\}} | \, \mathrm{do}(X = x)) = P(v) \frac{\sum_{x'} H(x')}{H(x)}, \tag{12}$$

where x' ranges over the domain of X, and $H(x')$ is defined as

$$H(x') = P(X = x' \mid \mathrm{An}(X) = v_{\mathrm{An}(X)}) \cdot \prod_{V_i \in Z} P(V_i = v_{V_i} \mid \mathrm{An}(V_i) = v_{\mathrm{An}(v_i)}),$$

where $v_X = x'$. (Here, for a vertex V_i, $\mathrm{An}(V_i)$ is the set of ancestors of V_i among the observed nodes V). Since each H is a product of at most $|V|$ conditional probabilities, and since conditional probabilities computed according to P and $\tilde{P}$ are 2ϵ-close, the values of $H(x')$ computed according to P and $\tilde{P}$ are $2|V|\epsilon$-close. Eq. 14 then gives

$$e^{-(4|V|+1)\epsilon} \leq \frac{\tilde{P}(V - \{X\} = v_{V-\{X\}} \mid \mathrm{do}(X = x))}{P(V - \{X\} = v_{V-\{X\}} \mid \mathrm{do}(X = x))} \leq e^{(4|V|+1)\epsilon}.$$

for every v and x. We therefore have

$$\kappa_{\mathbf{ID}(G, V-\{X\}, X)} \leq \lim_{\epsilon \downarrow 0} \frac{e^{(4|V|+1)\epsilon} - 1}{\epsilon} = 4|V| + 1. \qquad \square$$

4 Conclusion

In this paper, we gave an example of a class of semi-Markovian models in which the causal inference problem is highly ill-conditioned. However, Proposition 1.3 shows that at least some causal identification problems are not too badly conditioned.

An immediate open question therefore is to find an algorithm which can compute tight bounds for the condition number of a causal identification problem in a given semi-Markovian model. Since such an algorithm would operate only on the model (and not on the observed data), it can serve a guide for selecting between competing models which differ, e.g., in terms of which covariates are measured, *before* any data is collected: all else being equal, a model in which the causal inference problem to be solved is better conditioned and hence less susceptible to noise should be preferable.

The roots of causal identification in graphical models can be traced back to the setting of linear structural equation models, which were first studied in the seminal papers of Wright (1921, 1934) (see, e.g., Drton et al. (2011), Foygel et al. (2012) and Chen and Pearl (2014) for more recent results on identification in linear structural equation models). Not surprisingly, in contrast to the purely combinatorial identification procedures in the discrete case, the identification procedures for linear structural equation models are able to exploit the linear algebraic structure of the problem and often use, in addition to combinatorial considerations, algorithmic primitives from linear algebra and algebraic geometry as well (see e.g. Foygel et al. (2012) for an example). Condition numbers of the primitives themselves have been studied have been quite extensively, but exploring the condition number of causal identification in the setting of linear structural equation models remains open.

References

Peter Bürgisser and Felipe Cucker. *Condition: The Geometry of Numerical Algorithms*, volume 349 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2013.

Bryant Chen and Judea Pearl. Graphical tools for linear structural equation modeling, June 2014. URL http://www.cs.ucla.edu/pub/stat_ser/r432.pdf.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *Ann. Stat.*, 39(2):865–886, April 2011.

Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *Ann. Stat.*, 40(3):1682–1713, June 2012.

Joseph Y. Halpern. Axiomatizing causal reasoning. *J. Artif. Intell. Res.*, 12:317–337, May 2000.

Yimin Huang and Marco Valtorta. On the completeness of an identifiability algorithm for semi-Markovian models. *Ann. Math. Artif. Intell.*, 54(4):363–408, December 2008.

Manabu Kuroki and Masami Miyakawa. Identifiability criteria for causal effects of joint interventions. *J. Japan Stat. Soc.*, 29(2):105–117, 1999. doi: 10.14490/jjss1995.29.105.

William. S. Ohlemeyer. Closing statement in *Henley v. Philip Morris Inc.*, 1999. Case No. 995172, 3 February 1999, Superior Court of the State of California. Page 88 in the original, p. 42 in the digitization. Available at https://industrydocuments.library.ucsf.edu/tobacco/docs/#id=frxl0001. Accessed Mar. 1, 2016.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, December 1995.

Judea Pearl and James Robins. Probabilistic Evaluation of Sequential Plans from Causal Models with Hidden Variables. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, UAI'95, pages 444–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proc. 20th AAAI Conference on Artifical Intelligence*, pages 1219–1226. AAAI Press, July 2006.

Jin Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, University of California, Los Angeles, August 2002.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proc. 18th AAAI Conference on Artificial Intelligence*, pages 567–573. AAAI Press, 2002.

Sewall Wright. Correlation and causation. *J. Agric. Res.*, 20:557–585, 1921.

Sewall Wright. The method of path coefficients. *Ann. Math. Stat.*, 5:161–215, 1934.

A Proof of Proposition 1.1

In this section, we provide a proof of Proposition A.1 that was omitted from the main body of the paper. We follow the same notation as in the paper. However, the statement of the proposition is reproduced here for quick reference.

Proposition A.1. *Consider a semi-Markovian graph $G = (V, E, U, H)$ and a distribution $P(V, U)$ respecting it. Let $R = \{e_i = (A_i, V_i) \mid 1 \le i \le q\}$ be a set of k edges in $E \cup H$ such that e_i is ϵ_i-weak, with $\epsilon := \sum_{i=1}^{k} \epsilon_i$. Suppose that X, Y are disjoint subsets of V for which $\mathbb{P}(Y \mid \text{do}\, X)$ is not identifiable in G, but identifiable in $G' = (V, E \setminus R, U, H \setminus R)$. Then there exists a distribution $\tilde{P}(V, U)$ respecting G' such that*

$$-\epsilon \le \log \frac{\tilde{P}(V)}{P(V)} \le \epsilon, \quad \textit{and} \quad -\epsilon \le \log \frac{\tilde{P}(Y \mid \text{do}\, X)}{P(Y \mid \text{do}\, X)} \le \epsilon.$$

Note that $\tilde{P}(Y \mid \text{do}\, X)$ is computable (by the algorithms of Shpitser and Pearl (2006) and Huang and Valtorta (2008)) given $\tilde{P}(V)$, but $P(Y \mid \text{do}\, X)$ is not even uniquely determined given only the observed marginal $P(V)$.

Proof of Proposition A.1. We define $\tilde{P}$ on (V, U) by giving an explicit factorization which respects G' by construction. We first define $\tilde{P}(U) = P(U)$, so that P and $\tilde{P}$ agree when restricted to the hidden variables. The weakness of the edges removed from G to obtain G' only plays a part in defining the factorization on the visible nodes in V. Let B be a vertex in V, and let $A_1, A_2, \ldots, A_k$ be the (possibly empty) set of vertices in $U \cup V$ such that among the k edges removed from G to obtain G', those incident on B are $(A_1, B), (A_2, B), \ldots, (A_l, B)$. Let $\epsilon_{B_1}, \epsilon_{B_2}, \ldots, \epsilon_{B_l}$ be such that the edge (A_i, B) is ϵ_{B_i}-weak. Let $\Xi(B)$ denote the set of parents of B in $U \cup V$ disjoint from $\{A_1, A_2, \ldots, A_l\}$. We then define:

$$\tilde{P}(B = b \mid \Xi(B) = \xi) = P(B = b \mid \Xi(B) = \xi, A_i = a_i, 1 \leq i \leq l), \tag{13}$$

where b, ξ are values in the domain of B and $\Xi(B)$ respectively, and a_i are *arbitrary* values in the domain of the A_i. From the definition of weakness, we then have, for all b, ξ in the domain of B and $\Xi(B)$ respectively, and for all a_i' in the domain of the A_i:

$$\left|\log \tilde{P}(B = b \mid \Xi(B) = \xi) - \log P(B = b \mid \Xi(B) = \xi, A_i = a_i', 1 \leq i \leq l)\right| \leq \sum_{i=1}^{l} \epsilon_{B_i}. \tag{14}$$

We now define $\tilde{P}(V, U)$ in the standard way by factorizing in terms of the conditional probabilities defined in eq. (13): it then respects G' by construction. Using eq. (14), we have, for every u and v in the domain of U and V respectively,

$$\left|\log \tilde{P}(V = v, U = u) - \log P(V = v, U = u)\right| \leq \sum_{i=1}^{1} \epsilon_i = \epsilon. \tag{15}$$

The claim of the proposition comparing $P(V)$ and $\tilde{P}(V)$ now follows from the fact that $P(V = v)$ and $\tilde{P}(V = v)$ are both obtained by marginalizing P and $\tilde{P}$ over the domain of U, and the latter are ϵ-close by eq. (15). The proof of the claim comparing $P(X \mid \mathrm{do}(Y))$ and $\tilde{P}(X \mid \mathrm{do}\, Y)$ follows in exactly the same fashion starting from eq. (13), by using the factorizations of these quantities in terms of the conditional probabilities appearing in eq. (13). □