Condition Number Bounds for Causal Inference

Spencer L. Gordon[1] **Vinayak M. Kumar**[1] **Leonard J. Schulman**[1] **Piyush Srivastava**[2]

[1]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA.
[2]School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, Maharashtra, India.

Abstract

An important achievement in the field of causal inference was a complete characterization of when a causal effect, in a system modeled by a causal graph, can be determined uniquely from purely observational data. The identification algorithms resulting from this work produce exact *symbolic* expressions for causal effects, in terms of the observational probabilities. More recent work has looked at the *numerical* properties of these expressions, in particular using the classical notion of the *condition number*. In its classical interpretation, the condition number quantifies the sensitivity of the output values of the expressions to small numerical perturbations in the input observational probabilities. In the context of causal identification, the condition number has also been shown to be related to the effect of certain kinds of uncertainties in the *structure* of the causal graphical model.

In this paper, we first give an upper bound on the condition number for the interesting case of causal graphical models with small "confounded components". We then develop a tight characterization of the condition number of any given causal identification problem. Finally, we use our tight characterization to give a specific example where the condition number can be much lower than that obtained via generic bounds on the condition number, and to show that even "equivalent" expressions for causal identification can behave very differently with respect to their numerical stability properties.

1 INTRODUCTION

Scientists have long designed controlled experiments to determine cause-and-effect relationships between quantities of interest in systems they are studying. However, in specific applications, several factors, such as considerations of ethics and feasibility, may prevent the use of randomized controlled trials to determine the strength of cause-and-effect relationships. This led to an exploration of when and how it might be possible to measure cause-and-effect relationships from observational data alone. One well-studied approach to this question has been in the framework of *causal DAGs*, starting with the work of Pearl [1995]. The framework models parts of the system being studied, and domain-specific knowledge about absence of direct causal relationships between such parts, as a directed acyclic graph with both *observed* and *latent* vertices. The resulting formalization has been quite influential, and led initially to a systematization of the original question in terms of graph theoretic notions (such as *d-separation* and the *back-door criterion*).

After a decade of work on causal DAGs, papers by Shpitser and Pearl [2006, 2008] and Huang and Valtorta [2006, 2008] (see also Tian [2002]) culminated in a sound and complete algorithm (the **ID** algorithm) which decides, given a causal DAG G with observed variables V and latent variables with known direct influences, a set $X \subseteq V$ of variables upon which to intervene, and an effect set $Y \subseteq (V \setminus X)$, whether the effect of an intervention on X upon Y can be determined from the joint distribution over V. In the case that the desired interventional distribution is *identifiable*, the **ID** algorithm gives a *symbolic expression*, or formula, for the interventional distribution, denoted $P(Y \mid \mathbf{do}(X))$, in terms of observational distribution $P(V)$.

When using the **ID** algorithm in practice, the *numerical* values of the observed marginals $(P(V = \mathbf{v}))_{\mathbf{v} \in \Sigma^V}$ will be input into the ***ID** expression* (or formula) output by the **ID** algorithm to get a final numerical answer. Of course, in this setting, the numerical input will most likely *not* be the true observational distribution $P(V)$, but rather an empirical approximation $\tilde{P}(V)$. It therefore becomes important to understand the robustness of the **ID** expression output by the **ID** algorithm, to numerical perturbations of its input.

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1948–1957.

A classical notion of robustness of a function to perturbations of its inputs is the *condition number*, ubiquitous in numerical analysis (see, e.g, Bürgisser and Cucker [2013]). Informally, the condition number bounds the ratio between the *relative* error in the output of a function to the *relative* error in the input of that function (see Definition 1 for a precise definition). The importance of this notion of measuring sensitivity of errors in numerical analysis comes from the fact that the logarithm of the condition number of a function f captures the *loss of precision* when an algorithm employing fixed-precision floating-point arithmetic is used to compute f [Bürgisser and Cucker, 2013, Theorem O.3].

In the context of causal identification, the condition number has another desirable property: Schulman and Srivastava [2016] observed that it can also be used to study errors resulting from certain kinds of model mis-specifications. Roughly speaking, they show that if an edge $X \to Y$ is such that changes in X have small multiplicative effect on any $P(Y = y)$, then omission of the edge from the model has small effect on the numerical output of the causal identification formula, provided that the condition number is small. In addition to this, they also showed that on a certain class of instances for the **ID** algorithm studied by Tian and Pearl [2002], the condition number of the **ID** expression output by the algorithm is not too large ($O(n)$, where n is the number of vertices) for all input distributions. Further, they constructed an instance, for which, on a tailored input marginal distribution, the expression output by the **ID** algorithm has a "large" condition number ($\Omega\left(\exp(n^{0.4})\right)$). More recent work of Sankararaman et al. [2020a,b]) has focused on the condition number of a related (but different!) notion of causal identification in the setting of structural equation models [Wright, 1921, 1934].

Our Contribution Several questions regarding the condition number of causal identification in the setting of general causal DAGs remain open: What is the condition number for a given instance? Can one obtain more general upper bounds than the ones shown by Schulman and Srivastava [2016]? In this paper, we report progress on both these questions. First, in Theorem 4, we give a general condition number bound for causal DAGs in which all C-components are small (such models turn out to be interesting especially in the context of learning and testing: see Acharya et al. [2018]). In our main technical result, Theorem 5 (along with Corollaries 10 and 11), we use ideas from numerical analysis and linear programming duality to give a characterization of the condition number of causal identification. This characterization, while giving an explicit method for computing the condition number for any given instance, also allows us to demonstrate two important subtle points about the condition number of causal identification. The first of these, shown in Theorem 7, is that the condition number of a given **ID** expression can be *smaller* than that obtained using "generic" upper bounds like the one obtained by Schulman and Srivastava [2016] or in our own Theorem 4. The second, more important, point is that "equivalent" **ID** expressions for the *same* interventional distribution can have very different behavior with regards to condition number and, hence, numerical stability. This is described in the remarks following Theorem 7. Notably, both these subtleties already manifest in an *extremely* simple example of causal identification (see Figure 1).

Related Work As discussed above, the condition number of causal identification has been studied by Schulman and Srivastava [2016] in the setting of causal DAGs, and by Sankararaman et al. [2020a,b] in the setting of structural equation models. We are not aware of any other work on the numerical stability of causal inference.

While the notion of *relative error* is more natural in the study of numerical stability and condition numbers, the notion of statistical or *total variation distance* seems to arise more naturally in studies of finite-sample learning and testing problems associated with causal DAGs (see, e.g., Acharya et al. [2018] and Bhattacharyya et al. [2020]). While these latter works do not consider the stability of causal identification measured with respect to the total variation distance, we note that relative error is a more demanding notion of approximation, especially for rare events: $\mathrm{Rel}(P, \tilde{P}) \leq \epsilon$ (see Section 1.1 for a formal definition) implies that the total variation distance between P and $\tilde{P}$ is also at most $O(\epsilon)$, and in particular, $\tilde{P}(E) = 0$ for any event E for which $P(E) = 0$. Thus, the notion of stability guaranteed by a "low" condition number starts with requiring a more demanding notion of approximation for the input, but also in turn guarantees a stricter notion of approximation for the output.

Notation and Terminology for Graphical Models Our notation and terminology for graphical models is largely taken from the work of Shpitser and Pearl [2006] and Huang and Valtorta [2006]. We quickly summarize this notation here, but refer to the above papers for more complete descriptions. We denote sets of vertices in graphs, and the vertices themselves, using capital letters. Given a directed acyclic graph (DAG) G and a set of vertices S in G, $\mathrm{Pa}(S)$ denotes the set of parents of S in G, while $\mathrm{An}(Y)$ denotes the set of ancestors of S (including S). A *causal graphical model* is a set of *observable* random variables V, and a set of *latent* random variables U arranged as a DAG, such that no vertex in U has a parent. A subset S of observed vertices is said to be *C-connected* if there is a path (ignoring the directions of the edges) between every pair of vertices in S that only goes through vertices in U; a maximal C-connected set is a *C-component*.

For random variables corresponding to a set S of vertices, we use the corresponding small letter $\mathbf{s}$ to denote a generic tuple of values that the variables in S may assume. The boldface symbol $\mathbf{s}$ is used when S is not a singleton, while

the normal weight symbol s is used when S is a singleton. In the former case, symbols such as s_i are used to denote the component of $\mathbf{s}$ indexed by i. The manifold of probability distributions supported by the model is given in terms of probability kernels of the form $P(V_i = v_i \mid \mathrm{Pa}(V_i) = \mathbf{pa}(V_i))$, for each vertex V_i in $V \cup U$ (of course, $\mathrm{Pa}(v_i) = \varnothing$ when $v_i \in U$), so that the global probability distribution is $P(V = \mathbf{v}, U = \mathbf{u}) = \prod_{V_i \in V \cup U} P(V_i = v_i \mid \mathrm{Pa}(V_i) = \mathbf{pa}(V_i))$, where $\mathbf{pa}(V_i)$ denotes the partial assignment to the parents $\mathrm{Pa}(V_i)$ of V_i under the full assignment $\mathbf{v} \cup \mathbf{u}$. The *observed* distribution $P(V = \mathbf{v})$ is the marginal of this distribution on the observed vertices: $P(V = \mathbf{v}) = \sum_{\mathbf{u}} P(V = \mathbf{v}, U = \mathbf{u})$.

Making an *intervention* on a subset X of the observed vertices consists of fixing the values of those vertices to some setting $\mathbf{x}$. The *intervention distribution* induced by such an intervention on a set Y of observed vertices (disjoint from X) is denoted $P(Y = \mathbf{y} \mid \mathbf{do}(X = \mathbf{x}))$ and is defined by

$$\begin{aligned} &P(Y = \mathbf{y} \mid \mathbf{do}(X = \mathbf{x})) \\ &= \sum_{(\mathbf{v} \cup \mathbf{u}) \setminus (\mathbf{y} \cup \mathbf{x})} \prod_{V_i \in (V \cup U) \setminus X} P(V_i = v_i \mid \mathrm{Pa}(V_i) = \mathrm{pa}(\mathbf{V_i})), \end{aligned} \tag{1}$$

and this corresponds to considering a modified graph in which all the incoming edges into the vertices in X have been removed. The problem of *causal identification* is to obtain, whenever possible, a formula for the intervention distribution in eq. (1) in terms of the entries $P(V = \mathbf{v})$ of the observed distribution. When such a formula exists, we call it an **ID** expression for the intervention distribution in eq. (1). The **ID** algorithm finds such an expression whenever one exists [Shpitser and Pearl, 2006, Huang and Valtorta, 2006] (or terminates with a proof that no such expression exists, in which case the intervention distribution is *non-identifiable*): we reproduce the version of the **ID** algorithm given by Shpitser and Pearl [2006] in algorithm 1.

1.1 DESCRIPTION OF RESULTS

We begin by setting up the notation for the condition number in the context of causal identification. Our notation follows that of Bürgisser and Cucker [2013]. For positive real numbers a and a', we use $\mathrm{Rel}(a, a') := \frac{|a - a'|}{a}$ to denote the relative error incurred when using a' to approximate a. This generalizes to n-dimensional vectors v and v' (with positive entries) as $\mathrm{Rel}(v, v') := \max_{i=1}^{n} \mathrm{Rel}(v_i, v'_i)$. In this paper, we will be concerned only with probability distributions over finite sets, so we specialize the following presentation to this case. Let P be a probability distribution over a finite set Ω. We can then identify P with a unique point in the probability simplex Δ_Ω defined as

$$\Delta_\Omega := \{x \in \mathbb{R}^\Omega \mid x_\omega \geq 0 \ \forall \omega \in \Omega \text{ and } \sum_{\omega \in \Omega} x_\omega = 1\}.$$

For technical reasons, we will exclude the trivial probability distributions which assign weight 1 to a single point of the domain Ω. Thus, we focus on distributions P in which for every $\omega \in \Omega$, $P(\omega) < 1$. We denote the set of these distributions as Δ_Ω^+.

Now, given any function $f : \Delta_\Omega \to \mathbb{R}_+$, and $P \neq \tilde{P} \in \Delta_\Omega^+$, we define

$$\zeta_f(P, \tilde{P}) := \frac{\mathrm{Rel}(f(P), f(\tilde{P}))}{\mathrm{Rel}(P, \tilde{P})}. \tag{2}$$

Definition 1 (Condition number). Let f be as in the previous paragraph, and consider $P \in \Delta_\Omega^+$. For any $\epsilon > 0$, we define the set $B_\varepsilon(P) := \left\{\tilde{P} \in \Delta_\Omega \mid \mathrm{Rel}(P, P') \leq \epsilon\right\}$. The *condition number of f at P*, denoted $\kappa(f, P)$, is defined as

$$\kappa(f, P) := \lim_{\varepsilon \downarrow 0} \sup_{\tilde{P} \in B_\varepsilon(P)} \zeta(P, \tilde{P}). \tag{3}$$

The condition number of f over a subset $S \subseteq \Delta_\Omega^+$ is defined as $\kappa(f, S) := \sup_{P \in S} \kappa(f, P)$. When S is the relative interior of Δ_Ω^+, the quantity $\kappa(f, S)$ is called the *condition number of f*, and we denote it by $\kappa(f)$.

Remark 2. Note that the definitions of ζ_f and $\kappa(f, \cdot)$ given in eqs. (2) and (3) apply, exactly as stated, even when f is a vector valued function, as long as each coordinate of f is non-negative.

It would be helpful to record some natural properties that one may expect the condition number to inherit as a measure of "increase" in relative error. The following standard fact (specialized here to our setting above) describes upper bounds on the condition number sums, products and composition. We note that these upper bounds are not tight. For completeness, we give a proof of these bounds in Supplementary Material Section A.1.

Lemma 3. *Fix $f, g : \Delta_\Omega \to \mathbb{R}_+$ and $P \in \Delta_\Omega^+$. Assume that $f(P), g(P) > 0$ and that f, g are differentiable at P. Then, we have*

1. $\kappa(fg, P), \kappa(f/g, P) \leq \kappa(f, P) + \kappa(g, P)$.
2. $\kappa(f + g, P) \leq \max\{\kappa(f, P), \kappa(g, P)\}$.
3. *For $h : \Delta_{\Omega'}^+ \to \Delta_\Omega^+$, where Ω' is possibly different from Ω, and $P \in \Delta_{\Omega'}^+$, $\kappa(f \circ h, P) \leq \kappa(f, h(P)) \cdot \kappa(h, P)$.*

Our first result uses the above and the structure of the ID algorithm [Shpitser and Pearl, 2006, Huang and Valtorta, 2006] to prove a new general upper bound on the condition number of causal identification.

Theorem 4. *Let G be a causal graph with n observed vertices, in which each C-component is of size at most c. Let X, Y be disjoint subsets of observed vertices in G such that $P(Y \mid \mathbf{do}(X))$ is identifiable in G. Then, the condition*

*number of the **ID** expression for $P(Y \mid \mathbf{do}(X))$ output by the **ID** algorithm, for any strictly positive P, is at most $n \cdot \exp(O(c \log c))$.*

This theorem is proved in Section 4. However, one might suspect that the bounds given by the above theorem and the techniques underlying it are sub-optimal for specific instances. We therefore turn towards developing tools for obtaining tighter bounds on the condition number. Our main technical result, Theorem 5, gives an exact numerical description of the condition number of any rational function of a discrete probability distribution, in terms of the gradient of the function.

Theorem 5. *Fix $f : \Delta_\Omega^+ \to \mathbb{R}_+$ and $P \in \Delta_\Omega^+$. Assume that $f(P) > 0$ and f is differentiable at P, with gradient $\nabla f(P) = (d_\omega)_{\omega \in \Omega}$. Then, we have*

$$\kappa(f, P) = \frac{1}{f(P)} \min_{\gamma \in \mathbb{R}} \sum_{\omega \in \Omega} P(\omega) |d_\omega - \gamma|.$$

Remark 6. Note that the minimizing γ in the statement of the theorem is the median of the $(d_\omega)_{\omega \in \Omega}$ with respect to the probability distribution P on Ω.

The proof of this theorem is given in Section 2, and is based on some ideas from numerical analysis combined with an application of linear programming duality. In Section 2, we also give two corollaries (Corollaries 10 and 11) which give potentially more convenient expressions than the one given in Theorem 5, for studying the condition number.

Specializing to the case of causal identification, Theorem 5 already gives an algorithm, analogous to the **ID** algorithm, for estimating the condition number of causal identification: one computes (symbolically) the gradient of the causal identification expression returned by the **ID** algorithm, via an automatic differentiation procedure, and then plugs in numerical values into the expression for the condition number given by Theorem 5 or one of its corollaries (Corollaries 10 and 11). However, instead of pursuing this route, we instead exploit the characterization in these results to show that even in an extremely simple example of causal identification, the condition number (and therefore numerical stability) can show some very subtle behavior.

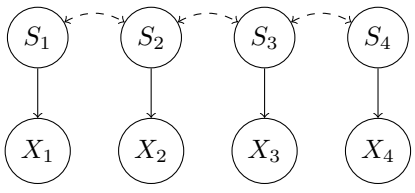

Figure 1: A simple example

The (very simple) example we consider is the following: we have a causal graph M_n with $2n$ observed vertices, labelled $S_1, S_2, \ldots S_n$ and $X_1, X_2, \ldots, X_n$. There is a directed edge $S_i \to X_i$ for each $1 \leq i \leq n$. In addition, there is a bidirected edge between S_i and S_{i+1} for each $1 \leq i \leq n-1$ (Figure 1 shows the case $n = 4$).[1] We denote by X the tuple $(X_1, X_2, \ldots, X_n)$, and by S the tuple $(S_1, S_2, \ldots, S_n)$. We assume further that each of the S_i and the X_i take values in $\{0, 1\}$. The definition of the intervention distribution shows that for a probability distribution P on this graph, we have

$$P(X = \mathbf{0} \mid \mathbf{do}(S = \mathbf{0})) = f(P) := \prod_{i=1}^{n} P(X_i = 0 \mid S_i = 0). \quad (4)$$

The above expression is a product of n terms, each of which is a quotient of two marginals of P. Applying the product and quotient rules (item 1) of Lemma 3 therefore implies that $\kappa(f, P) \leq 2n$, for any P. However, as the next theorem (proved in Section 3 using Theorem 5) shows, this upper bound in not tight even when P is the uniform distribution.

Theorem 7. *Consider the interventional distribution $P(\mathbf{X} = 0 \mid \mathbf{do}(S = \mathbf{0}))$ in the causal graph M_n defined above. There is a constant $a > 0$ (independent of n) such that the condition number $\kappa(f, P)$ for the expression f in eq. (4) for this interventional distribution is at most $a\sqrt{n}$, when P is the uniform distribution.*

Thus, the above example shows that even for a simple example, the condition number of a *given* causal identification expression can be much lower than that predicted by generic upper bounds like those using Lemma 3 and Theorem 4. We now turn to the second subtlety: the difference between condition numbers of *equivalent* causal identification expressions.

Condition Numbers of "Equivalent" Expressions For a particular instance of causal identification, there can be several **ID** expressions which are *equivalent* in the sense that they are all equal when evaluated on a distribution P that lies on the manifold of probability distributions allowed by the causal graph under study, but which are *different* as rational functions. On the other hand, the errors in the *perturbed* version $\tilde{P}$ of P that is provided to such an **ID** expression as an input arise due to phenomena such as rounding to a fixed finite precision, statistical noise, or as in one of the settings considered by Schulman and Srivastava [2016] that was described in the introduction, due to a mis-specification of the causal DAG. One cannot expect such disparate sources of errors to produce only such perturbed inputs $\tilde{P}$ that lie on the manifold allowed by the causal DAG under study. The possibility therefore arises that **ID** expressions that agree on the manifold, but that are different as rational functions, can

[1] Each bi-directed edge is shorthand for a new latent variable with no parents and directed edges to each endpoint of the bi-directed edge.

have very different behaviours when the input is perturbed, which would reflect in their condition numbers being very different. We illustrate this by continuing with the simple example considered in Theorem 7, for which we see using standard d-separation arguments that the following is an **ID** expression *equivalent* (in the above sense) to the one in eq. (4):

$$P(X = \mathbf{0} \mid \mathbf{do}(S = \mathbf{0})) = g(P) := P(X = \mathbf{0} \mid S = \mathbf{0}).$$

Again, f and g agree on every P that lies on the manifold of probability distributions allowed by the model M_n. However, their behaviour on *perturbed* versions of such a P can be widely different. Indeed, as we show in Section 3.1, the upper bound on $\kappa(f, P)$ when P is uniform, obtained in Theorem 7, is in fact tight up to constant factors: there is a constant $a' > 0$ such that for all large enough n, $\kappa(f, P) \geq a'\sqrt{n}$. On the other hand, since $g(P)$ is a ratio of two marginals of P, the sum and quotient rules (items 1 and 2) of Lemma 3 imply that $\kappa(g, P) \leq 2$ for all P!

We therefore see that even in such a simple example, the condition number, and therefore the numerical stability under perturbation, of equivalent **ID** expressions can be quite different. This further highlights the importance of the condition number: in an application, it can serve as a criterion for choosing between various equivalent causal identification expressions. We hasten to add, however, that the condition number need not be the only such criterion for choosing between equivalent intervention expressions: we discuss some important open questions in this direction in the conclusion section of the paper.

2 RATIONAL FUNCTIONS OF PROBABILITY DISTRIBUTIONS

Working towards a proof of Theorem 5, we now study the condition number of differentiable rational functions of discrete probability distributions. We first observe that given $P \in \Delta_\Omega^+$ and $\epsilon > 0$, any member $\tilde{P} \in B_\epsilon(P)$ can be written in terms of feasible solutions to a system of linear inequalities. Recall from Definition 1 that $B_\varepsilon(P) := \left\{\tilde{P} \in \Delta_\Omega \mid \mathrm{Rel}(P, P') \leq \epsilon\right\}$.

Observation 8. *Fix $P \in \Delta_\Omega^+$ and $\epsilon \in (0, 1)$. Then, $\tilde{P} \in B_\epsilon(P)$ if and only if, for all $\omega \in \Omega$,*

$$\tilde{P}(\omega) = P(\omega)(1 + a_\omega \varepsilon), \tag{5}$$

where the real vector $(a_\omega)_{\omega\in\Omega}$ is an element of the set $\mathcal{A}(\epsilon, P)$ defined by the linear inequalities:

$$\sum_{\omega\in\Omega} a_\omega P(\omega) = 0, \quad \textit{and} \tag{6}$$

$$\epsilon \cdot a_\omega P(\omega) \leq 1 - P(\omega), \quad \textit{for all } \omega \in \Omega, \textit{ and} \tag{7}$$

$$-1 \leq a_\omega \leq 1, \quad \textit{for all } \omega \in \Omega. \tag{8}$$

Further, when ϵ satisfies $(1 + \epsilon) \cdot P(\omega) \leq 1$ for all $\omega \in \Omega$, the constraint in eq. (7) can be dropped and we have $\mathcal{A}(P) := \mathcal{A}(0, P) = \mathcal{A}(\epsilon, P)$.

(Here, eq. (6) captures the condition that the entries of $\tilde{P}$ must also sum up to 1, while eqs. (7) and (8) capture the condition that the entries of $\tilde{P}$ must be in $[0, 1]$ with $\mathrm{Rel}(P, \tilde{P}) \leq \epsilon$.)

The above observation, though simple, allows us to considerably simplify the study of condition numbers in our setting. However, before doing so, we need to relate the condition number to the derivatives of $f : \Delta_\Omega^+ \to \mathbb{R}_+$. We now proceed to do so, using arguments similar to those in other numerical analysis applications (see, e.g., [Bürgisser and Cucker, 2013, Section 14.1]). We first introduce some notation. For a differentiable function f, we denote by $\nabla f(P)$ the gradient of f at the point P. For two n-dimensional vectors a and b, $a \circ b$ denotes the co-ordinate-wise (or Schur) product of a and b, i.e., the n-dimensional vector whose ith entry is $a_i b_i$, while $\langle a,\, b\rangle$ denotes the usual inner product.

Lemma 9. *Fix $f : \Delta_\Omega^+ \to \mathbb{R}_+$ and $P \in \Delta_\Omega^+$. Assume that $f(P) > 0$ and f is differentiable at P, with gradient $\nabla f(P)$. Then, $\kappa(f, P) = \max_{a\in\mathcal{A}(P)} \frac{\langle \nabla f(P),\, a\circ P\rangle}{f(P)}$.*

Proof. Given $\epsilon > 0$ and a real vector $a \in \mathcal{A}(\epsilon, P)$, we denote by $\tilde{P}_{a,\varepsilon}$ the probability distribution $\tilde{P}$ defined in Equation (5). In this proof, we will be interested in limits as $\epsilon \downarrow 0$, so we assume without loss of generality that all positive ϵ appearing in the proof are small enough that $(1 + \epsilon) \cdot P(\omega) \leq 1$ for all $\omega \in \Omega$ (which is possible due to our assumption that $P(\omega) < 1$ for all atoms $\omega \in \Omega$), so that $\mathcal{A}(\epsilon, P) = \mathcal{A}(P)$.

Since f is differentiable at P, we can define, for $v \neq 0$, $r(v) := (f(P + v) - f(P) - \langle \nabla f(P),\, v\rangle)/\left\|v\right\|_\infty$. The definition of the differentiability of f at P with gradient $\nabla f(P)$ can then be written as $\lim_{v\to 0} |r(v)| = 0$, so that we can also set $r(0) = 0$. Specializing to our setting, we have, for all small enough non-negative ϵ,

$$\begin{aligned} f(\tilde{P}_{a,\varepsilon}) = f(P) &+ \varepsilon \cdot \langle \nabla f(P),\, a \circ P\rangle \\ &+ \varepsilon r(\varepsilon \cdot (a \circ P)) \left\|a \circ P\right\|_\infty , \end{aligned} \tag{9}$$

where r satisfies $\lim_{v\to 0} |r(v)| = 0$. By definition, we have $\mathrm{Rel}(P, \tilde{P}_{a,\varepsilon}) = \epsilon \cdot \left\|a\right\|_\infty$. Using eq. (9) we then get

$$\begin{aligned} \zeta_f(P, \tilde{P}_{a,\epsilon}) = &\frac{1}{\left\|a\right\|_\infty \cdot f(P)} \\ &\cdot \left| \langle \nabla f(P),\, a \circ P\rangle + r\left(\epsilon \cdot (a \circ P)\right) \cdot \left\|a \circ P\right\|_\infty \right|. \end{aligned}$$

Now, we recall that every $a \in \mathcal{A}(P)$ satisfies $\left\|a\right\|_\infty \leq 1$, and also that $\lim_{v\to 0} |r(v)| = 0$. Using these, we therefore get that

$$\lim_{\epsilon\to 0} \sup_{a\in\mathcal{A}(P)} \left| r\left(\epsilon \cdot (a \circ P)\right) \cdot \left\| \frac{a}{\left\|a\right\|_\infty} \circ P \right\|_\infty \right| = 0.$$

$$\begin{aligned}\underset{\substack{(r_\omega, s_\omega)_{\omega\in\Omega},\\ \gamma\in\mathbb{R}}}{\text{minimize}} \quad & \frac{1}{f(P)}\sum_{\omega\in\Omega}(r_\omega + s_\omega) \\ \text{subject to} \quad & r_\omega - s_\omega + \gamma P(\omega) = d_\omega P(\omega) \quad \forall\omega\in\Omega, \\ & r_\omega, s_\omega \geq 0 \qquad\qquad\quad \forall\omega\in\Omega\end{aligned}$$

(2.a) Original dual linear program in the proof of Theorem 5

$$\begin{aligned}\underset{\substack{(s_\omega)_{\omega\in\Omega},\\ \gamma\in\mathbb{R}}}{\text{minimize}} \quad & \frac{1}{f(P)}\sum_{\omega\in\Omega}((d_\omega - \gamma)P(\omega) + 2s_\omega) \\ \text{subject to} \quad & s_\omega \geq \max\{0, -(d_\omega - \gamma)P(\omega)\} \quad \forall\omega\in\Omega.\end{aligned}$$

(2.b) Simplified dual linear program in the proof of Theorem 5

Figure 2: The linear programs used in the proof of Theorem 5

Using this along with Observation 8 and the above expression for $\zeta_f(P, \tilde{P}_{a,\epsilon})$, we get

$$\begin{aligned}\kappa(f,P) &:= \lim_{\varepsilon\downarrow 0}\sup_{\tilde{P}\in B_\varepsilon(P)}\zeta(P,\tilde{P}) \\ &= \lim_{\epsilon\to 0}\sup_{a\in\mathcal{A}(P)}\zeta_f(P,\tilde{P}_{a,\epsilon}), \text{ by Observation 8} \\ &= \frac{1}{f(P)}\max_{a\in\mathcal{A}(P)}\frac{1}{\|a\|_\infty}\cdot|\langle\nabla f(P),\, a\circ P\rangle|\,. \end{aligned} \tag{10}$$

The latter in turn equals $\max_{a\in\mathcal{A}(P)}\frac{\langle\nabla f(P), a\circ P\rangle}{f(P)}$. To see this, notice that the expression to be maximized in eq. (10) does not change in value when a is multiplied by any non-zero constant, so that we can restrict the maximization, without loss of generality, to $a \in \mathcal{A}(P)$ for which $\|a\|_\infty = 1$. Further, since $a \in \mathcal{A}(P)$ if and only if $-a \in \mathcal{A}(P)$, and the expression to be maximized has the same value for a and $-a$, we can further restrict the maximization to $a \in \mathcal{A}(\mathcal{P})$ for which $\langle\nabla f(P),\, a\circ P\rangle$ is non-negative. □

The utility of Lemma 9 is that, combined with Observation 8, it gives us a prescription for evaluating $\kappa(f,P)$ as a problem of maximizing $\frac{\nabla f(P)(a\circ P)}{f(P)}$ over all a in the polyhedral set $\mathcal{A}(P)$. We now proceed to use standard tools from the theory of linear programming to better understand this problem.

The Condition Number Linear Program We now prove our main technical theorem, Theorem 5.

Proof of Theorem 5. Recall from the statement of the theorem that the $(d_\omega)_{\omega\in\Omega}$ are the coordinates of the gradient of f. Applying Lemma 9, we see that $\kappa(f,P)$ is the solution to the following linear program:

$$\begin{aligned}\underset{(a_\omega)_{\omega\in\Omega}}{\text{maximize}} \quad & \frac{1}{f(P)}\sum_{\omega\in\Omega}a_\omega d_\omega P(\omega) \\ \text{subject to} \quad & -1 \leq a_\omega \leq 1 \quad \forall\omega\in\Omega, \\ & \sum_{\omega\in\Omega}a_\omega P(\omega) = 0.\end{aligned} \tag{11}$$

Note that the program in (11) is a feasible and bounded linear program (e.g., $a = \mathbf{0}$ is feasible), so that strong duality holds. By taking its dual, we therefore get the program in Figure 2.a which also has the objective value $\kappa(f,P)$. The variables r_ω in this dual program can be eliminated using the equality constraints, and the program therefore simplifies to the one in Figure 2.b. The claim then follows since for any real number A, we have $|A| = A + 2\max(0,-A)$. □

The above theorem has two immediate corollaries.

Corollary 10. *With the same notation and hypotheses as in Theorem 5, consider the gradient $\nabla(\log f)(P) = (\delta_\omega)_{\omega\in\Omega}$ of $\log f$ at P, and let M denote the median of the entries of this vector, with respect to the probability distribution $P(\omega)$. Then, $\kappa(f,P) = \sum_{\omega\in\Omega}P(\omega)\,|\delta_\omega - M|$.*

Proof. This follows from the observation that $\delta_\omega = \frac{d_\omega}{f(P)}$, and that the minimizing γ in the expression in Theorem 5 is the median of the vector $(d_\omega)_{\omega\in\Omega}$ with respect to the probability distribution P. □

Corollary 11. *With the same notation and hypotheses as in Corollary 10, we have $\kappa(f,P)^2 \leq \mathrm{Var}_{\omega\sim P}[\delta_\omega]$.*

Proof. We begin by rewriting the result of Corollary 10 as $\kappa(f,P)^2 = \min_M \mathbb{E}_{\omega\sim P}[|\delta_\omega - M|]^2 \leq \min_M \mathbb{E}_{\omega\sim P}[|\delta_\omega - M|^2] = \mathrm{Var}_{\omega\sim P}[\delta_\omega]$. Here, the inequality follows from Jensen's inequality, while the final equality is the definition of the variance (the minimizing M is $\mathbb{E}_{\omega\sim P}[\delta_\omega]$ in this definition). □

3 CONDITION NUMBER UPPER BOUNDS: AN INTERESTING EXAMPLE

In this section we prove Theorem 7, by applying Corollary 11 to the **ID** expression in eq. (4). Recall that we have a collection of $2n$ random variables $S_1, S_2, \ldots, S_n$ and $X_1, X_2, \ldots X_n$, each taking values in $\{0,1\}$. We denote by P their joint distribution on $\Omega = \{0,1\}^{2n}$, and are interested in the condition number of the function $f(P) := \prod_{i=1}^n P(X_i = 0 \mid S_i = 0)$. For notational simplicity, we use, for $1 \leq i \leq n$, $p_{i0} := P(X_i = 0, S_i = 0)$, $p_{i1} := P(X_i = 1, S_i = 0)$, $p_i := P(S_i = 0) = p_{i0} + p_{i1}$,

and $r_i := p_{i1}/p_{i0}$. Now, for any $\mathbf{x}, \mathbf{s} \in \{0,1\}^n$ and $b \in \{0,1\}$, we have

$$\begin{aligned} \frac{\partial p_{ib}}{\partial P(X=\mathbf{x}, S=\mathbf{s})} &= \mathbb{I}[x_i = b, s_i = 0], \text{ and} \\ \frac{\partial p_i}{\partial P(X=\mathbf{x}, S=\mathbf{s})} &= \mathbb{I}[s_i = 0] \end{aligned} \tag{12}$$

where $\mathbb{I} : 2^\Omega \to \{0,1\}$ is the indicator function for an event. From the definition of $f(P)$, eq. (12), and some algebra we get

$$\frac{\partial \log f(P)}{\partial P(X=\mathbf{x}, S=\mathbf{s})} = A_{\mathbf{xs}} - B_{\mathbf{xs}}, \tag{13}$$

where

$$A_{\mathbf{xs}} := \sum_{\substack{i \in [n]: \\ x_i = 0 \\ s_i = 0}} \frac{r_i}{p_i} \quad \text{and} \quad B_{\mathbf{xs}} := \sum_{\substack{i \in [n]: \\ x_i = 1 \\ s_i = 0}} \frac{1}{p_i}.$$

Combined with Corollary 11, eq. (13) then gives

$$\kappa(f,P)^2 \leq \operatorname*{Var}_{(\mathbf{x},\mathbf{s}) \sim P}[A_{\mathbf{xs}} - B_{\mathbf{xs}}]. \tag{14}$$

We now specialize to the setting of the theorem, and fix P to be the uniform distribution. When the pair $(\mathbf{x}, \mathbf{s})$ is sampled according to P, we define the random variables $D_i(\mathbf{x}, \mathbf{s})$, $1 \leq i \leq n$, such that

$$D_i(\mathbf{x}, \mathbf{s}) = \begin{cases} 1 & \text{when } x_i = 0,\ s_i = 0, \\ -1 & \text{when } x_1 = 1,\ s_i = 0, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

When P is uniform, the D_i are independent and identically distributed, and further, each of them have expectation 0 and variance $1/2$. Further, when P is uniform, we also have $p_i = \frac{1}{2}$ and $r_i = 1$ for all $1 \leq i \leq n$. Thus, interpreting $A_{\mathbf{xs}}$ and $B_{\mathbf{xs}}$ also as random variables when $(\mathbf{x}, \mathbf{s})$ is sampled from P, we get

$$A_{\mathbf{xs}} - B_{\mathbf{xs}} = 2 \sum_{i=1}^{n} D_i. \tag{16}$$

Substituting this into eq. (14) gives

$$\kappa(f,P)^2 \leq \operatorname*{Var}_{(\mathbf{x},\mathbf{s}) \sim P}[A_{\mathbf{xs}} - B_{\mathbf{xs}}] = 4 \operatorname*{Var}_{(\mathbf{x},\mathbf{s}) \sim P}[\sum_{i=1}^{n} D_i(\mathbf{x}, \mathbf{s})] \tag{17}$$

$$= 4 \sum_{i=1}^{n} \operatorname*{Var}_{(\mathbf{x},\mathbf{s}) \sim P}[D_i(\mathbf{x}, \mathbf{s})] = 2n. \tag{18}$$

Here, the equality in eq. (17) comes from eq. (16), while the two equalities in eq. (18) come, respectively, from the independence of the D_i when P is uniform, and from a direct calculation of their variance in this case, both of which were argued above. We therefore get $\kappa(f,P) \leq \sqrt{2n}$, and this completes the proof of Theorem 7.

3.1 TIGHTNESS OF THE UPPER BOUND

We now show that the upper-bound on the condition number derived above using Corollary 11 is tight up to constant factors. To do this, we start with the exact characterization derived in Corollary 10. To translate into the notation above, that corollary yields

$$\kappa(f,P) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{s}) \sim P}\left[|A_{\mathbf{xs}} - B_{\mathbf{xs}} - M|\right], \tag{19}$$

where M is the median of quantities $A_{\mathbf{xs}} - B_{\mathbf{xs}}$ when the pair $(\mathbf{x}, \mathbf{s})$ is sampled according to P. When P is uniform, as in the computation above, we see using the representation in eq. (16) that the distribution of $A_{\mathbf{xs}} - B_{\mathbf{xs}}$ is symmetric around zero. This implies that the median M is 0, and using the representation in eq. (16), eq. (19) simplifies to

$$\kappa(f,P) = 2\,\mathbb{E}\left[\left|\sum_{i=1}^{n} D_i\right|\right], \tag{20}$$

where the D_i, as defined in eq. (15), are i.i.d. random variables with mean 0 and variance $1/2$, whose distribution is symmetric around 0. The central limit theorem (see, e.g., Williams [1991, Theorem 18.4]) applied to the sum of the D_i then implies that for any constant $\alpha > 0$,

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}\left[\left|\sum_{i=1}^{n} D_i\right| > \alpha\sqrt{n/2}\right] \\ &= 2 \lim_{n\to\infty} \mathbb{P}\left[\sum_{i=1}^{n} D_i > \alpha\sqrt{n/2}\right] = 2\mathbb{P}\left[Z > \alpha\right], \end{aligned} \tag{21}$$

where $Z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable. (Here, the first equality comes from the fact that the distribution of the D_i is symmetric around 0, while the second equality is from the central limit theorem.) Now, we choose $\alpha > 0$ such that $\mathbb{P}\left[Z > \alpha\sqrt{2}\right] \geq 0.3$. Then, from eq. (21), there exists n_0 such that for all $n > n_0$, $\mathbb{P}\left[|\sum_{i=1}^{n} D_i| > \alpha\sqrt{n}\right] > 0.5$. It follows that for all $n > n_0$, $\mathbb{E}\left[|\sum_{i=1}^{n} D_i|\right] > 0.5\alpha\sqrt{n}$. Plugging this into eq. (20), we see that whenever $n \geq n_0$, $\kappa(f,P) \geq \alpha\sqrt{n}$, which shows that the upper-bound obtained in Theorem 7 is tight up to constant factors.

4 GENERIC UPPER BOUNDS ON THE CONDITION NUMBER

In this section we consider graphs in which the sizes of all C-components are bounded, and prove Theorem 4.

Proof of Theorem 4. For reference, we reproduce the **ID** algorithm of Shpitser and Pearl [2006] (see Algorithm 1).

[2] We give a brief glossary of the notation used in algorithm 1 in Supplementary Material Section A.2.

Algorithm 1 The **ID** algorithm (reproduced for reference, essentially verbatim, from Shpitser and Pearl [2006])[2]

function $\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G)$
1: if $\mathbf{x} = \varnothing$, return $\sum_{\mathbf{v} \setminus \mathbf{y}} P(\mathbf{v})$.
2: if $\mathbf{V} \neq \mathrm{An}(\mathbf{Y})_G$,
return $\mathbf{ID}(\mathbf{y}, \mathbf{x} \cap \mathrm{An}(\mathbf{Y})_G, P(\mathrm{An}(\mathbf{Y})), \mathrm{An}(\mathbf{Y})_G)$.
3: let $\mathbf{W} = (\mathbf{V} \setminus \mathbf{X}) \setminus \mathrm{An}(\mathbf{Y})_{G_{\overline{\mathbf{X}}}}$.
if $\mathbf{W} \neq \varnothing$, return $\mathbf{ID}(\mathbf{y}, \mathbf{x} \cup \mathbf{w}, P, G)$.
4: if $C(G \setminus \mathbf{X}) = \{S_1, \ldots, S_k\}$ (for $k \geq 2$),
return $\sum_{v \setminus (\mathbf{y} \cup \mathbf{x})} \prod_i \mathbf{ID}(s_i, \mathbf{v} \setminus s_i, P, G)$.
else if $C(G \setminus \mathbf{X}) = \{S\}$,
5: if $C(G) = \{G\}$, throw $\mathbf{FAIL}(G, S)$.
6: if $S \in C(G)$, return $\sum_{s \setminus \mathbf{y}} \prod_{V_i \in S} P(v_i \mid v_\pi^{(i-1)})$.
7: if $\exists S', S \subseteq S' \in C(G)$,
return $\mathbf{ID}(\mathbf{y}, \mathbf{x} \cap S', \prod_{V_i \in S'} P(V_i \mid V_\pi^{(i-1)} \cap S', v_\pi^{(i-1)} \setminus S', S')$.

We recall that the **ID** algorithm can be viewed as operating symbolically: on input $\mathbf{x}, \mathbf{y}$ and G, the algorithm either outputs **FAIL** and determines the required intervention to be non-identifiable, or otherwise outputs a rational function $\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G)$, in terms of entries of the observed distribution P, such that if the (exact) numerical value of such an observed distribution P is plugged in to the function $\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G)$, the resulting value is $P(Y = \mathbf{y} \mid \mathbf{do}(X = \mathbf{x}))$.

However, the **ID** algorithm produces this symbolic expression for the rational function $\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G)$ through a recursive algorithm. We use the bounds from Lemma 3 to carefully analyze the effects of various steps of this recursive algorithm.

We first consider the return statements on lines 1 and 6 of the algorithm, which do not lead to a recursive call. If the algorithm returns immediately on line 1, item 2 of Lemma 3, dealing with sums immediately gives

$$\kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G), P) \leq 1, \tag{22}$$

since each of the terms in the sum on line 1 trivially has condition number at most 1.

For line 6, we similarly apply the bounds for sums along with the product and quotient rules (item 1 of Lemma 3). First, the sum rule, combined with the quotient rule gives that the condition number for each of the factors in the product on line 6 is at most 2. The product rule (followed by the sum rule for the marginalization over the values in $\mathbf{s} \setminus \mathbf{y}$) then gives that if the algorithm returns directly on line 6 (without making any recursive calls), then

$$\kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G), P) \leq 2|S| \leq 2c, \tag{23}$$

where the last inequality uses that S, being C-connected, must be contained in a C-component of G, and must therefore be of size at most c.

We now consider the behavior of the algorithm under recursive calls. The simplest case is of the recursive call on line 3. In this case, there is no modification of P, and hence we directly get $\kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G), P) = \kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x} \cup \mathbf{w}, P, G), P)$. For the case of the recursive call on line 2, we note that the condition number of the map $P \mapsto P(\mathrm{An}(\mathbf{Y}))$ is at most 1: this follows exactly as in eq. (22), since each component of this map is a marginalization of P. Thus, applying item 3 of Lemma 3 gives (with $Q := P(\mathrm{An}(\mathbf{Y}))$).

$$\kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G), P) \leq \kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x} \cap \mathrm{An}(\mathbf{Y})_G, Q, \mathrm{An}(\mathbf{Y})_G), Q), \tag{24}$$

For analyzing the remaining recursive calls on lines 4 and 7 of the algorithm, it will be convenient to define the notion of a *full* instance. We say that an input instance $(\mathbf{y}, \mathbf{x}, P, G)$ to the **ID** algorithm is a *full* instance if the following two conditions are satisfied: 1. The set of observed vertices $T := G \setminus X$ is C-connected, and 2. For every vertex $u \in T \setminus Y$, there is a directed path from u to some vertex in Y that does *not* pass through any vertex in X.

We now observe two properties of full instances.

1. *The instances produced by the recursive calls on line 4 of the algorithm are full instances.* This follows because the S_i on line 4 are C-connected by definition, and since all the target vertices are already included in S_i (so that the second condition above is vacuously satisfied).
2. *All recursive calls from a full instance are to full instances.* Suppose $(\mathbf{y}, \mathbf{x}, P, G)$ is a full instance. The recursive calls on lines 2 and 7 are also easily seen to be full instances, since, by definition, all vertices outside $\mathbf{X}$ have directed paths to $\mathbf{Y}$ (that do not go through $\mathbf{X}$) in the original full instance. The same observation then implies that the recursive calls on lines 3 and 4 *cannot* occur for a full instance, since (i) $\mathbf{W}$ as defined on line 3 would be empty for full instances, and (ii) the number of C-components in $G \setminus \mathbf{X}$ would be at most 1 for a full instance, so that the conditions for the recursive calls on lines 3 and 4 are both false.

Note that the last point also implies that any recursive call on a full instance *cannot increase the size of the* $\mathbf{X}$ *set* (since the recursive calls on lines 3 and 4 are the only ones that can do so). Now, let $\xi(\ell, m) \geq 1$ denote the maximum of $\kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G), P)$ over all positive P and all full instances $(\mathbf{y}, \mathbf{x}, P, G)$ in which the number of vertices in $\mathbf{X}$ is at most m and the number of vertices outside $\mathbf{X}$ is at most ℓ (recall that by definition of a full instance, these latter set of vertices form a C-connected set in G).

Now consider an arbitrary instance $(\mathbf{y}, \mathbf{x}, P, G)$ for which the **ID** algorithm makes the recursive calls on line 4. Since all these calls are to full instances, and since each S_i on line 4 is C-connected and hence of size at most c, an application

of the product and sum rules (items 1 and 2) of Lemma 3 yields

$$\kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G), P) \leq n \cdot \xi(c, n). \quad (25)$$

Thus, in conjunction with eqs. (22) and (23), this shows that it suffices to understand the condition number of full instances. Now, as argued in point 2 above, only the return statements on lines 1, 2, 6, and 7 can occur for a full instance (i.e., the return statements on lines 3 and 4 cannot occur). From line 1 and eq. (22), we get that

$$\xi(a, 0) = 1 \text{ for all } 1 \leq a \leq c. \quad (26)$$

Equation (24) shows that the effect of line 2 on the condition number for a full instance is already built into the definition of ξ, as the latter is separately increasing in both its arguments. The effect of line 6 has already been considered in eq. (23). The same equation also shows that the condition number of the map $P \mapsto \prod_{V_i \in S'} P(V_i \mid V_\pi^{(i-1)} \cap S', v_\pi^{(i-1)} \setminus S', S')$ used in line 7 is at most $2c$. Note also that the set S' in line 7, being C-connected, is of size at most c. Further, for line 7 to execute, it must be the case that $X \cap S' \neq X$ (for otherwise, line 5 would execute and the algorithm would fail, contrary to our assumption that we have an identifiable instance). Thus, we have $|X \cap S'| \leq \min\{|X| - 1, c\}$. Using the composition rule (item 3) of Lemma 3, and combining the effects of lines 6 and 7, we therefore get

$$\xi(a, m) \leq \max\{2c, 2c \cdot \xi(a, \min\{m - 1, c\})\}, \quad (27)$$

for all $1 \leq a \leq c$, and $1 \leq m \leq n$. From eqs. (26) and (27), we therefore get $\xi(a, m) \leq (2c)^{c+1}$, for all $1 \leq a \leq c$, and $1 \leq m \leq n$. Combining with eqs. (22) to (25), this gives the claimed upper bound: $\kappa(\mathbf{ID}(\mathbf{y}, \mathbf{x}, P, G), P) \leq n \cdot (2c)^{c+1}$, when all C-components in G are of size at most c, and G has n observed vertices. □

5 CONCLUSION

This paper develops tools for understanding the numerical stability of causal inference, in terms of the condition number. Starting from the result that causal inference is not "too unstable" in the important special case when all C-components are of small size, these tools are then used to show that the condition number may behave in a non-trivial manner even in an example as simple as the one in Figure 1. In particular, the condition number may be much lower than what a simple upper-bound predicts, and can also depend crucially on which of several "equivalent" expressions are used for estimating the causal effect in question. As alluded to earlier, combined with techniques for automatic differentiation, the tools also give a procedure for algorithmically computing the condition number of any given instance, albeit with a running time that is polynomial in the size of the table of observed marginals (which is typically exponential in the size of the input graphical model).

These results lead to several important open problems. Perhaps the most obvious is to improve upon the generic algorithm for computing instance-specific condition numbers alluded to above. A more open-ended question is to develop a better understanding of how to choose between different equivalent **ID** expressions. The results of this paper point to the condition number being one of the possible criteria for this choice. However, we also emphasize that it is not the *only* criterion. For example, different expressions might also need to be evaluated in terms of the feasibility of accurate statistical estimation of some of the marginals or conditionals appearing in them. Similarly, previous work has also looked at the use of measurements of certain interventional distributions in identification expressions for other interventional expressions (see, e.g., Tian and Pearl [2000], Shpitser and Pearl [2008]) and also at methods for simplifications of **ID** expressions [Tikka and Karvanen, 2017, 2018]. It would be very interesting to combine the tools from this paper with the ideas developed in the above papers, in order to work out a notion of a "best" **ID** expression for a given instance.

Author Contributions

The authors are listed in alphabetical order of last names.

Acknowledgements

We thank anonymous reviewers at UAI for several helpful comments. VMK acknowledges support from the Stephen Adelman Memorial SURF endowment and NSF grant CCF-1909972. LJS acknowledges support from NSF grant CCF-1909972. PS acknowledges support from DAE, Govt. of India under project no. RTI4001, from the Ramanujan Fellowship of SERB, from Adobe Systems Incorporated via a gift to TIFR, and from Infosys foundation, through its support for the Infosys-Chandrasekharan virtual center for Random Geometry. The contents of this paper do not necessarily reflect the views of the funding agencies listed above.

References

Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *Proc. 32nd Conference on Neural Information Processing Systems (NeurIPS)*, pages 9469–9481, 2018. URL https://proceedings.neurips.cc/paper/2018/file/78631a4bb5303be54fa1cfdcb958c00a-Paper.pdf.

Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, Ashwin Maran, and N. Variyam Vinodchandran. Learning and sampling of atomic interventions from observations. In *Proc. 37th International Conference on*

Machine Learning (ICML), volume 119 of *PMLR*, pages 842–853, 2020. URL http://proceedings.mlr.press/v119/bhattacharyya20a.html.

Peter Bürgisser and Felipe Cucker. *Condition: The Geometry of Numerical Algorithms*. Springer, 2013.

Yimin Huang and Marco Valtorta. Pearl's Calculus of Intervention Is Complete. In *Proc. 22nd Uncertainty in Artificial Intelligence Conference (UAI)*, pages 217–224, 2006.

Yimin Huang and Marco Valtorta. On the completeness of an identifiability algorithm for semi-Markovian models. *Annals of Mathematics and Artificial Intelligence*, 54: 363–408, 2008.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, December 1995. URL http://biomet.oxfordjournals.org/content/82/4/669.

Karthik Abhinav Sankararaman, Anand Louis, and Navin Goyal. Stability of Linear Structural Equation Models of Causal Inference. In *Proc. 35th Uncertainty in Artificial Intelligence Conference (UAI)*, volume 115, pages 323–333. PMLR, August 2020a. URL http://proceedings.mlr.press/v115/sankararaman20a.html.

Karthik Abinav Sankararaman, Anand Louis, and Navin Goyal. Robust Identifiability in Linear Structural Equation Models of Causal Inference, July 2020b. URL http://arxiv.org/abs/2007.06869.

Leonard J. Schulman and Piyush Srivastava. Stability of Causal Inference. In *Proc. 32nd Uncertainty in Artificial Intelligence Conference (UAI)*, pages 666–675, June 2016. URL http://auai.org/uai2016/proceedings/papers/214.pdf.

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

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, June 2008. URL http://dl.acm.org/citation.cfm?id=1390681.1442797.

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

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. In *Proc. 16th Uncertainty in Artificial Intelligence Conference (UAI)*, pages 589–598, 2000.

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

Santtu Tikka and Juha Karvanen. Simplifying probabilistic expressions in causal inference. *Journal of Machine Learning Research*, 18(36):1–30, 2017. URL http://jmlr.org/papers/v18/16-166.html.

Santtu Tikka and Juha Karvanen. Enhancing identification of causal effects by pruning. *Journal of Machine Learning Research*, 18(194):1–23, 2018. URL http://jmlr.org/papers/v18/17-563.html.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

Sewall Wright. The Method of Path Coefficients. *The Annals of Mathematical Statistics*, 5:161–215, 1934. URL http://www.jstor.com/stable/2957502.