Causal Inference with Information Fields

Benjamin Heymann
Criteo AI Lab, Paris, France
b.heymann@criteo.com

Michel De Lara
CERMICS, École des Ponts, Marne-la-Vallée, France
michel.delara@enpc.fr

Jean-Philippe Chancelier
CERMICS, École des Ponts, Marne-la-Vallée, France
jpc@cermics.enpc.fr

Abstract

Inferring the potential consequences of an unobserved event is a fundamental scientific question. To this end, Pearl's celebrated do-calculus provides a set of inference rules to derive an interventional probability from an observational one. In this framework, the primitive causal relations are encoded as functional dependencies in a Structural Causal Model (SCM), which maps into a Directed Acyclic Graph (DAG) in the absence of cycles. In this paper, by contrast, we capture causality without reference to graphs or functional dependencies, but with information fields. The three rules of do-calculus reduce to a unique sufficient condition for conditional independence: the topological separation, which presents some theoretical and practical advantages over the d-separation. With this unique rule, we can deal with systems that cannot be represented with DAGs, for instance systems with cycles and/or 'spurious' edges. We provide an example that cannot be handled – to the extent of our knowledge – with the tools of the current literature.

1 Introduction

As the world shifts toward more and more data-driven decision-making, causal inference is taking more space in applied sciences, statistics and machine learning. This is because it allows for better, more robust decision-making, and provides a way to interpret the data that goes beyond correlation Pearl and Mackenzie [2018]. For instance, causal inference provides a language to describe and solve Simpson's paradox, which embodies the "correlation is not causation" principle as can be found in any "Statistics 101" basic course. The main concern in causal inference is to compute post-intervention probability distributions from observational data. For this purpose, graphical models are practical because they allow representing assumptions easily and benefit from an extensive scientific literature.

In his seminal work, Pearl builds on graphical models Cowell et al. [2006] to introduce the so-called do-calculus. Several extensions to this do-calculus have been proposed recently Winn [2012], Lattimore and Rohde [2019], Tikka et al. [2019a], Correa and Bareinboim [2020]. As asserted by Pearl, language is an important element in this research program Pearl [2010]. Causal graphical models move the focus from joint probability distributions to functional dependencies thanks to the Structural Causal Model (SCM) framework. By leveraging the concept of information sets, we bring a new, complementary view to the causal reasoning toolbox.

So, we introduce a general, unifying framework for causal inference that may be used for both recursive and nonrecursive systems Halpern [2000] (i.e. with and without cycles). But the cost for this conceptual generalization is a bit of abstraction: in what we propose, the structure is implicit, and there are no arrows. In particular, while DAGs modeling does not rely directly on random variables

Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020.

but on joint probability distributions (see Pearl [2011], footnote 3 or Peters et al. [2017] Appendix A), our approach requires going back to the classical primitives of probabilistic models: sample sets, σ-fields, measurable maps and random variables.

The present paper, however, has been written so that the main messages can be understood with the usual graphical concepts used in the field of causal inference. In particular, the notion of Topological Separation is explained for the specific case of DAGs, Theorem 5 and Examples 4, 5 and 6 should be readable without the concept of information field.

Roughly speaking the information fields are useful in this paper to introduce the notions of (1) well-posedness (Remark 2), (2) context specific independence (3) intervention variable and (4) conditional precedence. Moreover, they constitute a key technical concept for the proofs (see the companion paper Heymann et al. [2020]).

Related work and contributions. We extend the causal modeling toolbox thanks to two notions: information fields and topological separation. The concept of information field extends the expressiveness of the Structural Causal Model, and allows for instance to naturally encode context specific independence Tikka et al. [2019a]. The topological separation is practical because it just requires to check that two sets are disjoints (see the last three examples). By contrast, the d-separation requires to check that *all* the paths that connect two variables are blocked. Moreover, as its name suggests, the topological separation has a theoretical interpretation. Our main result is Theorem 5, which is a generalization of do-calculus that can be applied in particular to nonrecursive systems Bongers et al. [2020], and which subsumes several recent results. We pinpoint the novelty of our approach with our last example, a system with cycles where our framework identifies a probabilistic independence that the framework developed in Forré and Mooij [2020] (for cycles) does not.

For the sake of readability we state and illustrate some of our key results without proofs, which are provided in the companion paper Heymann et al. [2020] alongside other results. Section 2 provides a few reminders on probability theory, and then explains how one can move from the standard Structural Causal Model (SCM) to our proposal of Information Dependency Model (IDM) with the help of information fields. Section 3 introduces the notion of conditional precedence, which in particular allows us to encode intervention variables in the IDM. Section 4 contains the definition of Topological Separation. Section 5 contains our main result, which states that Topological Separation implies conditional independence. We then explain why this theorem subsumes several recent results and Pearl's do-calculus. We also provide an example for which a recently published paper Forré and Mooij [2020] on causality and cycles does not identify a conditional independence, but our framework does.

2 Information fields

We start with a few reminders from probability theory. A *σ-field* (henceforth sometimes referred to as *field*) over a set $\mathbb{D}$ is a subset $\mathcal{D} \subset 2^{\mathbb{D}}$, containing $\mathbb{D}$, and which is stable under complementation and countable union. The trivial σ-field over the set $\mathbb{D}$ is $\{\varnothing, \mathbb{D}\}$. The complete σ-field over the set $\mathbb{D}$ is $2^{\mathbb{D}}$. When $\mathcal{D}' \subset \mathcal{D}$ are two σ-fields over the set $\mathbb{D}$, we say that $\mathcal{D}'$ is a *subfield* of $\mathcal{D}$. If two sets $\mathbb{D}_1, \mathbb{D}_2$ are equipped with σ-fields $\mathcal{D}_1, \mathcal{D}_2$, we denote by $\mathcal{D}_1 \otimes \mathcal{D}_2$ the *product σ-field* on $\mathbb{D}_1 \times \mathbb{D}_2$ generated by the rectangles $\{D_1 \times D_2 | D_i \in \mathcal{D}_i\}$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathbb{U}, \mathcal{U})$ a set and a σ-field over this set. Probability theory defines a *random variable* as a measurable mapping from $(\Omega, \mathcal{F})$ to $(\mathbb{U}, \mathcal{U})$.

Let $\mathbb{A}$ be a finite set and $\Omega = \times_{a\in\mathbb{A}}\Omega_a$ a sample space and $\mathcal{F}_a$ be a σ-field over Ω_a. Let $\Pr_a$ a probability over $(\Omega_a, \mathcal{F}_a)$ and $\Pr = \bigotimes_{a\in\mathbb{A}} \Pr_a$. Let $(\mathbb{U}_a)_{a\in\mathbb{A}}$ be a collection of sets, with $\mathcal{U}_a$ being a σ-field over $\mathbb{U}_a$ for all a. We are interested in some random variables $(U_a)_{a\in\mathbb{A}}$ such that U_a is valued in $\mathbb{U}_a$.

It is standard to model causal hypotheses using **Structural Causal Models (SCMs)**Peters et al. [2017]. An SCM consists of a list of assignments $(\lambda_a)_{a\in A}$ alongside a parental mapping $P : \mathbb{A} \to 2^{\mathbb{A}}$ such that

$$U_a(\omega) = \lambda_a(U_{P(a)}(\omega), \omega_a) \quad \forall \omega \in \Omega \quad \forall a \in \mathbb{A} \tag{1}$$

where ω_a is the projection of ω on Ω_a.

To get the **graphical representation** of an SCM –in $(\mathbb{A}, \mathbb{A} \times \mathbb{A})$– we draw an arrow $a \to b$ whenever $a \in P(b)$. Usually, the graphical representation is assumed to be a DAG, which means that the parental mapping induces a partial order. We will not need this hypothesis here. More often than not, the reasoning is made on the graphical representation which is uniquely defined by the parental mapping, so that the assignements functions do not even need to be specified. For a given applied problem, the SCM is derived from expert knowledge, assumptions and data analysis methods. The SCM is a central tool in causal analysis but its graphical representation does not naturally account for situations such as Context Specific Independence Tikka et al. [2019a], where some edges are spurious.

We may wish for more flexibility, so we call *configuration space* the product space [1]

$$\mathbb{H} = \Omega \times \prod_{a \in \mathbb{A}} \mathbb{U}_a \ . \tag{2}$$

The *configuration field* $\mathcal{H} = \mathcal{F} \otimes \bigotimes_{a \in \mathbb{A}} \mathcal{U}_a$ is a σ-field over $\mathbb{H}$ (with $\mathcal{F} = \bigotimes_{a \in \mathbb{A}} \mathcal{F}_a$). We then extend the definition of SCM thanks to the following observation: we can express the SCM by saying that λ_a is a map from $\mathbb{H}$ to $\mathbb{U}_a$ ($\lambda_a : (\mathbb{H}, \mathcal{H}) \to (\mathbb{U}_a, \mathcal{U}_a)$) while imposing that λ_a "only depends on $U_{P(a)}$ and ω_a". It is standard (see [Dellacherie and Meyer, 1975, Chap. 1 p. 18]) in probability theory that such property is – under mild assumptions– equivalent to a measurability constraint on the random variable U_a. Hence (1) can be restated as

$$\lambda_a^{-1}(\mathcal{U}_a) \subset \mathcal{F}_a \otimes \bigotimes_{b \neq a} \{\varnothing, \Omega_b\} \otimes \bigotimes_{b \in P(a)} \mathcal{U}_b \otimes \bigotimes_{b \notin P(a)} \{\varnothing, \mathbb{U}_b\}, \tag{3}$$

or, with a slight abuse of notations that we will sometimes use throughout this presentation[2]

$$\lambda_a^{-1}(\mathcal{U}_a) \subset \mathcal{F}_a \otimes \bigotimes_{b \in P(a)} \mathcal{U}_b \ . \tag{4}$$

Informally, an information field is anything one may want to see on the right-hand side of Equation (4). For instance, consider the case where $\mathbb{A} = \{a, b, c\}$. If $\lambda_a^{-1}(\mathcal{U}_a) \subset \mathcal{F}_a \otimes \{\varnothing, \Omega_b\} \otimes \{\varnothing, \Omega_c\} \otimes \{\varnothing, \Omega_a\} \otimes \{\varnothing, \Omega_b\} \otimes \{\varnothing, \Omega_c\}$, that we abusively write $\lambda_a^{-1}(\mathcal{U}_a) \subset \mathcal{F}_a$, this means that $\lambda_a(\omega_a, \omega_b, \omega_c, u_a, u_b, u_c) = \lambda_a(\omega_a, \cancel{\omega_b}, \cancel{\omega_c}, \cancel{u_a}, \cancel{u_b}, \cancel{u_c})$ only depends on ω_a, that is, only depends on its own "source of uncertainty" (the field $\mathcal{F}_a$). If (abusively) $\lambda_b^{-1}(\mathcal{U}_b) \subset \mathcal{U}_c \otimes \mathcal{F}_a$, this means that $\lambda_b(\omega_a, \omega_b, \omega_c, u_a, u_b, u_c) = \lambda_b(\omega_a, \cancel{\omega_b}, \cancel{\omega_c}, \cancel{u_a}, \cancel{u_b}, u_c)$ only depends on (ω_a, u_c), that is, only depends on the uncertainty ω_a

(the field $\mathcal{F}_a$) and on the variable u_c (the field $\mathcal{U}_c$). More complex examples will be given later.

We thus extend the definition for SCMs. We propose the name *Information Dependency Model*

Definition 1 (Information Dependency Model)**.** *An Information Dependency Model is a collection $(\mathcal{I}_a)_{a \in \mathbb{A}}$ of subfields of $\mathcal{H}$ such that, for $a \in \mathbb{A}$, $\mathcal{I}_a \subset \mathcal{F}_a \otimes \bigotimes_{b \in \mathbb{A}} \mathcal{U}_b$. The subfield $\mathcal{I}_a$ is called the* ***information field*** *of a.*

The SCM is now defined by the **measurability property**

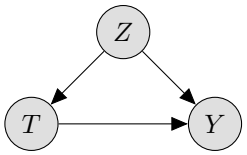
$$\lambda_a^{-1}(\mathcal{U}_a) \subset \mathcal{I}_a \quad \forall a \in \mathbb{A} \ . \tag{5}$$

Property (5) expresses in a very general way that the random variable U_a may only depend upon the information $\mathcal{I}_a$ available to the random variable. For a given applied problem, like for the SCM, the IDM can be derived from expert knowledge, assumptions and data analysis methods. In particular, any SCM can be mapped into an IDM.

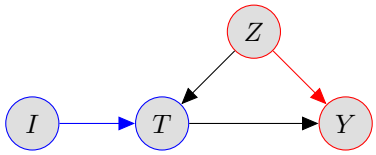
Example 1 (Common cause)**.** *First, to better understand how DAGs – and more generally SCMs can be modeled with information fields, we provide a detailed instance for a set of random variables that can be represented by the DAG in Figure 1.* ***Such an effort is not required in practice****, because the measurability properties are fully specified by the DAG for such a simple instance.*

[1] also called *hybrid space* Witsenhausen [1971], hence the $\mathbb{H}$ notation

[2] we omit the trivial fields in the product



(a) DAG.



(b) Modeling intervention with the intervention variable I.

Figure 1: Common cause.

Let $\mathbb{A} = \{Z, T, Y\}$. To simplify the exposition, we suppose that the values of each of the three random variables represented on the DAG belong to $\{0,1\}$. Then, $\mathbb{U}_Z = \mathbb{U}_T = \mathbb{U}_Y = \{0,1\}$, each equipped with the complete field $\mathcal{U}_Z = \mathcal{U}_T = \mathcal{U}_Y = \{\varnothing, \{0\}, \{1\}, \{0,1\}\}$. We take $\Omega = \{0,1\}^3$ as Nature set, equipped with the complete field $\mathcal{F} = 2^\Omega$ made of all subsets of Ω. We write $\Omega = \Omega_Z \times \Omega_T \times \Omega_Y$, where $\Omega_Z = \Omega_T = \Omega_Y = \{0,1\}$, and $\mathcal{F} = \mathcal{F}_Z \otimes \mathcal{F}_T \otimes \mathcal{F}_Y$, where $\mathcal{F}_Z = \mathcal{F}_T = \mathcal{F}_Y = \{\varnothing, \{0\}, \{1\}, \{0,1\}\}$. To represent, for instance, the arrows pointing to Y in the DAG in Figure 1 (as well as implicit assumptions about information on Nature), we require that the information field $\mathcal{I}_Y$ satisfies $\mathcal{I}_Y \subset \{\varnothing, \Omega_Z\} \otimes \{\varnothing, \Omega_T\} \otimes \mathcal{F}_Y \otimes \mathcal{U}_Z \otimes \mathcal{U}_T \otimes \{\varnothing, \mathbb{U}_Y\}$. This relation expresses that the information of Y depends at most on its own "source of uncertainty" (the field $\mathcal{F}_Y$) and on the decisions of both Z and T (the field $\mathcal{U}_Z \otimes \mathcal{U}_T$). Again, the effort of describing explicitly the information field is not required in the case of DAGs, because the mapping from DAGs to IDMs is trivial. On the other hand the IDM allows to express more sophisticated hypotheses.

Remark 2 (Solvability)**.** *When the system is recursive –i.e. when it admits a fixed causal ordering and can be represented by a DAG – there is no question of well-posedness. One can simulate a sample of random variables by first generating the variables that do not have ancestors, and then following the graph along their descendants. Such procedure is not possible for the more general case of nonrecursive systems, and we need an additional property to ensure that the system is well defined: we present the* ***solvability*** *property in the companion paper Heymann et al. [2020]. We need in particular to exclude cases such as self-information (that is $a \in P(a)$), and more generally case where the system of equations* (1) *could have several solutions (consider for instance $x := y$ and $y := x$) or no solution at all .*

3 Conditional precedence

In this section, we exploit the flexibility of the information field to extend the definition of precedence. For any $B \subset \mathbb{A}$, let $\mathcal{H}_B = \mathcal{F} \otimes \bigotimes_{b \in B} \mathcal{U}_b \subset \mathcal{H}$. In our extended definition of SCM (the Information Dependency Model), we do not specify a precedence relation: the primitives are the information fields, and the notion of precedence is deduced from those fields. For instance, the traditional *precedence relation* – or parental relation– on $\mathbb{A}$ writes

$$\mathcal{P}a = \bigcap_{B \in \mathbb{A}; \mathcal{I}_a \subset \mathcal{H}_B} B. \tag{6}$$

One can check that $P(a) = \mathcal{P}a$ when $P(a)$ is minimal: the relation $U_a(\omega) = \lambda_a(U_{P(a)}(\omega), \omega_a)$ implies that $\mathcal{I}_a \subset \mathcal{H}_{P(a)}$, moreover the minimality means that it is the smallest subset of $\mathbb{A}$ satisfying such constraint. So on a DAG, $b \in \mathcal{P}a$ means that there is an arrow from b to a.

Here is how the information field allows to extend the definition of precedence.

Definition 3 (Conditional Precedence)**.** *For any subset $H \subset \mathbb{H}$ of configurations, and any subset $W \subset \mathbb{A}$, we set*

$$\mathcal{P}_{W,H}a = \bigcap_{B \in \mathbb{A}; \mathcal{I}_a \cap H \subset \mathcal{H}_{B \cup W}} B\ , \tag{7}$$

and call it the precendence conditioned on (W, H).

Tikka et al. Tikka et al. [2019a] manage to summarize the three rules of do-calculus thanks to the notions of context specific independence and labeled DAGs. Our definition allows us to reproduce their approach.

Example 2 (Context Specific Independence)**.** *In order to model spurious edges, Tikka et al. Tikka et al. [2019a] rely on so-called* labeled DAGs *that can be turned into a context specific DAG by removing the arcs that are deactivated (spurious) in the context of interest. In the formalism that we propose, such context is represented by a subset of $\mathbb{H}$. Indeed, if we denote by $H \in \mathbb{H}$ the context for which an arc (a, b) is deactivated (in the language of Tikka et al. [2019a]), we encode this by the following two properties*

$$a \notin \mathcal{P}_{\varnothing,H} b \tag{8a}$$

$$a \in \mathcal{P}_{\varnothing,\mathbb{H}\setminus H} b \ . \tag{8b}$$

Such a property can be encoded in the information set of b $\mathcal{I}_b$.

For the reader familiar with Tikka et al. [2019a], it is then easy to guess how we are going to model intervention variables.

Example 3 (Intervention variables)**.** *To introduce the possibility to intervene on a variable, we use a simple procedure. Suppose we are interested in an intervention profile $\hat{\lambda}_Z$ for a subset $Z \subset \mathbb{A}$. For this purpose, we consider a new family of fields $\hat{\mathcal{I}}_z \subset \mathcal{H}$, for $z \in Z$ and we suppose that $\hat{\lambda}_Z$ is $\hat{\mathcal{I}}_z$-measurable, for any $z \in Z$. Then, we enrich the model as follows: (i) we introduce a new* intervention variable *I, equipped with $\Omega_I = \{0, 1\}$ and $\mathbb{U}_I = \{0, 1\}$, and who only has access to its private information in Ω_I; (ii) we straightforwardly adapt the information fields for $\mathbb{A} \setminus (Z \cup I)$ and the probability $\mathbb{P}$; (iii) we replace the information field $\mathcal{I}_z$ by $\big(\{0\} \otimes \mathcal{I}_z\big) \cup \big(\{1\} \otimes \hat{\mathcal{I}}_z\big)$, for $z \in Z$.*

More formally, we introduce the new model $\big(\tilde{\mathbb{A}}, (\tilde{\Omega}, \tilde{\mathcal{F}}), (\tilde{\mathbb{U}}_a, \tilde{\mathcal{U}}_a)_{a\in\tilde{\mathbb{A}}}, (\tilde{\mathcal{I}}_a)_{a\in\tilde{\mathbb{A}}}\big)$, where $\tilde{\mathbb{A}} = \mathbb{A} \cup \{I\}$, $\tilde{\Omega} = \Omega \times \{0, 1\}$, $\tilde{\mathbb{U}}_I = \{0, 1\}$, $\tilde{\mathbb{U}}_a = \mathbb{U}_a$ for any $a \in \mathbb{A}$, and

$$\tilde{\mathcal{I}}_a = \mathcal{I}_a \otimes \{\varnothing, \mathbb{U}_I\} \ , \quad \forall a \in \mathbb{A} \setminus Z \ , \tag{9a}$$

$$\tilde{\mathcal{I}}_z = \hat{\mathcal{I}}_z \otimes \mathcal{U}_I \ , \quad \forall z \in Z \ , \tag{9b}$$

$$\mathcal{I}_I = \bigotimes_{a\in\mathbb{A}} \{\varnothing, \Omega_a\} \otimes \big\{\varnothing, \{0\}, \{1\}, \{0, 1\}\big\} \otimes \bigotimes_{a\in\mathbb{A}} \{\varnothing, \mathbb{U}_a\} \ . \tag{9c}$$

We also extend the probability $\mathbb{P}$ as a product probability $\tilde{\mathbb{P}} = \mathbb{P} \otimes \mu$ on $\tilde{\Omega}$, where μ is a full support probability on $\{0, 1\}$.

4 Topological separation

Next we introduce the topological separation, which can be seen generalization of the d-separation.

For any subsets $B \subset \mathbb{A}$ and $B_j \subset \mathbb{A}$, $j = 1, \ldots, n$, we write $B_1 \sqcup \cdots \sqcup B_n = B$ when we have both $B_j \cap B_k = \varnothing$ for all $j \neq k$ and $B_1 \cup \cdots \cup B_n = B$. In addition, we denote by $\overline{B}^{W,H}$ the smallest subset of $\mathbb{A}$ that contains B and its own predecessors under $\mathcal{P}_{W,H}$. As explained in the companion paper Heymann et al. [2020] (see also Witsenhausen [1975]), $\overline{B}^{W,H}$ is the **topological closure** of B under a topology induced by $\mathcal{P}_{W,H}$.

Definition 4 (Topological Separation)**.** *Let $H \subset \mathbb{H}$ and $B, C, W \subset \mathbb{A}$. We say that B and C are (conditionally)* topologically separated *(w.r.t. (W, H)), denoted by $B \underset{t}{\perp\!\!\!\perp} C \mid (W, H)$, if there exists $W_B, W_C \subset W$ such that*

$$W_B \sqcup W_C = W \text{ and } \overline{B \cup W_B}^{W,H} \cap \overline{C \cup W_C}^{W,H} = \varnothing \ . \tag{10}$$

Observe that the condition is on the *existence* of a partition of the set of variables over which we want to condition.

On a DAG, when $H = \mathbb{H}$, we have topological separation of B and C with respect to W when there is a partition (W_B, W_C) of W such that the sets of ancestors of $B \cup W_B$ and $C \cup W_C$ – using $\mathcal{P}_{W,H}(a) = P(a) \setminus W$ – are disjoint.

Because it can be proved that d-separation and topological separation are equivalent on a DAG, we think this definition is very handy even for DAGs. Indeed, (1) the partition (W_B, W_C) can be derived mechanically (see companion paper Heymann et al. [2020]), (2) once the partition is given, it is usually much quicker to check that the ancestors sets are disjoints than checking that all the path between B and C are blocked by W.

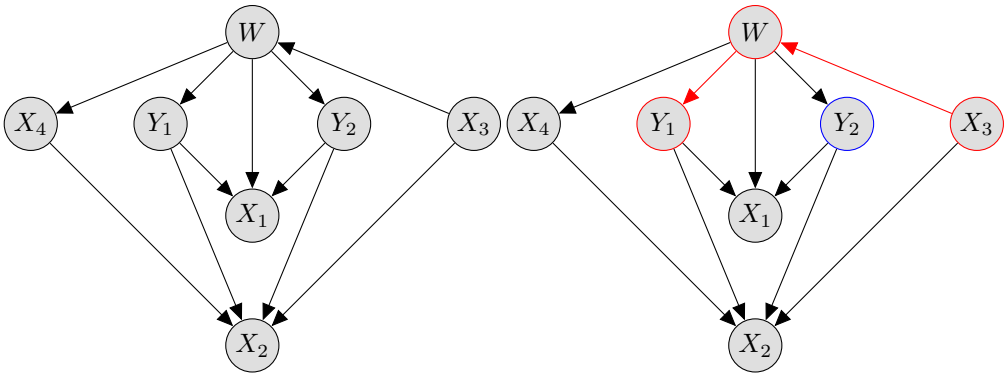
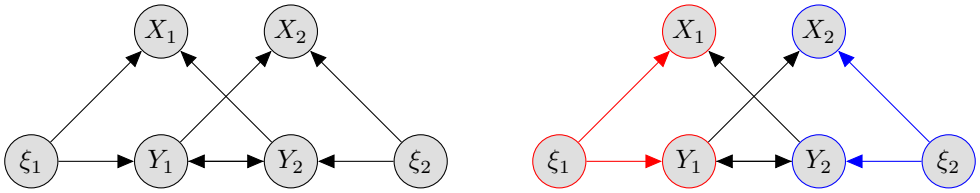


Figure 2: Topological separation is easy to check.



(a) Original graph.

(b) topological separation is checked on the initial graph $W_{X_i} = Y_i$.

Figure 3: Topological separation is easy to check: nonrecursive system.

Example 4 (Topological separation is easy to check: recursive system)**.** *The DAG in Figure 2 (left) illustrate why this notion is practical. If one want to check that Y_1 and Y_2 are d-separated by W, one need to check that **every path** that goes from Y_1 to Y_2 are blocked by W. By contrast, the topological separation can be checked visually on Figure 2 (right) by setting $W_{Y_1} = W$, $W_{Y_2} = \varnothing$ and checking that the red and blue sets are closed and do not intersect.*

Example 5 (Topological separation is easy to check: nonrecursive system)**.** *We display in Figure 3 a nonrecursive system for which we check (for X_1 and X_2 with respect to Y_1 and Y_2) our proposal of topological separation. This is – in our humble opinion – simpler to check than Forré et al. 's σ-separation Forré and Mooij [2020] because there are less intermediate steps.*

5 Independence and Do-calculus with Information fields

In this section, we suppose the random variables valued in finite sets for the sake of simplicity. We can now state our version of Pearl's three rules of do-calculus. The statement looks like a simple sufficient condition for conditional independence thanks to the fact that we encode the intervention variables in the information fields.

Theorem 5 (Do-calculus)**.**

$$Y \underset{t}{\perp\!\!\!\perp} Z \mid (W, H) \implies \Pr(U_Y | U_W, U_{\overline{Z}^{W,H}}, H) = \Pr(U_Y | U_W, H) \,. \tag{11}$$

We stress the conciseness of Theorem 5 — permitted by the notions introduced in this paper — as we now show that it implies the three rules of Pearl, as well as two recent results.

Proposition 6. *Rule 1 from Tikka et al. [2019a] can be deduced from Theorem 5. In particular, Theorem 5 subsumes Pearl's do-calculus from Pearl [1995].*

Remark 7. *In the same manner we could derive the 3 rules of Theorem 1 (do-calculus for stochastic interventions) in Correa and Bareinboim [2020].*

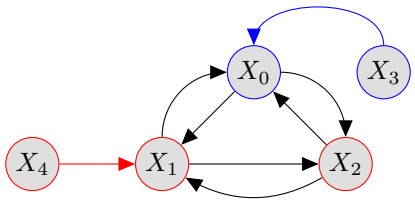


Figure 4: X_3 and X_4 are independent conditioned on (X_0, X_1, X_2) but not independent if we only condition on (X_0, X_1). Visual proof of topological separation:
$W_{X_4} = \{X_1, X_2\}$ and $W_{X_3} = \{X_0\}$

Example 6. *The last example is inspired by the work of Witsenhausen Witsenhausen [1971] on causality. This example provides arguments to explain why it is well posed (solvable). It is depicted in Figure 4 and corresponds to the following nonrecursive binary SCM (N_i are independent, binary noise variables, $\oplus$ is the XOR operator):*

$$X_0 = (X_1.(\neg X_2)) \oplus (N_0 \oplus X_3) \quad \textit{and} \quad X_1 = (X_2.(\neg X_0)) \oplus (N_1 \oplus X_4)$$
$$X_2 = (X_0.(\neg X_1)) \oplus N_2, \quad X_3 = N_3 \quad \textit{and} \quad X_4 = N_4.$$

*The random variables X_3 and X_4 are topologically separated by (X_0, X_1, X_2) –note that X_2 is needed –, **hence** X_3 **and** X_4 **are independent conditioned on** (X_0, X_1, X_2) but not independent if we only condition on (X_0, X_1).*

Observe that the intuition that we could equivalently replace X_0, X_1 and X_2 by a unique variable W is misleading: with such a change, we would get a collider $X_4 \to W \leftarrow X_3$ over which we are conditioning, which would make X_4 and X_3 non blocked with respect to W.

Let us try to apply the elegant recent result of Forré et al. (Theorem 5.2 from Forré and Mooij [2020]) on conditional independence in the presence of cycles. We first observe that the Directed Mixed Graph (DMG) induced by the Input/output Structural Causal Model (ioSCM) associated to our example (see Definition 2.3 and 5.1 ibid) looks like the graph of Figure 4. Second, we observe that X_0, X_1 and X_2 belong to the same strongly connected component S (see Forré and Mooij [2020]), in the sense that they are all ancestors and descendants of each other. Third, let us consider the walk $X_4 \to X_1 \leftarrow X_0 \leftarrow X_3$. Forré et al. provide a condition for a walk to be open in definition 4.2. By apply this definition, we see that $X_4 \to X_1 \leftarrow X_0 \leftarrow X_3$ is $\{X_0, X_1, X_2\}$-σ-open because:

- *$X_4 \to X_1 \leftarrow X_0$ satisfies the collider definition (4.2, (a) in Forré and Mooij [2020]) because $X_1 \in \{X_0, X_1, X_2\}$*
- *$X_1 \leftarrow X_0 \leftarrow X_3$ satisfies the left chain condition because $X_0 \in \{X_0, X_1, X_2\} \cap S$, where S is the strongly connected component of X_1.*

hence this walk is $\{X_0, X_1, X_2\}$-σ-open (see Definition 4.2 ibid). Hence it seems that Theorem 5.2 from Forré and Mooij [2020] could not be used to state that X_3 and X_4 are independent conditioned on (X_0, X_1, X_2).

We illustrate our theory with a numerical exact computation, taking the N_i as binomial variables of parameter 0.1. We solve the cycle by enumerating the 8 possible combinations of values for X_0, X_1 and X_2 and selecting the only admissible one. The results are shown in Table 1a and Table 1b.

This example illustrates the novelty of the IDM approach.

6 Discussion

In this paper, we simplify and generalize the do-calculus by leveraging the concept of information field. The do-calculus is reduced to one rule. We underline that the results come from the information structure, not the probability. Also, because our approach is not based on graphical models, our work provides a new proof of Pearl's original result. For most cases, one only needs to understand the notion of inverse image to work with information fields on top of SCMs and DAGs. In exchange,

Table 1: (Example 6)

(a) We check numerically the independence of X_3 and X_4 when conditioned on X_0, X_1 and X_2

			$\mathbb{P}(X_4 = 1 \mid X_0, X_1, X_2, X_3)$	
X_0	X_1	X_2	$X_3 = 0$	$X_3 = 1$
0	0	0	0.012	0.012
0	0	1	0.5	0.5
0	1	0	0.5	0.5
0	1	1	0.012	0.012
1	0	0	0.012	0.012
1	0	1	0.012	0.012
1	1	0	0.5	0.5
1	1	1	0.5	0.5

(b) We check numerically that X_3 and X_4 are not independent when conditioned on n X_0 and X_1

		$\mathbb{P}(X_4 = 1 \mid X_0, X_1, X_3)$	
X_0	X_1	$X_3 = 0$	$X_3 = 1$
0	0	0.023	0.023
0	**1**	**0.1**	**0.474**
1	0	0.012	0.012
1	1	0.5	0.5

information fields provide a compact, unifying and versatile language that brings new intuitions on the causal structure of the problem.

For instance, we illustrated why the Topological Separation is practical: once the partition of the conditioning variables known, checking that an intersection is empty is easier than checking a blocking condition on a collection of paths.

Also, we were able to recover a few recent results – how to handle spurious edges Tikka et al. [2019a], how to handle stochastic interventions Correa and Bareinboim [2020], how to handle cycles Forré and Mooij [2020]– , and we think that the Information Dependency Model is a powerful technical tool for investigating potential new extensions of already existing results in the field of causal inference. Moreover, many of those papers require the introduction of ad hoc frameworks. The Information Dependency Model is a good candidate to bring uniformity and consistency. It can be a temporary detour to introduce new notions, for instance the definition of Conditional Precedence 3 would have been harder to express with the SCM as primitive.

In addition, we presented and solved an example that cannot be handled easily with the current state of the literature.

Last, we mention that the notion of well-posedness we use was introduced in Witsenhausen [1975] half a century ago for another field of applied mathematics. It is interesting to observe that this notion could serve a new purpose in the field of causal inference.

Further work includes drawing connections with other research programs, such as Proposition 6 or questions related to identification Shpitser and Pearl [2006, 2008], Tikka et al. [2019b], using the framework developed in this paper. Also, it would be interesting to study the connections of this work with Bongers et al. [2020], Forré and Mooij [2020].

Acknowledgments

We would like to thank Alexandre Gilotte, Laure Alexandre, Eustache Diemert, Matthieu Martin, David Rohde, Thibaud Rahier, and Amélie Héliou and the anonymous reviewers for their feedback.

References

Judea Pearl and Dana Mackenzie. *The book of Why: the new science of cause and effect*. Basic Books, 2018.

Robert G Cowell, Philip Dawid, Steffen L Lauritzen, and David J Spiegelhalter. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer Science & Business Media, 2006.

John Winn. Causality with gates. In *Artificial Intelligence and Statistics*, pages 1314–1322, 2012.

Finnian Lattimore and David Rohde. Replacing the do-calculus with Bayes rule. *arXiv preprint arXiv:1906.07125*, 2019.

Santtu Tikka, Antti Hyttinen, and Juha Karvanen. Identifying causal effects via context-specific independence relations. In *Advances in Neural Information Processing Systems*, pages 2804–2814, 2019a.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *AAAI 2020 : The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Judea Pearl. The mathematics of causal relations. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures (P. Shrout, K. Keyes and K. Ornstein, eds.). Oxford University Press, Corvallis, OR*, pages 47–65, 2010.

Joseph Y Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12: 317–337, 2000.

Judea Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, second edition edition, 2011. ISBN 9780511803161. doi: 10.1017/CBO9780511803161.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Benjamin Heymann, Michel De Lara, and Jean-Philippe Chancelier. Causal inference with information fields (long version), 2020. preprint.

Stephan Bongers, Patrick Forré, Jonas Peters, Bernhard Schölkopf, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *arXiv preprint arXiv:1611.06221*, 2020.

Patrick Forré and Joris M. Mooij. Causal calculus in the presence of cycles, latent confounders and selection bias. volume 115 of *Proceedings of Machine Learning Research*, pages 71–80. PMLR, Tel Aviv, Israel, 22–25 Jul 2020. URL `http://proceedings.mlr.press/v115/forre20a.html`.

H. S. Witsenhausen. On information structures, feedback and causality. *SIAM J. Control*, 9(2): 149–160, May 1971.

Claude Dellacherie and Paul-André Meyer. *Probabilités et potentiel*. Hermann, Paris, 1975.

H. S. Witsenhausen. The intrinsic model for discrete stochastic control: Some open problems. In A. Bensoussan and J. L. Lions, editors, *Control Theory, Numerical Methods and Computer Systems Modelling*, volume 107 of *Lecture Notes in Economics and Mathematical Systems*, pages 322–335. Springer-Verlag, 1975.

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

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference, AAAI-06/IAAI-06*, volume 2, pages 1219–1226. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008.

Santtu Tikka, Antti Hyttinen, and Juha Karvanen. Causal effect identification from multiple incomplete data sources: A general search-based approach. *arXiv preprint arXiv:1902.01073*, 2019b.