Similarity Orders from Causal Equations

Johannes Marti and Riccardo Pinosio

ILLC, University of Amsterdam

Abstract. The purpose of this paper is to demonstrate that, contrary to the received wisdom, causal reasoning can be formalized wholly within the framework of Lewis' conditional logic. To this aim we simulate causal reasoning based on structural equations in Lewis' order semantics. This reduction is based on a formalization of an intuitive idea for computing relative similarity between worlds. Worlds are the more similar the more they satisfy the same relevant propositions, where relevance is a comparative notion represented by a preorder. In the context of causal reasoning this relevance order on propositions depends on the causal structure of the problem domain.

Keywords: causal reasoning, conditional logic, counterfactual conditionals, non-monotonic reasoning, similarity orders, structural equations

1 Introduction

In this paper we show how causal reasoning based on systems of structural equations can be embedded into the framework of the similarity order semantics for conditional logic.

The order semantics for conditional logic was developed in [4] in order to analyze counterfactual conditionals. On this a approach a relative similarity order over possible worlds is taken as basic and counterfactual conditionals are evaluated by a minimization procedure in this similarity order. With this approach it has proven to be difficult to account for counterfactual conditionals which rely on causal dependencies in the problem domain [12]. The difficulty is to give an account of how to determine a relative similarity order that captures these causal dependencies [5].

In artificial intelligence the framework based on systems of structural equations has been very successful as a formalization of causal reasoning. Pearl's book [8] is the standard treatment of this approach. Here we also find a semantics for a restricted class of counterfactual conditionals. This semantics is not prone to the kind of counterexamples that have been proposed against the similarity approach.

Pearl already notices [8, Section 7.4] a close relation between the semantics for counterfactual conditionals on systems of structural equations and the semantics on similarity orders. In [10] Schulz brings the approach to counterfactual conditionals using structural equations closer to premise semantics for conditional logic [13, 3], which is a framework essentially equivalent to the similarity approach [6]. In particular, by importing notions from premise semantics

into the setting of structural equations, Schulz extends the class of conditionals that can be evaluated. However, it remained an open question how to construct a relative similarity order that captures precisely the causal dependencies encoded in a system of structural equations:

> While – as we have seen – we can understand Pearl's system as an instantiation of the similarity approach, so far we do not know the exact nature of the similarity relation that would give us Pearl's interpretation of would have conditionals. Of course, it would be nice to have a reformulation of Pearl's theory in terms of similarity. But the only thing we can say so far is that it looks as if the relevant similarity relation differs clearly from what has been proposed in premise semantics. [9, p. 113]

We show how to construct a relative similarity order between possible worlds from a system of structural equations such that the truth of counterfactual conditionals is preserved.

For our construction we introduce the notion of a relevance order. A relevance order is a preorder with a proposition associated to every element in the order. This notion is motivated by the premise semantics for conditional logic where a set of relevant propositions is associated to every world. In the definition of relevance orders we take, following previous work [7], the relevant propositions to be world-independent. Moreover, our notion of a relevance order makes precise the idea from [5] that relevance is a matter of degree. We thus order propositions by comparative relevance instead of just having a set of relevant propositions.

The paper has the following structure. In Section 2 we briefly review the language and order semantics of conditional logic. In Section 3 we review the framework of systems of structural equations. In Section 4 we introduce the notion of a relevance order. Section 5 contains the construction of relative similarity orders from systems of structural equations and proves the preservation of true conditionals. In Section 6 we show that the framework of relevance orders can also account for backtracking counterfactual conditionals.

2 Conditional logic

In this section we present the syntax of conditional logic and review its semantics on relative similarity orders. The purpose is to fix the notation and clarify the setting. For an extensive technical treatment of conditional logic we refer to [14].

The language of conditional logic is the set of all formulas generated according to the following grammar:

$$\varphi ::= x \mid \varphi \wedge \varphi \mid \neg\varphi \mid \varphi \rightsquigarrow \varphi$$

where $x \in \mathsf{A}$ is an element from a fixed finite set $\mathsf{A} = \{x_0, x_1, \dots, x_{n-1}\}$ of atomic sentences. The Boolean connectives $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined in terms of $\neg$ and $\wedge$ in the usual way. Formulas of the form $\varphi \rightsquigarrow \psi$ are conditionals, where φ is the antecedent and ψ the consequent.

The semantics for conditional logic on relative similarity orders is based on a set W of worlds. We assume that the set of worlds W is the set of all Boolean valuations over the set of atomic sentences. This choice ensures the existence of enough possible worlds, which we need for the proof of our main result.

Definition 1. *A world $w : \mathsf{A} \to 2$ is a function which assigns to every variable x_k with $k < n$ a binary value $w_k = w(x_k)$. We write W for the set of all worlds.*

The order semantics is based on preorders, which are just reflexive and transitive relations. If $\leq$ is preorder on a set V we also write $w < v$ for $w \leq v$ and not $v \leq w$. Given a set $U \subseteq V$ we use $\mathsf{Min}(\leq, U) \subseteq U$ for the set of minimal elements of U in $\leq$, that is

$$\mathsf{Min}(\leq, U) = \{m \in U \mid \text{if } u \leq m \text{ then } m \leq u \text{ for all } u \in U\} .$$

The semantic structures in the order semantics are relative similarity orders.

Definition 2. *A relative similarity order $\leq$ over W is a tertiary relation on W such that $\leq_w$ is preorder for every $w \in W$.*

We think of $v \leq_w u$ as meaning that the world v is more similar to the actual world w than the world u.

The standard semantics of the conditional on relative similarity orders is defined using minimization.

Definition 3. *The semantics of a conditional $\varphi \rightsquigarrow \psi$ on a relative similarity order $\leq$ is given by:*

$$w, \leq \models \varphi \rightsquigarrow \psi \quad \textit{iff} \quad v, \leq \models \psi \textit{ for all } v \in \mathsf{Min}(\leq_w, A), \textit{where } A = \{v \in W \mid v, \leq \models \varphi\}$$

Intuitively, this clause says that $\varphi \rightsquigarrow \psi$ is true at a world w if ψ is true at all the φ-worlds that are maximally similar to w.

3 Causal reasoning with structural equations

In this section we review causal reasoning based on functional causal models. We are working within the extended version of Pearl's framework [8] introduced by Schulz in [10]. These extensions allow for a more general approach to the evaluation of conditionals than [8].

Again we assume a fixed set of atomic binary variables $\mathsf{A} = \{x_0, x_1, \ldots, x_{n-1}\}$ as given. These atomic variables represent the basic facts in the causal structure of the problem domain. We restrict our presentation to binary variables since these can be taken to be atomic sentences of conditional logic. However, the construction of this paper also works if the variables take values in any finite set. In that case one has to adapt the language of conditional logic by introducing atomic sentences expressing that a variable has a certain fixed value.

The causal structure of the problem domain is represented by a system of structural equations, which are called recursive causal models in [8, Definition 7.1.1] and dynamics in [10, Definition 1].

Definition 4. *A system of structural equations F is a set $F_{en} \subseteq \mathsf{A}$ and a function $F_k : 2^k \to 2$ for every number k such that $x_k \in F_{en}$. We call F_{en} the endogenous variables of F and define the set of exogenous variables as $F_{ex} = \mathsf{A} \setminus F_{en}$. If $F_{en} = \mathsf{A}$ then we call F complete.*

If $F_k : 2^k \to 2$ is a constant function for every $k \in F_{en}$ then we call F constant. Every consistent conjunction of literals φ induces a constant system of equations $\mathsf{S}(\varphi)$ such that $\mathsf{S}(\varphi)_{en}$ is the set of all variables occurring in φ and $\mathsf{S}(\varphi)_k$ is the constant function with value 1 if x_k occurs positively in φ and with value 0 otherwise.

We also call a system of structural equations just a system of equations. Intuitively, one thinks of a system of equations F as specifying, for every k such that $x_k \in F_{en}$, an equation

$$x_k = F_k(x_0, \ldots, x_{k-1}) \, .$$

This equation represents the causal dependence of the effect x_k on its causes, which are a subset of the variables $x_0, \ldots, x_{k-1}$. The following definition makes this subset of causes explicit.

Definition 5. *Let F be a system of structural equations. The variable $x_k \in F_{en}$ depends on the variable $x_l \in \mathsf{A}$ if $x_k \in F_{en}$ and*

$$F_k(x_0, \ldots, x_{l-1}, 0, x_{l+1}, \ldots, x_{k-1}) \neq F_k(x_0, \ldots, x_{l-1}, 1, x_{l+1}, \ldots, x_{k-1})$$

for some assignment of binary values to $x_0, \ldots, x_{l-1}, x_{l+1}, \ldots, x_{k-1}$. Define the parent relation P on A such that $x_l P x_k$ if x_k depends on x_l.

The graph determined by the parent relation P on A is called the causal diagram of F. In our setting it follows by definition that this graph is acyclic since the equation F_k determining the value of x_k depends only on the previous variables $x_0, \ldots, x_{k-1}$. For this reason we can define the causal diagram of F as a poset.

Definition 6. *The causal diagram $G(F) = (\mathsf{A}, \leq)$ associated to F is a poset over A where $\leq$ is the reflexive, transitive closure of the parent relation P.*

Our definition of a system of equations differs from Pearl's presentation in [8] where the equation defining the value of some variable can depend on the values of all other variables. Pearl then restricts his attention to recursive systems of equations, which are defined as those whose causal diagram is acyclic. One can show that any finite recursive system of equations can be put into the form of Definition 4; thus our setting is not more restrictive.

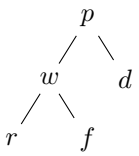
Our presentation allows us to exploit recursion on the natural numbers. If one uses the definition from [8] one has to resort to recursion over the parent relation, which is less familiar.

Example 1. Throughout this paper we use a slightly adapted version of an example from [11, p. 339]. In the example we consider five binary causal variables: r there is enough rain, f fertilizer is used, w the wheat crop is large, d there is high demand for wheat and p the wheat prize is high. The causal dependencies between these variables are that enough rain and the use of fertilizer causes

the wheat crop to be large, and that a small wheat crop or a high demand for wheat causes the wheat prize to be high. This problem domain is represented by the following system of equations F, where we use Boolean formulas to specify binary functions:

$$\begin{aligned} w &= r \wedge f \\ p &= \neg w \vee d \; . \end{aligned}$$

The causal diagram $G(F)$ of this system of equations is as follows:



If a world w satisfies all the causal laws represented by an equation in F then it is a solution to F. This motivates the following definition.

Definition 7. *A world w is a solution to a system of equations F if*

$$w_k = F(w_0, \ldots, w_{k-1}) \qquad \text{for all } k \text{ with } x_k \in F_{en} \; .$$

We write $[\![F]\!] \subseteq W$ for the the set of all solutions of F. We also use the notation $[\![x_k = G(x_0, \ldots, x_{k-1})]\!]$ for the set of solutions of the system of equations F with $F_{en} = \{x_k\}$ and $F_k = G$.

For a complete system of equations one can compute a unique solution by recursion on the natural numbers. This simplifies the presentation of [10] which relies on fixed points of logic programs.

Definition 8. *For a complete system of equations F define the world $\sigma(F) = w$ by a recursion on $k < n - 1$ such that $w_0 = F_0$ and $w_{k+1} = F_{k+1}(w_0, \ldots, w_k)$.*

The following proposition is proven by induction.

Proposition 1. *A system of equations F is complete if and only if it has a unique solution, which in this case is $\sigma(F)$.*

The following notion was introduced by [10].

Definition 9. *Let F be a system of equations. The basis $B^{F,w}$ of a world w is the complete system of equations for $B^{F,w}_{en} = \mathsf{A}$ where for $k < n$ we define*

$$B^{F,w}_k(x_0, \ldots, x_{k-1}) = \begin{cases} F_k(x_0, \ldots, x_{k-1}) & x_k \in F_{en} \text{ and } w_k = F_k(w_0, \ldots, w_{k-1}) \\ w(k) & \text{otherwise} \; . \end{cases}$$

The basis $B^{F,w}$ is a system of equations which differs minimally from F but has w as its unique solution. One can prove by a simple induction that:

Proposition 2. *The world w is the unique solution $\sigma(B^{F,w})$ of $B^{F,w}$.*

Intuitively, the basis $B^{F,w}$ keeps all the laws from F that are not violated at w and sets all other variables to their value in w by means of a constant equation.

The next definition captures intervention on a causal system by fixing the value of some variables on constant values. This corresponds to the definition of a submodel in [8].

Definition 10. *Let F and A be systems of equations such that A is constant. The intervention $F \mid A$ of F with A is defined to be the system of equations such that $(F \mid A)_{en} = F_{en} \cup A_{en}$ and for an k with $x_k \in F_{en} \cup A_{en}$*

$$(F \mid A)_k = \begin{cases} A_k & x_k \in A_{en} \\ F_k & x_k \in F_{en} \setminus A_{en} \end{cases} .$$

Note that $F \mid A$ is complete if F or A is complete.

We now define the semantics of the conditional on a system of equations.

Definition 11. *The semantics of a conditional $\varphi \rightsquigarrow \psi$, where φ is a consistent conjunction of literals, on a system of equations F is given by:*

$$w, F \models \varphi \rightsquigarrow \psi \quad \textit{iff} \quad \sigma(B^{F,w} \mid \mathsf{S}(\varphi)), F \models \psi .$$

Note that this clause is well-defined because $B^{F,w} \mid \mathsf{S}(\varphi)$ is complete since $B^{F,w}$ is complete.

Our definition of the semantics follows [10]. In the restricted case where φ contains only endogenous variables of F and w satisfies all the laws of F the above semantic clause for the conditional is equivalent to [8, Definition 7.1.5]. However, the semantics of [10] extends the semantics of [8] in two respects. First, the antecedent can contain variables that are exogenous in F. In [8] this is not possible because there interventions are only defined for endogenous variables. Second, conditionals can be evaluated at worlds which violate some of the laws represented in F. This works thanks to Schulz' notion of a basis, which deals with a violation of a law by an intervention. In [8] the world of evaluation only sets the values of exogenous variables in F. Hence it is implicitly assumed that it satisfies all the laws of F.

4 Relative similarity orders from relevance orders

In this section we introduce the notion of a relevance order and show how it can be used to construct a relative similarity order.

Definition 12. *A relevance order for a set of worlds W is a tuple $(D, \sqsubseteq, \mathsf{e})$ where D is a set whose elements we call descriptions, $\sqsubseteq$ is a preorder on D and $\mathsf{e} : D \to \mathcal{P}W$ is a function mapping descriptions to sets of worlds.*

We keep the standard terminology and take propositions to be just sets of worlds. However, we treat the description $d \in D$ in a similar fashion as the proposition $\mathsf{e}(d)$ determined by d. For instance we say that d is true at a world w if $w \in \mathsf{e}(d)$.

The notion of a relevance order allows us to rank propositions according to how important it is to keep their truth value constant when switching to counterfactual worlds. Propositions which are low in the order are considered more important. For example, mathematical theorems would intuitively count as more relevant than physical laws, which in turn are more relevant than particular facts.

In Definition 12 propositions are not ordered directly as sets of worlds but by means of descriptions having those propositions as extensions. This simplifies later proofs, since one does not need to verify that distinct elements in the order are really distinct as sets of worlds.

We now show how to construct a relative similarity order between worlds from a relevance order. The following technical notion is needed.

Definition 13. *A proposition $U \subseteq W$ is v, u-separating if either $v \in U$ and $u \notin U$, or $v \notin U$ and $u \in U$. Given a relevance order $(D, \sqsubseteq, \mathsf{e})$ we say that a description $d \in D$ is v, u-separating if the proposition $\mathsf{e}(d)$ is v, u-separating. We use $\mathsf{sep}(v, u) \subseteq D$ for the set of all v, u-separating descriptions. We use $\mathsf{sep}_w(v, u) \subseteq D$ for all the v, u-separating descriptions that are true at w.*

Now for the construction of the relative similarity order.

Definition 14. *The relative similarity order $(W, \leq)$ determined by a relevance order $(D, \sqsubseteq, \mathsf{e})$ for W is defined such that*

$$v \leq_w u \quad \textit{iff} \quad v \in \mathsf{e}(d) \textit{ for all } d \in \mathsf{Min}(\sqsubseteq, \mathsf{sep}_w(v, u)) \,.$$

Similarity of worlds to w is determined only by the relevant propositions which are true at w. Thus, a world is the more similar to w the more of these proposition it makes true. When comparing two worlds for similarity to w, we consider only the most relevant propositions true at w which can distinguish between the two worlds. In the special case where $\sqsubseteq$ is a total preorder, meaning that any two elements are required to be comparable, our clause reduces to the discrimin ordering of [1].

We need to verify that Definition 14 actually yields a relative similarity order.

Proposition 3. *The relation $\leq_w$ from Definition 14 is reflexive and transitive for every world $w \in W$.*

Proof. Reflexivity holds because there are no v, v-separating descriptions.

For transitivity assume that $v \leq_w u$ and $u \leq_w z$. We show that $v \leq_w z$. So pick any $d \in \mathsf{Min}(\sqsubseteq, \mathsf{sep}_w(v, z))$ and assume for a contradiction that $v \notin \mathsf{e}(d)$. Distinguish cases on whether $u \in \mathsf{e}(d)$.

If $u \in \mathsf{e}(d)$ then $d \in \mathsf{sep}_w(v, u)$ and because $v \leq_w u$ it follows that $v \in \mathsf{e}(d)$. This contradicts the assumption $v \notin \mathsf{e}(d)$.

If $u \notin \mathsf{e}(d)$ we distinguish cases on whether $z \in \mathsf{e}(d)$. If $z \in \mathsf{e}(d)$ then $d \in \mathsf{sep}_w(u, z)$ and by $u \leq_w z$ we get that $u \in \mathsf{e}(d)$, which is a contradiction. If $z \notin \mathsf{e}(d)$ then contrary to our assumption d would not be v, z-separating.

With a proof similar to the one of Propositions 3 one can show that the relative similarity order determined by a relevance order satisfies the weak centering axiom $w \leq_w v$ and the triangularity axiom $v \leq_w u \wedge u \leq_v w \to v \leq_u w$. It is an open question what further axioms are enforced by this construction.

5 Relative similarity orders from system of equations

This section contains the main technical result of this paper. We first define a ranking over propositions $R(F)$ and then prove that the semantics of the conditional in F is equivalent to its semantics on the relative similarity order determined by $R(F)$.

Definition 15. *For every system of equations F with causal diagram $G(F) = (\mathsf{A}, \leq)$ define a ranking over descriptions $R(F) = (D, \sqsubseteq, \mathsf{e})$. The set D of descriptions is given by*

$$\begin{aligned} &D \subseteq \mathsf{A} \times (2 + \{\star\}) \,, \\ &D = \{(x_k, a) \mid x_k \in \mathsf{A}, a \in 2\} \cup \{(x_k, \star) \mid x_k \in F_{en}\} \,. \end{aligned}$$

The preorder $\sqsubseteq$ on D is defined such that

$$(x_k, a) \sqsubseteq (x_l, b) \quad \textit{iff} \quad x_k < x_l \text{ in } G(F) \text{ or } (x_k = x_l \text{ and } (a = \star \text{ or } a = b)) \,.$$

We leave it to the reader to check that this is indeed reflexive and transitive. The evaluation $\mathsf{e} : D \to \mathcal{P}W$ is given by

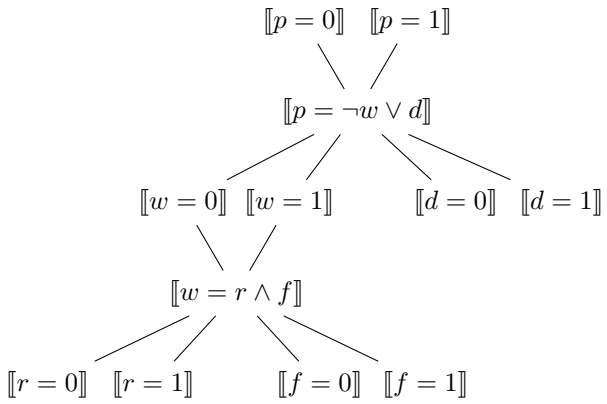
$$\begin{aligned} \mathsf{e}(x_k, a) &= [\![x_k = a]\!] \,, && x_k \in \mathsf{A}, a \in 2 \\ \mathsf{e}(x_k, \star) &= [\![x_k = F_k(x_0, \ldots, x_{k-1})]\!] \,. && x_k \in F_{en} \end{aligned}$$

One can obtain the ranking $R(F)$ from the causal diagram $G(F)$ by replacing all exogenous variables x_k with the antichain of two descriptions evaluating to $[\![x_k = 0]\!]$ and $[\![x_k = 1]\!]$ respectively, and all endogenous variables x_k with the following poset of descriptions, where the evaluations are displayed:

$$\begin{array}{ccc} [\![x_k = 0]\!] & & [\![x_k = 1]\!] \\ & \diagdown \quad \diagup & \\ & [\![x_k = F_k(x_0, \ldots, x_{k-1})]\!] & \end{array}$$

Moreover a description in the subposet of a variable x_k is smaller than a description in the subposet of another variable x_l exactly if $x_k < x_l$ in $G(F)$.

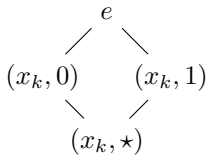
Example 2. For the system of equations F from Example 1 we obtain the following relevance order $R(F)$:



The idea behind the definition of $R(F)$ is that when evaluating counterfactual conditionals on a system of structural equations it is more important to keep the past than the future facts constant, and one rather gives a causal law up than any causes occurring in it. This order is reflected in the relevance ranking $R(F)$.

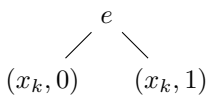
The following lemma states the crucial property of $R(F)$ which we exploit in the proof of Theorem 1.

Lemma 1. *Let F be a system of equations and consider two worlds w and v such that $w_l = v_l$ for all $l \neq k$ for some $k < n$. If $x_k \in F_{en}$ then any $d \in \mathsf{sep}(w, v)$ in $R(F)$ is an element in a suborder of $R(F)$ which has the following shape*



In other words: if $x_k \in F_{en}$ then any $d \in \mathsf{sep}(w, v)$ is either equal to $(x_k, 0)$, $(x_k, 1)$ or $(x_k, \star)$, or $(x_k, 0) \sqsubset d$ and $(x_k, 1) \sqsubset d$.

If $x_k \in F_{ex}$ then any $d \in \mathsf{sep}(w, v)$ in $R(F)$ is an element in a suborder of $R(F)$ which has the following shape



In other words: if $x_k \in F_{ex}$ then any $d \in \mathsf{sep}(w, v)$ is either equal to $(x_k, 0)$ or $(x_k, 1)$, or $(x_k, 0) \sqsubset d$ and $(x_k, 1) \sqsubset d$.

Proof. We show the case where $x_k \in F_{en}$ and leave the second similar case to the reader. We reason by contraposition.

First assume that $d = (x_l, a)$ where $x_l \neq x_k$ and $a \in 2$. Then certainly $d \notin \mathsf{sep}(w, v)$ because $\mathsf{e}(x_l, a) = [\![x_l = a]\!]$ and by assumption w and v agree on the value of x_l.

Now assume that $d = (x_l, \star)$ such that not $x_k \leq x_l$ in $G(F)$. First we have again by assumption that $w_l = v_l$ because $k \neq l$. Moreover F_l does not depend on x_k because otherwise x_k would be a parent of x_l and so $x_k \leq x_l$ in $G(F)$. So $F_l(w_0, \ldots, w_{l-1}) = F_l(v_0, \ldots, v_{l-1})$ again by the assumption that $w_m = v_m$ for all $m \neq k$. The facts that $w_l = v_l$ and that $F_l(w_0, \ldots, w_{l-1}) = F_l(v_0, \ldots, v_{l-1})$ entail that either both or none of w and v is in $[\![x_l = F_l(x_0, \ldots, x_{l-1})]\!] = \mathsf{e}(x_l, \star)$. Hence $d = (x_l, \star) \notin \mathsf{sep}(w, v)$. This concludes the proof.

We now prove our main result.

Theorem 1. *Let F and A be systems of equations such that A is constant. Denote by $\leq$ be the relative similarity order determined by $R(F)$. Then for every world $w \in W$:*

$$\mathsf{Min}(\leq_w, [\![A]\!]) = \{\sigma(B^{F,w} \mid A)\} .$$

Proof. Consider the world $s = \sigma(B^{F,w} \mid A)$ and take any world $z \in \mathsf{Min}(\leq_w, [\![A]\!])$. We show by an induction on $k < n$ that $z_k = s_k$.

It is sufficient to prove that $z_k = s_k$ on the assumption that $z_l = s_l$ for all $l < k$. This also covers the base case where $k = 0$ because then there are no $l < k$ and hence the assumption is trivially satisfied.

So pick any $k < n$ and assume that $z_l = s_l$ for all $l < k$. We want to show that $z_k = s_k$.

First consider the case where $x_k \in A_{en}$. In this case $s_k = A_k(s_0, \ldots, s_{k-1})$ by Definition 10 of $B^{F,w} \mid A$. Because $z \in \mathsf{Min}(\leq_w, [\![A]\!])$ and so $z \in [\![A]\!]$ we also have that $z_k = A_k(z_0, \ldots, z_{k-1})$. But $A_k(z_0, \ldots, z_{k-1}) = A_k(s_0, \ldots, s_{k-1})$ because A_k is a constant function.

In the other case where $x_k \notin A_{en}$ we have that $s_k = B^{F,w}_k(s_0, \ldots, s_{k-1})$. Again, we distinguish cases depending on the truth of $x_k \in F_{en}$ and $w_k = F_k(w_0, \ldots, w_{k-1})$.

If $x_k \in F_{en}$ and $w_k = F_k(w_0, \ldots, w_{k-1})$ then by Definition 9 of $B^{F,w}$ we have that $s_k = B^{F,w}_k(s_0, \ldots, s_{k-1}) = F_k(s_0, \ldots, s_{k-1})$. Assume for a contradiction that $z_k \neq F_k(s_0, \ldots, s_{k-1})$. By the induction hypothesis it follows that $z_k \neq F_k(z_0, \ldots, z_{k-1})$. We show that there is a world z' such that $z' \in [\![A]\!]$, $z' \leq_w z$ but not $z \leq_w z'$, which contradicts the assumption that $z \in \mathsf{Min}(\leq_w, [\![A]\!])$. The world z' is defined by setting $z'_l = z_l$ for all $l < n$ with $l \neq k$ and $z'_k = F_k(z'_0, \ldots, z'_{k-1})$. Since $x_k \notin A_{en}$, A is constant and $z \in [\![A]\!]$ it follows that $z' \in [\![A]\!]$. Now we have that $z', w \in [\![x_k = F_k(x_0, \ldots, x_{k-1})]\!]$ but $z \notin [\![x_k = F_k(x_0, \ldots, x_{k-1})]\!]$. Since $[\![x_k = F_k(x_0, \ldots, x_{k-1})]\!] = \mathsf{e}(x_k, \star)$ it follows by Lemma 1 that $(x_k, \star)$ is the only minimal z, z'-separating description that is true at w. So it follows by Definition 14 of $\leq_w$ that $z' \leq_w z$ but not $z \leq_w z'$.

In the other case we have that either $x_k \notin F_{en}$ or that $w_k \neq F_k(w_0, \ldots, w_{k-1})$. By Definition 9 it follows that $s_k = B^{F,w}(s_0, \ldots, s_{k-1}) = w_k$. Now assume for a

contradiction that $s_k \neq z_k$. We again construct a z' such that $z' \in [\![A]\!]$, $z' \leq_w z$ but not $z \leq_w z'$ contradicting the assumption that $z \in \mathsf{Min}(\leq_w, [\![A]\!])$. The world z' is defined by setting $z'_l = z_l$ for all $l < n$ with $l \neq k$ and $z'_k = w_k$. Since $x_k \notin A_{en}$, A is constant and $z \in [\![A]\!]$ it follows that $z' \in [\![A]\!]$. We now show that in $R(F)$ the description (x_k, w_k) is the only minimal z, z'-separating description that is true at w. For this we distinguish cases depending on whether $x_k \in F_{en}$.

First consider the case with $x_k \in F_{en}$. Then $w_k \neq F_k(w_0, \ldots, w_{k-1})$. We now consider the descriptions $(x_k, \star)$, (x_k, w_k) and (x_k, a) for $a \neq w_k$. We know that $w \notin [\![x_k = F_k(x_0, \ldots, x_{k-1})]\!] = \mathsf{e}(x_k, \star)$ and that $w \notin [\![x_k = a]\!] = \mathsf{e}(x_k, a)$. However, $[\![x_k = w_k]\!] = \mathsf{e}(x_k, w_k)$ is z, z'-separating and true at w. It follows by the first part of Lemma 1 that (x_k, w_k) is the only minimal z, z'-separating description that is true at w.

In the other case $x_k \notin F_{en}$ we only need to consider the descriptions (x_k, w_k) and (x_k, a) for $a \neq w_k$. By the same reasoning as in the previous case it follows that (x_k, w_k) is z, z'-separating and true at w and (x_k, a) fails to be true at w. So by the second part of Lemma 1 we get that (x_k, w_k) is the only minimal z, z'-separating description that is true at w.

From the fact that the only minimal z, z'-separating description true at w is (x_k, w_k) it follows that $z' \leq_w z$ but not $z \leq_w z'$ because $z' \in [\![x_k = w_k]\!]$ but $z \notin [\![x_k = w_k]\!]$. This concludes the proof of the theorem.

By unfolding Definition 3 and Definition 11 one now easily concludes that the truth of conditionals is preserved by construction of this paper.

Corollary 1. *Let F be systems of equations and denote by $\leq$ be the relative similarity order determined by $R(F)$. Consider a conditional $\varphi \rightsquigarrow \psi$ where φ is a consistent conjunction of literals. Then for every world $w \in W$*

$$w, F \models \varphi \rightsquigarrow \psi \quad \textit{iff} \quad w, \leq \models \varphi \rightsquigarrow \psi \,.$$

6 Backtracking counterfactual conditionals

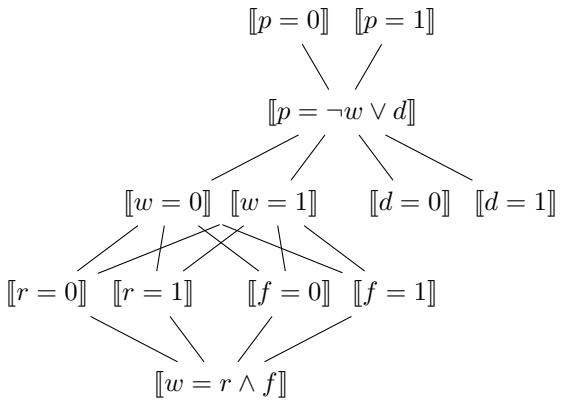
In this section we show that the framework of relevance orders is flexible enough to cope with backtracking counterfactual conditionals. A backtracking counterfactual conditional infers from an effect to a cause. This means that the antecedent of the conditional counterfactually assumes that a different effect than the one actually obtaining occurs, and the consequent reasons to a counterfactual cause. In Example 1, the conditional "if the wheat crop had been smaller last year, then there would have been either less rain or less fertilizer applied" is backtracking, since it reasons from the counterfactual effect of a smaller crop to the absence of one of the actual causes.

Backtracking does not arise in the evaluation of conditionals specified by the semantic clause in Definition 11. Take for instance the world $frw\bar{d}\bar{p}$ as the actual world. One can check that according to Definition 11 the above backtracking conditional $\neg w \rightsquigarrow \neg r \vee \neg f$ turns out false at this world. The reason for this is that intervening with the antecedent of the conditional cancels the causal

law leading from the consequent to the antecedent. The failure of backtracking also becomes obvious by inspecting the relevance order defined in Definition 15. There a causal law is deemed as less relevant than the causes appearing in it. Thus counterfactual worlds which violate the causal law will be more similar to the actual world than those which violate the causes.

Backtracking conditionals, however, seem to play a role in communication. In [5, p. 457] Lewis argues that there is an ambiguity between interpreting counterfactual conditionals with a standard, non backtracking or with a backtracking resolution. In conversation, different contextual factors can trigger the backtracking interpretation of counterfactual conditionals. We show that the ambiguity between these two interpretations can be accounted for as a choice between different relevance orders. We now show how to modify the relevance order from Example 2 to allow for backtracking in the evaluation of the conditional $\neg w \rightsquigarrow \neg r \vee \neg f$.

The most obvious modification of the original order from Example 2 to allow for backtracking is to make the causal law $w = r \wedge f$ more relevant than the causal facts $r = 0$, $r = 1$, $f = 0$ and $f = 1$ occurring in it. Thus one obtains the following order:



In the similarity order determined by this relevance order the conditional $\neg w \rightsquigarrow \neg r \vee \neg f$ is true at $u = frwd\bar{p}$. In particular, one can verify that $\mathsf{Min}(\leq_u, [\![w = 0]\!]) = \{\bar{f}r\bar{w}\bar{d}p, f\bar{r}\bar{w}\bar{d}p\}$. Note that we have two non logically equivalent minimal worlds. For this reason, both $\neg w \rightsquigarrow r$ and $\neg w \rightsquigarrow \neg r$ are false at u, which means that conditional excluded middle fails. This situation could never arise in the case of non backtracking counterfactual conditionals, as one can see from Theorem 1.

The example shows that the approach presented here can be adapted to deal with backtracking counterfactual conditionals. The relevance order given in the example, however, admits backtracking only over one particular causal law, namely $\neg w \rightsquigarrow \neg r \vee \neg f$. By a stepwise swapping of causal laws with causes, one can precisely determine how many steps one can backtrack along which laws. The limit case, allowing all backtracking, is determined by a relevance order in which all causal laws are strictly more relevant than all particular facts that occur

as causes or effects. Our theory does not determine how much backtracking is admissible, but leaves this choice to the modeler, depending on the application.

7 Conclusions and Further Work

In this paper we have presented a construction of relative similarity orders between possible worlds from systems of structural equations which preserves the truth of counterfactual conditionals. This shows that the framework of similarity orders can adequately model causal dependencies of a problem domain.

Our construction crucially depends on the notion of a relevance order over propositions. Relevance orders are a powerful tool to construct relative similarity orders, as they allow us to precisely control which propositions matter for the similarity of worlds. The treatment of backtracking in Section 6 gives an idea of the flexibility of the approach. It might thus be of interest to employ relevance orders in other settings where conditional logic is used. For instance, in belief revision they could be used to determine plausibility orders from rankings over evidence. A technical problem is to characterize the class of relative similarity orders which arise from relevance orders and to axiomatize its conditional logic.

A natural question is whether there is an inverse construction reading off a system of structural equations from a relative similarity order between worlds. One would thus define causality by means of relative similarity orders. This would be a semantic analog to Lewis' definition of causality, which is widely considered to be defective.

It has been shown in [2] that certain principles of conditional logic which are valid over similarity orders can be falsified by cyclic systems of equations. It might be worthwhile to see whether the techniques from this paper can be adapted to this case by giving up the assumption that relevance and similarity orders are transitive.

References

1. Sylvie Coste-Marquis, Jérôme Lang, Paolo Liberatore, and Pierre Marquis. Expressive power and succinctness of propositional languages for preference representation. In Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, editors, *KR*, pages 203–212. AAAI Press, 2004.
2. Joseph Y. Halpern. From causal models to counterfactual structures. *Review of Symbolic Logic*, 6(2):305–322, 2013.
3. Angelika Kratzer. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, 10(2):201–216, 1981.
4. David Lewis. *Counterfactuals*. Blackwell Publishers, 1973.
5. David Lewis. Counterfactual dependence and time's arrow. *Noûs*, 13(4):455–476, 1979.
6. David Lewis. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, 10(2):217–234, 1981.
7. Johannes Marti and Riccardo Pinosio. Topological semantics for conditionals. In Vít Punčochář and Petr Švarný, editors, *The Logica Yearbook 2013*. College Publications, to appear.

8. Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
9. Katrin Schulz. *Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals*. PhD thesis, University of Amsterdam, 2007.
10. Katrin Schulz. "If you'd wiggled A, then B would've changed" - Causality and counterfactual conditionals. *Synthese*, 179(2):239–251, 2011.
11. Herbert A. Simon and Nicholas Rescher. Cause and counterfactual. *Philosophy of Science*, 33(4):323–340, 1966.
12. Pavel Tichý. A counterexample to the Stalnaker-Lewis analysis of counterfactuals. *Philosophical Studies*, 29(4):271–273, 1976.
13. Frank Veltman. Prejudices, presuppositions, and the theory of counterfactuals. In Jeroen Groenendijk and Martin Stokhof, editors, *Amsterdam Papers in Formal Grammar*, volume 1, pages 248–282. Centrale Interfaculteit, Universiteit van Amsterdam, 1976.
14. Frank Veltman. *Logics for Conditionals*. PhD thesis, University of Amsterdam, 1985.