

---

# FACTORED SPACE MODELS: TOWARDS CAUSALITY BETWEEN LEVELS OF ABSTRACTION

---

**Scott Garrabrant\***

Machine Intelligence Research Institute,  
Independent  
scott.garrabrant@gmail.com

**Matthias G. Mayer\***

Machine Intelligence Research Institute,  
Independent  
matthias.georg.mayer@gmail.com

**Magdalena Wache\***

Machine Intelligence Research Institute,  
Principles of Intelligent Behaviour in  
Biological and Social Systems  
magdalena-wache@mailbox.org

**Leon Lang**

University of Amsterdam  
l.lang@uva.nl

**Sam Eisenstat**

Machine Intelligence Research Institute  
sam@intelligence.org

**Holger Dell**

Goethe University Frankfurt, Germany,  
IT University of Copenhagen, Denmark  
hold@itu.dk

## ABSTRACT

Causality plays an important role in understanding intelligent behavior, and there is a wealth of literature on mathematical models for causality, most of which is focused on causal graphs. Causal graphs are a powerful tool for a wide range of applications, in particular when the relevant variables are known and at the same level of abstraction. However, the given variables can also be unstructured data, like pixels of an image. Meanwhile, the causal variables, such as the positions of objects in the image, can be arbitrary deterministic functions of the given variables. Moreover, the causal variables may form a hierarchy of abstractions, in which the macro-level variables are deterministic functions of the micro-level variables. Causal graphs are limited when it comes to modeling this kind of situation. In the presence of deterministic relationships there is generally no causal graph that satisfies both the Markov condition and the faithfulness condition. We introduce *factored space models* as an alternative to causal graphs which naturally represent both probabilistic and deterministic relationships at all levels of abstraction. Moreover, we introduce *structural independence* and establish that it is equivalent to statistical independence in every distribution that factorizes over the factored space. This theorem generalizes the classical soundness and completeness theorem for  $d$ -separation.

**Keywords** factored space models · deterministic causality · abstraction · structural independence

## 1 Introduction

Learning causal relationships plays a central role in human intelligence [24] [12], and learning a causal model is necessary for any system that can generalize out of distribution [21]. Therefore, a rigorous mathematical theory of causality is key to understanding intelligent behavior.

There is a large body of literature about the mathematical foundations of causality [17], a highly successful area of research, most of which is based on causal graphs. An important feature of causal graphs is that they can encode statistical independence in their structure via *d-separation* [15]. It is commonly assumed that a graph  $G$  is a useful model

---

\*Equal contribution

for a probability distribution  $P$  if it is a *perfect map* of  $P$ , that is, if d-separation in  $G$  implies statistical independence in  $P$  (Markov condition), and the converse (faithfulness condition) [25].

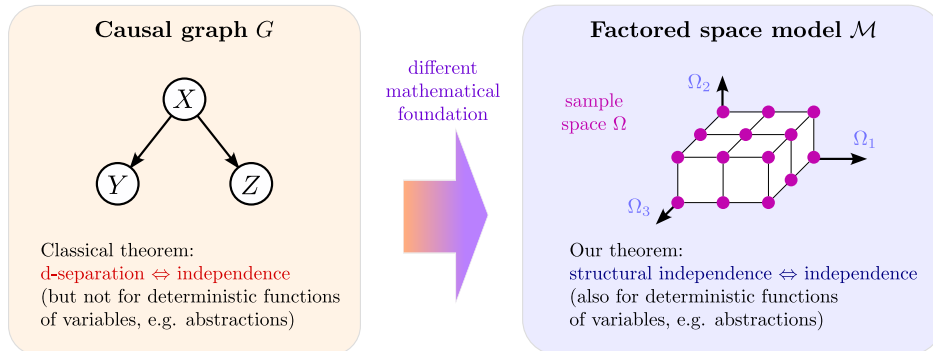


Figure 1: Summary of our results. In a causal graph, the standard criterion for independence is d-separation. However, d-separation is undefined for variables that are deterministic functions of other variables, such as  $A := X + Y + Z$ . To address this issue, we introduce factored space models (FSMs) as an alternative to causal graphs. FSMs are based on expressing the sample space as a Cartesian product, and can be visualized as a hyper-rectangle. FSMs allow us to define *structural independence* as an independence criterion which is also defined for deterministic functions of variables. Further, we establish a theorem that generalizes the soundness and completeness of d-separation [15] to structural independence.

However, as we show in Section 3, when variables are deterministically related, there is generally no perfect map with these variables. This is an important limitation because deterministic relationships appear in many applications. For example, mathematical relationships, equations in physics, and computer programs all contain deterministic relationships between variables. In particular, deterministic relationships are inherent when the causal variables are unknown, such as in *causal representation learning* [23]. For example, if the given variables are the pixels of an image, then the causal variables such as the positions of objects in the image can be arbitrary deterministic functions of the given variables. Moreover, when we model a system at *different levels of abstraction*, there are deterministic relationships between the detailed, micro-level variables and the abstract, macro-level variables. For example, the temperature of gas in a container is a deterministic function of the kinetic energy of all the individual gas particles in the container.

We introduce a framework in which works naturally with deterministic relationships, and treats functions of variables on equal footing. Our main contributions are the following:

1. **Factored space models (FSMs)**, an alternative to causal graphs. FSMs naturally represent all random variables, including deterministic functions of other variables, rather than distinguishing those variables that appear in a particular graph. FSMs are a reframing of *finite factored set models* [8], and they are based on representing the sample space as a Cartesian product, as depicted in Figure 1.
2. **Structural independence**, a criterion that characterizes statistical independence.
3. **Generalizing the soundness and completeness of d-separation**. We establish Theorem 6.2, which generalizes the soundness and completeness theorem for d-separation [15] to variables that can be deterministically related.

**Overview.** The paper proceeds as follows. In section Section 3, we show the limitations of causal graphs when it comes to modeling deterministic relationships. In Section 4, we introduce the FSM framework, prove that structural independence forms a semigraphoid, and show how to construct an FSM from a causal graph. In Section 6, we prove the soundness and completeness of structural independence. Finally, in Section 7, we discuss the implications of our results and provide directions for future work.

## 2 Related work

In this section we review the related work on causal abstractions, and on representing deterministic relationships between variables. We also compare FSMs with causal graphs.

## 2.1 Causal abstractions

The *causal abstractions* framework [3] [1] [22] [27] shares our broad aim, that of understanding relationships between variables not well modeled by causal graphs, and in particular when different variables reside at different levels of abstraction. In the causal abstractions framework, different levels of abstractions are conceived as different causal models. *Transformations* between such models are then used to relate them.

In our work, we seek to provide a language that can simultaneously do the work of causal models and d-separation, and also express the sort of relations between variables that can occur when we speak of microscale and macroscale variables together. For example, we may describe a gas both on a very detailed level in terms of the kinetic energy of each gas particle, and on a more abstract level in terms of the overall temperature of the gas. In particular, a representation for different levels of abstractions should include deterministic relations—we may want to model a macroscale variable as being determined by the microscale variables. One way to regard the division that exists in causal abstractions framework between the causal models on the one hand and the transformations on the other is that it stems from an inability to place variables that stand in such deterministic relations together in one causal graph. This is because no such graph can be a perfect map once we account for these deterministic relations, as we show in Section 3. We do not develop this here, but we are interested in the application of such ideas to neural network interpretability. The causal abstraction framework has been used in this way [10] [9].

## 2.2 Representing deterministic relationships

In a causal graph, the nodes are random variables, and the independences between those variables are represented by d-separation of the nodes. To represent independences between deterministic functions of variables, we need to represent these functions of variables as their own nodes. However, as illustrated in Section 3, when we allow variables that are functions of other variables, d-separation becomes incomplete. There have been various attempts to model causality with variables that are deterministically related. There are approaches like Causal Constraints models [4] and an axiomatization of causal models with constraints [2], which extend causal models to include variables that are deterministically related by a constraint.

However, these approaches only consider an additional set of functional constraints on a fixed set of given variables rather than modeling any deterministic function of variables as its own variable. Moreover, both of these approaches do not provide a criterion for independence in the same way that d-separation provides an independence criterion nodes in a causal graph. In contrast, structural independence on an FSM, our analog of d-separation, is a criterion for independence that correctly represents independence even when variables are deterministically related.

There is work which does consider the independence between deterministically related variables [16], which provides a causal discovery method based on independence of functions of variables. This type of independence between functions of variables is what we represent in FSMs. However, the existing work only applies to the special case of 2 binary variables, and does not provide a general model for how to represent such independences. In contrast, FSMs can model situations with an arbitrary number of discrete variables. The closest to an independence criterion for variables which can have deterministic relationships is D-separation [11] (note the capital ‘D’ to distinguish from d-separation). But just like for d-separation there are simple counterexamples for variables that are independent but not D-separated, as shown in Section 3.

## 2.3 Causal graphs and FSMs

Like a causal graph, an FSM models a set of probability distributions on some variables, and implies certain conditional independence relations between them. Other research has also examined ways of representing probability distributions. In particular, the factors of an FSM can be regarded as factors of a *factor graph* [26]. FSMs generalize causal graphs; we can construct an FSM from a causal graph in order to represent the same independence properties, as we demonstrate in Section 5.2. Here, the factors are independent, which is in line with the spirit of causal modeling, as expressed by the principle of independent mechanisms [20].

### 2.3.1 Soundness and completeness of d-separation

The theory of causal graphs provides us with the d-separation criterion [15], which characterizes which conditional independence relations are implied by a causal graph. Given three sets of variables  $X, Y, Z$  corresponding to sets of nodes of a causal graph,  $X$  and  $Y$  are conditionally independent given  $Z$  in all probability distributions that factorize according to the graph if and only if  $X$  and  $Y$  are d-separated by  $Z$  in the graph.

We prove a more general version of this theorem, in which d-separation in a graph  $G$  is replaced by *structural independence* in a *factored space*. Our theorem is more general because in the classic theorem,  $X$ ,  $Y$ , and  $Z$  are nodes in a graph, while in our version they can be arbitrary variables, including deterministic functions of variables such as  $X + Y$ . In particular, when an FSM is constructed from a causal graph, the classic theorem follows from our theorem.

### 3 Causal graphs cannot capture deterministic relationships

Methods based on causal graphs are highly successful in modeling independence, intervention and counterfactuals. However, as we illustrate in this section, when there is a deterministic relationship between random variables, there is generally no *perfect map* of the probability distribution. That is, there is no graph  $G$  such that d-separation in  $G$  is equivalent to statistical independence.

Let  $\vec{X} := (X_1, \dots, X_n)$  be a vector-valued random variable, and let  $Y := \frac{1}{n} \sum_{i=1}^n X_i$  be the mean of  $\vec{X}$ . One can think of  $Y$  as an abstraction of  $X$ . Further, let  $Z$  be a variable that is causally downstream of  $Y$ , for example  $Z := Y + N$  wherein  $N$  is independent noise. Then  $\vec{X}$  and  $Z$  are independent given  $Y$  (denoted  $\vec{X} \perp\!\!\!\perp Z \mid Y$ ), since when  $Y$  is known,  $\vec{X}$  provides no further information about  $Z$ . Moreover,  $Y \perp\!\!\!\perp Z \mid \vec{X}$  holds, since knowing  $\vec{X}$  fully determines  $Y$ , so  $Z$  provides no further information about  $Y$ . Furthermore, we have  $Z \not\perp\!\!\!\perp (Y, \vec{X})$ .

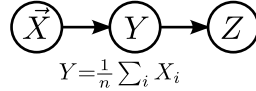


Figure 2: This graph violates the faithfulness condition:  $Y$  and  $Z$  are not d-separated by  $\vec{X}$ , despite  $Y \perp\!\!\!\perp Z \mid \vec{X}$ .

These three independence statements together contradict the *intersection axiom* [19] of d-separation ( $\text{d-sep}(X, Y \mid Z) \wedge \text{d-sep}(X, Z \mid Y) \Rightarrow \text{d-sep}(X, (Y, Z))$ ). Therefore, there is no graph in which the d-separations are equivalent to the statistical independencies. In particular, the graph in Figure 2 is not a perfect map, because it does not reflect that  $Y \perp\!\!\!\perp Z \mid \vec{X}$  holds.

Even *D-separation* [11], which extends d-separation to include nodes that are deterministic functions of their parents, does not cover this case.

### 4 Factored space models

In this section, we formally introduce the factored space model (FSM) framework. We define the building blocks of the framework, including *derived variables*, *history*, *structural independence* and *structural time*.

**Derived variables.** Random variables are a central object of study in statistics and causality. Formally, a *random variable* on a sample space  $\Omega$  is a measurable function  $X : \Omega \rightarrow \text{Val}(X)$  from  $\Omega$  to a value space  $\text{Val}(X)$ . When we speak of variables in this paper, we always mean discrete random variables. Any subset of  $\Omega$  is called *event*.

Our contribution is a better way to represent variables that are deterministically related. That is, some variables are a deterministic function of other variables. We define a variable being a deterministic function of another variable as follows:

**Definition 4.1** (Derived Variable). *Let  $X : \Omega \rightarrow \text{Val}(X)$  and  $Y : \Omega \rightarrow \text{Val}(Y)$  be two random variables, and let  $C \subseteq \Omega$  be an event. Then, we say that  $Y$  is derived from  $X$  on  $C$ , or that it is a deterministic function of  $X$  on  $C$  if there is a function  $f : \text{Val}(X) \rightarrow \text{Val}(Y)$ , such that for all  $\omega \in C$ , it holds that  $Y(\omega) = f(X(\omega))$ . We also write this as  $X \triangleright_C Y$ . We write  $X \triangleright Y$  as a shorthand for  $X \triangleright_\Omega Y$ , and say that  $Y$  is derived from  $X$  or a deterministic function of  $X$ .*

Given multiple variables  $X_1, \dots, X_n$ , we define the variable  $(X_1, \dots, X_n) : \Omega \rightarrow \text{Val}(X_1) \times \dots \times \text{Val}(X_n)$  by  $\omega \mapsto (X_1(\omega), \dots, X_n(\omega))$ . Then, if  $Y \triangleleft (X_1, \dots, X_n)$ , we say that  $Y$  is a deterministic function of the variables  $X_1, \dots, X_n$ .

**Operations on indexed families.** As we make heavy use of indexed families, we introduce some operations on indexed families and probability distributions over sets of indexed families. Formally, an indexed family  $a$  with the index set  $I$ , written as  $a = (a_i)_{i \in I}$  is a function  $f_a : I \rightarrow T$  from an index set  $I$  to a target set  $T$  of possible values. However, rather than being viewed as a function,  $a$  is treated as a collection of indexed elements  $a_i := f_a(i)$ . For example, when we write  $x \in a$ , that means there is an  $i \in I$  with  $x = a_i$ . We define the following operations.

1. **Projection.** Let  $a = (a_i)_{i \in I}$  be an indexed family. Let  $j \in I$ , and  $J \subseteq I$ . Then the *projection* of  $a$  to  $j$  is defined as  $\pi_j(a) := a_j$ . The projection of  $a$  to  $J$  is defined as  $\pi_J(a) := (a_j)_{j \in J}$ . Further, let  $A$  be a set of indexed families which are all defined over the same index set  $I$ . Then, we also define  $\pi_j(A) := \{\pi_j(a) \mid a \in A\}$ . Analogously,  $\pi_J(A) := \{\pi_J(a) \mid a \in A\}$ . (Note that when  $J$  is empty,  $\pi_J(A)$  is a singleton with the empty family as its only element.) We also use the notation  $A_j := \pi_j(A)$  and  $A_J := \pi_J(A)$  as a shorthand.
2. **Merge.** Let  $a = (a_j)_{j \in J}$ ,  $b = (b_k)_{k \in K}$  be two indexed families over disjoint index sets  $J$  and  $K$ . Then  $a \cdot b := (c_i)_{i \in J \cdot K}$  with  $c_i := a_i$  if  $i \in J$  and  $c_i := b_i$  if  $i \in K$ . Note that the merge  $a \cdot b$  is similar to a concatenation of vectors. A concatenation of vectors can be seen as the special case of  $a \cdot b$  when  $J \cdot K$  is ordered and  $j < k$  for all  $j \in J$  and  $k \in K$ . Unlike concatenation, merging is commutative.
3. **Cartesian product.** Let  $(A_i)_{i \in I}$  be a family of sets  $A_i$  over the index set  $I$ . Then its *Cartesian product* is  $\times_{i \in I} A_i := \{(a_i)_{i \in I} \mid \forall i \in I : a_i \in A_i\}$ , which is a singleton if  $I$  is empty. For a set  $B$  of indexed families over  $J$ , and a set  $C$  of indexed families over  $K$ , wherein  $J$  and  $K$  are disjoint, we write  $B \times C := \{b \cdot c \mid b \in B, c \in C\}$ . Note that unlike the Cartesian product over vectors, the Cartesian product over families is commutative because it uses the merge rather than concatenation. Also note that the projection of a Cartesian product is again a Cartesian product. That is, if  $A = \times_{i \in I} A_i$ , then the projection  $A_J$  equals  $\times_{i \in J} A_i$ .

Using these definitions, we now define factored spaces and factored space models.

#### 4.1 Factored spaces and factored space models

We now define two central mathematical structures of our framework — *factored spaces* and *factored space models*.

**Definition 4.2** (Factored Space). *A finite sample space  $\Omega$  is called a factored space if there is a finite index set  $I$  and finite sets  $\Omega_i$  for all  $i \in I$  such that  $\Omega = \times_{i \in I} \Omega_i$  holds. The sets  $\Omega_i$  are called the factors of  $\Omega$ . Moreover, the random variables  $U_i : \Omega \rightarrow \Omega_i$  with  $U_i(\omega) = \pi_i(\omega)$  are called the background variables of  $\Omega$ , and we write  $U = (U_i)_{i \in I}$ .*

The elements of  $\Omega$  are indexed families of the form  $\omega = (\omega_i)_{i \in I}$  with  $\omega_i \in \Omega_i$ . When using factored spaces for modeling probability distributions, we consider those distributions that *factorize* over  $\Omega$ .

**Definition 4.3** (Factorizing Distribution). *Let  $P$  be a probability distribution on a factored space  $\Omega$ . We say that  $P$  factorizes over  $\Omega$  if  $P(\omega) = \prod_{i \in I} P(\omega_i)$  holds for all  $\omega \in \Omega$ , wherein  $P(\omega_i) := P(\pi_i^{-1}(\omega_i))$ .*

Note that  $P$  factorizes over  $\Omega$  if and only if all background variables  $U_i$  are mutually independent under  $P$ . As illustrated in Figure 3, one can think of the factors as axes, and the sample space as a hyper-rectangle along those axes. We remark that  $U : \Omega \rightarrow \Omega$  is the identity function on  $\Omega$ , so any variable  $X : \Omega \rightarrow \text{Val}(X)$  is a deterministic function of  $U$ . This is analogous to every variable in a structural causal model [17] being a deterministic function of the independent background variables  $U$ .

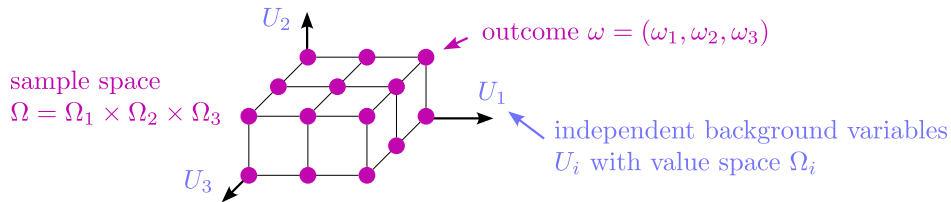


Figure 3: Factored space with three factors.

Next, we formally define the notion of a *factored space model* for a probability distribution  $P$  on a finite set  $\text{Obs}$  of possible observations.

**Definition 4.4** (Factored Space Model). *Let  $\text{Obs}$  be a finite set and let  $P$  be a distribution on  $\text{Obs}$ . Furthermore, let  $O : \Omega \rightarrow \text{Obs}$  be a random variable on a factored space  $\Omega$ . Then we say that the tuple  $\mathcal{M} := (\Omega, O)$  is a factored space model for  $P$  if there is a distribution  $P^\Omega$  that factorizes over  $\Omega$  and satisfies  $P^\Omega(O = o) = P(o)$  for all  $o \in \text{Obs}$ .*

We call a variable  $X$  on  $\Omega$  *observed* if it is a deterministic function of  $O$ , and *unobserved* otherwise. We remark that every probability distribution  $P$  over  $\text{Obs}$  has a trivial factored space model with a single factor (where  $\Omega = \text{Obs}$  and  $|I| = 1$ ), but it may have many other factored space models.

## 4.2 History

In the following, we build up to a notion of *structural independence* of variables on factored space models which is an analog of d-separation and which applies to arbitrary variables. To this end, we first define the *history*  $\mathcal{H}(X | C)$  of a random variable  $X$  given an event  $C \subseteq \Omega$ . The letter  $C$  indicates that this is the event we condition on. We would like to define the history such that it is the set of indices  $i$  of the background variables  $U_i$  that  $X$  depends on if we condition on  $C$  in a distribution that factorizes over  $\Omega$ . One might think that the set of background variables that  $X$  depends on should be defined as the smallest set of variables  $U_J$  such that  $U_J \triangleright_C X$ . However, this condition is not enough, since conditioning on  $C$  may make the background variables dependent. To see this, consider the following example.

Let  $\Omega_1 := \Omega_2 := \{0, 1\}$  be the outcomes of two independent coin flips, and let  $P$  be the joint distribution over the factored space  $\Omega = \Omega_1 \times \Omega_2$ . Then, the background variables  $U_1$  and  $U_2$  are the results of first and second coin respectively. Since the coins are independent,  $P$  factorizes over  $\Omega$ . Further, let  $C := \{00, 11\}$  be the event that both coins have the same result. Note that given  $C$ , we have that  $U_1$  and  $U_2$  are dependent in  $P$ . Thus, the set of background variables that  $U_1$  depends on given  $C$ , is  $\{U_1, U_2\}$ , rather than just  $\{U_1\}$ . When  $C$  is such that it does not introduce a dependence between  $U_J$  and  $U_{I \setminus J}$  for  $J \subseteq I$ , we say that  $J$  *disintegrates*  $C$ .

**Definition 4.5** (Disintegration). *Let  $\Omega = \times_{i \in I} \Omega_i$  be a factored space, let  $C \subseteq \Omega$  be an event and let  $J \subseteq I$ . Then  $J$  disintegrates  $C$  if  $C = C_J \times C_{I \setminus J}$ .*

Note that in the previous example,  $\{1, 2\}$  trivially disintegrates  $C$  but  $\{1\}$  does not disintegrate  $C$ . We can now formally define the history.

**Definition 4.6** (History, Generation). *Let  $\Omega = \times_{i \in I} \Omega_i$  be a factored space with the background variables  $U$ . Let  $J \subseteq I$ , let  $X: \Omega \rightarrow \text{Val}(X)$ , and let  $C \subseteq \Omega$ . Then  $J$  generates  $X$  given  $C$  if  $U_J \triangleright_C X$  and  $C = C_J \times C_{I \setminus J}$ . The history  $\mathcal{H}(X | C)$  of  $X$  given  $C$  is the intersection of all  $J \subseteq I$  that generate  $X$  given  $C$ . That is,*

$$\mathcal{H}(X | C) := \bigcap \{J \subseteq I : J \text{ generates } X \text{ given } C\}. \quad (1)$$

Note that, for any event  $C$ , we have  $C = C_I \times C_\emptyset$  and thus the index set  $I$  trivially disintegrates  $C$ . Moreover, for any variable  $X$ , the index set  $I$  trivially generates  $X$  because  $U_I$  is the identity function on  $\Omega$  and  $X$  is a function  $\Omega \rightarrow \text{Val}(X)$ , so  $X$  is derived from  $U_I$  on  $C$  in the sense of Definition 4.1. Therefore, the set on the right side of (1) contains at least the set  $I$ .

**Lemma 4.7** (History is minimal generating set). *For a factored space  $\Omega = \times_{i \in I} \Omega_i$ , let  $X: \Omega \rightarrow \text{Val}(X)$ , and let  $C \subseteq \Omega$ . Then,  $\mathcal{H}(X | C)$  is the unique minimal set which generates  $X$  given  $C$ .*

*Proof (sketch).* We first show that generation is closed under intersection, see Lemma A.2 in the appendix. It follows that  $\mathcal{H}(X | C)$  generates  $X$  given  $C$ . As  $I$  is finite,  $\mathcal{H}(X | C)$  is the unique minimum of  $\{J \subseteq I \mid J \text{ generates } X \text{ given } C\}$ . Therefore,  $\mathcal{H}(X | C)$  can be equivalently defined as *the minimal set that generates  $X$  given  $C$* .  $\square$

**Shorthand notation for the history.** For an event  $A \subseteq \Omega$ , we write  $\mathcal{H}(A | C)$  as a shorthand for  $\mathcal{H}(1_A | C)$ , wherein  $1_A$  is the variable that is 1 for  $\omega \in A$ , and 0 otherwise. When  $x$  and  $y$  are values of the random variables  $X$  and  $Y$ , we use the notation  $\mathcal{H}(x | C)$ ,  $\mathcal{H}(X | y)$ ,  $\mathcal{H}(A | y)$  or  $\mathcal{H}(x | y)$ . Here,  $x$  is a shorthand for the event  $\{\omega \in \Omega \mid X(\omega) = x\}$  which is a common abbreviation when denoting probabilities, such as  $P(x | y)$ . We write  $\mathcal{H}(X)$ ,  $\mathcal{H}(A)$  and  $\mathcal{H}(x)$  as a shorthand for the unconditional history  $\mathcal{H}(X | \Omega)$ ,  $\mathcal{H}(A | \Omega)$  and  $\mathcal{H}(x | \Omega)$  respectively.

We remark that the disintegration condition is vacuous for the unconditional history, as  $\Omega$  always satisfies  $\Omega = \Omega_J \times \Omega_{I \setminus J}$ .

**History of joint variables.** The following lemma formalizes the fact that if a variable  $X: \Omega \rightarrow \text{Val}(X)$  depends on exactly the positions  $J_1 \subseteq I$  in  $\Omega$  and a variable  $Y: \Omega \rightarrow \text{Val}(Y)$  depends on exactly the positions  $J_2 \subseteq I$ , then the joint variable  $(X, Y)$  depends on exactly the positions  $J_1 \cup J_2$ .

**Lemma 4.8** (History of joint variable). *Let  $X$  and  $Y$  be variables defined on a factored space  $\Omega = \times_{i \in I} \Omega_i$ . For all events  $C \subseteq \Omega$ , we have*

$$\mathcal{H}((X, Y) | C) = \mathcal{H}(X | C) \cup \mathcal{H}(Y | C).$$

The proof is a straightforward application of Lemma 4.7 and can be found in Appendix A.2. Next, we introduce a lemma which establishes the relationship between the history of a variable  $X$  and the histories of its values  $x \in \text{Val}(X)$ .

**Lemma 4.9** (History of a variable is the union of the histories of events). *Let  $X$  be a random variable defined on a factored space  $\Omega$ . For all events  $C \subseteq \Omega$ , we have*

$$\mathcal{H}(X | C) = \bigcup_{x \in \text{Val}(X)} \mathcal{H}(x | C).$$

The proof is an inductive application of Lemma 4.8 and can be found in Appendix A.3.

### 4.3 Structural independence

In the following, we define two variables  $X : \Omega \rightarrow \text{Val}(X)$  and  $Y : \Omega \rightarrow \text{Val}(Y)$  to be structurally independent if they depend on disjoint factors of the factored space  $\Omega = \times_{i \in I} \Omega_i$ . We use our notion of history to formalize this definition and extend it to the situation of conditioning on a third variable. This is our analog of d-separation in Bayes nets.

**Definition 4.10** (Structural Independence). *Let  $X$ ,  $Y$ , and  $Z$  be random variables in a factored space  $\Omega$ . Then,  $X$  and  $Y$  are structurally independent in  $\Omega$ , denoted as  $X \perp^\Omega Y$ , if  $\mathcal{H}(X) \cap \mathcal{H}(Y) = \emptyset$ . Moreover,  $X$  and  $Y$  are structurally independent given  $Z$ , denoted as  $X \perp^\Omega Y | Z$ , if we have*

$$\mathcal{H}(X | z) \cap \mathcal{H}(Y | z) = \emptyset \text{ for all } z \in \text{Val}(Z).$$

We choose the name *structural independence* because it represents those statistical independences which come from a structure in the process that generated the distribution. For example, when throwing two fair coins  $X_1$  and  $X_2$ , then  $X_1$  is statistically independent of the XOR variable  $X_\oplus = X_1 \oplus X_2$ , because  $P(x_1, x_\oplus) = 0.25 = P(x_1)P(x_\oplus)$  holds for all  $x_1, x_\oplus \in \{0, 1\}$ . However, this independence is not structural because it relies on the exact parameters of the process. If one coin is slightly unfair and  $P(X_1 = 0) = 51\%$ , then  $X_1$  and  $X_\oplus$  are not independent anymore. In contrast,  $X_1$  and  $X_2$  are still independent, no matter the distribution of  $X_1$  and  $X_2$ . In that sense, the independence of  $X_1$  and  $X_2$  is a property of the process which generates a class of distributions rather than a property of a particular distribution. The formal connection between statistical independence and structural independence is established in Section 6, where we prove that  $X$  and  $Y$  are independent given  $Z$  in all product probability distributions over  $\Omega$  if and only if  $X$  and  $Y$  are structurally independent given  $Z$  in  $\Omega$ .

### 4.4 Structural time

Let  $X$  and  $Y$  be two random variables. In the following, we formalize a notion of  $X$  being *before*  $Y$  in the sense that  $X$  is always determined before  $Y$  in the process that generated  $X$  and  $Y$ .

**Definition 4.11** (Structural Time). *The structural time  $\leq^\Omega$  of a factored space  $\Omega$  compares two variables  $X$  and  $Y$  on  $\Omega$ . We say that  $X$  is before  $Y$  in  $\Omega$ , denoted  $X \leq^\Omega Y$ , if  $\mathcal{H}(X) \subseteq \mathcal{H}(Y)$ . We say that  $X$  is strictly before  $Y$ , denoted  $X <^\Omega Y$  if  $\mathcal{H}(X) \subsetneq \mathcal{H}(Y)$ .*

We can interpret the history  $\mathcal{H}(X)$  as the set of *sources of randomness* that  $X$  depends on. If  $\mathcal{H}(X) \subseteq \mathcal{H}(Y)$  holds, then  $X$  depends on a subset of the sources of randomness that  $Y$  depends on, which fits with the intuition that  $X$  is determined before  $Y$ . More precisely, in any process in which the values of  $U_i$  become known one after the other, the value of  $X$  will always be determined before the value of  $Y$ , or at the same time.

In Section 5.2 we show that when we construct an FSM from a causal graph, then for variables that are nodes in a causal graph, the structural time is equivalent to the ancestor relation. Structural time is more general than the ancestor relation, since it is defined on all variables on  $\Omega$ , not only the ones which are nodes in the graph.

In the following lemma, we show that the structural time can be equivalently expressed in terms of structural independence. We hold that by this lemma the name *structural time* is philosophically justified, but this topic is outside the scope of this paper.

**Lemma 4.12** (Structural time and structural independence). *Let  $X$  and  $Y$  be random variables on  $\Omega$ . Then  $X$  is structurally before  $Y$  if and only if  $Y \perp^\Omega Z \Rightarrow X \perp^\Omega Z$  holds for all variables  $Z$  on  $\Omega$ .*

The lemma follows from the definition of structural independence, and the proof can be found in Appendix A.4.

## 5 Bayesian networks and factored space models

In this section, we compare factored space models (FSMs) with Bayesian Networks. In Section 5.1 we compare the properties of d-separation and structural independence in terms of the graphoid axioms they fulfill. In Section 5.2 we construct an FSM from a causal graph, and show that node variables in the graph are d-separated if and only if they are structurally independent in the corresponding FSM.

## 5.1 Graphoid and semigraphoid axioms

Structural independence is our analog of d-separation. In this section, we compare the properties of d-separation, which forms a *compositional graphoid* and structural independence, which forms a *compositional semigraphoid*. This comparison also provides an intuition as to why structural independence represents deterministic relationships correctly while d-separation does not.

**Definition 5.1** (Compositional Semigraphoid). A semigraphoid is a set of triplets  $(X, Y, Z)$ , usually denoted as  $X \perp Y \mid Z$ , that satisfy the symmetry, decomposition, weak union, and contraction axioms, as listed in Table 1. A graphoid [19] is a semigraphoid that also satisfies the intersection axiom. A compositional graphoid or semigraphoid additionally satisfies the composition axiom.

Axioms (Semigraphoid: 1.-4., Graphoid: 1.-5.)			Indep. (Probability)	d-Sep. (DAG)	Str. Indep. (FSM)
1. Symmetry:	$X \perp Y \mid W$	$\Rightarrow Y \perp X \mid W$	✓	✓	✓
2. Decomposition:	$X \perp Y, Z \mid W$	$\Rightarrow X \perp Y \mid W$	✓	✓	✓
3. Weak Union:	$X \perp Y, Z \mid W$	$\Rightarrow X \perp Z \mid Y, W$	✓	✓	✓
4. Contraction:	$(X \perp Y \mid W) \wedge (X \perp Z \mid Y, W)$	$\Rightarrow X \perp Y, Z \mid W$	✓	✓	✓
5. Intersection:	$(X \perp Y \mid Z, W) \wedge (X \perp Z \mid Y, W) \wedge (Y \neq Z)$	$\Rightarrow X \perp Y, Z \mid W$	✗	✓	✗
6. Composition:	$(X \perp Y \mid W) \wedge (X \perp Z \mid W)$	$\Rightarrow X \perp Y, Z \mid W$	✗	✓	✓

green: same as independence  
red: different from independence

Table 1: Comparison of **independence** in a probability distribution, **d-separation** in a directed acyclic graph (DAG), and **structural independence** on an FSM with regard to the graphoid and composition axioms. Independence satisfies 1.-4., which means it forms a *semigraphoid* [18], d-separation satisfies 1.-6., which means it forms a *compositional graphoid*. Structural independence satisfies 1.-4. and 6., which means it forms a *compositional semigraphoid*. Note that in terms of these axioms, structural independence is closer to independence than d-separation, as d-separation differs from independence for both the intersection and the composition axiom, while for structural independence the only difference is the composition axiom.

**Proposition 5.2.** *Structural independence is a compositional semigraphoid.*

The proof is straightforward and we defer it to Appendix B.1.

**Composition.** The composition axiom  $(X \perp Y \mid W) \wedge (X \perp Z \mid W) \Rightarrow (X \perp Y, Z \mid W)$  illustrates that structural independence is not the same as independence: Structural independence satisfies this axiom intuitively, because if  $X$  and  $Y$  are influenced by disjoint sources of randomness, and  $X$  and  $Z$  are influenced by disjoint sources of randomness, then  $X$  and  $Y, Z$  are influenced by disjoint sources of randomness as well. To see that statistical independence does not satisfy the composition axiom, consider the following example: Let  $X$  be a message,  $Y$  be a string of uniformly distributed bits, and let  $Z$  be their bitwise XOR:  $Z = X \oplus Y$ . Then  $X$  and  $Y$  are independent, and  $X$  and  $Z$  are independent, but  $X$  and  $(Y, Z)$  are dependent, so composition does not hold. This situation could be modeled with an FSM as follows. The variables  $X$  and  $Y$  are structurally independent, but as  $Z$  is computed from  $X$ , the histories of  $X$  and  $Z$  overlap and  $X$  and  $Z$  are structurally dependent, so composition is not violated. Independence is about the question “Does  $X$  provide information about  $Y$ ?” while structural independence is about the question “Are  $X$  and  $Y$  influenced by the same source of randomness?”.

**Intersection.** The intersection axiom  $(X \perp Y \mid Z, W) \wedge (X \perp Z \mid Y, W) \wedge (Y \neq Z) \Rightarrow (X \perp Y, Z \mid W)$  is particularly interesting. Statistical independence does not in general satisfy the intersection axiom, though intersection does hold in those cases where  $P(x, y, z, w)$  is never zero [18]. However, we want to allow  $P(x, y, z, w) = 0$ . In particular, if any variable is a deterministic function of another variable, there must be values  $x, y, z, w$  with  $P(x, y, z, w) = 0$ . The fact that d-separation satisfies intersection means that d-separation cannot represent the independence relations among a set of variables, some of which are deterministically related. It is therefore an important property of structural independence that the intersection axiom does *not* hold.

In the next section, we continue our comparison of d-separation and structural independence by constructing an FSM from a causal graph.



## 5.2 From Bayesian networks to factored space models

In this section, we construct a factored space from a directed acyclic graph (DAG), and we construct an FSM from a Bayesian network. We show that a distribution  $P$  factorizing over a DAG is equivalent to our construction being a factored space model of  $P$ . We also show that in our construction, structural independence applied to node variables is equivalent to d-separation, and structural time applied to node variables is equivalent to the ancestor relationship. We also introduce the notion of a *perfect map*, and use it to show that factored space models are more expressive than Bayesian networks. We start by formally introducing Bayesian networks.

**Bayesian networks.** Let  $G = (V, E, \text{Val})$  be a directed acyclic graph (DAG), wherein  $V$  is the vertex set, and  $E$  is the edge set. Moreover, each vertex  $v \in V$  is associated with a set of possible values  $\text{Val}_v$  with at least 2 elements, and  $\text{Val} = \times_{v \in V} \text{Val}_v$  is the set of all combinations of values for  $V$ .

For a vertex  $v$ , we write  $\text{pa}(v) \subseteq V$  for the set of parents of  $v$ , and  $\text{an}(v)$  for the set of ancestors of  $v$ . If  $P$  is a distribution over  $\text{Val}$  and  $x_v \in \text{Val}_v$  is a value of  $v$ , we write  $P(x_v)$  as a shorthand for the probability  $P(\{(y_u)_{u \in V} \in \text{Val}_V : x_v = y_v\})$ . We say that  $P$  *factorizes* over  $G$ , if for all  $x = (x_v)_{v \in V} \in \text{Val}$ , we have

$$P(x) = \prod_{v \in V} P(x_v \mid x_{\text{pa}(v)}). \quad (2)$$

Following the notation in [15], a pair  $\mathcal{B} = (G, P)$  where  $P$  factorizes over  $G$  is called Bayesian network or Bayes net. In the following, we show how to construct a factored space  $\Omega$  and a factored space model  $\mathcal{M}^G$  from  $G$ . Factors in this construction are closely related to factor graphs [26].

### 5.2.1 Our construction of a factored space model from a Bayesian network

Given the DAG  $G = (V, E, \text{Val})$ , we now construct a factored space model  $\mathcal{M}^G = (\Omega, O)$  for the observation space  $\text{Obs} = \text{Val}$ . We first construct a factored space  $\Omega$ . To this end, we define the index set  $I$  via  $I := \bigcup_{v \in V} I_v$ , where for all  $v \in V$ , we let

$$I_v := \{(v, x_{\text{pa}(v)}) : x_{\text{pa}(v)} \in \text{Val}_{\text{pa}(v)}\}. \quad (3)$$

In other words, the index set  $I$  is partitioned into the sets  $I_v$ , and each element of  $I_v$  corresponds to a possible value for the vertices in  $\text{pa}(v)$ . If  $v$  is a root, we remark that  $\text{pa}(v) = \emptyset$  holds and that  $\text{Val}_{\text{pa}(v)}$  contains the empty tuple  $()$  as its only element. Finally, for each  $i = (v, x_{\text{pa}(v)}) \in I$ , we define the set  $\Omega_i$  via  $\Omega_i := \text{Val}_v$ . Then  $\Omega := \times_{i \in I} \Omega_i$  is the factored space constructed from  $G$ .

We now define the random variable  $O: \Omega \rightarrow \text{Val}$ . To this end, we first define  $X = (X_v)_{v \in V}$ , wherein  $X_v: \Omega \rightarrow \text{Val}_v$  is a random variable associated with each node  $v$ . We define  $X_v$  recursively as

$$X_v(\omega) := \omega_{(v, X_{\text{pa}(v)}(\omega))}. \quad (4)$$

In particular, for roots  $v$  we have  $X_v(\omega) = \omega_{(v, ())} \in \text{Val}_v$ . Since  $G$  is acyclic,  $X_v$  is well-defined also for non-roots  $v$ . Note that  $\text{Val}(X) = \text{Val}$  and  $\text{Val}(X_v) = \text{Val}_v$ . We call  $\mathcal{M}^G = (\Omega, O)$  with  $O = X$  the *factored space model constructed from  $G$* . This concludes the construction.

To show that  $\mathcal{M}^G$  is in fact a factored space model of a distribution  $P$  over  $\text{Val}$  in the sense of Definition 4.4 when  $P$  factorizes over  $G$ , we first state the following lemma about the relationship between  $G$  and  $\mathcal{M}^G$ .

**Lemma 5.3.** *Let  $\Delta^*(G)$  be the set of distributions on  $\text{Val}$  that factorize over  $G$  and let  $\Delta^\otimes(\Omega)$  be the set of distributions on  $\Omega$  that factorize over  $\Omega$ . Let  $\tau: \Delta^\otimes(\Omega) \rightarrow \Delta^*(G)$  be the function defined as follows for all  $P^\Omega \in \Delta^\otimes(\Omega)$  and  $x \in \text{Val}$ :*

$$\tau(P^\Omega)(x) = P^\Omega(X = x).$$

*Then  $\tau$  is bijective and its inverse  $\tau^{-1}$  satisfies the following for all  $P \in \Delta^*(G)$  and  $\omega \in \Omega$ :*

$$\tau^{-1}(P)(\omega) = \prod_{(v, x_{\text{pa}(v)}) \in I} P(\omega_{v, x_{\text{pa}(v)}} \mid x_{\text{pa}(v)}).$$

The proof can be found in Appendix B.2.

### 5.2.2 Properties of our construction

In this section, we fix a DAG  $G = (V, E, \text{Val})$ , the corresponding factored space model  $\mathcal{M}^G = (\Omega, O)$  constructed from  $G$ , and the node variables  $X_v$  for  $v \in V$ . We show that our construction preserves factorization, d-separation, and the ancestor relation. We start with the factorization property.

**Proposition 5.4** (Factorization property). *For all distribution  $P$  over  $\text{Val}$ , we have that  $P$  factorizes over  $G$  if and only if  $\mathcal{M}^G$  is a factored space model of  $P$ .*

*Proof.* For the forward implication, let  $P$  factorize over  $G$ , that is  $P \in \Delta^*(G)$ . Then, with  $\iota$  defined as in Lemma 5.3, let  $P^\Omega := \iota(P)$ . By Lemma 5.3,  $\tau$  is the inverse of  $\iota$ . Therefore, for all  $x \in \text{Val}$ , we have that  $P(x) = \tau(P^\Omega)(x) = P^\Omega(X^{-1}(x)) = P^\Omega(X = x)$ . Since  $P^\Omega$  also factorizes over  $\Omega$  by the definition of  $\iota$ ,  $\mathcal{M}^G$  is an FSM of  $P$  in the sense of Definition 4.4.

For the backward implication, let  $\mathcal{M}^G$  be an FSM of  $P$ . That is, there is a distribution  $P^\Omega$  that factorizes over  $\Omega$  such that  $P^\Omega(X = x) = P(x)$ . That means  $P = \tau(P^\Omega)$ . By Lemma 5.3,  $P^\Omega = \iota(P)$ , and therefore  $P$  factorizes over  $G$  by the definition of  $\iota$ .  $\square$

Next, we show how d-separation in  $G$  is preserved as structural independence in  $\mathcal{M}^G$ .

**Proposition 5.5** (d-separation). *Let  $V_1, V_2, V_3 \subseteq V$  be three sets of nodes in  $G$ . Then  $V_1$  and  $V_2$  are d-separated given  $V_3$  in  $G$  if and only if  $X_{V_1}$  and  $X_{V_2}$  are structurally independent given  $X_{V_3}$  in  $\mathcal{M}^G$ .*

We defer the straightforward proof to Appendix B.3.

Finally, we show how the ancestor relation in  $G$  is preserved as the structural time relation between the node variables.

**Proposition 5.6** (Ancestor relation). *Let  $v_1, v_2 \in V$  be two distinct nodes in the graph  $G$ . Then  $v_1$  is an ancestor of  $v_2$  in  $G$  if and only if  $X_{v_1} <^\Omega X_{v_2}$  holds.*

*Proof.* The history of  $X_v$  is  $\mathcal{H}(X_v) = \bigcup_{u \in \text{an}(v) \cup \{v\}} I_u$ . Since  $I_v$  is nonempty, we have that  $v_1$  is an ancestor of  $v_2$   $\iff \text{an}(v_1) \cup \{v_1\} \subsetneq \text{an}(v_2) \cup \{v_2\} \iff \bigcup_{v \in \text{an}(v_1) \cup \{v_1\}} I_v \subsetneq \bigcup_{v \in \text{an}(v_2) \cup \{v_2\}} I_v \iff \mathcal{H}(X_{v_1}) \subsetneq \mathcal{H}(X_{v_2})$ , which is exactly the definition of  $X_{v_1} <^\Omega X_{v_2}$ .  $\square$

### 5.2.3 Factored space models are more expressive than Bayesian networks

A DAG whose d-separations capture the independences of a distribution is called a *perfect map* [15]. We now define our analog. That is, an FSM whose structural independences capture the independences of a distribution. With this definition, we can show that FSMs are more expressive than DAGs, in that there are distributions  $P$  such that there is no perfect map DAG of  $P$  while there is a perfect map FSM of  $P$ .

**Definition 5.7** (Perfect map). *Let  $G$  be a DAG with nodes  $V$  and value space  $\text{Val}(X)$ . Further, let  $\mathcal{M} = (\Omega, O)$  be a factored space model. Let  $P$  be a distribution over some observation space  $\text{Obs}$ . Then we say that*

1.  $G$  is a perfect map of  $P$  if  $\text{Val}(X) = \text{Obs}$  and for all sets of nodes  $V_1, V_2, V_3 \subseteq V$ , the sets  $V_1$  and  $V_2$  are d-separated given  $V_3$  in  $G$  if and only if  $X_{V_1}$  and  $X_{V_2}$  are independent given  $X_{V_3}$  in  $P$ .
2.  $\mathcal{M}$  is a perfect map of  $P$  with regard to a family of variables  $X = (X_w)_{w \in W}$  with  $X_w: \Omega \rightarrow \text{Obs}$  if  $\mathcal{M}$  is a factored space model of  $P$  and for all sets of variables  $X_{W_1}, X_{W_2}, X_{W_3} \subseteq X$ , we have that  $X_{W_1}$  and  $X_{W_2}$  are independent given  $X_{W_3}$  in  $P$  if and only if  $X_{W_1}$  and  $X_{W_2}$  are structurally independent given  $X_{W_3}$  in  $\Omega$ .

Perfect maps are often considered to be the graphs that correspond to a distribution. The following proposition shows that factored spaces are more expressive than DAGs.

**Proposition 5.8** (Perfect maps of graphs and factored spaces). *Let  $P$  be a distribution over some observation space  $\text{Obs}$  of the form  $\text{Obs} = \times_{v \in V} \text{Val}(X_v)$ . Then, it holds that*

1. If there is a DAG  $G$  with nodes  $V$  that is a perfect map of  $P$ , then there is a factored space model  $\mathcal{M} = (\Omega, X)$ , with  $X = (X_v)_{v \in V}$  that is also a perfect map of  $P$  with regard to  $X$ .
2. If there is a factored space model  $\mathcal{M}$  that is a perfect map of  $P$  with regard to  $(X_v)_{v \in V}$ , there may still be no DAG  $G$  with nodes  $V$  that is a perfect map of  $P$ .

We defer the proof to Appendix B.4. Proposition 5.8 illustrates that factored space models are more expressive than Bayesian networks. We can construct a factored space model  $\mathcal{M}^G$  for a DAG  $G$ , which is a perfect map for the same distributions as  $G$ . However, it is not possible in general to construct a DAG  $G$  from a factored space model  $\mathcal{M}$  in such a way that  $G$  is a perfect map for the same distributions as  $\mathcal{M}$ .

In the following section, we show that the soundness and completeness theorem of d-separation [15] generalizes to this additional expressiveness. That is, we show the soundness and completeness of structural independence.

## 6 Soundness and Completeness of Structural Independence

This section contains our main result: random variables are conditionally independent in all probability distributions that factorize over a factored space if and only if they are structurally independent. To state the theorem formally, we first review the definition of conditional independence.

**Definition 6.1.** *Let  $P$  be a probability distribution over a sample space  $\Omega$ , and let  $A, B, C \subseteq \Omega$  be events. We say that  $A$  and  $B$  are independent given  $C$ , denoted  $A \perp\!\!\!\perp^P B \mid C$ , if the following holds:*

$$P(A \cap C)P(B \cap C) = P(A \cap B \cap C)P(C). \quad (5)$$

*If  $x, y, z$  are values of the random variables  $X, Y, Z$ , then we write  $x \perp\!\!\!\perp^P y \mid z$  if  $P(x, z)P(y, z) = P(x, y, z)P(z)$  holds. For random variables  $X, Y, Z$ , we write  $X \perp\!\!\!\perp^P Y \mid Z$  if  $x \perp\!\!\!\perp^P y \mid z$  holds for all values  $x, y, z$  of  $X, Y, Z$ .*

Equation (5) is equivalent to the more common definition  $P(A \mid C)P(B \mid C) = P(A \cap B \mid C)$  if  $P(C) > 0$ . If  $P(C) = 0$ , our convention is that  $A \perp\!\!\!\perp^P B \mid C$  holds and indeed Equation (5) is trivially true.

### 6.1 Main theorem

For random variables  $X, Y, Z$  over a factored space  $\Omega$ , how does probabilistic independence  $X \perp\!\!\!\perp^P Y \mid Z$  (Definition 6.1) relate to structural independence  $X \perp\!\!\!\perp^\Omega Y \mid Z$  (Definition 4.10)? We answer this question as follows: Structural independence implies probabilistic independence for all distributions  $P$  that factorize over the factored space (soundness). Conversely, probabilistic independence in all distributions  $P$  that factorize over the factored space implies structural independence (completeness).

**Theorem 6.2** (Soundness and Completeness of Structural Independence). *Let  $X, Y, Z$  be random variables on a factored space  $\Omega = \times_{i \in I} \Omega_i$ . Then the following statements are equivalent.*

- (i)  $X$  and  $Y$  are structurally independent given  $Z$  in  $\Omega$ .
- (ii)  $X \perp\!\!\!\perp^P Y \mid Z$  holds for all probability distributions  $P$  that factorize over  $\Omega$ .

In order to prove Theorem 6.2, we first prove the soundness and completeness of structural independence for *events* in  $\Omega$ , instead of random variables. It is then easy to lift these results to random variables. For brevity, in what follows we write  $\Delta^\otimes(\Omega)$  for the set of all distributions that factorize over  $\Omega$ . The soundness direction ((i)  $\implies$  (ii)) of Theorem 6.2 for events is given by the following lemma.

**Lemma 6.3** (Soundness for Events). *Let  $A, B, C \subseteq \Omega$  be events in a factored space  $\Omega = \times_{i \in I} \Omega_i$ . If the histories satisfy  $\mathcal{H}(A \mid C) \cap \mathcal{H}(B \mid C) = \emptyset$ , then  $A \perp\!\!\!\perp^P B \mid C$  holds for all  $P \in \Delta^\otimes(\Omega)$ .*

The proof is a straightforward transformation of definitions and can be found in Appendix C.1. The completeness direction ((ii)  $\implies$  (i)) of Theorem 6.2 for events is given by the following lemma.

**Lemma 6.4** (Completeness for Events). *Let  $A, B, C \subseteq \Omega$  be events in a factored space  $\Omega = \times_{i \in I} \Omega_i$ . If  $A \perp\!\!\!\perp^P B \mid C$  holds for all  $P \in \Delta^\otimes(\Omega)$ , then the histories satisfy  $\mathcal{H}(A \mid C) \cap \mathcal{H}(B \mid C) = \emptyset$ .*

The proof can be found in Appendix C.3 and is the most technical contribution of this paper. In a sense that we make formally precise, we prove that the history contains exactly those indices of  $I$  that are probabilistically relevant for  $A$  given  $C$ . We show that, if  $A \perp\!\!\!\perp^P B \mid C$  holds for all distributions  $P \in \Delta^\otimes(\Omega)$ , then no index  $i \in I$  can be relevant to both  $A$  given  $C$  and to  $B$  given  $C$ . Since the history contains exactly the relevant indices, this implies  $\mathcal{H}(A \mid C) \cap \mathcal{H}(B \mid C) = \emptyset$ , which establishes Lemma 6.4. Using Lemmas 6.3 and 6.4, the proof of Theorem 6.2 is straightforward.

*Proof of Theorem 6.2.* For all random variables  $X, Y, Z$  on  $\Omega$ , we have:

$$\begin{aligned}
& X \perp^\Omega Y \mid Z \\
\iff & \forall z \in \text{Val}(Z) : \mathcal{H}(X \mid z) \cap \mathcal{H}(Y \mid z) = \emptyset && \text{(Definition 4.10)} \\
\iff & \forall z \in \text{Val}(Z) : \bigcup_{x \in \text{Val}(X)} \mathcal{H}(x \mid z) \cap \bigcup_{y \in \text{Val}(Y)} \mathcal{H}(y \mid z) = \emptyset && \text{(Lemma 4.9)} \\
\iff & \forall (x, y, z) \in \text{Val}(X) \times \text{Val}(Y) \times \text{Val}(Z) : \mathcal{H}(x \mid z) \cap \mathcal{H}(y \mid z) = \emptyset \\
\iff & \forall P \in \Delta^\otimes(\Omega) : \forall (x, y, z) \in \text{Val}(X) \times \text{Val}(Y) \times \text{Val}(Z) : x \perp\!\!\!\perp^P y \mid z && \text{(Lemmas 6.3 and 6.4)} \\
\iff & \forall P \in \Delta^\otimes(\Omega) : X \perp\!\!\!\perp^P Y \mid Z. && \text{(Definition 6.1)}
\end{aligned}$$

This completes the proof.  $\square$

## 6.2 Extension: Strong completeness

Our main theorem, Theorem 6.2, states that the structural independences  $X \perp^\Omega Y \mid Z$  in a factored space  $\Omega$  are exactly those triples  $(X, Y, Z)$  which satisfy the conditional independences  $X \perp\!\!\!\perp^P Y \mid Z$  for *all* distributions  $P \in \Delta^\otimes(\Omega)$ . We now strengthen this result by showing that it is sufficient to exhibit the conditional independences  $X \perp\!\!\!\perp^P Y \mid Z$  “locally” for all distributions  $P$  from *any non-empty open subset*  $S \subseteq \Delta^\otimes(\Omega)$ . Here, we view the set  $\Delta^\otimes(\Omega)$  as a subset of  $\mathbb{R}^\Omega$  with the Euclidean topology.

To strengthen the completeness result, we first show that if a conditional independence holds locally around some distribution  $Q$ , then it holds globally on  $\Delta^\otimes(\Omega)$ . We write  $d : \Delta^\otimes(\Omega)^2 \rightarrow \mathbb{R}$  for the Euclidean distance on  $\Delta^\otimes(\Omega) \subseteq \mathbb{R}^\Omega$ .

**Lemma 6.5** (From local to global conditional independence). *Let  $X, Y, Z$  be random variables in a factored space  $\Omega = \times_{i \in I} \Omega_i$ . Let  $Q \in \Delta^\otimes(\Omega)$  be a distribution. If  $X \perp\!\!\!\perp^{Q'} Y \mid Z$  holds for all  $Q'$  with  $d(Q, Q') < \varepsilon$ , then  $X \perp\!\!\!\perp^P Y \mid Z$  holds for all  $P \in \Delta^\otimes(\Omega)$ .*

The proof is based on the fact that the independence statement has the form of a polynomial, and any polynomial which is zero in an  $\varepsilon$ -environment must be the zero polynomial. It can be found in Appendix C.2.

The following proposition strengthens completeness: It is a priori not necessary that the probabilistic independence holds for *all* distributions that factorize over  $\Omega$ . Rather, if a probabilistic independence holds for any non-empty open set of distributions that factorize over  $\Omega$ , then this already implies structural independence.

**Proposition 6.6** (Strong Completeness). *Let  $X, Y, Z$  be random variables in a factored space  $\Omega$ . If there exists a nonempty open set  $S \subseteq \Delta^\otimes(\Omega)$  with  $X \perp\!\!\!\perp^P Y \mid Z$  for all  $P \in S$ , then  $X \perp^\Omega Y \mid Z$ .*

*Proof.* The proposition follows directly from Theorem 6.2 and Lemma 6.5.  $\square$

## 7 Discussion and conclusion

We have introduced the factored space model framework, and defined *structural independence*, which models statistical independence between variables. Structural independence improves upon d-separation in that it is applicable to variables with deterministic as well as probabilistic relationships. This improvement is useful for modeling systems at different levels of abstraction, which introduces deterministic relationships. We think that having an independence criterion that is applicable *between* levels of abstraction, rather than only within one level of abstraction, is a very natural representation. More speculatively, we believe that it may help us understand self-referencing systems because there may not even be discrete levels of abstraction. For example, some language models can already be described as thinking about their own thinking process [13] [6], which in some sense means that high-level summaries of thoughts are causally influencing object-level thoughts.

Further, we think that a theory of causality that applies across layers of abstraction may be a useful tool to understand agents using the *intentional stance*, that is as systems that are best described in terms of their goals [5], or as systems whose actions are caused by their model of the consequences [14] [7].

### Author contributions.

The contributions of the authors include, but are not limited to, the following:

- **S. Garrabrant** originally conceived of and developed the FSM framework, and found the first proof of the soundness and completeness theorem, as presented in [8].
- **M. G. Mayer** reframed the ideas from [8] to the formalism presented in this paper and developed the more elegant proof of the soundness and completeness theorem presented in Appendix C.
- **M. Wache** clarified the relationship between Bayes Nets and FSM, and wrote most of this paper, including developing the overall structure, making connections to the existing literature, developing examples, creating figures, and formalizing proofs.
- **L. Lang** developed the exposition of the proof for the soundness and completeness theorem.
- **S. Eisenstat** developed precursor ideas and examples that grew into the theory of FSMs, and collaborated with Scott Garrabrant to work out the ideas further.
- **H. Dell** provided guidance and regular feedback and edited the paper in detail.

## References

- [1] Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. “Approximate Causal Abstractions”. In: *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*. Ed. by Amir Globerson and Ricardo Silva. Vol. 115. Proceedings of Machine Learning Research. AUAI Press, 2019, pp. 606–615.
- [2] Sander Beckers, Joseph Halpern, and Christopher Hitchcock. “Causal models with constraints”. In: *Conference on Causal Learning and Reasoning*. PMLR. 2023, pp. 866–879.
- [3] Sander Beckers and Joseph Y. Halpern. “Abstracting Causal Models”. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*. AAAI Press, 2019, pp. 2678–2685. DOI: 10.1609/AAAI.V33I01.33012678.
- [4] Tineke Blom, Stephan Bongers, and Joris M Mooij. “Beyond structural causal models: Causal constraints models”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 585–594.
- [5] Daniel C. Dennett. *The intentional stance*. The MIT Press, 1989. ISBN: 978-0-262-54053-7.
- [6] Aniket Didolkar et al. “Metacognitive Capabilities of LLMs: An Exploration in Mathematical Problem Solving”. In: *arXiv preprint arXiv:2405.12205* (2024).
- [7] Tom Everitt et al. “Agent Incentives: A Causal Perspective”. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*. AAAI Press, 2021, pp. 11487–11495. DOI: 10.1609/AAAI.V35I13.17368.
- [8] Scott Garrabrant. *Temporal Inference with Finite Factored Sets*. 2021. arXiv: 2109.11513 [cs.AI]. URL: <https://arxiv.org/abs/2109.11513>.
- [9] Atticus Geiger, Chris Potts, and Thomas Icard. “Causal abstraction for faithful model interpretation”. In: *arXiv preprint arXiv:2301.04709* (2023).
- [10] Atticus Geiger et al. “Causal abstractions of neural networks”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 9574–9586.
- [11] Dan Geiger, Thomas Verma, and Judea Pearl. “Identifying independence in Bayesian networks”. In: *Networks* 20.5 (1990), pp. 507–534.
- [12] Alison Gopnik and Henry M Wellman. “Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory.” In: *Psychological bulletin* 138.6 (2012), p. 1085.
- [13] Saurav Kadavath et al. “Language models (mostly) know what they know”. In: *arXiv preprint arXiv:2207.05221* (2022).
- [14] Zachary Kenton et al. “Discovering agents”. In: *Artif. Intell.* 322 (2023), p. 103963. DOI: 10.1016/J.ARTINT.2023.103963.
- [15] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009. ISBN: 978-0-262-27738-9.
- [16] Ciarán M. Lee and Robert W. Spekkens. “Causal Inference via Algebraic Geometry: Feasibility Tests for Functional Causal Structures with Two Binary Observed Variables”. In: *Journal of Causal Inference* 5.2 (2017), p. 20160013. DOI: doi:10.1515/jci-2016-0013.
- [17] Judea Pearl. *Causality*. Cambridge University Press, 2009. ISBN: 978-0-521-89560-6.
- [18] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988. ISBN: 978-1-55860-479-7.

- [19] Judea Pearl and Azaria Paz. “Graphoids: Graph-Based Logic for Reasoning about Relevance Relations or When would  $x$  tell you more about  $y$  if you already know  $z$ ?” In: *Advances in Artificial Intelligence II, Seventh European Conference on Artificial Intelligence, ECAI 1986, Brighton, UK, July 20-25, 1986, Proceedings*. Ed. by Benedict du Boulay, David C. Hogg, and Luc Steels. North-Holland, 1986, pp. 357–363.
- [20] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. ISBN: 978-0-262-03731-0.
- [21] Jonathan Richens and Tom Everitt. “Robust agents learn causal world models”. In: *International Conference on Learning Representations (2024)*.
- [22] Paul K. Rubenstein et al. “Causal Consistency of Structural Equation Models”. In: *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. Ed. by Gal Elidan, Kristian Kersting, and Alexander Ihler. AUAI Press, 2017.
- [23] Bernhard Schölkopf et al. “Toward Causal Representation Learning”. In: *Proc. IEEE* 109.5 (2021), pp. 612–634. DOI: 10.1109/JPROC.2021.3058954.
- [24] Steven A Sloman and David Lagnado. “Causality in thought”. In: *Annual review of psychology* 66.1 (2015), pp. 223–247.
- [25] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. The MIT Press, 2000. ISBN: 978-0-262-28415-8.
- [26] Ian H Witten and Eibe Frank. “Data mining: practical machine learning tools and techniques with Java implementations”. In: *Acm Sigmod Record* 31.1 (2002), pp. 76–77.
- [27] Fabio Massimo Zennaro. “Abstraction between Structural Causal Models: A Review of Definitions and Properties”. In: *CoRR* abs/2207.08603 (2022). DOI: 10.48550/ARXIV.2207.08603. arXiv: 2207.08603.

## A Proofs for Section 4

### A.1 History as the minimal generating partition

In this section, we prove that the history  $\mathcal{H}(X | C)$  is the unique minimal partition that generates  $X$  given  $C$ . We start by proving that disintegration is closed under intersection and union. For a subset  $J \subseteq I$ , we use  $\bar{J}$  as a shorthand for  $I \setminus J$ .

**Lemma A.1** (Disintegration closed under intersection and union). *In a factored space  $\Omega = \times_{i \in I} \Omega_i$ , let  $C \subseteq \Omega$  and  $J, K \subseteq I$  such that  $J$  disintegrates  $C$  and  $K$  also disintegrates  $C$ . Then,  $J \cap K$  disintegrates  $C$ , and  $J \cup K$  disintegrates  $C$  as well.*

*Proof.* Let  $J, K \subseteq I$  be such that  $J$  disintegrates  $C$  and  $K$  disintegrates  $C$ . That is, we have  $C = C_J \times C_{\bar{J}}$  and  $C = C_K \times C_{\bar{K}}$ . We claim that  $C = C_{J \cap K} \times C_{\overline{J \cap K}}$  and  $C = C_{J \cup K} \times C_{\overline{J \cup K}}$  hold as well.

We first make the following observation:

(\*) For all  $L \subseteq I$ , we have  $C \subseteq C_L \times C_{\bar{L}}$ .

To see this, let  $c \in C$ , then we have  $c = c_L \cdot c_{\bar{L}}$ , and both  $c_L \in C_L$  and  $c_{\bar{L}} \in C_{\bar{L}}$  hold as required. Then (\*) implies  $C \subseteq C_{J \cap K} \times C_{\overline{J \cap K}}$  and  $C \subseteq C_{J \cup K} \times C_{\overline{J \cup K}}$ , so it remains to prove the reverse inclusions.

To this end, we let  $C^\times := C_{J \cap K} \times C_{\overline{J \cap K}} \times C_{J \cap \bar{K}} \times C_{\overline{J \cap \bar{K}}}$  and claim  $C^\times \subseteq C$ . Let  $c \in C^\times$ . By the definition of  $C^\times$ , there exist  $c^1, c^2, c^3, c^4 \in C$  such that the following holds:

$$\begin{aligned} c &= c^1_{J \cap K} \cdot c^2_{\overline{J \cap K}} \cdot c^3_{J \cap \bar{K}} \cdot c^4_{\overline{J \cap \bar{K}}} = (c^1_J)_{J \cap K} \cdot (c^2_{\bar{J}})_{\overline{J \cap K}} \cdot (c^3_J)_{J \cap \bar{K}} \cdot (c^4_{\bar{J}})_{\overline{J \cap \bar{K}}} \\ &= (c^1_J \cdot c^2_{\bar{J}})_{(J \cap K) \cup (\overline{J \cap K})} \cdot (c^3_J \cdot c^4_{\bar{J}})_{(J \cap \bar{K}) \cup (\overline{J \cap \bar{K}})} = (c^1_J \cdot c^2_{\bar{J}})_K \cdot (c^3_J \cdot c^4_{\bar{J}})_{\bar{K}}. \end{aligned}$$

Since  $J$  disintegrates  $C$  and  $c^1, c^2 \in C$ , we have  $c^1_J \cdot c^2_{\bar{J}} \in C$ . Analogously,  $c^3_J \cdot c^4_{\bar{J}} \in C$ . From this, and the fact that  $K$  disintegrates  $C$ , we obtain  $c \in C$ . Therefore, we have  $C^\times \subseteq C$ .

By (\*), we have  $C \subseteq C_{J \cap K} \times C_{\overline{J \cap K}} \subseteq C^\times$  and  $C \subseteq C_{J \cup K} \times C_{\overline{J \cup K}} \subseteq C^\times$ . Combined with  $C^\times \subseteq C$ , we have proved  $C = C_{J \cap K} \times C_{\overline{J \cap K}}$  and  $C = C_{J \cup K} \times C_{\overline{J \cup K}}$  as required.  $\square$

**Lemma A.2** (Generation closed under intersection). *In a factored space  $\Omega = \times_{i \in I} \Omega_i$ , let  $X: \Omega \rightarrow \text{Val}(X)$ , let  $C \subseteq \Omega$ , and let  $J, K \subseteq I$  be such that  $J$  generates  $X$  given  $C$  and  $K$  generates  $X$  given  $C$ . Then,  $J \cap K$  generates  $X$  given  $C$  as well.*

*Proof.* Let  $J, K \subseteq I$  be such that  $J$  generates  $X$  given  $C$  and  $K$  generates  $X$  given  $C$ . In particular,  $J$  and  $K$  both disintegrate  $C$ . By Lemma A.1, it follows that  $J \cap K$  also disintegrates  $C$ . As both  $J$  and  $K$  generate  $X$  given  $C$ , we have  $U_J \triangleright_C X$  and  $U_K \triangleright_C X$ . We claim that  $U_{J \cap K} \triangleright X$ . Let  $\omega, \omega' \in C$  s.t.  $\omega_{J \cap K} = \omega'_{J \cap K}$ . Since  $J$  generates  $X$  given  $C$ , we have  $\omega^1 = \omega_J \cdot \omega'_{\bar{J}} \in C$  and  $X(\omega) = X(\omega^1)$ . Since  $J \cup K$  disintegrates  $C$ , we have  $\omega^2 = \omega_{J \cap K} \cdot \omega'_{I \setminus (J \cap K)} \in C$ . Because  $\omega^1$  and  $\omega^2$  agree on  $J$ , we have  $X(\omega^1) = X(\omega^2)$ . Now, since  $K$  generates  $X$  given  $C$ , we have  $\omega^2 = \omega_{J \cap K} \cdot \omega'_{I \setminus (J \cap K)} \in C$  and  $X(\omega^1) = X(\omega^2)$ . Since  $\omega_{J \cap K} = \omega'_{J \cap K}$ , we have  $\omega^2 = \omega$  and therefore  $X(\omega) = X(\omega')$   $\square$

Using Lemma A.2, we now prove Lemma 4.7, which we restate here for convenience.

**Lemma 4.7** (History is minimal generating set). *For a factored space  $\Omega = \times_{i \in I} \Omega_i$ , let  $X: \Omega \rightarrow \text{Val}(X)$ , and let  $C \subseteq \Omega$ . Then,  $\mathcal{H}(X | C)$  is the unique minimal set which generates  $X$  given  $C$ .*

*Proof.* By definition,  $\mathcal{H}(X | C) = \bigcap \{J \subseteq I \mid J \text{ generates } X \text{ given } C\}$ . As generation is closed under intersection by Lemma A.2, it follows that  $\mathcal{H}(X | C)$  generates  $X$  given  $C$ . As  $I$  is finite,  $\mathcal{H}(X | C)$  is the unique minimum of  $\bigcap \{J \subseteq I \mid J \text{ generates } X \text{ given } C\}$ .  $\square$

## A.2 History of joint variable

We prove Lemma 4.8, which we restate for convenience.

**Lemma 4.8** (History of joint variable). *Let  $X$  and  $Y$  be variables defined on a factored space  $\Omega = \times_{i \in I} \Omega_i$ . For all events  $C \subseteq \Omega$ , we have*

$$\mathcal{H}((X, Y) | C) = \mathcal{H}(X | C) \cup \mathcal{H}(Y | C).$$

*Proof of Lemma 4.8.* We first prove  $\mathcal{H}((X, Y) | C) \subseteq \mathcal{H}(X | C) \cup \mathcal{H}(Y | C)$ . We begin by observing that  $\mathcal{H}(Y | C) \cup \mathcal{H}(Y | C)$  disintegrates  $C$  because disintegration is closed under union by Lemma A.1. Moreover, by definition of the history, we have  $U_{\mathcal{H}(X|C)} \triangleright_C X$  and  $U_{\mathcal{H}(Y|C)} \triangleright_C Y$ . Thus, we have  $(U_{\mathcal{H}(X|C)}, U_{\mathcal{H}(Y|C)}) \triangleright_C (X, Y)$ , which is equivalent to  $U_{\mathcal{H}(X|C) \cup \mathcal{H}(Y|C)} \triangleright_C (X, Y)$ . Together with the fact that  $\mathcal{H}(X | C) \cup \mathcal{H}(Y | C)$  disintegrates  $C$ , this means that  $\mathcal{H}(X | C) \cup \mathcal{H}(Y | C)$  generates  $(X, Y)$  given  $C$ . By Lemma 4.7, the history is the smallest generating set. Therefore, we have  $\mathcal{H}((X, Y) | C) \subseteq \mathcal{H}(X | C) \cup \mathcal{H}(Y | C)$ .

Now we prove the reverse inclusion, that is,  $\mathcal{H}(X | C) \cup \mathcal{H}(Y | C) \subseteq \mathcal{H}((X, Y) | C)$ . By definition of the history, we have  $U_{\mathcal{H}((X,Y)|C)} \triangleright_C (X, Y)$ , and therefore also  $U_{\mathcal{H}((X,Y)|C)} \triangleright_C X$  and  $U_{\mathcal{H}((X,Y)|C)} \triangleright_C Y$ . Together with the fact that  $\mathcal{H}((X, Y) | C)$  disintegrates  $C$  by definition, this means that  $\mathcal{H}((X, Y) | C)$  generates  $X$  given  $C$ , and  $\mathcal{H}((X, Y) | C)$  also generates  $Y$  given  $C$ . As the history is the smallest generating set by Lemma 4.7, we have  $\mathcal{H}(X | C) \subseteq \mathcal{H}((X, Y) | C)$  and  $\mathcal{H}(Y | C) \subseteq \mathcal{H}((X, Y) | C)$ , and therefore also  $\mathcal{H}(X | C) \cup \mathcal{H}(Y | C) \subseteq \mathcal{H}((X, Y) | C)$ . This completes the proof.  $\square$

## A.3 History as union of histories of events

Here we prove Lemma 4.9, which we restate for convenience.

**Lemma 4.9** (History of a variable is the union of the histories of events). *Let  $X$  be a random variable defined on a factored space  $\Omega$ . For all events  $C \subseteq \Omega$ , we have*

$$\mathcal{H}(X | C) = \bigcup_{x \in \text{Val}(X)} \mathcal{H}(x | C).$$

*Proof.* We first recall that  $\mathcal{H}(x | C)$  is a shorthand for  $\mathcal{H}(1_x | C)$ , where  $1_x$  is the indicator random variable defined via  $1_x(\omega) = 1$  if  $X(\omega) = x$  and  $1_x(\omega) = 0$  otherwise. We observe that each  $1_x$  is a deterministic function of  $X$  and that  $X$  is a deterministic function of the joint variable  $(1_x)_{x \in \text{Val}(X)}$ . In particular, we have  $X \triangleright_C (1_x)_{x \in \text{Val}(X)} \triangleright_C X$ . By transitivity of  $\triangleright_C$ , this implies  $\mathcal{H}(X | C) = \mathcal{H}((1_x)_{x \in \text{Val}(X)} | C)$ . Since  $\text{Val}(X)$  is finite, we can inductively apply Lemma 4.8 to arrive at  $\mathcal{H}((1_x)_{x \in \text{Val}(X)} | C) = \bigcup_{x \in \text{Val}(X)} \mathcal{H}(1_x | C) = \bigcup_{x \in \text{Val}(X)} \mathcal{H}(x | C)$ . Thus, the claim follows.  $\square$

## A.4 Structural time is inclusion

Here we prove Lemma 4.12, which we restate for convenience.

**Lemma 4.12** (Structural time and structural independence). *Let  $X$  and  $Y$  be random variables on  $\Omega$ . Then  $X$  is structurally before  $Y$  if and only if  $Y \perp^\Omega Z \Rightarrow X \perp^\Omega Z$  holds for all variables  $Z$  on  $\Omega$ .*

*Proof.* Let  $X$  and  $Y$  be random variables in a factored space  $\Omega$ . We prove the two directions of the equivalence separately.

“ $\Rightarrow$ ”: Suppose that  $X$  is structurally before  $Y$ . By Definition 4.11, this means  $\mathcal{H}(X) \subseteq \mathcal{H}(Y)$ . We claim that  $Y \perp^\Omega Z \Rightarrow X \perp^\Omega Z$  holds for all variables  $Z$  on  $\Omega$ . To this end, let  $Z$  be any variable on  $\Omega$  and suppose  $Y \perp^\Omega Z$  holds. By Definition 4.10, we have  $\mathcal{H}(Y) \cap \mathcal{H}(Z) = \emptyset$ . By  $\mathcal{H}(X) \subseteq \mathcal{H}(Y)$ , this implies  $\mathcal{H}(X) \cap \mathcal{H}(Z) = \emptyset$  and thus  $X \perp^\Omega Z$  as claimed.

“ $\Leftarrow$ ”: Suppose that  $Y \perp^\Omega Z \Rightarrow X \perp^\Omega Z$  holds for all variables  $Z$  on  $\Omega$ . We claim that  $\mathcal{H}(X) \subseteq \mathcal{H}(Y)$  holds. To this end, let  $i \in \mathcal{H}(X)$  be arbitrary. We choose  $Z = U_i$ , which implies  $\mathcal{H}(Z) = \{i\}$ . Since  $\mathcal{H}(X) \cap \mathcal{H}(Z) = \{i\} \neq \emptyset$ , we have  $X \not\perp^\Omega Z$ . By applying the contrapositive  $X \not\perp^\Omega Z \not\Rightarrow Y \not\perp^\Omega Z$  of the assumption, this implies  $Y \not\perp^\Omega Z$ , which implies  $\mathcal{H}(Y) \cap \mathcal{H}(Z) \neq \emptyset$  and thus  $i \in \mathcal{H}(Y)$ . Since this holds for all  $i \in \mathcal{H}(X)$ , we have  $\mathcal{H}(X) \subseteq \mathcal{H}(Y)$  as claimed, and so  $X$  is structurally before  $Y$ .  $\square$



## B Proofs for Section 5

### B.1 Compositional semigraphoid

Here we prove Proposition 5.2, which states that structural independence is a compositional semigraphoid. We first establish the composition axiom.

**Lemma B.1** (Composition axiom). *Let  $X, Y, Z$ , and  $W$  be random variables in a factored space  $\Omega = \times_{i \in I} \Omega_i$ . If  $X \perp^\Omega Y \mid W$  and  $X \perp^\Omega Z \mid W$  hold, then  $X \perp^\Omega (Y, Z) \mid W$  holds.*

*Proof.* We start by assuming that the premise  $X \perp^\Omega Y \mid W$  and  $X \perp^\Omega Z \mid W$  holds. Let  $w \in \text{Val}(W)$  be arbitrary. By Definition 4.10, we have  $\mathcal{H}(X \mid w) \cap \mathcal{H}(Y \mid w) = \emptyset$  and  $\mathcal{H}(X \mid w) \cap \mathcal{H}(Z \mid w) = \emptyset$ , which implies  $\mathcal{H}(X \mid w) \cap (\mathcal{H}(Y \mid w) \cup \mathcal{H}(Z \mid w)) = \emptyset$ . Together with Lemma 4.8, we obtain  $\mathcal{H}(X \mid w) \cap \mathcal{H}((Y, Z) \mid w) = \emptyset$ . As this holds for all  $w \in \text{Val}(W)$ , we arrive at  $X \perp^\Omega (Y, Z) \mid W$ , which is exactly the conclusion of the composition axiom.  $\square$

The remaining four axioms can be shown directly in a straightforward manner, but we instead use the soundness and completeness theorem for a short proof.

**Proposition 5.2.** *Structural independence is a compositional semigraphoid.*

*Proof.* We need to show axioms 1–4 and 6 in Table 1. Axiom 6 (composition) is true by Lemma B.1. Moreover, Axioms 1–4 follow directly from Theorem 6.2 and the fact that statistical independence is a semigraphoid [18].  $\square$

### B.2 Distribution over DAG and corresponding FSM

Here we provide a proof for Lemma 5.3 which establishes the relationship between a DAG  $G = (V, E, \text{Val})$  and the FSM  $\mathcal{M}^G = (\Omega, X)$  constructed from  $G$ .

First, we prove a preparatory lemma.

**Lemma B.2.** *Let  $G = (V, E, \text{Val})$  be a DAG, let  $\mathcal{M}^G = (\Omega, X)$  be the factored space model constructed from  $G$ . Let  $P^\Omega$  factorize over  $\Omega$ . Then we have*

$$P^\Omega(X = x) = \prod_{v \in V} P^\Omega(U_{(v, x_{\text{pa}(v)})} = x_v).$$

*Proof.* We claim that for any fixed  $x \in \text{Val}$ , we have  $\{X = x\} = \{\omega \in \Omega : \forall v \in V : U_{(v, x_{\text{pa}(v)})}(\omega) = x_v\}$ . We show the inclusion in both directions.

' $\subseteq$ ': Let  $\omega \in \Omega$ ,  $v \in V$ , and let  $X(\omega) = x$ . Recall that  $X$  is defined in (4) as  $X_v(\omega) := \omega_{(v, X_{\text{pa}(v)}(\omega))}$ , so we have that  $x_v = X_v(\omega) = U_{(v, x_{\text{pa}(v)})}(\omega)$ . Therefore,  $\omega \in \{\omega \in \Omega : \forall v \in V : U_{(v, x_{\text{pa}(v)})}(\omega) = x_v\}$ .

' $\supseteq$ ': Let  $\omega \in \Omega$  such that  $\forall v \in V : U_{(v, x_{\text{pa}(v)})}(\omega) = x_v$ . We show that  $X_v(\omega) = x_v$  for all  $v \in V$  by structural induction over  $G$ . If  $v$  is a root node, then  $\text{pa}(v) = \emptyset$ , and by (4) we have  $X_v(\omega) = \omega_{(v, ())} = U_{(v, ())}(\omega) = x_v$ . Induction step: Let  $v \in V$  be arbitrary and  $X_{v'}(\omega) = x_{v'}$  for all  $v' \in \text{pa}(v)$ . Then,  $X_v(\omega) = U_{(v, X_{\text{pa}(v)}(\omega))}(\omega) = U_{(v, x_{\text{pa}(v)})}(\omega) = x_v$ .

We conclude  $P^\Omega(X = x) = P^\Omega(\forall v \in V : U_{(v, x_{\text{pa}(v)})} = x_v) = \prod_{v \in V} P^\Omega(U_{(v, x_{\text{pa}(v)})} = x_v)$ .  $\square$

Now we prove Lemma 5.3, which we restate here for convenience.

**Lemma 5.3.** *Let  $\Delta^*(G)$  be the set of distributions on  $\text{Val}$  that factorize over  $G$  and let  $\Delta^\otimes(\Omega)$  be the set of distributions on  $\Omega$  that factorize over  $\Omega$ . Let  $\tau : \Delta^\otimes(\Omega) \rightarrow \Delta^*(G)$  be the function defined as follows for all  $P^\Omega \in \Delta^\otimes(\Omega)$  and  $x \in \text{Val}$ :*

$$\tau(P^\Omega)(x) = P^\Omega(X = x).$$

*Then  $\tau$  is bijective and its inverse  $\tau^{-1}$  satisfies the following for all  $P \in \Delta^*(G)$  and  $\omega \in \Omega$ :*

$$\tau^{-1}(P)(\omega) = \prod_{(v, x_{\text{pa}(v)}) \in I} P(\omega_{(v, x_{\text{pa}(v)})} \mid x_{\text{pa}(v)}).$$

*Proof.* First we must establish that  $\tau$  is a function from  $\Delta^\otimes(\Omega) \rightarrow \Delta^*(G)$ . Let  $P^\Omega \in \Delta^\otimes(\Omega)$  and  $P := \tau(P^\Omega)$ , we claim that  $P$  factorizes over  $G$ . Indeed, suppose for convenience that  $1, 2, \dots, n \in V$  is a topological ordering of the vertices of  $G$ . Then for all  $x \in \text{Val}$ , we have the following:

$$P(x) = P^\Omega(X = x) = P^\Omega\left(\bigwedge_{v=1}^n X_v = x_v\right) = \prod_{v=1}^n P^\Omega(X_v = x_v \mid X_1 = x_1 \wedge \dots \wedge X_{v-1} = x_{v-1}).$$

Here, the last equality follows from the definition of conditional probability. Recall from (4) that  $X$  is defined via  $X_v(\omega) = \omega_{(v, X_{\text{pa}(v)}(\omega))}$ . Since  $1, \dots, n$  is a topological ordering of  $V$ , by conditioning on  $X_1 = x_1 \wedge \dots \wedge X_{v-1} = x_{v-1}$ , we have  $X_v(\omega) = \omega_{(v, x_{\text{pa}(v)})}$ . Thus, we can continue the calculation as follows:

$$= \prod_{v=1}^n P^\Omega(X_v = x_v \mid X_{\text{pa}(v)} = x_{\text{pa}(v)}) = \prod_{v \in V} P(x_v \mid x_{\text{pa}(v)}).$$

The first equality follows by inductively applying (4):  $X_v$  is a deterministic function of  $X_{\text{pa}(v)}$  and the values  $\omega_{(v, x_{\text{pa}(v)})}$  alone. Therefore, by induction on  $u \in V = \{1, \dots, n\}$ , for all  $u \in V$ , we have that  $X_u$  is a function of the entries  $\omega_{(u', x_{\text{pa}(u')})}$  for  $u' \leq u$ . Under the conditioning  $X_{\text{pa}(v)} = x_{\text{pa}(v)}$ , this shows that  $X_v$  is conditionally independent in  $P^\Omega$  from all  $X_u$  with  $u < v$  and  $u \notin \text{pa}(v)$ . The ultimate equality follows from the definition of  $P$ . The equation shows that  $P$  factorizes over  $G$ , and thus  $P \in \Delta^*(G)$ .

It remains to argue that  $\tau$  is bijective. To do so, it suffices to show that the given function  $\tau^{-1}$  is a function from  $\Delta^*(G) \rightarrow \Delta^\otimes(\Omega)$  and that  $\tau^{-1}$  is indeed the inverse of  $\tau$ . We first observe that the given function  $\tau^{-1}$  is a function from  $\Delta^*(G) \rightarrow \Delta^\otimes(\Omega)$ , because  $\tau^{-1}(P)$  is by definition a product of distributions over  $\Omega_i$  for all  $i \in I$ . Finally, we must show that  $\tau^{-1}$  is the inverse of  $\tau$ , so let  $P^\Omega \in \Delta^\otimes(\Omega)$  and  $P \in \Delta^*(G)$ . We claim that  $\tau^{-1}(\tau(P^\Omega)) = P^\Omega$  and  $\tau(\tau^{-1}(P)) = P$  holds. Indeed, for all  $\omega \in \Omega$ , we have the following:

$$\begin{aligned} \tau^{-1}(\tau(P^\Omega))(\omega) &= \prod_{(v, x_{\text{pa}(v)}) \in I} \tau(P^\Omega)(\omega_{v, x_{\text{pa}(v)}} \mid x_{\text{pa}(v)}) && \text{(by definition of } \tau^{-1}\text{)} \\ &= \prod_{(v, x_{\text{pa}(v)}) \in I} P^\Omega(X_v = \omega_{v, x_{\text{pa}(v)}} \mid X_{\text{pa}(v)} = x_{\text{pa}(v)}) && \text{(by definition of } \tau\text{)} \\ &= \prod_{(v, x_{\text{pa}(v)}) \in I} P^\Omega(U_{v, X_{\text{pa}(v)}}(\omega) = \omega_{v, x_{\text{pa}(v)}} \mid X_{\text{pa}(v)} = x_{\text{pa}(v)}) && \text{(by definition of } X_v\text{)} \\ &= \prod_{(v, x_{\text{pa}(v)}) \in I} P^\Omega(U_{v, x_{\text{pa}(v)}}(\omega) = \omega_{v, x_{\text{pa}(v)}}) && \text{(by independence)} \\ &= \prod_{(v, x_{\text{pa}(v)}) \in I} P^\Omega(\omega_{v, x_{\text{pa}(v)}}) = P^\Omega(\omega). && \text{(since } P^\Omega \text{ factorizes)} \end{aligned}$$

Further, we have

$$\begin{aligned} \tau(\tau^{-1}(P))(x) &= \tau^{-1}(P)(X = x) && \text{(by definition of } \tau\text{)} \\ &= \prod_{v \in V} \tau^{-1}(P)(U_{(v, x_{\text{pa}(v)})} = x_v) && \text{(by Lemma B.2)} \\ &= \prod_{v \in V} \tau^{-1}(P)(U_{(v, x_{\text{pa}(v)})} = x_v \mid X_{\text{pa}(v)} = x_{\text{pa}(v)}) && \text{(by independence)} \\ &= \prod_{v \in V} \tau^{-1}(P)(X_v = x_v \mid X_{\text{pa}(v)} = x_{\text{pa}(v)}) && \text{(by definition)} \\ &= \prod_{v \in V} P(x_v \mid x_{\text{pa}(v)}) && \text{(by definition of } \tau^{-1}\text{)} \\ &= P(x) && \text{(since } P \text{ factorizes)} \end{aligned}$$

Thus,  $\tau$  is a bijective function with inverse  $\tau^{-1}$ . This concludes the proof.  $\square$

### B.3 Equivalence of d-separation and structural independence for nodes

Here we prove Proposition 5.5, which states that d-separation for nodes in a graph  $G$  holds if and only if the corresponding variables are structurally independent in the FSM  $\mathcal{M}^G$  constructed from  $G$ . This proof is an application of Theorem 6.2, but it is also possible to prove Proposition 5.5 directly without Theorem 6.2.

**Proposition 5.5** (d-separation). *Let  $V_1, V_2, V_3 \subseteq V$  be three sets of nodes in  $G$ . Then  $V_1$  and  $V_2$  are d-separated given  $V_3$  in  $G$  if and only if  $X_{V_1}$  and  $X_{V_2}$  are structurally independent given  $X_{V_3}$  in  $\mathcal{M}^G$ .*

*Proof.* By Lemma 5.3, for any distribution  $P$  which factorizes over  $G$ , there is a  $P^\Omega$  over  $\Omega$  with  $P(x) = P^\Omega(X = x)$ . Lemma 5.3 also implies that for any distribution  $P^\Omega$  over  $\Omega$  that factorizes over  $\mathcal{M}^G$  there is a distribution  $P$  which factorizes over  $G$  with  $P(x) = P^\Omega(X = x)$ . Therefore, with Definition 6.1, it follows that for all values  $x_{V_1}, x_{V_2}, x_{V_3} \in \text{Val}_{V_1} \times \text{Val}_{V_2} \times \text{Val}_{V_3}$

$$\begin{aligned} x_{V_1} \perp\!\!\!\perp^P x_{V_2} \mid x_{V_3} \text{ for all distributions } P \text{ that factorize over } G \\ \iff \\ x_{V_1} \perp\!\!\!\perp^{P^\Omega} x_{V_2} \mid x_{V_3} \text{ for all distributions } P^\Omega \text{ that factorize over } \Omega. \end{aligned} \quad (6)$$

We have that

$$\begin{aligned} & V_1 \text{ and } V_2 \text{ are d-separated given } V_3 \text{ in } G. \\ \iff & \text{(i)} \quad X_{V_1} \perp\!\!\!\perp^P X_{V_2} \mid X_{V_3} \text{ for all distributions } P \text{ that factorize over } G. \\ \iff & \text{(ii)} \quad X_{V_1} \perp\!\!\!\perp^{P^\Omega} X_{V_2} \mid X_{V_3} \text{ for all distributions } P^\Omega \text{ that factorize over } \mathcal{M}^G. \\ \iff & \text{(iii)} \quad X_{V_1} \perp\!\!\!\perp^\Omega X_{V_2} \mid X_{V_3}. \end{aligned}$$

Wherein (i) is true by the soundness and completeness of d-separation [15], (ii) follows from (6), and (iii) is true by Theorem 6.2. This concludes the proof.  $\square$

#### B.4 Perfect maps of DAGs and factored spaces

Here we prove Proposition 5.8, which states that if there is a DAG that is a perfect map of a distribution  $P$ , then there is a factored space model that is also a perfect map of  $P$ , but not vice versa.

*Proof of Proposition 5.8.* To prove 1., let  $G$  be a DAG that is a perfect map of  $P$ . We want to prove that  $\mathcal{M}^G = (\Omega, X)$  is a perfect map of  $P$  with regard to  $X$ . Let  $X_{V_1}, X_{V_2}, X_{V_3} \subseteq X$  be arbitrary. Then, by Proposition 5.5,  $X_{V_1}$  and  $X_{V_2}$  are structurally independent given  $X_{V_3}$  in  $\Omega$  if and only if they are d-separated in  $G$ . Since  $G$  is a perfect map, this is equivalent to  $X_{V_1}$  and  $X_{V_2}$  being independent given  $X_{V_3}$ .

To prove 2., consider the following counterexample. Let  $\Omega = \Omega_1 = \{0, 1, 2\}$ , let  $P$  be an arbitrary distribution over  $\Omega$ , and let  $X = (X_1, X_2)$  with  $X_1 = U_1$ , and  $X_2 = [U_1 > 0]$ . Since  $X_2$  is a deterministic function of  $X_1$ , we have the following independence and dependence relations:

nontrivial	trivial independencies	trivial dependencies
$X_2 \perp\!\!\!\perp X_2 \mid X_1$	$X_1 \perp\!\!\!\perp X_1 \mid X_1$	$X_1 \not\perp\!\!\!\perp X_1$
$X_1 \not\perp\!\!\!\perp X_1 \mid X_2$	$X_2 \perp\!\!\!\perp X_2 \mid X_2$	$X_2 \not\perp\!\!\!\perp X_2$
$X_1 \not\perp\!\!\!\perp X_2$	$X_1 \perp\!\!\!\perp X_1 \mid \{X_1, X_2\}$	$X_1 \not\perp\!\!\!\perp \{X_1, X_2\}$
	$X_2 \perp\!\!\!\perp X_2 \mid \{X_1, X_2\}$	$X_2 \not\perp\!\!\!\perp \{X_1, X_2\}$ ,

as well as the independencies which follow directly from the listed independencies. To see that  $\mathcal{M} = (\Omega, X)$  is a perfect map of  $P$ , note that the histories are  $\mathcal{H}(X_1) = \mathcal{H}(X_2) = \{1\}$ ,  $\mathcal{H}(X_1 \mid X_2) = \{1\}$ , and  $\mathcal{H}(X_2 \mid X_1) = \emptyset$ . One can easily verify that the above independencies correspond to the structural independencies. In particular  $X_2 \perp\!\!\!\perp^\Omega X_2 \mid X_1$ . However, there is no graph with nodes  $\{v_1, v_2\}$  such that  $v_2$  is d-separated from itself given  $v_1$ , since the path from  $v_2$  to itself has zero edges and can only be blocked by itself.  $\square$

For a more complex example when there is no graph that is a perfect map for a distribution, see Section 3.

## C Proofs for Section 6

Throughout this section, let  $\Omega = \times_{i \in I} \Omega_i$  be a factored space. Let us state some standard definitions.

**Definition C.1** (Outer product). *Let  $J, K \subseteq I$  be disjoint. Let  $P_J$  and  $P_K$  be distributions over  $\Omega_J$ , and  $\Omega_K$ , respectively. Then the outer product of  $P_J$  and  $P_K$  is the distribution  $P_J \otimes P_K$  over  $\Omega_{J \cup K}$  with  $(P_J \otimes P_K)(\alpha) := P_J(\alpha_J)P_K(\alpha_K)$  for all  $\alpha \in \Omega_{J \cup K}$ .*

**Definition C.2** (Marginal distribution). *Let  $P$  be an arbitrary distribution over  $\Omega$ , and let  $J \subseteq I$ . Then the marginal distribution of  $P$  in  $J$  is given by  $P_J = P \circ U_J^{-1}$ . For  $i \in I$ , we write  $P_i := P_{\{i\}}$ .*

Note that if  $P$  factorizes over  $\Omega$ , then it is the outer product of its marginal distributions in  $i \in I$ . That is,  $P = \bigotimes_{i \in I} P_i$ . In that case, the marginal distribution is  $P_J = \bigotimes_{i \in J} P_i$ . For a finite set  $S$ , we also write  $\Delta(S)$  for the set of all distributions on  $S$ . Then the set  $\Delta^\otimes(\Omega)$  of all distributions that factorize over  $\Omega$  satisfies  $\Delta^\otimes(\Omega) = \{\bigotimes_{i \in I} P_i \mid \forall i \in I: P_i \in \Delta(\Omega_i)\}$ .

### C.1 Proof of soundness for events

In this section, we prove Lemma 6.3, the soundness of structural independence for events. We first characterize derived variables (Definition 4.1) in an alternative way as follows.

**Lemma C.3.** *Let  $X: \Omega \rightarrow \text{Val}(X)$  and  $Y: \Omega \rightarrow \text{Val}(Y)$  be two random variables and  $C \subseteq \Omega$ . Then the following statements are equivalent:*

- (i)  $X \triangleright_C Y$ .
- (ii) For all  $\omega, \omega' \in C$ , we have that  $X(\omega) = X(\omega')$  implies  $Y(\omega) = Y(\omega')$ .

*Proof.* (i)  $\implies$  (ii): By  $X \triangleright_C Y$ , there is a function  $f: \text{Val}(X) \rightarrow \text{Val}(Y)$  with  $Y(\omega) = f(X(\omega))$  for all  $\omega \in C$ . For all  $\omega, \omega' \in C$ , if  $X(\omega) = X(\omega')$ , then  $Y(\omega) = f(X(\omega)) = f(X(\omega')) = Y(\omega')$ , so (ii) holds.

(ii)  $\implies$  (i): We construct  $f: \text{Val}(X) \rightarrow \text{Val}(Y)$  as follows: For  $x \in \text{Val}(X)$ , if there exists  $\omega \in C$  with  $X(\omega) = x$ , define  $f(x) := Y(\omega)$ . By (ii),  $f(x)$  does not depend on the choice of  $\omega \in C$  with  $X(\omega) = x$ . Let  $f(x)$  be arbitrary if there is no  $\omega \in C$  with  $X(\omega) = x$ . By construction, we conclude  $Y(\omega) = f(X(\omega))$  for all  $\omega \in C$ , and thus (i) holds.  $\square$

Recall from Definition 4.6 that  $J$  generates  $X$  given  $C$  if  $J$  disintegrates  $C$  and  $U_J \triangleright_C X$  holds. We naturally extend this definition to events  $A \subseteq \Omega$  in that  $J$  generates  $A$  given  $C$  if  $J$  generates  $1_A$  given  $C$ , where  $1_A$  is the indicator random variable that is 1 for all  $\omega \in A$  and 0 otherwise. In the following lemma, we characterize Definition 4.6 differently.

**Lemma C.4** (Alternative characterization of generation). *Let  $\Omega = \times_{i \in I} \Omega_i$  be a factored space, let  $A, C \subseteq \Omega$ , let  $J \subseteq I$  be a set with complement  $\bar{J} = I \setminus J$ , and suppose that  $J$  disintegrates  $C$ . The following statements are equivalent:*

- (i)  $J$  generates  $A$  given  $C$ .
- (ii) For all  $\omega, \omega' \in C$  with  $\omega_J = \omega'_J$ , we have  $\omega \in A$  if and only if  $\omega' \in A$ .
- (iii)  $A \cap C = (A \cap C)_J \times C_{\bar{J}}$ .

Moreover, the history  $\mathcal{H}(A \mid C)$  is the smallest set  $J$  that disintegrates  $C$  and that satisfies  $A \cap C = (A \cap C)_J \times C_{\bar{J}}$ .

*Proof.* (i)  $\implies$  (ii): Let  $\omega, \omega' \in C$  with  $\omega_J = \omega'_J$ . This means  $U_J(\omega) = U_J(\omega')$ , and by (i) we have  $U_J \triangleright_C 1_A$ . By Lemma C.3, we obtain  $1_A(\omega) = 1_A(\omega')$ . Thus, (ii) holds.

(ii)  $\implies$  (iii): The inclusion  $A \cap C \subseteq (A \cap C)_J \times C_{\bar{J}}$  holds by definition of the Cartesian product. For the other inclusion, we have  $(A \cap C)_J \times C_{\bar{J}} \subseteq C_J \times C_{\bar{J}} = C$  since  $J$  disintegrates  $C$ . Thus, it remains to show that  $(A \cap C)_J \times C_{\bar{J}} \subseteq A$  holds. Let  $\omega \in (A \cap C)_J \times C_{\bar{J}}$ . Then  $\omega = b_J \cdot c_{\bar{J}}$  with  $b \in (A \cap C)_J$  and  $c \in C_{\bar{J}}$ . We have  $b, \omega \in C$  and  $b_J = \omega_J$ , which by (ii) implies that  $b \in A$  holds if and only if  $\omega \in A$  holds. Since  $b \in A$ , we conclude  $\omega \in A$ , which was to be shown. This proves (iii).

(iii)  $\implies$  (i): Let  $\omega, \omega' \in C$  and assume  $U_J(\omega) = U_J(\omega')$ . In order to establish (i), it suffices to show that  $1_A(\omega) = 1_A(\omega')$  holds, since Lemma C.3 then implies  $U_J \triangleright_C 1_A$  as required. For reasons of symmetry, it is enough to show that  $\omega \in A$  implies  $\omega' \in A$ . Assume  $\omega \in A$ , so  $\omega \in (A \cap C)_J \times C_{\bar{J}}$ . We obtain

$$\omega' = \omega'_J \cdot \omega'_{\bar{J}} = \omega_J \cdot \omega'_{\bar{J}} \in (A \cap C)_J \times C_{\bar{J}} = A \cap C,$$

wherein the last step uses (iii). This shows  $\omega' \in A$ . Thus, (i) holds.

Finally, recall from Lemma 4.7 that the history  $\mathcal{H}(A \mid C)$  is the smallest set  $J$  that generates  $A$  given  $C$ . Together with (i)  $\implies$  (iii), this proves the final claim.  $\square$

We are now ready to prove the soundness of structural independence for events.

**Lemma 6.3** (Soundness for Events). *Let  $A, B, C \subseteq \Omega$  be events in a factored space  $\Omega = \times_{i \in I} \Omega_i$ . If the histories satisfy  $\mathcal{H}(A | C) \cap \mathcal{H}(B | C) = \emptyset$ , then  $A \perp\!\!\!\perp^P B | C$  holds for all  $P \in \Delta^\otimes(\Omega)$ .*

*Proof.* Suppose that  $\mathcal{H}(A | C) \cap \mathcal{H}(B | C) = \emptyset$  holds. Let  $P \in \Delta^\otimes(\Omega)$  be an arbitrary distribution that factorizes over  $\Omega$ . We need to prove  $A \perp\!\!\!\perp^P B | C$ .

Let  $J := \mathcal{H}(A | C)$ . By  $\mathcal{H}(A | C) \cap \mathcal{H}(B | C) = \emptyset$ , we have  $\mathcal{H}(B | C) \subseteq \bar{J} = I \setminus J$ . Since  $J$  generates  $A$  given  $C$  by Lemma 4.7, Lemma C.4 implies

$$A \cap C = (A \cap C)_J \times C_{\bar{J}}.$$

We now claim that  $\bar{J}$  generates  $B$  given  $C$ . First,  $\bar{J}$  disintegrates  $C$  since  $J$  disintegrates  $C$ . Since  $\mathcal{H}(B | C) \subseteq \bar{J}$  holds and  $\mathcal{H}(B | C)$  generates  $B$  given  $C$ , we obtain  $U_{\bar{J}} \triangleright_C U_{\mathcal{H}(B|C)} \triangleright_C 1_B$ , and so  $\bar{J}$  generates  $B$  given  $C$ . Thus, Lemma C.4 also implies

$$B \cap C = (B \cap C)_{\bar{J}} \times C_J = C_J \times (B \cap C)_{\bar{J}},$$

wherein we use commutativity of the Cartesian product over indexed families in the last step. Moreover, we observe the following identity:

$$\begin{aligned} A \cap B \cap C &= (A \cap C) \cap (B \cap C) \\ &= ((A \cap C)_J \times C_{\bar{J}}) \cap (C_J \times (B \cap C)_{\bar{J}}) \\ &= ((A \cap C)_J \cap C_J) \times (C_{\bar{J}} \cap (B \cap C)_{\bar{J}}) \\ &= (A \cap C)_J \times (B \cap C)_{\bar{J}}. \end{aligned}$$

Finally, using that  $P$  factorizes over the factors and the previous identities, we obtain

$$\begin{aligned} P(A \cap C) \cdot P(B \cap C) &= P((A \cap C)_J \times C_{\bar{J}}) \cdot P(C_J \times (B \cap C)_{\bar{J}}) \\ &= P_J((A \cap C)_J) \cdot P_{\bar{J}}(C_{\bar{J}}) \cdot P_J(C_J) \cdot P_{\bar{J}}((B \cap C)_{\bar{J}}) \\ &= P((A \cap C)_J \times (B \cap C)_{\bar{J}}) \cdot P(C_J \times C_{\bar{J}}) \\ &= P(A \cap B \cap C) \cdot P(C). \end{aligned}$$

This is exactly the definition of  $A \perp\!\!\!\perp^P B | C$ . □

## C.2 Local-to-global principle

Here we prove the local-to-global principle (Lemma 6.5). We first prove a simple preparatory lemma.

**Lemma C.5.** *Let  $\Omega = \times_{i \in I} \Omega_i$  be a factored space and  $Q, P \in \Delta^\otimes(\Omega)$ . For  $\lambda \in [0, 1]$ , define  $R_i^\lambda = (1 - \lambda)Q_i + \lambda P_i$  and  $R^\lambda = \otimes_{i \in I} R_i^\lambda$ . Let  $A \subseteq \Omega$  be an event. Then the function  $\lambda \mapsto R^\lambda(A)$  is a polynomial of degree at most  $|I|$  in  $\lambda$ .*

*Proof.* We have

$$R^\lambda(A) = \sum_{a \in A} R^\lambda(a) = \sum_{a \in A} \prod_{i \in I} R_i^\lambda(a_i) = \sum_{a \in A} \prod_{i \in I} \left( (1 - \lambda)Q_i(a_i) + \lambda P_i(a_i) \right).$$

Thus,  $R^\lambda(A)$  is a polynomial of degree at most  $|I|$ . □

Now we prove Lemma 6.5, the local-to-global principle, which states that if there is an  $\varepsilon$ -ball  $B$  of distributions  $Q'$  such that  $X \perp\!\!\!\perp^{Q'} Y | Z$  for all  $Q' \in B$ , then  $X \perp\!\!\!\perp^P Y | Z$  for all  $P \in \Delta^\otimes(\Omega)$ .

*Proof of Lemma 6.5.* Let  $B \subseteq \Delta^\otimes(\Omega)$  be the set of all  $Q' \in \Delta^\otimes(\Omega)$  with  $d(Q, Q') < \varepsilon$ . Let  $\varepsilon$  be such that  $X \perp\!\!\!\perp^{Q'} Y | Z$  for all  $Q' \in B$ . Let  $P \in \Delta^\otimes(\Omega)$ ,  $x \in \text{Val}(X)$ ,  $y \in \text{Val}(Y)$ , and  $z \in \text{Val}(Z)$  be arbitrary. We need to show  $x \perp\!\!\!\perp^P y | z$ , or equivalently  $P(x, z)P(y, z) = P(x, y, z)P(z)$ .

With the parameter  $\lambda \in [0, 1]$ , we now define a distribution  $R^\lambda$  that interpolates between  $Q$  and  $P$  as follows: For all  $i \in I$ , we define  $R_i^\lambda = (1 - \lambda)Q_i + \lambda P_i$ , and we let  $R^\lambda = \otimes_{i \in I} R_i^\lambda$ . Then we have  $R^0 = Q$  and  $R^1 = P$ . Note that the map  $\lambda \mapsto R^\lambda$  is continuous. Thus,  $R^\lambda \rightarrow Q$  as  $\lambda \rightarrow 0^+$ . As  $B$  is an open set around  $Q$  and  $R^\lambda \rightarrow Q$  as  $\lambda \rightarrow 0^+$ , there are then infinitely many  $\lambda > 0$  with  $R^\lambda \in B$ .

Now, consider the function

$$f(\lambda) = R^\lambda(x, z)R^\lambda(y, z) - R^\lambda(x, y, z)R^\lambda(z),$$

which by Lemma C.5 is a polynomial of degree at most  $2|I|$ . For infinitely many  $\lambda > 0$ , we have  $R^\lambda \in B$ , and thus  $f(\lambda) = 0$  by construction of  $B$ . Since  $f$  is a polynomial with infinitely many roots, it must be identically zero. Thus

$$P(x, z)P(y, z) - P(x, y, z)P(z) = f(1) = 0,$$

that is,  $x \perp\!\!\!\perp^P y \mid z$ . This concludes the proof.  $\square$

### C.3 Proof of completeness for events

Here we prove the completeness of structural independence for events (Lemma 6.4). We introduce further notation for this section:

- $\perp\!\!\!\perp^\otimes$ : For  $A, B, C \subseteq \Omega$ , we write  $A \perp\!\!\!\perp^\otimes B \mid C$  if  $A \perp\!\!\!\perp^P B \mid C$  holds for all  $P \in \Delta^\otimes(\Omega)$ .
- $\Delta_C^\otimes(\Omega)$ : For an event  $C \subseteq \Omega$ , let  $\Delta_C^\otimes(\Omega) := \{P \in \Delta^\otimes(\Omega) : P(C) > 0\}$ . That is, the set of distributions  $P$  that factorize over  $\Omega$  and assign a positive probability to  $C$ .
- $\Delta_{C,i}^\otimes(\Omega)$ : For  $i \in I$ , let  $\Delta_{C,i}^\otimes(\Omega) := \{(P, Q) : P, Q \in \Delta_C^\otimes(\Omega) \text{ and } P_j = Q_j \text{ for all } j \in I \setminus \{i\}\}$  denote the pairs of distributions in  $\Delta_C^\otimes(\Omega)$  that can only differ in  $i$ .

We aim to show that  $A \perp\!\!\!\perp^\otimes B \mid C$  implies  $\mathcal{H}(A \mid C) \cap \mathcal{H}(B \mid C) = \emptyset$ .

A key ingredient to the proof is the concept of *irrelevant* factors, which form the *cohistory*.

**Definition C.6** (irrelevant factors, cohistory). *Let  $A, C \subseteq \Omega$ , and let  $(P, Q) \in \Delta_{C,i}^\otimes(\Omega)$ . We say that  $i \in I$  is  $(P, Q)$ -irrelevant to  $A$  given  $C$  if  $P(A \mid C) = Q(A \mid C)$ . A factor  $i \in I$  is (globally) irrelevant to  $A$  given  $C$ , if it is  $(P, Q)$ -irrelevant for all  $(P, Q) \in \Delta_{C,i}^\otimes(\Omega)$ . The cohistory of  $A$  given  $C$ , denoted  $\text{Cohistory}(A \mid C)$  is the set of all  $i \in I$  that are irrelevant to  $A$  given  $C$ .*

We say  $i$  is *relevant* if it is not irrelevant, and  $(P, Q)$ -*relevant* if it is not  $(P, Q)$ -irrelevant to  $A$  given  $C$ . Intuitively,  $i$  is relevant to  $A$  given  $C$ , when changing  $P$  only in the factor  $i$  can change  $P(A \mid C)$ . Note that global irrelevance of  $i$  implies the  $(P, Q)$ -irrelevance of  $i$  for all  $(P, Q) \in \Delta_{C,i}^\otimes$ , while global relevance only implies  $(P, Q)$ -relevance for at least one pair  $(P, Q) \in \Delta_{C,i}^\otimes$ .

Recall that for a set  $J \subseteq I$ , we write  $\bar{J} = I \setminus J$  for the complement of  $J$  in  $I$ . We will show the following two lemmas, which combined result in Lemma 6.4:

**Lemma C.7.** *Let  $A, B, C \subseteq \Omega$  with  $A \perp\!\!\!\perp^\otimes B \mid C$ . Then  $\text{Cohistory}(A \mid C) \cup \text{Cohistory}(B \mid C) = I$ .*

**Lemma C.8.** *Let  $A, C \subseteq \Omega$ . We have  $\text{Cohistory}(A \mid C) = \overline{\mathcal{H}(A \mid C)}$ .*

*Proof of Lemma 6.4.* Lemma C.7 states that  $A \perp\!\!\!\perp^\otimes B \mid C$  implies  $\text{Cohistory}(A \mid C) \cup \text{Cohistory}(B \mid C) = I$ . By Lemma C.8, this is equivalent to  $\mathcal{H}(A \mid C) \cap \mathcal{H}(B \mid C) = \emptyset$ .  $\square$

Thus, our task reduces to proving Lemmas C.7 and C.8, to which we devote two separate subsections.

#### C.3.1 Proof of Lemma C.7

Let  $A, B, C \subseteq \Omega$ . and assume  $A \perp\!\!\!\perp^\otimes B \mid C$ . We want to show that  $\text{Cohistory}(A \mid C) \cup \text{Cohistory}(B \mid C) = I$ , which means that for all  $i \in I$ ,  $i$  is globally irrelevant to one of  $A$  or  $B$  given  $C$ . Before we show this, we show a weaker claim, the mutual exclusion principle. It states that for any given pair of distributions  $(P, Q) \in \Delta_{C,i}^\otimes(\Omega)$ ,  $i$  cannot be  $(P, Q)$ -relevant to both  $A$  given  $C$  and  $B$  given  $C$ . However, the mutual exclusion principle makes no claim that  $i$  is globally irrelevant to  $A$  given  $C$  or to  $B$  given  $C$ . We later strengthen the mutual exclusion principle to show the claim about global irrelevance.

**Lemma C.9** (Mutual exclusion principle). *Let  $A, B, C \subseteq \Omega$  with  $A \perp\!\!\!\perp^\otimes B \mid C$ . Let  $i \in I$ , and let  $(P, Q) \in \Delta_{C,i}^\otimes(\Omega)$ . Then  $i$  cannot both be  $(P, Q)$ -relevant to  $A$  given  $C$  and  $(P, Q)$ -relevant to  $B$  given  $C$ .*

*Proof.* We define the distribution  $R$  as  $R(\omega) := (P(\omega) + Q(\omega))/2$ . Since for  $j \neq i$ ,  $P$  and  $Q$  have identical factors, we have  $R_j = P_j = Q_j$ . Further,  $R$  has the factor  $R_i$  with  $R_i(\omega_i) = (P_i(\omega_i) + Q_i(\omega_i))/2$  in  $i$ . Consequently,  $R \in \Delta^{\otimes}(\Omega)$ . By our assumption,  $A \perp\!\!\!\perp^R B \mid C$  holds. We will show that this implies that  $P(A \mid C) = Q(A \mid C)$  or  $P(B \mid C) = Q(B \mid C)$  holds.

The independence  $A \perp\!\!\!\perp^R B \mid C$  means that

$$R(A \cap C) \cdot R(B \cap C) = R(A \cap B \cap C) \cdot R(C).$$

By substituting the definition of  $R$  and multiplying both sides by 4, we obtain

$$(P(A \cap C) + Q(A \cap C)) \cdot (P(B \cap C) + Q(B \cap C)) = (P(A \cap B \cap C) + Q(A \cap B \cap C)) \cdot (P(C) + Q(C)).$$

The left side of the equation multiplies out to

$$P(A \cap C) \cdot P(B \cap C) + Q(A \cap C) \cdot Q(B \cap C) + P(A \cap C) \cdot Q(B \cap C) + Q(A \cap C) \cdot P(B \cap C)$$

and the right side multiplies out to

$$P(A \cap B \cap C) \cdot P(C) + Q(A \cap B \cap C) \cdot Q(C) + P(A \cap B \cap C) \cdot Q(C) + Q(A \cap B \cap C) \cdot P(C).$$

As  $A \perp\!\!\!\perp^P B \mid C$  and  $A \perp\!\!\!\perp^Q B \mid C$ , the first two terms on each side cancel out, and we have

$$P(A \cap C) \cdot Q(B \cap C) + Q(A \cap C) \cdot P(B \cap C) = P(A \cap B \cap C) \cdot Q(C) + Q(A \cap B \cap C) \cdot P(C).$$

By definition of  $\Delta_C^{\otimes}(\Omega)$ ,  $P(C) > 0$  and  $Q(C) > 0$ , and we can divide both sides of the equation by  $P(C) \cdot Q(C)$ .

$$\begin{aligned} P(A \mid C) \cdot Q(B \mid C) + Q(A \mid C) \cdot P(B \mid C) &= P(A \cap B \mid C) + Q(A \cap B \mid C) \\ &= P(A \mid C) \cdot P(B \mid C) + Q(A \mid C) \cdot Q(B \mid C), \end{aligned}$$

where we transformed the right-hand side further by again using the independencies for  $P$  and  $Q$ . By subtracting the right-hand side from the equation, and then factoring the left-hand side of the result, we obtain

$$(P(A \mid C) - Q(A \mid C)) \cdot (Q(B \mid C) - P(B \mid C)) = 0.$$

Thus, at least one of the factors needs to be zero, which means  $P(A \mid C) = Q(A \mid C)$  or  $P(B \mid C) = Q(B \mid C)$ .  $\square$

To strengthen the mutual exclusion principle to global irrelevance via an interpolation argument, we first prove the following simple lemma.

**Lemma C.10.** *Let  $P, P' \in \Delta_C^{\otimes}(\Omega)$ . For  $\lambda \in [0, 1]$ , define  $P_j^\lambda = (1 - \lambda)P_j + \lambda P'_j$  and  $P^\lambda = \bigotimes_{j \in I} P_j^\lambda$ . Then  $P^\lambda \in \Delta_C^{\otimes}(\Omega)$ .*

*Proof.*  $P^\lambda \in \Delta^{\otimes}(\omega)$  by definition. It remains to show that  $P^\lambda(C) > 0$ . We have

$$\begin{aligned} P^\lambda(C) &= \sum_{c \in \mathcal{C}} \prod_{j \in I} \left( (1 - \lambda)P_j(c_j) + \lambda P'_j(c_j) \right) \\ &\geq \sum_{c \in \mathcal{C}} \left( \prod_{j \in I} (1 - \lambda)P_j(c_j) + \prod_{j \in I} \lambda P'_j(c_j) \right) \\ &= \sum_{c \in \mathcal{C}} (1 - \lambda)^{|I|} P(c) + \sum_{c \in \mathcal{C}} \lambda^{|I|} P'(c) \\ &= (1 - \lambda)^{|I|} P(C) + \lambda^{|I|} P'(C) \\ &> 0, \end{aligned}$$

wherein the first step uses the same expansion as the proof of Lemma C.5. For the first inequality, we multiply out the product over  $j$ , and remove all the summands that contain both  $P_j(c_j)$ , and  $P'_j(c_j)$ , that is summands of the form  $(1 - \lambda)P_j(c_j) \cdot \lambda P'_j(c_j) \cdot \dots$ , which are all non-negative.  $\square$

Now we prove Lemma C.7, which states that  $\text{Cohistory}(A \mid C) \cup \text{Cohistory}(B \mid C) = I$ .

*Proof of Lemma C.7.* Let  $i \in I$ . We now show that  $i \in \text{Cohistory}(A | C) \cup \text{Cohistory}(B | C)$ . Since the cohistory is the set of irrelevant factors, we need to show that if  $i$  is relevant to  $A$  given  $C$ , then  $i$  is irrelevant to  $B$  given  $C$ .

Assume  $i$  is relevant to  $A$  given  $C$ . That means there is a tuple  $(P, Q) \in \Delta_{C,i}^{\otimes}(\Omega)$  with  $P(A | C) \neq Q(A | C)$ , i.e.,  $P(A | C) - Q(A | C) \neq 0$ . We view  $\Delta_{C,i}^{\otimes}(\Omega) \subset (\Delta_C^{\otimes}(\Omega))^2 \subset \mathbb{R}^{2|\Omega|}$  as a topological space with the induced standard topology. Since the function  $(P', Q') \mapsto P'(A | C) - Q'(A | C)$  is continuous, there is an open set  $S \subseteq \Delta_{C,i}^{\otimes}(\Omega)$  with  $(P, Q) \in S$  and with  $P'(A | C) \neq Q'(A | C)$  for all  $(P', Q') \in S$ . Consequently, the mutual exclusion principle (Lemma C.9) guarantees that

$$P'(B | C) = Q'(B | C) \text{ for all } (P', Q') \in S. \quad (7)$$

In the rest of the proof, we argue that (7) extends to all  $(P', Q') \in \Delta_{C,i}^{\otimes}(\Omega)$ , which precisely means that  $i \in \text{Cohistory}(B | C)$ . We prove this with a similar strategy to Lemma 6.5.

Let  $(P', Q') \in \Delta_{C,i}^{\otimes}(\Omega)$  be arbitrary. We interpolate between  $(P, Q)$  and  $(P', Q')$  as follows. Let  $P_j^\lambda := (1 - \lambda)P_j + \lambda P'_j$  and  $P^\lambda := \bigotimes_{j \in I} P_j^\lambda$ , and let  $Q^\lambda$  be defined analogously. Then Lemma C.10 shows that  $P^\lambda, Q^\lambda \in \Delta_C^{\otimes}(\Omega)$ . One can also easily verify that  $P_j^\lambda = Q_j^\lambda$  for all  $j \neq i$ , and so  $(P^\lambda, Q^\lambda) \in \Delta_{C,i}^{\otimes}(\Omega)$ .

Now, consider the function  $f : [0, 1] \rightarrow \mathbb{R}$  given by

$$f(\lambda) := P^\lambda(B \cap C)Q^\lambda(C) - Q^\lambda(B \cap C)P^\lambda(C).$$

By Lemma C.5, this function is a polynomial in  $\lambda$ . We have  $(P^\lambda, Q^\lambda) \rightarrow (P, Q)$  for  $\lambda \rightarrow 0^+$ , which means there are infinitely many  $\lambda > 0$  for which  $(P^\lambda, Q^\lambda) \in S$ . For such  $\lambda$ , we have  $P^\lambda(B | C) = Q^\lambda(B | C)$  by (7). Multiplying with  $P^\lambda(C) \cdot Q^\lambda(C)$ , we have that  $P^\lambda(B \cap C)Q^\lambda(C) = Q^\lambda(B \cap C)P^\lambda(C)$ . So, for all  $\lambda$  with  $(P^\lambda, Q^\lambda) \in S$  we have  $f(\lambda) = 0$ . Thus, the polynomial  $f$  has infinitely many roots and must be the zero polynomial. Consequently, we have  $f(1) = 0$ , and thus  $P'(B | C) = Q'(B | C)$ . Since  $P'$  and  $Q'$  are arbitrary,  $i$  is globally irrelevant to  $B$  given  $C$ . This concludes the proof.  $\square$

### C.3.2 Proof of Lemma C.8

Let  $A, C \subseteq \Omega$ . In this section, we show that  $\text{Cohistory}(A | C) = \overline{\mathcal{H}(A | C)}$ . Since  $\overline{\text{Cohistory}}$  is the set of relevant factors, this means the history contains exactly the relevant factors. The direction “ $i$  is relevant to  $A$  given  $C \implies i \in \mathcal{H}(A | C)$ ” is easy, while the direction “ $i \in \mathcal{H}(A | C) \implies i$  is relevant to  $A$  given  $C$ ” is much more difficult.

**Proof sketch for Lemma C.8.** We know that  $\mathcal{H}(A | C)$  is the smallest set that generates  $A$  given  $C$ , so it suffices to prove that the set of relevant factors generates  $A$  given  $C$ . This involves showing that the set of relevant factors disintegrates  $C$ , Lemma C.20. To prove Lemma C.20, we establish  $C$  as the support of a distribution that factorizes into two components. Establishing that factorization relies on Lemma C.19, which states that the background variables of the cohistory are independent of the background variables of its complement given  $C$ . Then, the final proof directly verifies the generation property by using the alternative characterization of generation established in Lemma C.4. All other lemmas are preparatory.

**Further notation for conditional independence.** In Definition 6.1, we introduced the notation  $X \perp\!\!\!\perp^P Y | Z$  for random variables  $X, Y, Z$ , and  $A \perp\!\!\!\perp^P B | C$  for events  $A, B, C$ . In what follows, we also mix these notations. For example,  $X \perp\!\!\!\perp^P Y | C$  means  $x \perp\!\!\!\perp^P y | C$  for all  $x \in \text{Val}(X)$  and  $y \in \text{Val}(Y)$ , where  $x$  and  $y$  are treated as the events  $X^{-1}(x)$  and  $Y^{-1}(y)$ , respectively. Similarly,  $X \perp\!\!\!\perp^P B | C$  means  $x \perp\!\!\!\perp^P B | C$  for all  $x \in \text{Val}(X)$ . Finally, we also write  $\perp\!\!\!\perp^{\otimes}$  if the independence holds for all  $P \in \Delta^{\otimes}(\Omega)$ , for example  $X \perp\!\!\!\perp^{\otimes} B | C$ .

We start by considering the *support* of a distribution  $P$  over  $\Omega$ , namely the event  $\text{supp}(P) \subseteq \Omega$  such that  $P(\omega) > 0$  for all  $\omega \in \text{supp}(P)$ . We obtain the following lemma.

**Lemma C.11.** *Let  $J \subseteq I$  and let  $P, Q \in \Delta(\Omega)$  be two distributions over  $\Omega$ , and let  $C \subseteq \Omega$ . Then we have*

1. *If  $\text{supp}(P) \subseteq \text{supp}(Q)$ , and  $P(C) > 0$ , then  $Q(C) > 0$ .*
2. *If  $P = P_J \otimes P_{\bar{J}}$ , then,  $\text{supp}(P) = \text{supp}(P_J) \times \text{supp}(P_{\bar{J}})$ .*
3. *If  $P = P_J \otimes P_{\bar{J}}$ ,  $Q = Q_J \otimes Q_{\bar{J}}$ ,  $\text{supp}(P_J) \subseteq \text{supp}(Q_J)$ , and  $P(C) > 0$ , then  $Q(C) > 0$ .*

Note that  $P$  and  $Q$  do not have to factorize over  $\Omega$  here.



*Proof of Lemma C.11.* 1. Let  $\text{supp}(P) \subseteq \text{supp}(Q)$ . We now prove the contraposition. If  $Q(C) = 0$ , then  $C \cap \text{supp}(Q) = \emptyset$  by definition of the support. It follows that  $C \cap \text{supp}(P) = \emptyset$ , and thus,  $P(C) = 0$ .

2. Let  $P = P_J \otimes P_{\bar{J}}$ . For all  $\alpha \in \Omega_J$  and  $\beta \in \Omega_{\bar{J}}$ , we have

$$\begin{aligned} \alpha \cdot \beta \in \text{supp}(P) &\iff 0 < P(\alpha \cdot \beta) = P_J(\alpha)P_{\bar{J}}(\beta) \\ &\iff P_J(\alpha) > 0 \text{ and } P_{\bar{J}}(\beta) > 0 \\ &\iff \alpha \in \text{supp}(P_J) \text{ and } \beta \in \text{supp}(P_{\bar{J}}) \\ &\iff \alpha \cdot \beta \in \text{supp}(P_J) \times \text{supp}(P_{\bar{J}}). \end{aligned}$$

3. Let  $P = P_J \otimes P_{\bar{J}}$ , and  $Q = Q_J \otimes P_{\bar{J}}$ , let  $\text{supp}(P_J) \subseteq \text{supp}(Q_J)$ , and let  $P(C) > 0$ . By 2., we have

$$\text{supp}(P) = \text{supp}(P_J) \times \text{supp}(P_{\bar{J}}) \subseteq \text{supp}(Q_J) \times \text{supp}(P_{\bar{J}}) = \text{supp}(Q). \quad (8)$$

Since  $P(C) > 0$ , it follows from 1. that  $Q(C) > 0$ . This concludes the proof.  $\square$

We now use Lemma C.11 to show the following lemma.

**Lemma C.12.** *Let  $A, C \subseteq \Omega$  be events and  $P, Q \in \Delta_C^{\otimes}(\Omega)$ . Further, let  $P_j = Q_j$  for all  $j$  that are relevant to  $A$  given  $C$ . Then  $P(A | C) = Q(A | C)$ .*

*Proof.* Essentially, the proof is a progressive application of the definition of irrelevance. We change one factor at a time to transform  $P$  into  $Q$ . However, as the fact that  $j$  is irrelevant to  $A$  given  $C$  only makes a statement about distributions in  $\Delta_C^{\otimes}(\Omega)$ , we need to prove that the intermediate distributions are all in  $\Delta_C^{\otimes}(\Omega)$ .

Let  $J := \text{Cohistory}(A | C)$ . Let  $Q$  be such that  $Q_J$  is strictly positive. (Later we strengthen the claim to arbitrary distributions  $Q \in \Delta_C^{\otimes}(\Omega)$ .) Let  $P_j = Q_j$  for all  $j$  that are relevant to  $A$  given  $C$ . Since  $J$  is finite, we can express it as  $J = \{j_1, \dots, j_q\}$  with mutually distinct elements  $j_k$ . We now progressively replace factors in  $P$  until it is equal to  $Q$ . We do this by defining a sequence  $P^0, P^1, \dots, P^q$  of distributions, wherein  $P^0 = P$  and  $P^q = Q$ . Then we show that  $P(A | C)^{k-1} = P(A | C)^k$  for all  $k \in \{1, \dots, q\}$ . We define  $P^n$  recursively as follows.

1. For all  $i \in I$ , set  $P_i^0 := P_i$ .
2. For all  $k \in \{1, \dots, q\}$  and  $j \in I$ , set  $P_j^k := \begin{cases} P_j^{k-1} & \text{if } j \neq j_k \\ Q_j & \text{if } j = j_k. \end{cases}$
3. Set  $P^k = \bigotimes_{i \in I} P_i^k$ .

Note that by construction, we have that  $P^q = Q$ , and also  $P^{k-1}$  and  $P^k$  can only differ in  $j_k$ . Next we show by induction that for all  $k \in \{1, \dots, q\}$ , we have  $P^k(C) > 0$ . First, we note that since  $P^0 = P$ , we have  $P^0(C) > 0$ . We now assume that  $P^{k-1}(C) > 0$  (induction assumption). Since we assumed that  $Q_J$  is strictly positive,  $P_{j_k}^k = Q_{j_k}$  is strictly positive. As  $P_{I \setminus \{j_k\}}^k = P_{I \setminus \{j_k\}}^{k-1}$ , we have that  $\text{supp}(P^k) \supseteq \text{supp}(P^{k-1})$  by Lemma C.11. Therefore,  $P^{k-1}(C) > 0$  implies  $P^k(C) > 0$ . That means,  $P^k \in \Delta_C^{\otimes}(\Omega)$  for all  $k \in \{1, \dots, q\}$ . As all  $j_k$  are irrelevant to  $A$  given  $C$  by definition, we have that

$$P(A | C) = P^0(A | C) = P^1(A | C) = \dots = P^q(A | C) = Q(A | C).$$

This proves the lemma for those  $Q$  for which  $Q_J$  is strictly positive. We now prove the general case. Let  $Q \in \Delta_C^{\otimes}(\Omega)$  be arbitrary. Let  $Q'$  be a probability distribution with  $Q'_{\bar{J}} = P_{\bar{J}} = Q_{\bar{J}}$  such that  $Q'_J$  is strictly positive, for example  $Q'_J$  could be chosen as the uniform distribution. From Lemma C.11 part 2, it follows that  $\text{supp}(Q') \supseteq \text{supp}(Q)$ . Together with  $Q(C) > 0$ , this implies  $Q'(C) > 0$  by Lemma C.11, part 1. Then  $P(A | C) = Q'(A | C)$ . By the same argument with  $Q$  instead of  $P$ , we obtain  $Q(A | C) = Q'(A | C)$ , and so  $P(A | C) = Q(A | C)$ . This concludes the proof.  $\square$

Now we prove a simple preparatory lemma, namely that mixed independence satisfies the following decomposition graphoid property.

**Lemma C.13** (Decomposition of mixed independence). *Let  $Y, Z$  be two random variables on  $\Omega$  and let  $B, C \subseteq \Omega$ . Then,*

$$B \perp\!\!\!\perp^P (Y, Z) | C \implies B \perp\!\!\!\perp^P Y | C.$$

**Corollary C.14.** *For  $J' \subseteq J \subseteq I$ , the independence  $B \perp\!\!\!\perp^P U_J | C$  implies  $B \perp\!\!\!\perp^P U_{J'} | C$ .*

*Proof.* Let  $B \perp^P (Y, Z) \mid C$ . Let  $y \in \text{Val}(Y)$ . Then we have

$$\begin{aligned} P(B \cap C) \cdot P(Y^{-1}(y) \cap C) &= \sum_{z \in \text{Val}(Z)} P(B \cap C) \cdot P((Y, Z)^{-1}(y, z) \cap C) \\ &= \sum_{z \in \text{Val}(Z)} P(B \cap (Y, Z)^{-1}(y, z) \cap C) \cdot P(C) \quad (\text{by } B \perp^P (Y, Z) \mid C) \\ &= P(B \cap Y^{-1}(y) \cap C) \cdot P(C). \end{aligned}$$

That shows  $B \perp^P Y \mid C$ . The corollary follows from  $U_J = (U_{J'}, U_{J \setminus J'})$ .  $\square$

**Lemma C.15.** *Let  $J \subseteq I$ , and let  $P$  be a distribution over  $\Omega$  with  $P = P_J \otimes P_{\bar{J}}$ . Further, let  $A_J \subseteq \Omega_J$ , and  $A_{\bar{J}} \subseteq \Omega_{\bar{J}}$ . Then we have  $P(A_J \times A_{\bar{J}}) = P_J(A_J) \cdot P_{\bar{J}}(A_{\bar{J}})$ .*

*Proof.*

$$\begin{aligned} P_J(A_J) \cdot P_{\bar{J}}(A_{\bar{J}}) &= \sum_{\omega_J \in A_J} P_J(\omega_J) \sum_{\omega_{\bar{J}} \in A_{\bar{J}}} P_{\bar{J}}(\omega_{\bar{J}}) \\ &= \sum_{\substack{\omega_J \in A_J \\ \omega_{\bar{J}} \in A_{\bar{J}}}} P_J(\omega_J) P_{\bar{J}}(\omega_{\bar{J}}) \\ &= \sum_{\omega_J \cdot \omega_{\bar{J}} \in A_J \times A_{\bar{J}}} P(\omega_J \cdot \omega_{\bar{J}}) \quad (P = P_J \otimes P_{\bar{J}}) \\ &= P(A_J \times A_{\bar{J}}) \end{aligned}$$

$\square$

For  $J \subseteq I$ ,  $\alpha \in \Omega_J$  and  $D \subseteq \Omega$ , we define  $D^\alpha$  as

$$D^\alpha := \{\omega \in D : \omega_J = \alpha\} = D \cap U^{-1}(\alpha) \quad (9)$$

We obtain the following lemma.

**Lemma C.16.** *Let  $P$  be a distribution that factorizes over  $\Omega$ , let  $J \subseteq I$ ,  $\alpha \in \Omega_J$ , and  $D \subseteq \Omega$ . Further, let  $\delta_\alpha \in \Delta(\Omega_J)$  be the delta distribution at  $\alpha$ , which is the distribution that assigns probability 1 to  $\alpha$ . Then we have the following identities:*

1.  $P(D^\alpha) = P_J(\alpha) \cdot P_{\bar{J}}(D_{\bar{J}}^\alpha)$ .
2.  $(\delta_\alpha \otimes P_{\bar{J}})(D) = P_{\bar{J}}(D_{\bar{J}}^\alpha)$ .

*Proof.* Part 1 follows from  $D^\alpha = \{\alpha\} \times D_{\bar{J}}^\alpha$  and from Lemma C.15. For part 2, we note

$$(\delta_\alpha \otimes P_{\bar{J}})(D) = \sum_{\omega \in D} (\delta_\alpha \otimes P_{\bar{J}})(\omega) = \sum_{\omega \in D} \delta_\alpha(\omega_J) P_{\bar{J}}(\omega_{\bar{J}}) = \sum_{\omega \in D^\alpha} P_{\bar{J}}(\omega_{\bar{J}}) = \sum_{\omega_{\bar{J}} \in D_{\bar{J}}^\alpha} P_{\bar{J}}(\omega_{\bar{J}}) = P_{\bar{J}}(D_{\bar{J}}^\alpha).$$

$\square$

Ultimately, the goal is to show that  $\overline{\mathcal{H}(A \mid C)} = \text{Cohistory}(A \mid C)$  (Lemma C.8). First, we show that  $U_{\overline{\mathcal{H}(A \mid C)}}$  and  $U_{\text{Cohistory}(A \mid C)}$  have similar independence properties, namely that  $U_J \perp^{\otimes} U_{\bar{J}} \mid C$  holds both for  $J = \mathcal{H}(A \mid C)$  (Lemma C.17) and for  $J = \text{Cohistory}(A \mid C)$  (Lemma C.19). Note that Lemma C.17 is not necessary for our proof of Lemma C.8, but rather serves as a motivation why Lemma C.19 is a useful step in the direction of showing that  $\overline{\mathcal{H}(A \mid C)} = \text{Cohistory}(A \mid C)$ .

**Lemma C.17.** *Let  $J = \mathcal{H}(A \mid C)$ . Then we have  $U_J \perp^{\otimes} U_{\bar{J}} \mid C$ .*

*Proof.* Let  $P \in \Delta^{\otimes}(\Omega)$ ,  $\alpha \in \Omega_J$  and  $\beta \in \Omega_{\bar{J}}$ . We first observe that  $\{U_J = \alpha\} \cap (C_J \times C_{\bar{J}}) = (\{\alpha\} \cap C_J) \times C_{\bar{J}}$ . Since  $C = C_J \times C_{\bar{J}}$  by definition of the history, we have

$$\begin{aligned} P(U_J^{-1}(\alpha) \cap C) P(U_{\bar{J}}^{-1}(\beta) \cap C) &= P\left(\left(\{\alpha\} \cap C_J\right) \times C_{\bar{J}}\right) P\left(\left(\{\beta\} \cap C_{\bar{J}}\right) \times C_J\right) \\ &= P_J(\{\alpha\} \cap C_J) P_{\bar{J}}(C_{\bar{J}}) P_{\bar{J}}(\{\beta\} \cap C_{\bar{J}}) P_J(C_J) && \text{(Lemma C.15)} \\ &= P(\{\alpha \cdot \beta\} \cap C) P(C) && \text{(Lemma C.15)} \\ &= P(U_J^{-1}(\alpha) \cap U_{\bar{J}}^{-1}(\beta) \cap C) P(C). \end{aligned}$$

This precisely means that  $U_J \perp\!\!\!\perp^{\otimes} U_{\bar{J}} \mid C$ . □

In Lemma C.19, we show that this independence also holds with  $\text{Cohistory}(A \mid C)$  in place of  $\mathcal{H}(A \mid C)$ . We first show that the independence holds for  $A$  and  $U_{\text{Cohistory}(A \mid C)}$ .

**Lemma C.18.** *Let  $A, C \subseteq \Omega$ , and let  $J = \text{Cohistory}(A \mid C)$ . Then  $A \perp\!\!\!\perp^{\otimes} U_J \mid C$ .*

*Proof.* Let  $P \in \Delta^{\otimes}(\Omega)$  and  $\alpha \in \Omega_J$ . Let  $B := U_J^{-1}(\alpha)$ . We need to show  $A \perp\!\!\!\perp^P B \mid C$ , i.e.,  $P(A \cap C) \cdot P(B \cap C) = P(A \cap B \cap C) \cdot P(C)$ . Note that if  $P(B \cap C) = 0$ , then both sides of the equation are zero and the independence holds. So we can assume that  $P(B \cap C) > 0$ .

Let  $\delta_{\alpha} \in \Delta^{\otimes}(\Omega_J)$  be the delta distribution at  $\alpha$ . For this proof, we would like to apply Lemma C.12 to the distributions  $(\delta_{\alpha} \otimes P_{\bar{J}})$ , and  $P$ . The conditions we need in order to apply Lemma C.12 are

1.  $P \in \Delta^{\otimes}(\Omega)$
2.  $(\delta_{\alpha} \otimes P_{\bar{J}}) \in \Delta^{\otimes}(\Omega)$
3.  $(\delta_{\alpha} \otimes P_{\bar{J}})_i = P_i$  for all factors  $i$  that are relevant to  $A$  given  $C$ .

1. is satisfied since  $P \in \Delta^{\otimes}(\Omega)$  by our assumption, and  $P(C) > 0$  because  $P(B \cap C) > 0$ . For 2., we need to show that  $(\delta_{\alpha} \otimes P_{\bar{J}}) \in \Delta^{\otimes}(\Omega)$  and  $(\delta_{\alpha} \otimes P_{\bar{J}})(C) > 0$ . First, we note that  $\delta_{\alpha} = \bigotimes_{i \in J} \delta_{\alpha_i}$  since for any  $\alpha' \in \Omega_J$  we have

$$\prod_{i \in J} \delta_{\alpha_i}(\alpha'_i) = 1 \iff \alpha'_i = \alpha_i \text{ for all } i \in J \iff \alpha' = \alpha \iff \delta_{\alpha}(\alpha') = 1. \quad (10)$$

Since  $P$  factorizes over  $\Omega$ ,  $P_{\bar{J}}$  also factorizes, and thus  $(\delta_{\alpha} \otimes P_{\bar{J}}) \in \Delta^{\otimes}(\Omega)$ .

Next, we show that  $(\delta_{\alpha} \otimes P_{\bar{J}})(C) > 0$ . Since  $B = U_J^{-1}(\alpha)$ ,  $B \cap C = \{\omega \in C : \omega_J = \alpha\} = C^{\alpha}$ . By Lemma C.16 part 1, it follows that

$$P(B \cap C) = P(C^{\alpha}) = P_J(\alpha) \cdot P_{\bar{J}}(C_{\bar{J}}^{\alpha}). \quad (11)$$

Since  $P(B \cap C) > 0$ , we also have  $P_{\bar{J}}(C_{\bar{J}}^{\alpha}) > 0$ . Further, by Lemma C.16 part 2., we have that  $(\delta_{\alpha} \otimes P_{\bar{J}})(C) = P_{\bar{J}}(C_{\bar{J}}^{\alpha}) > 0$ . Therefore, we have  $(\delta_{\alpha} \otimes P_{\bar{J}}) \in \Delta^{\otimes}(\Omega)$ , and the second condition for Lemma C.12 is satisfied.

For proving 3., we observe that the set of factors that are relevant to  $A$  given  $C$  is exactly  $\bar{J}$ , since  $J$  is the cohistory of  $A$  given  $C$ . Further, since  $(\delta_{\alpha} \otimes P_{\bar{J}}) = \delta_{\alpha} \otimes (\bigotimes_{i \in \bar{J}} P_i)$ , we have  $(\delta_{\alpha} \otimes P_{\bar{J}})_i = P_i$  for all  $i \in \bar{J}$ , and 3. is satisfied.

We can now apply Lemma C.12, and obtain  $P(A \mid C) = (\delta_{\alpha} \otimes P_{\bar{J}})(A \mid C)$ . The right-hand side can be further transformed as follows (using Lemma C.16, part 2 in step (\*)):

$$(\delta_{\alpha} \otimes P_{\bar{J}})(A \mid C) = \frac{(\delta_{\alpha} \otimes P_{\bar{J}})(A \cap C)}{(\delta_{\alpha} \otimes P_{\bar{J}})(C)} \stackrel{(*)}{=} \frac{P_{\bar{J}}((A \cap C)_{\bar{J}}^{\alpha})}{P_{\bar{J}}(C_{\bar{J}}^{\alpha})} = \frac{P_{\bar{J}}(A_{\bar{J}}^{\alpha} \cap C_{\bar{J}}^{\alpha})}{P_{\bar{J}}(C_{\bar{J}}^{\alpha})} = P_{\bar{J}}(A_{\bar{J}}^{\alpha} \mid C_{\bar{J}}^{\alpha}).$$

So overall, we have

$$P(A \mid C) = P_{\bar{J}}(A_{\bar{J}}^{\alpha} \mid C_{\bar{J}}^{\alpha}). \quad (12)$$

Combining these results, we obtain

$$\begin{aligned} P(A \mid C) \cdot P(B \cap C) &= P_{\bar{J}}(A_{\bar{J}}^{\alpha} \mid C_{\bar{J}}^{\alpha}) \cdot P_J(\alpha) \cdot P_{\bar{J}}(C_{\bar{J}}^{\alpha}) && \text{(11) and (12)} \\ &= P_J(\alpha) \cdot P_{\bar{J}}(A_{\bar{J}}^{\alpha} \cap C_{\bar{J}}^{\alpha}) \\ &= P_J(\alpha) \cdot P_{\bar{J}}((A \cap C)_{\bar{J}}^{\alpha}) && \text{(distributivity of projection and (9))} \end{aligned}$$

$$\begin{aligned}
&= P((A \cap C)^\alpha) && \text{(Lemma C.16 part 1)} \\
&= P(A \cap C \cap U_{\bar{J}}^{-1}(\alpha)) && (9) \\
&= P(A \cap B \cap C). && \text{(definition of } B)
\end{aligned}$$

Multiplying both sides with  $P(C)$ , it follows that  $A \perp\!\!\!\perp^P B \mid C$ . This concludes the proof.  $\square$

**Lemma C.19.** *Let  $A, C \subseteq \Omega$  and  $J = \text{Cohistory}(A \mid C)$  be the cohistory of  $A$  given  $C$ . Then  $U_J \perp\!\!\!\perp^\otimes U_{\bar{J}} \mid C$ .*

*Proof.* Let  $\alpha \in \Omega_J$  and let  $B := U_J^{-1}(\alpha)$ . We need to show that  $B \perp\!\!\!\perp^\otimes U_{\bar{J}} \mid C$ . Recall that by Lemma C.18, we have  $A \perp\!\!\!\perp^\otimes U_{\text{Cohistory}(A \mid C)} \mid C$ , so we have  $A \perp\!\!\!\perp^\otimes B \mid C$ . By Lemma C.7, it follows that  $\text{Cohistory}(A \mid C) \cup \text{Cohistory}(B \mid C) = I$ , and therefore  $\bar{J} \subseteq \text{Cohistory}(B \mid C)$ . Moreover, Lemma C.18 also implies that  $B \perp\!\!\!\perp^\otimes U_{\text{Cohistory}(B \mid C)} \mid C$ . Since  $\bar{J} \subseteq \text{Cohistory}(B \mid C)$ , Corollary C.14 implies that  $B \perp\!\!\!\perp^\otimes U_{\bar{J}} \mid C$ . This concludes the proof.  $\square$

**Lemma C.20.** *Let  $A, C \subseteq \Omega$  and  $J = \text{Cohistory}(A \mid C)$ . Then  $J$  disintegrates  $C$ , i.e.,  $C = C_J \times C_{\bar{J}}$ .*

*Proof.* Let  $P \in \Delta^\otimes(\Omega)$  be any strictly positive distribution, for example the uniform distribution on each factor. Define  $Q(D) := P(D \mid C)$  for any event  $D \in \Omega$ . Then  $Q$  is a distribution on  $\Omega$  that may *not* necessarily factorize over all  $i \in I$ . However, we show that it factors into two components  $Q_J$  and  $Q_{\bar{J}}$  with  $\text{supp}(Q_J) = C_J$  and  $\text{supp}(Q_{\bar{J}}) = C_{\bar{J}}$ . Let  $\alpha \in \Omega_J, \beta \in \Omega_{\bar{J}}$  as before. Further, let  $Q_J$  and  $Q_{\bar{J}}$  be the marginal distributions of  $Q$  as in Definition C.2, which means that we have  $Q_J(\alpha) = Q(U_J^{-1}(\alpha))$  and  $Q_{\bar{J}}(\beta) = Q(U_{\bar{J}}^{-1}(\beta))$ . Then  $Q$  is the outer product  $Q = Q_J \otimes Q_{\bar{J}}$ , because

$$\begin{aligned}
Q(\alpha \cdot \beta) &= P(\alpha \cdot \beta \mid C) && \text{(definition of } Q) \\
&= P(U_J^{-1}(\alpha) \cap U_{\bar{J}}^{-1}(\beta) \mid C) \\
&= P(U_J^{-1}(\alpha) \mid C) \cdot P(U_{\bar{J}}^{-1}(\beta) \mid C) && \text{(Lemma C.19)} \\
&= Q_J(\alpha) \cdot Q_{\bar{J}}(\beta).
\end{aligned}$$

Since  $P$  is strictly positive and  $Q = P(\cdot \mid C)$ , we have  $\text{supp}(Q) = C$ . We also have  $\text{supp}(Q_J) = C_J$ , and analogously  $\text{supp}(Q_{\bar{J}}) = C_{\bar{J}}$ , by the following derivation.

$$\begin{aligned}
\alpha \in \text{supp}(Q_J) &\iff 0 < Q_J(\alpha) = Q(U_J^{-1}(\alpha)) \\
&\iff 0 < P(U_J^{-1}(\alpha) \mid C) = \frac{P(U_J^{-1}(\alpha) \cap C)}{P(C)} \\
&\iff P(U_J^{-1}(\alpha) \cap C) > 0 \\
&\iff \emptyset \neq U_J^{-1}(\alpha) \cap C && (P \text{ strictly positive}) \\
&\iff \alpha \in C_J.
\end{aligned}$$

Since  $Q = Q_J \otimes Q_{\bar{J}}$ , Lemma C.11 implies that  $\text{supp}(Q) = \text{supp}(Q_J) \times \text{supp}(Q_{\bar{J}})$ . Since  $\text{supp}(Q) = C$ ,  $\text{supp}(Q_J) = C_J$ , and  $\text{supp}(Q_{\bar{J}}) = C_{\bar{J}}$ , we have  $C = C_J \times C_{\bar{J}}$ , which concludes the proof.  $\square$

We are now ready to prove Lemma C.8, which states that  $\overline{\mathcal{H}(A \mid C)} = \text{Cohistory}(A \mid C)$ .

*Proof of Lemma C.8.* Let  $J := \mathcal{H}(A \mid C)$ . We need to show  $\bar{J} = \text{Cohistory}(A \mid C)$ . For the inclusion  $\bar{J} \subseteq \text{Cohistory}(A \mid C)$ , let  $i \in \bar{J}$ . Let  $(P, Q) \in \Delta_{\mathcal{C}, i}^\otimes(\Omega)$ . Since  $J$  generates  $A$  given  $C$  by Lemma 4.7, Lemma C.4 implies that  $A \cap C = (A \cap C)_J \times C_{\bar{J}}$ . Further, by definition of the history, we have  $C = C_J \times C_{\bar{J}}$ . We obtain:

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{P((A \cap C)_J \times C_{\bar{J}})}{P(C_J \times C_{\bar{J}})} \stackrel{(*)}{=} \frac{P_J((A \cap C)_J) P_{\bar{J}}(C_{\bar{J}})}{P_J(C_J) P_{\bar{J}}(C_{\bar{J}})} = \frac{P_J((A \cap C)_J)}{P_J(C_J)}.$$

In the step  $(*)$ , we use Lemma C.15 and the fact that  $P$  factorizes over  $\Omega$ . Similarly, we get

$$Q(A \mid C) = \frac{Q_J((A \cap C)_J)}{Q_J(C_J)}.$$

Since  $(P, Q) \in \Delta_{C,i}^{\otimes}(\Omega)$ , and  $i \notin J$  by our assumption, we have  $P_J = Q_J$ , and thus  $P(A | C) = Q(A | C)$ . Thus,  $i \in \text{Cohistory}(A | C)$ .

Finally, we show the other inclusion:  $\mathcal{H}(A | C) \subseteq \overline{\text{Cohistory}(A | C)} =: K$ . By Lemma C.4, the history  $\mathcal{H}(A | C)$  is the smallest set  $J$  which satisfies  $A \cap C = (A \cap C)_J \times C_{\bar{J}}$  and  $C = C_J \times C_{\bar{J}}$ . Thus, it suffices to show that  $K$  has these properties. By Lemma C.20, we know that  $C = C_K \times C_{\bar{K}}$ . Thus, what remains is to show that  $A \cap C = (A \cap C)_K \times C_{\bar{K}}$ .

For the inclusion from left to right, let  $\omega \in A \cap C$ . It follows that  $\omega_K \in (A \cap C)_K$ . Since  $\omega \in C$ , we have  $\omega_{\bar{K}} \in C_{\bar{K}}$ , and thus  $\omega = \omega_K \cdot \omega_{\bar{K}} \in (A \cap C)_K \times C_{\bar{K}}$ .

For the inclusion from right to left, let  $\omega \in A \cap C$  and  $\alpha \in C_{\bar{K}}$ . It suffices to show that  $\omega_K \cdot \alpha \in A \cap C$ , or equivalently  $\delta_{\omega_K \cdot \alpha}(A \cap C) = 1$ . To show this, we would like to apply Lemma C.12 to  $\delta_\omega$  and  $\delta_{\omega_K \cdot \alpha}$ , so we need to show that both are in  $\Delta_C^{\otimes}(\Omega)$ .

By (10), we have  $\delta_\omega, \delta_{\omega_K \cdot \alpha} \in \Delta_C^{\otimes}(\Omega)$ . Next we show that both assign positive probability to  $C$ . Since  $\omega \in A \cap C$ , we have  $\omega \in C$  and thus  $\delta_\omega(C) = 1 > 0$ . Since  $\omega_K \in (A \cap C)_K$  so  $\omega_K \in C_K$ , we have  $\delta_{\omega_K}(C_K) = 1$ . Since  $\alpha \in C_{\bar{K}}$ , we have  $\delta_\alpha(C_{\bar{K}}) = 1$ . It follows that

$$\delta_{\omega_K \cdot \alpha}(C) = \delta_{\omega_K \cdot \alpha}(C_K \times C_{\bar{K}}) = \delta_{\omega_K}(C_K) \cdot \delta_\alpha(C_{\bar{K}}) = 1 \cdot 1 = 1 > 0.$$

Thus,  $\delta_\omega, \delta_{\omega_K \cdot \alpha}$  are two distributions in  $\Delta_C^{\otimes}(\Omega)$  with the same factors in  $K = \overline{\text{Cohistory}(A | C)}$ . From Lemma C.12, it follows that  $\delta_\omega(A | C) = \delta_{\omega_K \cdot \alpha}(A | C)$ , and therefore

$$\delta_{\omega_K \cdot \alpha}(A \cap C) = \frac{\delta_{\omega_K \cdot \alpha}(A \cap C)}{\delta_{\omega_K \cdot \alpha}(C)} = \delta_{\omega_K \cdot \alpha}(A | C) = \delta_\omega(A | C) = \frac{\delta_\omega(A \cap C)}{\delta_\omega(C)} = 1.$$

Therefore,  $\omega_K \cdot \alpha \in A \cap C$ . This shows that  $A \cap C = (A \cap C)_K \times C_{\bar{K}}$ . Therefore,  $\mathcal{H}(A | C) \subseteq K$ . This concludes the proof of Lemma C.8, and thus of the completeness of structural independence for events, Lemma 6.4, and consequently the main theorem, Theorem 6.2.  $\square$