Proceedings of Machine Learning Research 323:1–30, 2026 5th Conference on Causal Learning and Reasoning

A fine-grained look at causal effects in causal spaces

Junhyung Park JUN.PARK@INF.ETHZ.CH

Yuqing Zhou YUQZHOU@STUDENT.ETHZ.CH

ETH Zürich, Switzerland

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

The notion of causal effect is fundamental across many scientific disciplines. Traditionally, quantitative researchers have studied causal effects at the level of variables; for example, how a certain drug dose (W) causally affects a patient's blood pressure (Y). However, in many modern data domains, the raw variables—such as pixels in an image or tokens in a language model—do not have the semantic structure needed to formulate meaningful causal questions. In this paper, we offer a more fine-grained perspective by studying causal effects at the level of *events*, drawing inspiration from probability theory, where core notions such as independence are first given for events and σ-algebras, before random variables enter the picture. Within the measure-theoretic framework of causal spaces, a recently introduced axiomatisation of causality, we first introduce several binary definitions that determine whether a causal effect is present, as well as proving some properties of them linking causal effect to (in)dependence under an intervention measure. Further, we provide quantifying measures that capture the strength and nature of causal effects on events, and show that we can recover the common measures of treatment effect as special cases.

Keywords: Causality, causal effects, causal spaces

1. Introduction

The notion of *cause*, or *causal effect*, is central to the human psychology, as well as numerous academic domains. It is deeply engrained in everyday language via constructions such as "if", "so" or "because" (Kahneman, 2011). It also forms the backbone of virtually all scientific enquiries; examples of causal statements include "gravity causes apples to fall" (Cohen, 1981), "CO_2 emission is causing climate change" (Lee et al., 2023) and "vaccination reduces the risk of fatality from Covid-19" (Ledford et al., 2020). The goal of this paper is to give formal mathematical definitions of causal effects, in many nuanced forms.

Causality is rarely deterministic. When a government introduces a policy, it may have rough ideas or hopes about its impact, but it cannot know exactly what will occur as a result; similarly when a company runs an advertising campaign, when a doctor prescribes a drug, or when a user prompts a language model. Thus, naturally, a crucial aspect of causality is *randomness*. The latter concept is axiomatised by *probability spaces* (Kolmogorov, 1933), grounded in measure theory and forming the backbone of *probability theory*. However, probability spaces do not, by themselves, encode any causality. Recently, an analogous axiomatic framework of causality called *causal spaces* was introduced (Park et al., 2023). It was shown that causal spaces generalise and unify popular frameworks of causality, such as the *structural causal models* (SCMs) (Pearl, 2009) and the *potential outcomes* (Imbens and Rubin, 2015). We use causal spaces as the basis for our definitions of causal effects.

© 2026 J. Park & Y. Zhou.

Existing approaches (cf. Section 1.1 & Appendix A) define causal effects at the level of *variables*, tacitly assumed to be semantically meaningful (e.g., "salary" following a policy, "blood pressure" after a medical intervention). As a key distinction, we define causal effects first on *events*, then extend to σ-algebras. In many modern data domains, the raw variables, such as pixels in images or tokens in language models, often lack the right semantics for meaningful causal statements. However, we can define events such as "there is a cat in this image", "this image is in greyscale", "this text is about politics" or "this text is in Chinese" to which an image or a piece of text either belongs or does not belong, based on which causal effects can be analysed in a semantically meaningful way. Further, the extension to σ-algebras facilitates graded, rather than binary, encoding of outcome characteristics.

The paper is broadly organised in two parts.

- In Section 2, we give binary definitions of causal effects, as well as their conditional and post-intervention variants. The definitions specify when an outcome or an event has a causal effect on an event or a σ-algebra.
- In Section 3, we propose ways of capturing the nature and strength of causal effects.

Our work has parallels with how the notion of (conditional) independence develops in probability theory. A binary notion of independence is first given for events, via $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, and then for σ-algebras. Following these, many quantities are proposed to capture the nature and strength of (conditional) dependence—see Appendix A.4.

In this paper, we do not consider identification, hypothesis tests of causal effects, or estimating the quantifying measures, from finite samples: our focus is solely on population-level quantities.

1.1. Related works

Treatment effect We recall the *average treatment effect* (ATE) (Rubin, 1974) and its many variants. These are the most widely used definitions of causal effects, and are most often given in the *potential outcomes framework* (Imbens and Rubin, 2015; Hernán and Robins, 2020).[1] Let $W \in \{0, 1\}$ be a binary treatment variable, and Y_0 and Y_1 be potential outcome variables corresponding to control ($W = 0$) and treatment ($W = 1$) respectively. Then the ATE is

$$\text{ATE} = \mathbb{E}[Y_1 - Y_0].$$

The *conditional* average treatment effect (CATE) (Rosenbaum and Rubin, 1983) was proposed to take *treatment effect heterogeneity* into account, i.e., the effect of treatment can differ across sub-populations. It is defined with respect to a conditioning set of variables $\mathbf{X}$:

$$\text{CATE}(\boldsymbol{x}) = \mathbb{E}[Y_1 - Y_0 \mid \mathbf{X} = \boldsymbol{x}].$$

Many variants have been studied, including, but not limited to, the distributional treatment effect (DTE) (Heckman et al., 1997), conditional distributional treatment effect (CDTE) (Park et al., 2021; Kallus and Oprescu, 2023), individual treatment effect (ITE) (Rubin, 1974), local average treatment effect (LATE) (Imbens and Angrist, 1994), average treatment

[1] It can also be done in the *structural causal models* (SCM) framework (Peters et al., 2017, p.112).

effect on the treated (ATT) (Heckman and Robb Jr, 1985), quantile treatment effect (QTE) (Firpo, 2007) and marginal average treatment effect (MATE) (Heckman and Vytlacil, 2005). As mentioned earlier, common to these definitions is that causal effects are defined for *variables* rather than *events*. They also *quantify* magnitudes of effects from the outset, rather than first providing binary, "effect/no effect" notions. We review their precise definitions, as well as some other discussions on related works, in Appendix A.

Causal abstraction A complementary line of work is causal abstraction (Beckers and Halpern, 2019; Buchholz et al., 2024), which likewise aims to study causal effects at a semantic level rather than at the level of "raw" variables. There is a particularly high level of interest in this technique in the context of *mechanistic interpretability* (Geiger et al., 2025). In that literature, one typically defines an explicit abstraction map (or representation) α from a low-level model/state space to a higher-level set of abstract variables, and then defines and compares causal effects in the abstracted space. In this way, causal abstraction can talk about the causal influence of variables that are not the original coordinates (pixels, tokens, etc.), but instead live in a chosen or learned representational space. Our approach shares the same motivation—capturing causality at the right semantic granularity—but differs in what must be postulated: we do not require selecting an abstraction function α or constructing a separate abstract causal model. Instead, we define causal effects directly in the original causal space on events and on σ-algebras, so that "high-level" questions can be posed as events (or collections of events) without leaving the underlying space or committing to a particular representation map.

1.2. Preliminaries & notations

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space. For each event $A \in \mathcal{H}$, we write $\mathbf{1}_A : \Omega \to \{0, 1\}$ for the indicator, and for each outcome $\omega \in \Omega$, we write $\delta_\omega : \mathcal{H} \to \{0, 1\}$ for the delta measure.

We use the somewhat uncommon notation of conditional measures from Çinlar (2011). The conditional measure of A given a σ-algebra $\mathcal{G}$ is the $\mathcal{G}$-measurable random variable $\mathbb{P}_{\mathcal{G}}(\cdot, A)$. Similarly, for an event $G \in \mathcal{H}$, we denote the conditional measure of A given G by $\mathbb{P}_G(A)$, which is $\frac{\mathbb{P}(G \cap A)}{\mathbb{P}(G)}$ if $\mathbb{P}(G) > 0$ and undefined if $\mathbb{P}(G) = 0$. Note that $\mathbb{P}_{\mathcal{G}}(\cdot, A)$ and $\mathbb{P}_G(A)$ are different objects—the former is a random variable, and the latter a number.

We now formally introduce the causal space framework (Park et al., 2023). We require that the measurable space be in product form. Let T be the index set of the product. Then taking, for each $t \in T$, a measurable space $(\Omega_t, \mathcal{E}_t)$, we have the product measurable space

$$(\Omega, \mathcal{H}) = \otimes_{t \in T}(\Omega_t, \mathcal{E}_t) = (\times_{t \in T}\Omega_t, \otimes_{t \in T}\mathcal{E}_t).$$

For $S \in \mathcal{P}(T)$, where $\mathcal{P}(T)$ denotes the power set of T, we write $\mathcal{H}_S$ for the sub-σ-algebra of $\mathcal{H}$ generated by measurable rectangles $\times_{t \in T} A_t$, where $A_t \in \mathcal{E}_t$ differs from Ω_t only for finitely many $t \in S$. We also write $\Omega_S = \times_{s \in S}\Omega_s$, and for each $\omega = (\omega_t)_{t \in T}$, we write $\omega_S = (\omega_s)_{s \in S} \in \Omega_S$, where $\omega_s \in \Omega_s$.

A causal space (Park et al., 2023, Definition 2.2) is defined as follows.

Definition 1 (Causal space) *A* causal space *is the quadruple* $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$*, where* $(\Omega, \mathcal{H})$ *is a measurable space with the above product structure,* $\mathbb{P}$ *is a probability measure on* $(\Omega, \mathcal{H})$ *and* $\mathbb{K} = \{K_S : S \subseteq T\}$*, called the* causal mechanism*, is a collection of transition probability*

kernels K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$, called the causal kernel on $\mathcal{H}_S$. *This means that each K_S is a function $K_S : \Omega \times \mathcal{H} \to [0,1]$, such that*

- *for each $\omega \in \Omega$, the function $B \mapsto K_S(\omega, B) : \mathcal{H} \to [0,1]$ is a probability measure, and*
- *for each $B \in \mathcal{H}$, the function $\omega \mapsto K_S(\omega, B) : \Omega \to [0,1]$ is $\mathcal{H}_S$-measurable.*

The causal kernels must satisfy the following axioms:

(i) for all $A \in \mathcal{H}$ and $\omega \in \Omega$, we have $K_\emptyset(\omega, A) = \mathbb{P}(A)$;

(ii) for all $\omega \in \Omega$, $A \in \mathcal{H}$ and $B \in \mathcal{H}_S$, we have $K_S(\omega, A \cap B) = \mathbf{1}_B(\omega) K_S(\omega, A)$.

The probability measure $\mathbb{P}$ is the *observational measure*, and $\mathbb{K}$ encodes the causal information. Intuitively, for each $S \subseteq T$ and $\omega \in \Omega$, the measure $K_S(\omega, \cdot)$ describes the distribution we get on the whole space after we forcing the coordinates indexed by S to take the value ω_S. Axioms (i) and (ii) formalise this intuition by recovering the observational measure in the absence of intervention and enforcing consistency with the imposed value on the intervened coordinates. This is analogous to the interventional distribution in an SCM, although the two frameworks encode causality differently: SCMs do so through structural equations, whereas causal spaces take $\{K_S\}_{S \subseteq T}$ as primitive objectives. We now introduce the corresponding notion of *interventions* (Park et al., 2023, Definition 2.3):

Definition 2 (Intervention) *Let us take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$ and a measure $\mathbb{Q}$ on $(\Omega, \mathcal{H}_U)$. An* intervention on $\mathcal{H}_U$ via $\mathbb{Q}$ *yields a new causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q})})$, where the* intervention measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ *is a probability measure on $(\Omega, \mathcal{H})$ defined, for $A \in \mathcal{H}$, by*

$$\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A) = \int \mathbb{Q}(d\omega) K_U(\omega, A)$$

and $\mathbb{K}^{\mathrm{do}(U,\mathbb{Q})} = \{K_S^{\mathrm{do}(U,\mathbb{Q})} : S \subseteq T\}$ is the intervention causal mechanism, *where*

$$K_S^{\mathrm{do}(U,\mathbb{Q})}(\omega_S, A) = \int \mathbb{Q}(d\omega'_{U \setminus S}) K_{S \cup U}((\omega_S, \omega'_{U \setminus S}), A).$$

Hence, causal kernels of the original causal space encode what the new measure and causal kernels will be after an intervention. We denote by $\mathbb{E}^{\mathrm{do}(U,\mathbb{Q})}$ the expectation with $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$.

1.3. Running example

To illustrate our theory, we give a simple running example to accompany the definitions throughout the paper. In Appendix D, we present another example with a language model.

Example 1 (Travel insurance) *Suppose that we want to model how much money a traveller has to pay towards accident-related costs, depending on the danger level of the trip and whether or not the traveller takes out insurance. The insurance costs 30 CHF, and if an accident happens, the traveller has to pay 1000 CHF. We take the outcome sets*

$$\Omega_{\mathrm{Dan}} = \{N, L, H\}, \quad \Omega_{\mathrm{Ins}} = \{Y, N\}, \quad \Omega_{\mathrm{Pay}} = \{0, 30, 1000\},$$

		$(\omega_{\text{Dan}}, \omega_{\text{Ins}})$						
	$\mathbb{P}$	(N,Y)	(N,N)	(L,Y)	(L,N)	(H,Y)	(H,N)	Sum
	0	0	0.09	0	0.345	0	0.04875	0.48375
ω_{Pay}	30	0.01	0	0.35	0	0.15	0	0.51
	1000	0	0	0	0.005	0	0.00125	0.00625
	Sum	0.01	0.09	0.35	0.35	0.15	0.05	1

Table 1: The observational measure $\mathbb{P}$ on the danger of the trip, whether or not the traveller gets insurance and how much the traveller pays towards accident-related costs.

		$(\omega_{\text{Dan}}, \omega_{\text{Ins}})$						
	$K_{\text{Ins}}(Y,\cdot)$	(N,Y)	(N,N)	(L,Y)	(L,N)	(H,Y)	(H,N)	Sum
	0	0	0	0	0	0	0	0
ω_{Pay}	30	0.05	0	0.65	0	0.3	0	1
	1000	0	0	0	0	0	0	0
	Sum	0.05	0	0.65	0	0.3	0	1

Table 2: The causal kernel $K_{\text{Ins}}(Y,\cdot)$ corresponding to making the traveller buy insurance.

where Y and N stands for Yes/No, N, L and H stands for No/Low/High danger. The observational measure $\mathbb{P}$ is given in Table 1, and the causal kernels $K_{\text{Ins}}(Y,\cdot)$ and $K_{\text{Ins}}(N,\cdot)$ corresponding to making the traveller buy or not buy insurance are given in Tables 2 and 3.

Table 2 indicates that, after an intervention to make the traveller buy insurance, the payment is always 30 CHF and the insurance status is always "Yes". The probabilities of the traveller choosing the danger level to be "No", "Low" and "High" become 0.05, 0.65 and 0.3 respectively. Since the corresponding observational probabilities in Table 1 are 0.10, 0.70 and 0.20, the intervention makes higher danger levels slightly more likely.

Table 3 indicates that, after an intervention forcing the traveller not to buy insurance, the payment is 0 CHF with probability 0.985 and 1000 CHF with probability 0.015, while the payment 30 CHF occurs with probability 0. Likewise, the insurance status is now always "No". The probabilities of the danger levels to be "No", "Low" and "High" are 0.30, 0.60 and 0.10 respectively. Compared with Table 1, this indicates that forcing the traveller not to buy insurance slightly reduces the danger levels.

2. Binary definitions

In this section, we give binary, yes/no definitions of causal effects. In Section 2.1, we give the vanilla definition of causal effects, of whether the marginal observational measure $\mathbb{P}$ changes as a result of an intervention. In Section 2.2, we define causal effects conditioned either on an event $G \in \mathcal{H}$ or a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{H}$, i.e., whether the conditional measures $\mathbb{P}_G$ and $\mathbb{P}_{\mathcal{G}}$ change. In Section 2.3, we define post-intervention causal effects, which characterises the effect of an additional intervention after a first one has already been applied.

$K_{\text{Ins}}(N,\cdot)$		$(\omega_{\text{Dan}},\omega_{\text{Ins}})$ (N,Y)	(N,N)	(L,Y)	(L,N)	(H,Y)	(H,N)	Sum
	0	0	0.3	0	0.59	0	0.095	0.985
ω_{Pay}	30	0	0	0	0	0	0	0
	1000	0	0	0	0.01	0	0.005	0.015
	Sum	0	0.3	0	0.6	0	0.1	1

Table 3: The causal kernel $K_{\text{Ins}}(N,\cdot)$ corresponding to making the traveller uninsured.

Due to space constraints, we only focus on *active* causal effects and their negations in the main body, and defer the full trichotomy into no, active and dormant effects to Appendix B.

2.1. Causal effect

The following is the most basic definition of (active) causal effects.

Definition 3 (Active causal effect) *Let us take a causal space $(\Omega,\mathcal{H},\mathbb{P},\mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$ and events $A, B \in \mathcal{H}$.*

(i) If $K_U(\omega, A) = \mathbb{P}(A)$, then we say that ω has no active U-causal effect *on A.*

(ii) If $K_U(\omega, A) = \mathbb{P}(A)$ for all $\omega \in B$, then we say that B has no active U-causal effect *on A.*

For a sub-σ-algebra $\mathcal{F}$ of $\mathcal{H}$, we say that ω (resp. B) has no active U-causal effect on $\mathcal{F}$ if ω (resp. B) has no active U-causal effect on any $A \in \mathcal{F}$.

When the intervention set U is clear from context, we also simply say that ω_U (or ω) has an active causal effect on A.

Park et al. (2023, Definition B.2) gave a coarser version of the above, defining $\mathcal{H}_U$ to have no active causal effect on A if $K_U(\omega, A) = \mathbb{P}(A)$ for all $\omega \in \Omega$, coinciding with Definition 3(ii) when $B = \Omega$. Definition 3 allows a more fine-grained analysis of causal effects—the quantities having active causal effects are individual outcomes and events, rather than entire σ-algebras (or random variables). This is necessary in many modern data domains; for example, in a language model, instead of saying that "the prompt has a causal effect on the output text being about sports" (which is trivially true), we want to identify specific (sets of) prompts that have this causal effect.

Note that, depending on the intervention set, it is possible for an outcome $\omega \in \Omega$ to have an active U-causal effect on an event A, yet for the same outcome to have no active V-causal effect on the same event A. Note also that there is no *control* level, unlike the treatment effects defined in Section 1.1. Instead, the causal kernels $K_U(\omega,\cdot)$ are compared against the observational measure $\mathbb{P}(\cdot)$.

In the following result, we show that no active causal effect means independence under the intervention measure (see Definition 2). The proof is in Appendix E.

Lemma 4 *Let us take a causal space $(\Omega,\mathcal{H},\mathbb{P},\mathbb{K})$, an intervention set $U \subseteq T$ and an event $A \in \mathcal{H}$. If Ω has no active U-causal effect on A, then with respect to the intervention measure $\mathbb{P}^{\text{do}(U,\mathbb{Q})}$ with any measure $\mathbb{Q}$ on $(\Omega,\mathcal{H}_U)$, the event A and the σ-algebra $\mathcal{H}_U$ are independent.*

Let us now return to our running example to illustrate Definition 3, by showing that buying insurance has a causal effect on the probability of the traveller paying 1000 CHF.

Example 2 *Write $A = \{\omega_{\text{Pay}} = 1000\}$ for the event that the traveller pays 1000 CHF. From Tables 1, 2 and 3, the marginal probabilities of A under the observational state and after interventions to force the traveller to buy/not buy insurance are respectively*

$$\mathbb{P}(A) = 0.00625, \qquad K_{\text{Ins}}(Y, A) = 0 \neq \mathbb{P}(A), \qquad K_{\text{Ins}}(N, A) = 0.015 \neq \mathbb{P}(A).$$

Hence, both outcomes $\omega_{\text{Ins}} = Y$ and $\omega_{\text{Ins}} = N$ have active causal effects on the event A.

2.2. Conditional causal effect

Conditional causal effect is an important concept, as causal effects often differ across subpopulations (*treatment effect heterogeneity*) (Heckman et al., 1997). For example, the injection of the prompt "Answer in a politically neutral tone" in a language model will have a marked causal effect on the output if earlier parts of the prompt were on politics, but little to no effect otherwise. In the classical definitions of conditional treatment effects (e.g., CATE or CDTE, cf. Section 1.1 & Appendix A), one conditions on a *variable*, which is again assumed to be semantically meaningful. In this section, we introduce binary definitions of conditional causal effects, conditioning on general *events* G and *σ-algebras* $\mathcal{G}$.

In this section, we write $\mathbb{P}^{\text{do}(U,\delta_\omega)}(\cdot) = K_U(\omega, \cdot)$ interchangeably, the former preferred when we want to condition on this measure, to avoid having two subscripts. It is trivial from the definition of interventions (Definition 2) that these two measures are equal. We first present the simpler definition of (active) causal effects conditioned on an event G.

Definition 5 (Conditional active causal effect I) *Take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$, events $A, B, G \in \mathcal{H}$ and a σ-algebra $\mathcal{F}$.*

(i) If $\mathbb{P}(G) > 0$, $\mathbb{P}^{\text{do}(U,\delta_\omega)}(G) > 0$ and $\mathbb{P}_G^{\text{do}(U,\delta_\omega)}(A) = \mathbb{P}_G(A)$, then we say that ω has no active U-causal effect *on A* conditioned on G.

(ii) If no $\omega \in B$ has an active U-causal effect on A conditioned on G, then we say that B has no active U-causal effect on A conditioned on G.

We say that ω (resp. B) has no active U-causal effect on $\mathcal{F}$ conditioned on G if ω (resp. B) has no active U-causal effect on any $A \in \mathcal{F}$ conditioned on G.

The slight complication in Definition 5 is due to the fact that conditional probabilities given G are only well-defined if G has positive measure. If we condition on the full event Ω, then we recover Definition 3. Otherwise, we require G to have positive measure under both $\mathbb{P}$ and $\mathbb{P}^{\text{do}(U,\delta_\omega)}$, before we compare $\mathbb{P}_G(A)$ and $\mathbb{P}_G^{\text{do}(U,\delta_\omega)}(A)$. Just as conditional independence cannot be determined if the conditioning event has zero measure, the existence of conditional active causal effect given G is undefined if G has zero measure under either $\mathbb{P}$ or $\mathbb{P}^{\text{do}(U,\delta_\omega)}$.

The more involved case is when we condition on a σ-algebra $\mathcal{G}$, rather than an event G. We need to be careful when conditioning an interventional measure $\mathbb{P}^{\text{do}(U,\delta_\omega)}$ on a σ-algebra $\mathcal{G}$, because there are two outcomes in play: one for the conditioning and another

for the intervention. Explicitly, given an event $A \in \mathcal{H}$, using ω for intervention and $\tilde{\omega}$ for conditioning, we have the function

$$(\omega, \tilde{\omega}) \mapsto \mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\delta_\omega)}(\tilde{\omega}, A).$$

For a fixed ω, the map $\tilde{\omega} \mapsto \mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\delta_\omega)}(\tilde{\omega}, A)$ is $\mathcal{G}$-measurable, and is defined $\mathbb{P}^{\mathrm{do}(U,\delta_\omega)}$-almost surely. With this in mind, we define conditional (active) causal effect given $\mathcal{G}$.

Definition 6 (Conditional active causal effect II) *Take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$, events $A, B \in \mathcal{H}$ and sub-σ-algebras $\mathcal{G}, \mathcal{F} \subseteq \mathcal{H}$.*

(i) Suppose that $\mathbb{P}$ and $\mathbb{P}^{\mathrm{do}(U,\delta_\omega)}$ are mutually absolutely continuous on $\mathcal{G}$. We say that ω has no active U-causal effect *on A conditioned on $\mathcal{G}$ if $\mathbb{P}_{\mathcal{G}}(\tilde{\omega}, A) = \mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\delta_\omega)}(\tilde{\omega}, A)$ for $\mathbb{P}$- (or equivalently, $\mathbb{P}^{\mathrm{do}(U,\delta_\omega)}$-)almost all $\tilde{\omega} \in \Omega$.*

(ii) If no $\omega \in B$ has an active U-causal effect on A conditioned on $\mathcal{G}$, then we say that B has no active U-causal effect on A conditioned on $\mathcal{G}$.

We say that ω (resp. B) has no active U-causal effect on $\mathcal{F}$ conditioned on $\mathcal{G}$ if ω (resp. B) has no active U-causal effect on any $A \in \mathcal{F}$ conditioned on $\mathcal{G}$.

The complication in Definition 6 arises from the fact that conditional measures given $\mathcal{G}$ are only well defined almost surely. However, in essence, Definition 6 is again just a direct conditional analogue of Definition 3; indeed, if the conditioning σ-algebra is trivial, i.e., $\mathcal{G} = \{\emptyset, \Omega\}$, then we recover Definition 3. When $\mathcal{G}$ is not trivial, $\mathbb{P}$ and $\mathbb{P}^{\mathrm{do}(U,\delta_\omega)}$ are required to be mutually absolutely continuous on $\mathcal{G}$, before comparing $\mathbb{P}_{\mathcal{G}}(\cdot, A)$ and $\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\delta_\omega)}(\cdot, A)$. If this were not the case, then in regions with zero measure under one measure and positive measure under the other, the existence of an active causal effect cannot be determined. Note that mutual absolute continuity is only required to hold on $\mathcal{G}$, not on the full σ-algebra $\mathcal{H}$.

The following result is a conditional analogue of Lemma 4: it shows that no active conditional causal effect means conditional independence under the intervention measure. The proof requires some gymnastics with monotone convergence theorem, and is in Appendix E.

Proposition 7 *Take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, events $A, G \in \mathcal{H}$ and a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{H}$. If Ω has no active U-causal effect on A conditioned on G (resp. $\mathcal{G}$), then with respect to the intervention measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ with any $\mathbb{Q}$ on $(\Omega, \mathcal{H}_U)$, the event A and the σ-algebra $\mathcal{H}_U$ are conditionally independent given G (resp. $\mathcal{G}$).*

Let us return to the running example again. According to Definition 5, we show that, conditioned on the event that the trip has no danger, buying insurance has no causal effect on the probability of the traveller paying 1000 CHF. However, given that the trip is highly dangerous, buying insurance does have a causal effect on the same event.

Example 3 *Recall that we had $A = \{\omega_{\mathrm{Pay}} = 1000\}$ for the event that the traveller pays 1000 CHF. Write $G = \{\omega_{\mathrm{Dan}} = N\}$ and $H = \{\omega_{\mathrm{Dan}} = H\}$ for the events that the danger level of the trip is non-existent and high. From the first two columns of Tables 1 and 2, the*

conditional probabilities of A the observational measure and after interventions to force the traveller to buy insurance given that the trip has no danger are

$$\mathbb{P}_G(A) = 0, \qquad \mathbb{P}_G^{\mathrm{do}(\mathrm{Ins},Y)}(A) = 0.$$

This means that, conditioned on G, $\omega_{\mathrm{Ins}} = Y$ has no causal effect on A, unlike the marginal case. However, conditioned on H, the last two columns of Tables 1 and 2 tell us

$$\mathbb{P}_H(A) = \frac{0.00125}{0.2} = 0.00625, \quad \mathbb{P}_H^{\mathrm{do}(\mathrm{Ins},Y)}(A) = 0 \neq \mathbb{P}_H(A),$$

so $\omega_{\mathrm{Ins}} = Y$ does have a causal effect on A conditioned on H.

2.3. Post-intervention causal effect

As a final variant of (active) causal effects, we present the definition of *post-intervention* causal effects. This definition identifies cases in which, once an intervention takes place on $\mathcal{H}_V$, no further intervention on $\mathcal{H}_U$ has an effect on A. This can be due to $\mathcal{H}_V$ being the *mediator* of the causal effect of $\mathcal{H}_U$ on A, or due to an intervention on $\mathcal{H}_V$ creating a different environment in which interventions on $\mathcal{H}_U$ have no active causal effect on A.

Definition 8 (Post-intervention causal effect) *Let us take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, subsets $U, V \subseteq T$, an outcome $\omega \in \Omega$ and events $A, B \in \mathcal{H}$.*

(i) If $K_{U\cup V}((\omega_U, \omega'_{V\setminus U}), A) = K_V((\omega_{U\cap V}, \omega'_{V\setminus U}), A)$ for all $\omega'_{V\setminus U} \in \Omega_{V\setminus U}$, then we say that ω has no active U-causal effect on A after intervening on $\mathcal{H}_V$.

(ii) If no $\omega \in B$ has an active U-causal effect on A after intervening on $\mathcal{H}_V$, then we say that B has no active U-causal effect on A after intervening on $\mathcal{H}_V$.

For a sub-σ-algebra $\mathcal{F}$ of $\mathcal{H}$, we say that ω (resp. B) has no active U-causal effect on $\mathcal{F}$ after intervening on $\mathcal{H}_V$ if ω (resp. B) has no active U-causal effect on any $A \in \mathcal{F}$ after intervening on $\mathcal{H}_V$.

Note that this is different to conditional causal effect—indeed, that conditioning and intervening are different is the whole premise of any theory of causality. The intuition behind this definition becomes clearer when we give the interventional causal kernels $K_U^{\mathrm{do}(V,\mathbb{Q})}$ explicitly in the following result, which is proved in Appendix E.

Proposition 9 *Let us take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, subsets $U, V \subseteq T$ and an event $A \in \mathcal{H}$. Suppose that ω has no active U-causal effect on A after intervening on $\mathcal{H}_V$.*

(i) If $U \cap V = \emptyset$, then in the causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(V,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(V,\mathbb{Q})})$ obtained via an intervention on $\mathcal{H}_V$ with any measure $\mathbb{Q}$ on $(\Omega, \mathcal{H}_V)$, we have that ω has no active U-causal effect on A, i.e., $K_U^{\mathrm{do}(V,\mathbb{Q})}(\omega, A) = \mathbb{P}^{\mathrm{do}(V,\mathbb{Q})}(A)$.

(ii) In the causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q})})$ obtained after an intervention on $\mathcal{H}_U$ with any $\mathbb{Q}$, we have that ω continues to have no active U-causal effect on A after intervening on $\mathcal{H}_V$.

We can see in part (i) that, indeed, once we intervene on $\mathcal{H}_V$, then ω no longer has an active causal effect on A. Definition 8 captures this property based on the causal kernels in the original causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$. If we let $V = \emptyset$, then we recover the original definition of (active) causal effects in Definition 3, since, by Axiom 1(i), we have $K_\emptyset(\omega, A) = \mathbb{P}(A)$.

3. Quantifying measures of causal effects

In this section, we move from existence of causal effects to their quantification by proposing *mean effect* and *maximum effect* scores. In Section 3.1, we consider causal effects on events, and in Section 3.2, we quantify causal effects on σ-algebras via a difference measure between distributions. We focus only on quantifying the vanilla marginal effects (corresponding to Definition 3), but it is clear that the same ideas and tools can be extended to quantifying conditional and post-intervention causal effects.

3.1. Quantifying causal effects on events

We first quantify causal effects on events. We compare the observational probability $\mathbb{P}$ with the interventional measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ through some scale function f, whose role is to specify the level of emphasis on changes near the boundaries (rare or near-certain events).

Definition 10 (Mean effect score) *Given an intervention set $U \subseteq T$, an interventional measure $\mathbb{Q}$ on $(\Omega, \mathcal{H}_U)$, and an event $A \in \mathcal{H}$, the* mean effect score *of an intervention on U via $\mathbb{Q}$ associated with a scale function $f : [0,1] \to [-\frac{1}{2}, \frac{1}{2}]$ (see below) is defined as*

$$\mathcal{S}_f^{\mathrm{do}(U,\mathbb{Q})}(A) = f(\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A)) - f(\mathbb{P}(A)).$$

This score yields a signed measure of the causal effect of intervening on U for event A with the measure $\mathbb{Q}$. The sign indicates whether the intervention raises or lowers $P(A)$, and f determines how changes are weighted across different baseline levels.

The scale function f should be chosen to reflect the needs of applications, in particular how strongly one wishes to weight probability changes near 0 or 1 compared with changes around $\frac{1}{2}$. We require f to be non-decreasing and symmetric around $\frac{1}{2}$, with $f(0) = -\frac{1}{2}$, $f(\frac{1}{2}) = 0$, and $f(1) = \frac{1}{2}$. These constraints satisfy three desiderata: (i) the score reflects both direction and magnitude of the shift between $\mathbb{P}(A)$ and $K_U(\omega, A)$; (ii) it is normalised to $[0,1]$; and (iii) it treats A and A^c symmetrically over $[0,1]$. If one regards a given absolute probability change as equally consequential regardless of the original $\mathbb{P}(A)$, a natural choice is the linear map

$$f_1(x) = x - \tfrac{1}{2},$$

in which case $\mathcal{S}_f^{\mathrm{do}(U,\mathbb{Q})}(A) = \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A) - \mathbb{P}(A)$ (raw difference). If increases from virtually impossible to merely rare (e.g., $0 \to 0.05$) should count more than equal-sized increases around 1/2 (e.g., $0.50 \to 0.55$), we prefer f with a small slope around 1/2 and large slopes near the boundaries. One possible such choice is

$$f_2(x) \;=\; \frac{\sinh(x - \frac{1}{2})}{2\,\sinh(\frac{1}{2})},$$

which satisfies the boundary and symmetry conditions and has derivative

$$f_2'(x) = \frac{\cosh(x - 1/2)}{2\sinh(1/2)},$$

minimal at $x = \frac{1}{2}$ and increasing as x approaches 0 or 1.

Let us go back to our running example to illustrate the mean effect score.

Example 4 *Recall again that* $A = \{\omega_{\text{Pay}} = 1000\}$*. Let* $\mathbb{Q}^Y, \mathbb{Q}^N$ *be measures such that* $\mathbb{Q}^Y_{\text{Ins}} = \delta_Y$ *and* $\mathbb{Q}^N_{\text{Ins}} = \delta_N$*. Using Tables 1, 2 and 3, the mean effect scores for* f_1 *and* f_2 *with* $\mathbb{Q}^Y$ *are respectively*

$$\mathcal{S}_{f_1}^{\text{do}(\text{Ins},\mathbb{Q}^Y)}(A) = -0.00625, \qquad \mathcal{S}_{f_2}^{\text{do}(\text{Ins},\mathbb{Q}^Y)}(A) \approx -0.00675.$$

The negative values of $\mathcal{S}_{f_1}^{\text{do}(\text{Ins},\mathbb{Q}^Y)}(A)$ *and* $\mathcal{S}_{f_2}^{\text{do}(\text{Ins},\mathbb{Q}^Y)}(A)$ *indicate that buying insurance has the effect of reducing the probability of* A*, with* f_2 *putting more emphasis on the change than the raw difference. Of course, the difference is mild, but the scale function can be chosen to emphasise the effects on the boundaries more dramatically. With* $\mathbb{Q}^N$*, the scores are*

$$\mathcal{S}_{f_1}^{\text{do}(\text{Ins},\mathbb{Q}^N)}(A) = 0.00875, \qquad \mathcal{S}_{f_2}^{\text{do}(\text{Ins},\mathbb{Q}^N)}(A) \approx 0.00942.$$

These scores indicate not only that the causal effect of not buying insurance is an increase in the probability of A*, but also that the magnitude of the effects is greater than that of* $\mathbb{Q}^Y$*.*

While the mean effect score averages the impact of the intervention with respec to $\mathbb{Q}$, one may also be interested in the largest deviation induced by the intervention within a given event. This leads to the following definition of *maximum effect score*.

Definition 11 (Maximum effect score) *Given an intervention set* $U \subseteq T$*, and events* $A \in \mathcal{H}$*,* $B \in \mathcal{H}_U$*, let us denote the outcome in* B *which leads to the maximal change in the probability of* A *under intervention on* $\mathcal{H}_U$ *as*

$$\omega_{\max} = \arg\max_{\omega \in B} \left| f(K_U(\omega, A)) - f(\mathbb{P}(A)) \right|.$$

Then, the maximum effect score *associated with a scale function* f *is defined by*

$$\mathcal{S}_f^{U,\max_B}(A) = f(K_U(\omega_{\max}, A)) - f(\mathbb{P}(A)).$$

3.2. Quantifying causal effects on σ-algebras

In this section, we consider causal effects on σ-algebras $\mathcal{F} \subseteq \mathcal{H}$, instead of individual events. In this case, we are not comparing two numbers, but a pair of distributions, which means that we need a way of comparing distributions.

Let $D_{\mathcal{F}}$ be a measure of difference between two probability measures on $\mathcal{F}$. We do not, in general, require $D_{\mathcal{F}}$ to be a metric, but we require $D_{\mathcal{F}}(\cdot,\cdot)$ to take values in a normed vector space $(F, \|\cdot\|_F)$. In particular, it can take negative values and need not be a single value. In general, it can be multi-dimensional or even infinite-dimensional, and should be chosen according to the needs in practice, such as weighted sum of different aspects of the distributions, or the kernel mean embedding (Muandet et al., 2017).

We first define the mean effect score on σ-algebras.

Definition 12 (Mean effect score II) *Given a sub-*σ*-algebra* $\mathcal{F}$ *of* $\mathcal{H}$*, an intervention set* $U \subseteq T$*, and an interventional measure* $\mathbb{Q}$ *on* $(\Omega, \mathcal{H}_U)$*, the mean effect score associated with* $D_{\mathcal{F}}$ *is defined as*

$$\mathcal{S}_{D_{\mathcal{F}}}^{\text{do}(U,\mathbb{Q})} = D_{\mathcal{F}}(\mathbb{P}^{\text{do}(U,\mathbb{Q})}, \mathbb{P}).$$

In words, $\mathcal{S}_{D_\mathcal{F}}^{\mathrm{do}(U,\mathbb{Q})}$ compares the $D_\mathcal{F}$-difference between the intervention measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ and the observational measure $\mathbb{P}$. In the following example, we illustrate the mean effect score with a particular choice of $D_\mathcal{F}$, namely, the difference in means. We show that, by considering the causal effect of a treatment in the causal space obtained by the control, and by choosing the difference measure $D_\mathcal{F}$ as the difference in means, we recover the ATE.

Example 5 (ATE) *For measures μ and ν on $\mathcal{F}$, define $D_\mathcal{F}$ as the difference in means:*

$$D_\mathcal{F}(\mu,\nu) = \int \omega\mu(d\omega) - \int \omega\nu(d\omega).$$

Let $W \in \{0,1\}$ be a binary treatment variable, and Y the outcome variable. Let $\mathcal{F} = \sigma(Y)$. Consider an intervened causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(W,\delta_0)}, \mathbb{K}^{\mathrm{do}(W,\delta_0)})$, obtained by intervening with $W = 0$. In this new causal space, the measure $\mathbb{P}^{\mathrm{do}(W,\delta_0)}$ is the law of Y under no treatment. Now let us consider intervening further on $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(W,\delta_0)}, \mathbb{K}^{\mathrm{do}(W,\delta_0)})$, with the measure $\mathbb{Q} = \delta_1$. Then, by Definition 2, we have

$$K_W^{\mathrm{do}(W,\delta_0)}(1,\cdot) = K_W(1,\cdot),$$

which implies

$$(\mathbb{P}^{\mathrm{do}(W,\delta_0)})^{\mathrm{do}(W,\delta_1)} = \mathbb{P}^{\mathrm{do}(W,\delta_1)},$$

i.e., sequential interventions on the same treatment variable W with $W = 0$ first then $W = 1$ is the same as intervening with $W = 1$ in the original causal space.

Hence, the difference-in-means mean effect score corresponding to the intervention on $\mathcal{H}_W$ with $\mathbb{Q} = \delta_1$ in the new causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(W,\delta_0)}, \mathbb{K}^{\mathrm{do}(W,\delta_0)})$ is

$$\mathcal{S}_{D_\mathcal{F}}^{\mathrm{do}(U,\mathbb{Q})} = \mathbb{E}_{Y\sim(\mathbb{P}^{\mathrm{do}(W,\delta_1)})^{\mathrm{do}(W,\delta_1)}}[Y] - \mathbb{E}_{Y\sim\mathbb{P}^{\mathrm{do}(W,\delta_0)}}[Y] = \mathbb{E}_{Y\sim\mathbb{P}^{\mathrm{do}(W,\delta_1)}}[Y] - \mathbb{E}_{Y\sim\mathbb{P}^{\mathrm{do}(W,\delta_0)}}[Y],$$

precisely recovering the ATE with our definitions. By considering conditional causal effects with the same $D_\mathcal{F}$, it is straightforward to derive the CATE as a special case as well.

Finally, let us give an analogue of Definition 11, i.e., the maximum effect score on σ-algebras. Recall that the difference measure $D_\mathcal{F}$ takes values in a normed space $(F, \|\cdot\|_F)$.

Definition 13 (Maximum effect score II) *Given a sub-σ-algebra $\mathcal{F}$ of $\mathcal{H}$, an intervention set $U \subseteq T$ and an event $B \in \mathcal{H}_U$, let us denote the outcome in B leading to the maximal change in the distributional difference under intervention on $\mathcal{H}_U$ as*

$$\omega_{\max} = \arg\max_{\omega\in B} \|D_\mathcal{F}(\mathbb{P}^{\mathrm{do}(U,\delta_\omega)}, \mathbb{P})\|_F.$$

Then the maximum effect score associated with a measure $D_\mathcal{F}$ is defined as

$$\mathcal{S}_{D_\mathcal{F}}^{U,\max} = D_\mathcal{F}(\mathbb{P}^{\mathrm{do}(U,\delta_{\omega_{\max}})}, \mathbb{P}).$$

The maximal distributional change is defined with respect to the norm in the normed space $(F, \|\cdot\|_F)$. If $F = \mathbb{R}$, then this simply boils down to the magnitude of the difference, and if F is a reproducing kernel Hilbert space, then it is the corresponding norm in this space.

We return to our running example one last time to illustrate the distributional causal effects on the σ-algebra generated by the payment variable.

Example 6 *Intuitively, purchasing insurance increases the traveller's expected payment, but at the same time dramatically reduces the variance of this payment. Our definition can be used to verify this intuition. Consider the causal effect on* $\mathcal{F} = \mathcal{H}_{\text{Pay}}$ *of intervening on insurance with the delta measure* $\mathbb{Q}_{\text{Ins}} = \delta_Y$*, i.e., the causal effect of buying insurance. Let* $D^m_{\mathcal{F}}$ *and* $D^v_{\mathcal{F}}$ *be the difference in the means and in the variances respectively. From the last column of Tables 1 and 2, the mean effect scores are respectively*

$$\mathcal{S}^{\text{do}(U,\mathbb{Q})}_{D^m_{\mathcal{F}}} = 8.45, \qquad \mathcal{S}^{\text{do}(U,\mathbb{Q})}_{D^v_{\mathcal{F}}} = -6244.5975.$$

This shows that, by buying insurance, the traveller loses 8.45 *CHF on average, but completely removes the probability of having to pay* 1000 *CHF, reflected in the big reduction in the variance. This demonstrates that, when considering causal effects on distributions on* σ*-algebras* $\mathcal{F}$*, it is important to consider different aspects of the causal effect.*

4. Conclusion

In this paper, we developed a measure-theoretic account of causal effects. Rather than starting from random variables, we define causal effects on events (and then on σ-algebras) within the causal space framework. This shift is motivated by (i) modern domains such as images and language models, where the raw variables (pixels, tokens) are often not the right semantic units for causal questions, and (ii) by drawing an analogy with the development of central notions such as independence in probability theory.

Our first contribution is a family of binary (yes/no) notions of causal effect that can be stated at a fine granularity (individual outcomes ω or events B affecting an event A or a σ-algebra $\mathcal{F}$), together with conditional and post-intervention variants. We also established structural links to (conditional) independence under interventional measures.

Our second contribution is to move from existence to quantification. For event-level targets we proposed mean and maximum effect scores that compare $\mathbb{P}(A)$ and $\mathbb{P}^{\text{do}(U,\mathbb{Q})}(A)$ through a user-chosen scale function, allowing the scoring to emphasize changes near the boundaries (rare or near-certain events) or to treat absolute probability changes uniformly. For σ-algebras, we proposed mean and maximum scores based on a generic "difference" functional $D_{\mathcal{F}}$ valued in a normed space, enabling multi-aspect and even infinite-dimensional summaries (e.g., via kernel mean embeddings). Importantly, standard treatment-effect objects—such as the ATE—are recovered as special cases of these definitions.

Finally, we stress that our focus here is on population-level definitions rather than inference: we did not develop identification results, estimation procedures or hypothesis tests for these effect notions from finite samples. Natural next steps include (i) identifying conditions under which the proposed quantities are estimable from observational or interventional data, (ii) designing tests for event-level and σ-algebra–level effects.

Acknowledgments

We are very grateful to Tobias Wegel, Florian Dorner, Julius von Kügelgen, Luigi Gresele, Fanny Yang and Thomas Icard for helpful discussions. JP is supported by SNSF Grant 218343.

References

Syed Mumtaz Ali and Samuel D Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

Sander Beckers and Joseph Y Halpern. Abstracting Causal Models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 2678–2685, 2019.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of Structural Causal Models with Cycles and Latent Variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Simon Buchholz, Junhyung Park, and Bernhard Schölkopf. Products, Abstractions and Inclusions of Causal Spaces. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 430–449, 2024.

Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

I Bernard Cohen. Newton's Discovery of Gravity. *Scientific American*, 244(3):166–181, 1981.

Ryan Cotterell, Anej Svete, Clara Meister, Tianyu Liu, and Li Du. Formal Aspects of Language Modeling. *arXiv preprint arXiv:2311.04329*, 2023.

Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. A Measure-Theoretic Characterization of Tight Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9744–9770, 2023.

Li Du, Holden Lee, Jason Eisner, and Ryan Cotterell. When is a Language Process a Language Model? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11083–11094, 2024.

Robin J Evans. Graphs for Margins of Bayesian Networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

Robin J Evans. Margins of Discrete Bayesian Networks. *The Annals of Statistics*, 46(6A): 2623–2656, 2018.

Sergio Firpo. Efficient Semiparametric Estimation of Quantile Treatment Effects. *Econometrica*, 75(1):259–276, 2007.

Antonio F Galvao and Liang Wang. Uniformly Semiparametric Efficient Estimation of Treatment Effects with a Continuous Treatment. *Journal of the American Statistical Association*, 110(512):1528–1542, 2015.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

James J Heckman and Richard Robb Jr. Alternative Methods for Evaluating the Impact of Interventions: An Overview. *Journal of econometrics*, 30(1-2):239–267, 1985.

James J Heckman and Edward Vytlacil. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738, 2005.

James J Heckman, Jeffrey Smith, and Nancy Clements. Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. *The review of economic studies*, 64(4):487–535, 1997.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Guido W Imbens. The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika*, 87(3):706–710, 2000.

Guido W Imbens and Joshua D Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge university press, 2015.

Daniel Kahneman. *Thinking, Fast and Slow*. Macmillan, 2011.

Nathan Kallus and Miruna Oprescu. Robust and Agnostic Learning of Conditional Distributional Treatment Effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6037–6060. PMLR, 2023.

Maurice Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 1938.

Andrei N Kolmogorov. *Foundations of the Theory of Probability*. NY: Chelsea Publishing Co, 1933.

Heidi Ledford, David Cyranoski, and Richard Van Noorden. COVID Vaccines: What Scientists Now Want to Know. *Nature*, 588(7837):205–6, 2020.

Hoesung Lee, Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barret, et al. IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland. Technical report, Intergovernmental Panel on Climate Change (IPCC), 2023.

Emily C Martin and Rebecca A Betensky. Testing Quasi-Independence of Failure and Truncation Times via Conditional Kendall's tau. *Journal of the American Statistical Association*, 100(470):484–492, 2005.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Junhyung Park and Krikamol Muandet. A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings. In *Advances in neural information processing systems*, volume 33, pages 21247–21259, 2020.

Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression. In *International conference on machine learning*, pages 8401–8412. PMLR, 2021.

Junhyung Park, Simon Buchholz, Bernhard Schölkopf, and Krikamol Muandet. A Measure-Theoretic Axiomatisation of Causality. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 28510–28540, 2023.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Karl Pearson. VII. Mathematical Contributions to the Theory of Evolution.—III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, pages 253–318, 1896.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Paul R Rosenbaum and Donald B Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of educational Psychology*, 66(5):688, 1974.

Friedrich Schmid and Rafael Schmidt. Multivariate Conditional Versions of Spearman's rho and Related Measures of Tail Dependence. *Journal of multivariate analysis*, 98(6): 1123–1140, 2007.

Claude E Shannon. A Mathematical Theory of Communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Charles Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Ziqiao Wang and Yongyi Mao. Generalization Bounds via Conditional f-Information. In *Advances in Neural Information Processing Systems*, volume 37, pages 52159–52188, 2024.

Aaron D Wyner. A Definition of Conditional Mutual Information for Arbitrary Ensembles. *Information and Control*, 38(1):51–59, 1978.

George Udny Yule. On the Theory of Correlation for Any Number of Variables, Treated by a New System of Notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 79(529):182–193, 1907.

Appendix A. More details on related works

In this section, we extend Section 1.1 to present the definitions of all the various definitions of treatment effects used in the potential outcomes framework.

A.1. Binary treatment

We first introduce the potential outcomes framework in the case of a binary treatment, as in Section 1.1. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space, where we interpret $\omega \in \Omega$ as "units", e.g., patients in a study. Let $W : \Omega \to \{0, 1\}$ be a binary treatment variable, $\mathbf{X}$ a (vector of) covariate variable(s), and Y, Y_0 and Y_1 the observed and potential outcomes under control and treatment respectively. The most elementary notion of causal effect is at the level of the *individual* ω. The *individual treatment effect* (ITE) is defined as

$$\mathrm{ITE}(\omega) = Y_1(\omega) - Y_0(\omega).$$

This is clearly unidentifiable from observational data, since we do not observe both Y_1 and Y_0 for the same unit.

As presented in Section 1.1, the *average treatment effect* (ATE) is defined as

$$\mathrm{ATE} = \mathbb{E}[Y_1 - Y_0].$$

The ATE is identifiable from observations, and there are many ways of estimating this quantity using observational data.

Next is the *distributional treatment effect* (DTE), which is of interest if we are interested in the impact of the treatment beyond the expectation. Let P_{Y_0} and P_{Y_1} be the distributions of Y_0 and Y_1 respectively, and let D be some metric between distributions. Then the DTE associated with D is given by

$$\mathrm{DTE}_D = D(P_{Y_0}, P_{Y_1}).$$

Here, the choice of D depends on the aspect of the distributions that we aim to capture. For example, if D simply captures the difference in the means of the two distributions, then we recover the ATE. If it is designed to capture the difference in the quantiles, then we obtain the *quantile treatment effect*.

The conditional versions of the above two definitions are the *conditional average treatment effect* (CATE) and the *conditional distributional treatment effect* (CDTE):

$$\mathrm{CATE}(\boldsymbol{x}) = \mathbb{E}[Y_1 - Y_0 \mid \mathbf{X} = \boldsymbol{x}]$$
$$\mathrm{CDTE}_D(\boldsymbol{x}) = D(P_{Y_1|\mathbf{X}=\boldsymbol{x}}, P_{Y_0|\mathbf{X}=\boldsymbol{x}}).$$

The *average treatment effect on the treated* (ATT) is used when we only care about the treatment effect on those who were actually treated. It is defined as

$$\mathrm{ATT} = \mathbb{E}[Y_1 - Y_0 \mid W = 1].$$

The *local average treatment effect* (LATE), or the average treatment effect on the compliers, is used when there is the issue of imperfect compliance, and is defined as

$$\mathrm{LATE} = \mathbb{E}[Y_1 - Y_0 \mid W_1 > W_0].$$

The *marginal average treatment effect* (MATE) is used when the resistance to treatment is quantified via a variable called V, and is defined at a particular resistance level v as

$$\text{MATE} = \mathbb{E}[Y_1 - Y_0 \mid V = v], \qquad v \in (0, 1).$$

It is easy to extend the ATT, LATE and MATE to take into account treatment effect heterogeneity by conditioning on the covariates $\mathbf{X}$, or to distributional effects rather than average.

A.2. Continuous treatment

If the treatment variable W is multi-valued or even continuous instead of being binary, then instead of having a well-defined control level $W = 0$, one is free to choose any two treatment levels to compare. Let w_1 and w_0 be two possible values of W. Then the ATE, also called the *dose-response function* (Imbens, 2000), between w_1 and w_0 is defined as

$$\text{ATE}(w_0, w_1) = \mathbb{E}[Y_{w_1} - Y_{w_0}],$$

with the analogous definitions for DTE, CATE, CDTE, etc. One can also consider the *derivative effect* (Galvao and Wang, 2015) at a particular value w of treatment:

$$\tau(w) = \frac{d\mathbb{E}[Y_w]}{dw}$$

to capture the effect of a small (infinitesimal) change in the treatment. One is also free to condition this on the covariates $\mathbf{X}$ to take treatment effect heterogeneity into account.

A.3. Total causal effect

The total causal effect in the SCM framework (Peters et al., 2017, p.91, Definition 6.12) is slightly different, in that the baseline is not some control treatment, but the observational distribution. This is more aligned with our approach, but it is still only defined at the level of random variables. It is defined as follows.

> There is a total causal effect from X to Y if and only if X and Y are not independent under the interventional distribution obtained after intervening on X to change it to a random variable $\tilde{N}_X$.

Pearl (2009, p.70, Definition 3.2.1) defines the "causal effect" as simply the interventional distribution, which is not the view that we adopt in this work.

A.4. Parallels with independence

As mentioned in the introduction (Section 1), the way we build up the theory of causal effects is analogous to how the theory of independence is developed in probability theory. First, a binary notion of independence is given, on the level of events. Given a probability space $(\Omega, \mathcal{H}, \mathbb{P})$, we say that events $A, B \in \mathcal{H}$ are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Then, we say that σ-algebras $\mathcal{F}, \mathcal{G} \subseteq \mathcal{H}$ are independent if, for all $A \in \mathcal{F}$ and all $B \in \mathcal{G}$, A and B are independent. Finally, we say that two random variables are independent if the σ-algebras they generate are independent.

Let us now consider conditioning. Let $G \in \mathcal{H}$ be an event, and $\mathcal{G} \subseteq \mathcal{H}$ a σ-algebra. Then we say that A and B are conditionally independent given G if $\mathbb{P}(G) > 0$ and

$$\mathbb{P}_G(A \cap B) = \mathbb{P}_G(A)\mathbb{P}_G(B).$$

Similarly, we say that A and B are conditionally independent given $\mathcal{G}$ if, for $\mathbb{P}$-almost all $\omega \in \Omega$,

$$\mathbb{P}_{\mathcal{G}}(\omega, A \cap B) = \mathbb{P}_{\mathcal{G}}(\omega, A)\mathbb{P}_{\mathcal{G}}(\omega, B).$$

While (conditional) independence has a crisp, universal definition, the nature/strength of (conditional) dependence has no single canonical scalar accepted across contexts. Instead, many measures exist for different purposes, for example the Pearson correlation coefficient (Pearson, 1896; Yule, 1907) for linear dependence, Kendall's τ (Kendall, 1938; Martin and Betensky, 2005) or Spearman's ρ (Spearman, 1904; Schmid and Schmidt, 2007) for rank dependence, and mutual information (Shannon, 1948; Wyner, 1978), f-divergences (Ali and Silvey, 1966; Wang and Mao, 2024) or Hilbert-Schmidt (conditional) independence criterion (Gretton et al., 2005; Park and Muandet, 2020) for general, distributional dependence.

Appendix B. No & dormant causal effect

In the main body of the paper, we only treated active causal effects (Definition 3), which compares the causal kernel $K_U(\omega, A)$ with the observational measure $\mathbb{P}$. In this section, we extend the definitions given in Section 2 to the stronger notion of *no causal effect* and *dormant causal effect*.

We start with the vanilla version of no causal effect.

Definition 14 (No causal effect) *Let us take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$, events $A, B \in \mathcal{H}$ and a sub-σ-algebra $\mathcal{F} \subseteq \mathcal{H}$.*

(i) If $K_S((\omega_{S\cap U}, \omega'_{S\setminus U}), A) = K_{S\setminus U}(\omega'_{S\setminus U}, A)$ for all $S \in \mathcal{P}(T)$ and all $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$, then we say that ω has no U-causal effect on A.

(ii) If no $\omega \in B$ has a U-causal effect on A, then we say that B has no U-causal effect on A.

We say that ω (resp. B) has no U-causal effect on $\mathcal{F}$ if ω (resp. B) has no U-causal effect on any $A \in \mathcal{F}$.

The notion of *no causal effect* is stronger than *no active causal effect*—indeed, it is easy to show that if ω has no causal effect on A, then it does not have an active causal effect, but not vice versa. The case in which ω has a causal effect but not an active causal effect is precisely the definition of *dormant causal effect*, defined in the following.

Definition 15 (Dormant causal effect) *Let us take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$, events $A, B \in \mathcal{H}$ and a sub-σ-algebra $\mathcal{F} \subseteq \mathcal{H}$.*

(i) If ω has a U-causal effect on A (in the sense of Definition 14) and does not have an active U-causal effect on A (in the sense of Definition 3), we say that ω has a dormant U-causal effect on A.

(ii) If B does not have an active U-causal effect on A but there exists some $\omega \in B$ that has a dormant U-causal effect on A, we say that B has a dormant U-causal effect on A.

We say that ω (resp. B) has a dormant U-causal effect on $\mathcal{F}$ if ω (resp. B) has no active U-causal effect on any $A \in \mathcal{F}$ but has a dormant U-causal effect on some $A \in \mathcal{F}$.

In other words, the causal effect is dormant if intervening on U alone leaves $\mathbb{P}(A)$ unchanged, yet there exists some other set S such that intervening on $U \cup S$ differs from intervening on S alone.

Next are the conditional versions. Recall that we had the conditional version of active causal effects in Section 2.2. We continue to use the notation $\mathbb{P}^{\mathrm{do}(U,\delta_\omega)}(\cdot)$ for $K_U(\omega, \cdot)$. We first condition on an event $G \in \mathcal{H}$, as we did in Definition 5.

Definition 16 (Conditional no causal effect I) *Take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$, events $A, B, G \in \mathcal{H}$ and a sub-σ-algebra $\mathcal{F} \subseteq \mathcal{H}$.*

(i) For each $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$, write $\omega' = (\omega'_{S\setminus U}, \omega_{T\setminus(S\setminus U)}) \in \Omega$ for the outcome obtained by replacing the $S \setminus U$ components of ω with $\omega'_{S\setminus U}$.

If, for all $S \subseteq T$ and all $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$, we have $\mathbb{P}^{\mathrm{do}(S,\delta_{\omega'})}(G) > 0$, $\mathbb{P}^{\mathrm{do}(S\setminus U,\delta_{\omega'})}(G) > 0$ and $\mathbb{P}_G^{\mathrm{do}(S,\delta_{\omega'})}(A) = \mathbb{P}_G^{\mathrm{do}(S\setminus U,\delta_{\omega'})}(A)$, then we say that ω has no U-causal effect on A conditioned on G.

(ii) If no $\omega \in B$ has a U-causal effect on A conditioned on G, then we say that B has no U-causal effect on A conditioned on G.

We say that ω (resp. B) has no U-causal effect on $\mathcal{F}$ conditioned on G if ω (resp. B) has no U-causal effect on any $A \in \mathcal{F}$ conditioned on G.

In Definition 16(i), we require G to have positive measure under $\mathbb{P}^{\mathrm{do}(S,\delta_{\omega'})}$ and $\mathbb{P}^{\mathrm{do}(S\setminus U,\delta_{\omega'})}$ for *all* $S \subseteq T$ and *all* $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$. Otherwise, the existence of conditional causal effect cannot be determined.

Next, we give the conditional version of dormant causal effect, given an event $G \in \mathcal{H}$.

Definition 17 (Conditional dormant effect I) *Let us take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$, events $A, B, G \in \mathcal{H}$ and a sub-σ-algebra $\mathcal{F} \subseteq \mathcal{H}$.*

(i) If ω has a U-causal effect on A conditioned on G (in the sense of Definition 16) and does not have an active U-causal effect on A conditioned on G (in the sense of Definition 5), we say that ω has a dormant U-causal effect *on A given G.*

(ii) If B does not have an active U-causal effect on A conditioned on G but there exists some $\omega \in B$ that has a dormant U-causal effect on A conditioned on G, we say that B has a dormant U-causal effect on A conditioned on G.

We say that ω (resp. B) has a dormant U-causal effect on $\mathcal{F}$ conditioned on G if ω (resp. B) has no active U-causal effect on any $A \in \mathcal{F}$ conditioned on G but has a dormant U-causal effect on some $A \in \mathcal{F}$ conditioned on G.

We also give the corresponding definitions when we condition on a σ-algebra $\mathcal{G}$.

Definition 18 (Conditional no causal effect II) *Take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$, events $A, B \in \mathcal{H}$ and sub-σ-algebras $\mathcal{F}, \mathcal{G} \subseteq \mathcal{H}$.*

(i) For each $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$, write $\omega' = (\omega'_{S\setminus U}, \omega_{T\setminus(S\setminus U)}) \in \Omega$ for the outcome obtained by replacing the $S \setminus U$ components of ω with $\omega'_{S\setminus U}$.

If, for all $S \subseteq T$ and all $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$, we have that $\mathbb{P}^{\mathrm{do}(S,\delta_{\omega'})}$ and $\mathbb{P}^{\mathrm{do}(S\setminus U,\delta_{\omega'})}$ are absolutely continuous with respect to each other on $\mathcal{G}$, and $\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(S,\delta_{\omega'})}(\tilde{\omega}, A) = \mathbb{P}_{\mathcal{G}}^{\mathrm{do}(S\setminus U,\delta_{\omega'})}(\tilde{\omega}, A)$ for $\mathbb{P}^{\mathrm{do}(S,\delta_{\omega'})}$-almost all $\tilde{\omega} \in \Omega$, then we say that ω has no U-causal effect on A conditioned on $\mathcal{G}$.

(ii) If no $\omega \in B$ has a U-causal effect on A conditioned on $\mathcal{G}$, then we say that B has no U-causal effect on A conditioned on $\mathcal{G}$.

We say that ω (resp. B) has no u-causal effect on $\mathcal{F}$ conditioned on $\mathcal{G}$ if ω (resp. B) has no active U-causal effect on any $A \in \mathcal{F}$ conditioned on $\mathcal{G}$.

Just like in Definition 6, we require the measures $\mathbb{P}^{\mathrm{do}(S,\delta_{\omega'})}$ and $\mathbb{P}^{\mathrm{do}(S\setminus U,\delta_{\omega'})}$ to be absolutely continuous on $\mathcal{G}$, but here, for *all* $S \subseteq T$ and *all* $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$. Otherwise, the existence of conditional causal effect cannot be determined.

Next, we give the conditional version of dormant causal effect, given a σ-algebra $\mathcal{G} \in \mathcal{H}$.

Definition 19 (Conditional dormant effect II) *Take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, an outcome $\omega \in \Omega$, events $A, B \in \mathcal{H}$ and sub-σ-algebras $\mathcal{F}, \mathcal{G} \subseteq \mathcal{H}$.*

(i) If ω has a U-causal effect on A conditioned on $\mathcal{G}$ (in the sense of Definition 18) and does not have an active U-causal effect on A conditioned on $\mathcal{G}$ (in the sense of Definition 6), we say that ω has a dormant U-causal effect *on A given $\mathcal{G}$.*

(ii) If B does not have an active U-causal effect on A conditioned on $\mathcal{G}$ but there exists some $\omega \in B$ that has a dormant U-causal effect on A conditioned on $\mathcal{G}$, we say that B has a dormant U-causal effect on A given $\mathcal{G}$.

We say that ω (resp. B) has a dormant U-causal effect on $\mathcal{F}$ conditioned on $\mathcal{G}$ if ω (resp. B) has no active U-causal effect on any $A \in \mathcal{F}$ conditioned on $\mathcal{G}$ but has a dormant U-causal effect on some $A \in \mathcal{F}$ conditioned on $\mathcal{G}$.

In the following remark, we collect some straightforward yet important properties of conditional causal effects.

Remark 20

(a) *If ω has no U-causal effect on A conditioned on G, then letting $S = U$ in Definition 18(i), we can see that*

$$\mathbb{P}_G^{\mathrm{do}(U,\delta_\omega)}(A) = \mathbb{P}_G^{\mathrm{do}(\emptyset,\delta_\omega)}(A) = \mathbb{P}_G(A).$$

According to Definition 6, we have that ω has no active U-causal effect on A conditioned on G. Hence, ω cannot have both no U-causal effect and active U-causal effect on A conditioned on G.

Similarly, letting $S = U$ in Definition 18, we have a direct contradiction with the definition of active causal effects given in Definition 6. Hence, ω cannot have both no U-causal effect and active U-causal effect on A conditioned on $\mathcal{G}$.

(b) *For any $U \subseteq T$, no $\omega \in \Omega$ has a U-causal effect on any $G \in \mathcal{H}$ conditioned on G. Indeed, for any $S \subseteq T$ and any $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$, we have*

$$\mathbb{P}_G^{\mathrm{do}(S,\delta_{\omega'})}(G) = 1 = \mathbb{P}_G^{\mathrm{do}(S\setminus U,\delta_{\omega'})}(G).$$

For any $U \in \mathcal{P}(T)$, no $\omega \in \Omega$ has a U-causal effect on any event $A \in \mathcal{G}$ conditioned on $\mathcal{G}$. Indeed, for any event $A \in \mathcal{G}$, any $S \subseteq T$ and any $\omega'_{S\setminus U} \in \Omega_{S\setminus U}$, we have

$$\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(S,\delta_{\omega'})}(\tilde{\omega}, A) = \mathbf{1}_A(\tilde{\omega}) = \mathbb{P}_{\mathcal{G}}^{\mathrm{do}(S\setminus U,\delta_{\omega'})}(\tilde{\omega}, A)$$

for $\mathbb{P}^{\mathrm{do}(S,\delta_{\omega'})}$-all $\tilde{\omega} \in \Omega$.

(c) *Let $U \in \mathcal{P}(T)$ and $\mathcal{F}_1, \mathcal{F}_2$ be sub-σ-algebras of $\mathcal{H}$. If $\mathcal{F}_1 \subseteq \mathcal{F}_2$ and ω has no U-causal effect on $\mathcal{F}_2$ conditioned on G (or $\mathcal{G}$), then it is clear that ω has no U-causal effect on $\mathcal{F}_1$ conditioned on G (or $\mathcal{G}$).*

Finally, we give the post-intervention versions of no and dormant causal effects. Recall that the active post-intervention causal effect was defined in Definition 8.

Definition 21 (No post-intervention causal effect) *Take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, subsets $U, V \in \mathcal{P}(T)$, an outcome $\omega \in \Omega$, events $A, B \in \mathcal{H}$ and a σ-algebra $\mathcal{F} \subseteq \mathcal{H}$.*

(i) *For each $S \subseteq T$ and $\omega'_{(S\cup V)\setminus(U\setminus V)} \in \Omega$, write $\omega' = (\omega'_{(S\cup V)\setminus(U\setminus V)}, \omega_{T\setminus(S\cup V)\setminus(U\setminus V)}) \in \Omega$ for the outcome obtained by replacing the $(S \cup V) \setminus (U \setminus V)$ components of ω with $\omega'_{(S\cup V)\setminus(U\setminus V)}$.*
If

$$K_{S\cup V}(\omega', A) = K_{(S\cup V)\setminus(U\setminus V)}(\omega', A)$$

for all $S \subseteq T$ and all $\omega'_{(S\cup V)\setminus(U\setminus V)} \in \Omega_{(S\cup V)\setminus(U\setminus V)}$, then we say that ω has no U-causal effect on A after intervening on $\mathcal{H}_V$.

(ii) *If no $\omega \in B$ has a U-causal effect on A after intervening on $\mathcal{H}_V$, then we say that B has no U-causal effect on A after intervening on $\mathcal{H}_V$.*

We say that ω (resp. B) has no U-causal effect on $\mathcal{F}$ after intervening on $\mathcal{H}_V$ if ω (resp. B) has no U-causal effect on any $A \in \mathcal{F}$ after intervening on $\mathcal{H}_V$.

Definition 22 (Dormant post-intervention causal effect) *Let us take a causal space* $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$*, subsets* $U, V \subseteq T$*, an outcome* $\omega \in \Omega$*, events* $A, B \in \mathcal{H}$ *and a sub-*σ*-algebra* $\mathcal{F} \subseteq \mathcal{H}$*.*

(i) If ω *has a* U*-causal effect on* A *after intervening on* $\mathcal{H}_V$ *(in the sense of Definition 21) and does not have an active* U*-causal effect on* A *after intervening on* $\mathcal{H}_V$ *(in the sense of Definition 8), we say that* ω_U *has a* dormant U-causal effect *on* A *after intervening on* $\mathcal{H}_V$*.*

(ii) If B *does not have an active* U*-causal effect on* A *after intervening on* $\mathcal{H}_V$ *but there exists some* $\omega \in B$ *that has a dormant* U*-causal effect on* A *after intervening on* $\mathcal{H}_V$*, we say that* B *has a dormant* U*-causal effect on* A *after intervening on* $\mathcal{H}_V$*.*

We say that ω *(resp.* B*) has a* U*-causal effect on* $\mathcal{F}$ *after intervening on* $\mathcal{H}_V$ *if* ω *(resp.* B*) has no active* U*-causal effect on any* $A \in \mathcal{F}$ *after intervening on* $\mathcal{H}_V$ *but has a dormant* U*-causal effect on some* $A \in \mathcal{F}$ *after intervening on* $\mathcal{H}_V$*.*

Appendix C. Marginalisation

The notion of marginal causal spaces has not yet appeared in previous works (for marginalisations of SCMs, see Evans (2016, 2018); Bongers et al. (2021)), but it is a simple extension of marginal probability spaces. We formally define it here.

Let T and $\tilde{T}$ be index sets such that $T \subseteq \tilde{T}$. Suppose that we have two causal spaces $\mathcal{C}$ and $\tilde{\mathcal{C}}$, with index sets T and $\tilde{T}$ respectively:

$$\mathcal{C} = (\times_{t \in T} \Omega_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K}), \qquad \tilde{\mathcal{C}} = (\times_{t \in \tilde{T}} \Omega_t, \otimes_{t \in \tilde{T}} \mathcal{E}_t, \tilde{\mathbb{P}}, \tilde{\mathbb{K}}),$$

where $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$ and $\tilde{\mathbb{K}} = \{\tilde{K}_S : S \in \mathcal{P}(\tilde{T})\}$ are the causal mechanisms. The measurable spaces $(\Omega_t, \mathcal{E}_t)$ for $t \in T$ are required to be the same for both $\mathcal{C}$ and $\tilde{\mathcal{C}}$. Write $\Omega = \times_{t \in T} \Omega_t$, $\tilde{\Omega} = \times_{t \in \tilde{T}} \Omega_t$ and $\Omega^{\setminus} = \times_{t \in \tilde{T} \setminus T} \Omega_t$, so that $\Omega \times \Omega^{\setminus} = \tilde{\Omega}$. Similarly, write $\mathcal{H} = \otimes_{t \in T} \mathcal{E}_t$ and $\tilde{\mathcal{H}} = \otimes_{t \in \tilde{T}} \mathcal{E}_t$.

Recall that the probability measure $\mathbb{P}$ is a marginalisation of $\tilde{\mathbb{P}}$ if, for all $A \in \mathcal{H}$, we have $\mathbb{P}(A) = \tilde{\mathbb{P}}(A \times \Omega^{\setminus})$. We say that the causal mechanism $\mathbb{K}$ is a marginalisation of $\tilde{\mathbb{K}}$ if, for any $A \in \mathcal{H}$ and any $S \in \mathcal{P}(T)$, we have $K_S(\omega_S, A) = \tilde{K}_S(\omega_S, A \times \Omega^{\setminus})$ for all $\omega_S \in \Omega_S$, where we recall that $\Omega_S = \times_{t \in S} \Omega_t$. In other words, each of the causal kernels in the smaller causal space $\mathcal{C}$ is a marginalisation of the corresponding causal kernel in the larger causal space $\tilde{\mathcal{C}}$, for each $\omega_S \in \Omega_S$. We say that $\mathcal{C}$ is a marginalisation of $\tilde{\mathcal{C}}$ if $\mathbb{P}$ is a marginalisation of $\tilde{\mathbb{P}}$ and $\mathbb{K}$ is a marginalisation of $\tilde{\mathbb{K}}$.

Now, let us return to causal effects in Section 2 and Appendix B.

Remark 23

(i) The definitions of no causal effect (Definitions 14, 16, 18 and 21) are not *preserved by marginalisation, since they depend on all possible subsets* S *that one can intervene on. Specifically, suppose that there is some* $\tilde{S} \in \mathcal{P}(\tilde{T})$ *in the larger causal space* $\tilde{\mathcal{C}}$ *such that* $\tilde{K}_{\tilde{S}}(\omega, A) \neq \tilde{K}_{\tilde{S} \setminus U}(\omega, A)$ *for some* $\omega \in \Omega$*, but the marginalisation is such that* $\tilde{S} \subsetneq \mathcal{P}(T)$*. Suppose further that we do have* $K_S(\omega, A) = K_{S \setminus U}(\omega, A)$ *for all* $S \in \mathcal{P}(T)$ *and all* $\omega \in \Omega$*. Then* ω_U *has no causal effect on* A *in the smaller causal space* $\mathcal{C}$ *but does have a causal effect on* A *in the larger causal space* $\tilde{\mathcal{C}}$*.*

(ii) However, it can be trivially seen that

(a) if ω_U has a causal effect on A in the smaller causal space $\mathcal{C}$, then it will always have a causal effect on A in the larger causal space $\tilde{\mathcal{C}}$;

(b) if ω_U has no causal effect on A in the larger causal space $\tilde{\mathcal{C}}$, then it will always have no causal effect on A in the smaller causal space $\mathcal{C}$.

(iii) Active causal effect (Section 2) is preserved by marginalisation. Indeed, if, for an event $A \in \mathcal{H}$, we have $\tilde{K}_U(\omega, A \times \Omega^{\backslash}) \neq \tilde{\mathbb{P}}(A \times \Omega^{\backslash})$ for some $\omega \in \Omega$, then by the definition of marginalisation, $K_U(\omega, A) \neq \mathbb{P}(A)$, and vice versa.

Appendix D. Language model example

We follow the measure-theoretic, stochastic process view of language models (Cotterell et al., 2023; Meister et al., 2023; Du et al., 2023, 2024). Let $\mathcal{V}$ be a (finite) vocabulary of tokens, equipped with the power set σ-algebra $2^{\mathcal{V}}$. This vocabulary includes the *end-of-sequence* token `EOS`. We consider two copies of the product measurable space $(\Omega^P, \mathcal{H}^P) = \otimes_{n\in\mathbb{N}}(\mathcal{V}, 2^{\mathcal{V}})$ and $(\Omega^O, \mathcal{H}^O) = \otimes_{n\in\mathbb{N}}(\mathcal{V}, 2^{\mathcal{V}})$ to which prompt and output text respectively belong. We suppose that the weights of the language model are frozen, and henceforth do not write the dependence on the weights explicitly.

We consider the causal effects of the prompt on the probability over the output text. In language models, causal queries are often about properties of the generated text, such as topic or style, rather than about individual tokens. A variable-level definition can be unnatural in this setting. The raw variables are tokens, while the causal questions are often about higher-level properties of the whole output. Our event-level definitions allows such questions to be formulated directly as events in the output space, without requiring the introduction of pre-defined variables.

We have the causal kernel K_P, which is a transition probability kernel from $(\Omega^P, \mathcal{H}^P)$ into $(\Omega^O, \mathcal{H}^O)$, i.e., for each prompt $\boldsymbol{p} \in \Omega^P$, we have a probability measure $K_P(\boldsymbol{p}, \cdot)$ on the output text space $(\Omega^O, \mathcal{H}^O)$ to obtain a language process (Meister et al., 2023, Definition 2.2). Further, we assume that the weights of the model are such that, for any prompt $\boldsymbol{p}$, the process is *tight* (Du et al., 2024, Definition 3.2), i.e., the stopping time $\tau_{\texttt{EOS}}$, defined as the first time the end-of-sequence token `EOS` is reached, is almost surely finite: $K_P(\boldsymbol{p}, \{\tau_{\texttt{EOS}} < \infty\}) = 1$.

Then the observational measure $\mathbb{P}$ on $(\Omega^O, \mathcal{H}^O)$ is given via a probability measure $\mathbb{P}^P$ on the prompt space $(\Omega^P, \mathcal{H}^P)$: for any $A \in \mathcal{H}^O$, we have

$$\mathbb{P}(A) = \int \mathbb{P}^P(d\boldsymbol{p}) K_P(\boldsymbol{p}, A).$$

Now, along with the notions introduced in the main body, we have all the ingredients for the study of causal effects of the prompt on the output text in language models.

Consider events $A, A_C, A_P \in \mathcal{H}^O$, defined as

$$\begin{aligned} A &= \{\text{the generated text is about politics}\} \\ A_C &= \{\text{the generated text is written in a politically conservative tone}\} \\ A_P &= \{\text{the generated text is written in a politically progressive tone}\}, \end{aligned}$$

with the relationships $A_C, A_P \subseteq A$, $A_C \cap A_P = \emptyset$ and $A_C \cup A_P \subsetneq A$. With respect to the observational marginal measure $\mathbb{P}$, all of these events have negligible probabilities, i.e. $\mathbb{P}(A) \ll 1$, since, among all possible texts, the observed proportion of political texts is very small.

Suppose that $B_P \in \mathcal{H}^P$ is an event in the prompt space consisting of prompts that ask the model to write in a politically progressive tone. Then for each $\boldsymbol{p} \in B_P$, we will have $K_P(\boldsymbol{p}, A_P) \gg \mathbb{P}(A_P)$, i.e., the probability that the generated text is written in a politically progressive tone will be much higher than the observational probability for this event, so B_P will have a large causal effect on the event A_P. However, it will have negligible, if any, causal effect on the event A_C, since the probability of A_C was negligible anyway.

Now, suppose that $B \in \mathcal{H}^P$ is an event that consists of all prompts that ask for a political output; in particular, we have $B_P \subsetneq B$. Then the conditional probability $\mathbb{P}_B(A)$ will be close to 1, and $\mathbb{P}_B(A_P)$ and $\mathbb{P}_B(A_C)$ will also be no longer negligible. Now, if we intervened with a prompt $\boldsymbol{p}$ in B_P, then $\mathbb{P}_B^{\mathrm{do}(P,\delta_{\boldsymbol{p}})} = \mathbb{P}^{\mathrm{do}(P,\delta_{\boldsymbol{p}})}$, since $\boldsymbol{p} \in B_P \subset B$, and $\mathbb{P}_B(A_C) \gg \mathbb{P}_B^{\mathrm{do}(P,\delta_{\boldsymbol{p}})}(A_C)$, since asking the model to produce a politically progressive text will significantly reduce the probability of the event A_C that the output text is written in a conservative tone.

This example illustrates why an event-level notion of causal effect is useful. In language model settings, the causal question of interest are typically higher-level properties of the generated text, rather than the individual tokens. By formulating causal effects directly on events in the output space, our framework makes it possible to study such semantic effects.

Appendix E. Proofs

Lemma 4. *Let us take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$ and an event $A \in \mathcal{H}$. If Ω has no active U-causal effect on A, then with respect to the intervention measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ with any measure $\mathbb{Q}$ on $(\Omega, \mathcal{H}_U)$, the event A and the σ-algebra $\mathcal{H}_U$ are independent.*

Proof Since Ω has no active causal effect on A, we have $K_U(\omega, A) = \mathbb{P}(A)$ for all $\omega \in \Omega$. Then see that, for any $B \in \mathcal{H}_U$, the intervention measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ on the intersection $A \cap B$ is

$$\begin{aligned}
\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A \cap B) &= \int \mathbb{Q}(d\omega) K_U(\omega, A \cap B) && \text{by Definition 2} \\
&= \int \mathbb{Q}(d\omega) \mathbf{1}_B(\omega) K_U(\omega, A) && \text{by Axiom 1(ii), since } B \in \mathcal{H}_U \\
&= \int \mathbb{Q}(d\omega) \mathbf{1}_B(\omega) \mathbb{P}(A) && \text{since } K_U(\omega, A) = \mathbb{P}(A) \text{ for all } \omega \in B \\
&= \mathbb{Q}(B)\mathbb{P}(A).
\end{aligned}$$

On the individual events A and B, they are

$$\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A) = \int \mathbb{Q}(d\omega) K_U(\omega, A) = \int \mathbb{Q}(d\omega)\mathbb{P}(A) = \mathbb{P}(A)$$

and by Axiom 1(ii),

$$\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(B) = \int \mathbb{Q}(d\omega) K_U(\omega, B) = \int \mathbb{Q}(d\omega)\mathbf{1}_B(\omega) = \mathbb{Q}(B).$$

Hence, we have

$$\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A \cap B) = \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(B),$$

i.e., A and B are independent under the measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$. Since $B \in \mathcal{H}_U$ was arbitrary, we conclude that $\mathcal{H}_U$ and A are independent under $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$. ■

Proposition 7. *Take a causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, an intervention set $U \subseteq T$, events $A, G \in \mathcal{H}$ and a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{H}$. If Ω has no active U-causal effect on A conditioned on G (resp. $\mathcal{G}$), then with respect to the intervention measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ with any $\mathbb{Q}$ on $(\Omega, \mathcal{H}_U)$, the event A and the σ-algebra $\mathcal{H}_U$ are conditionally independent given G (resp. $\mathcal{G}$).*

Proof We present the proof in two parts, the first part (i) in the case of conditioning on an event G, and the second part (ii) in the more involved case of conditioning on a σ-algebra $\mathcal{G}$.

(i) Recall that G has positive measure under both $\mathbb{P}$ and $\mathbb{P}^{\mathrm{do}(U,\delta_\omega)}$, so that the conditional probabilities $\mathbb{P}_G$ and $\mathbb{P}_G^{\mathrm{do}(U,\delta_\omega)}$ are well-defined. Since Ω has no active causal effect on A conditioned on G, we have, for all $\omega \in \Omega$,

$$\mathbb{P}_G(A) = \mathbb{P}_G^{\mathrm{do}(U,\delta_\omega)}(A) = \frac{K_U(\omega, A \cap G)}{K_U(\omega, G)}. \tag{*}$$

Hence see that, for any $B \in \mathcal{H}_U$,

$$\begin{aligned}
\mathbb{P}_G^{\mathrm{do}(U,\mathbb{Q})}(A \cap B) &= \frac{1}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} \int \mathbb{Q}(d\omega) K_U(\omega, A \cap B \cap G) && \text{by Definition 2} \\
&= \frac{1}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} \int \mathbb{Q}(d\omega) \mathbf{1}_B(\omega) K_U(\omega, A \cap G) && \text{by Axiom 1(ii)} \\
&= \frac{1}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} \int \mathbb{Q}(d\omega) \mathbf{1}_B(\omega) \mathbb{P}_G(A) K_U(\omega, G) && \text{by (*)} \\
&= \frac{\mathbb{P}_G(A)}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} \int \mathbb{Q}(d\omega) \mathbf{1}_B(\omega) K_U(\omega, G) \\
&= \frac{\mathbb{P}_G(A)}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} \int \mathbb{Q}(d\omega) K_U(\omega, G \cap B) && \text{by Axiom 1(ii)} \\
&= \mathbb{P}_G(A) \frac{\mathbb{P}^{\mathrm{do}(U,Q)}(G \cap B)}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} && \text{by Definition 2} \\
&= \mathbb{P}_G(A) \mathbb{P}_G^{\mathrm{do}(U,\mathbb{Q})}(B).
\end{aligned}$$

But here, see that

$$\begin{aligned}
\mathbb{P}_G^{\mathrm{do}(U,\mathbb{Q})}(A) &= \frac{1}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} \int \mathbb{Q}(d\omega) K_U(\omega, A \cap G) \\
&= \frac{1}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} \int \mathbb{Q}(d\omega) \mathbb{P}_G(A) K_U(\omega, G) && \text{by (*)} \\
&= \frac{\mathbb{P}_G(A)}{\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G)} \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(G) \\
&= \mathbb{P}_G(A).
\end{aligned}$$

Hence, combining with above, we have

$$\mathbb{P}_G^{\mathrm{do}(U,\mathbb{Q})}(A \cap B) = \mathbb{P}_G^{\mathrm{do}(U,\mathbb{Q})}(A)\mathbb{P}_G^{\mathrm{do}(U,\mathbb{Q})}(B),$$

meaning A and B are conditionally independent given G under the intervention measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$. Since $B \in \mathcal{H}_U$ was arbitrary, A and $\mathcal{H}_U$ are conditionally independent given G under the measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$.

(ii) Let us first see that, since Ω has no causal effect on A conditioned on $\mathcal{G}$, we have, for all $H \in \mathcal{G}$ and all $\omega \in \Omega$,

$$\mathbb{E}^{\mathrm{do}(U,\delta_\omega)}\left[\mathbf{1}_H \mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\delta_\omega)}(\cdot, A)\right] = \mathbb{E}^{\mathrm{do}(U,\delta_\omega)}\left[\mathbf{1}_H \mathbb{P}_{\mathcal{G}}(\cdot, A)\right]. \tag{**}$$

Then for any $H \in \mathcal{G}$,

$$\begin{aligned}
\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A \cap H) &= \int \mathbb{Q}(d\omega) K_U(\omega, A \cap H) && \text{Definition 2} \\
&= \int \mathbb{Q}(d\omega) \mathbb{P}^{\mathrm{do}(U,\delta_\omega)}(A \cap H) \\
&= \int \mathbb{Q}(d\omega) \mathbb{E}^{\mathrm{do}(U,\delta_\omega)}\left[\mathbf{1}_H \mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\delta_\omega)}(\cdot, A)\right] && \text{tower property} \\
&= \int \mathbb{Q}(d\omega) \mathbb{E}^{\mathrm{do}(U,\delta_\omega)}\left[\mathbf{1}_H \mathbb{P}_{\mathcal{G}}(\cdot, A)\right] && \text{by (**)} \\
&= \int \mathbb{Q}(d\omega) \int \mathbb{P}^{\mathrm{do}(U,\delta_\omega)}(d\omega') \mathbf{1}_H(\omega') \mathbb{P}_{\mathcal{G}}(\omega', A) \\
&= \int \mathbb{Q}(d\omega) \int K_U(\omega_U, d\omega') \mathbf{1}_H(\omega') \mathbb{P}_{\mathcal{G}}(\omega', A) \\
&= \mathbb{E}^{\mathrm{do}(U,\mathbb{Q})}\left[\mathbf{1}_H \mathbb{P}_{\mathcal{G}}(\cdot, A)\right].
\end{aligned}$$

Hence, for $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$-almost all $\omega' \in \Omega$, we have $\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\omega', A) = \mathbb{P}_{\mathcal{G}}(\omega', A)$.

Now, by Çinlar (2011, p.10, Theorem I.2.17), a positive function is measurable if and only if it is the limit of an increasing sequence of positive simple functions. Since the map $\omega' \mapsto \mathbb{P}_{\mathcal{G}}(\omega', A) : \Omega \to [0,1]$ is a $\mathcal{G}$-measurable function, we can write it as a limit of positive $\mathcal{G}$-simple functions, say $(f_n)_{n\in\mathbb{N}}$ with

$$f_n = \sum_{i_n=1}^{m_n} b_{i_n} \mathbf{1}_{H_{i_n}}$$

where $b_{i_n} \in \mathbb{R}$ and $H_{i_n} \in \mathcal{G}$. Then for any $H \in \mathcal{G}$ and any $B \in \mathcal{H}_U$,

$$\begin{aligned}
&\mathbb{E}^{\mathrm{do}(U,\mathbb{Q})}\left[\mathbf{1}_H \mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\cdot, A \cap B)\right] \\
&= \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A \cap B \cap H) && \text{tower property} \\
&= \int \mathbb{Q}(d\omega) K_U(\omega, A \cap B \cap H) && \text{Definition 2} \\
&= \int \mathbb{Q}(d\omega) \mathbf{1}_B(\omega) K_U(\omega, A \cap H) && \text{Axiom 1(ii)}
\end{aligned}$$

$$
\begin{aligned}
&= \int \mathbb{Q}(d\omega)\mathbf{1}_B(\omega)\mathbb{P}^{\mathrm{do}(U,\delta_\omega)}(A\cap H) \\
&= \int \mathbb{Q}(d\omega)\mathbf{1}_B(\omega)\mathbb{E}^{\mathrm{do}(U,\delta_\omega)}\left[\mathbf{1}_H\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\delta_\omega)}(\cdot,A)\right] && \text{tower property} \\
&= \int \mathbb{Q}(d\omega)\mathbf{1}_B(\omega)\mathbb{E}^{\mathrm{do}(U,\delta_\omega)}\left[\mathbf{1}_H\mathbb{P}_{\mathcal{G}}(\cdot,A)\right] && \text{by (**)} \\
&= \int \mathbb{Q}(d\omega)\mathbf{1}_B(\omega)\int \mathbb{P}^{\mathrm{do}(U,\delta_\omega)}(d\omega')\mathbf{1}_H(\omega')\left(\lim_{n\to\infty} f_n(\omega')\right) \\
&= \lim_{n\to\infty}\int \mathbb{Q}(d\omega)\mathbf{1}_B(\omega)\int K_U(\omega,d\omega')\mathbf{1}_H(\omega')f_n(\omega') && \text{monotone convergence} \\
&= \lim_{n\to\infty}\sum_{i_n=1}^{m_n} b_{i_n}\int \mathbb{Q}(d\omega)\mathbf{1}_B(\omega)K_U(\omega,H\cap H_{i_n}) \\
&= \lim_{n\to\infty}\sum_{i_n=1}^{m_n} b_{i_n}\int \mathbb{Q}(d\omega)K_U(\omega,H\cap H_{i_n}\cap B) && \text{Axiom 1(ii)} \\
&= \lim_{n\to\infty}\sum_{i_n=1}^{m_n} b_{i_n}\int \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega)\mathbf{1}_B(\omega)\mathbf{1}_H(\omega)\mathbf{1}_{H_{i_n}}(\omega) \\
&= \int \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega)\mathbf{1}_B(\omega)\mathbf{1}_H(\omega)\lim_{n\to\infty}\sum_{i_n=1}^{m_n} b_{i_n}\mathbf{1}_{H_{i_n}}(\omega) && \text{monotone convergence} \\
&= \int \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega)\mathbf{1}_B(\omega)\mathbf{1}_H(\omega)\mathbb{P}_{\mathcal{G}}(\omega,A) \\
&= \mathbb{E}^{\mathrm{do}(U,\mathbb{Q})}\left[\mathbf{1}_B\mathbf{1}_H\mathbb{P}_{\mathcal{G}}(\cdot,A)\right] \\
&= \mathbb{E}^{\mathrm{do}(U,\mathbb{Q})}\left[\mathbf{1}_H\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\cdot,B)\mathbb{P}_{\mathcal{G}}(\cdot,A)\right],
\end{aligned}
$$

where the last line follows from the definition of the conditional probability measure $\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\cdot,B)$, since the map $\omega'\mapsto\mathbf{1}_H(\omega')\mathbb{P}_{\mathcal{G}}(\omega',A)$ is $\mathcal{G}$-measurable. But we saw earlier that for $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$-almost all $\omega'\in\Omega$, we have $\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\omega',A)=\mathbb{P}_{\mathcal{G}}(\omega',A)$. Hence, for all $H\in\mathcal{G}$, we have

$$
\mathbb{E}^{\mathrm{do}(U,\mathbb{Q})}\left[\mathbf{1}_H\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\cdot,A\cap B)\right]=\mathbb{E}^{\mathrm{do}(U,\mathbb{Q})}\left[\mathbf{1}_H\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\cdot,B)\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\cdot,A)\right].
$$

Hence, for $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$-almost all $\omega'\in\Omega$, we have

$$
\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\omega',A\cap B)=\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\omega',B)\mathbb{P}_{\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\omega',A),
$$

in other words, A and B are conditionally independent given $\mathcal{G}$ under the measure $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$. Since $B\in\mathcal{H}_U$ was arbitrary, we have that $\mathcal{H}_U$ and A are conditionally independent given $\mathcal{G}$ under $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$, as required.

■

Proposition 9. *Let us take a causal space $(\Omega,\mathcal{H},\mathbb{P},\mathbb{K})$, subsets $U,V\subseteq T$ and an event $A\in\mathcal{H}$. Suppose that ω has no active U-causal effect on A after intervening on $\mathcal{H}_V$.*

(i) If $U\cap V=\emptyset$, then in the causal space $(\Omega,\mathcal{H},\mathbb{P}^{\mathrm{do}(V,\mathbb{Q})},\mathbb{K}^{\mathrm{do}(V,\mathbb{Q})})$ obtained via an intervention on $\mathcal{H}_V$ with any measure $\mathbb{Q}$ on $(\Omega,\mathcal{H}_V)$, we have that ω has no active U-causal effect on A, i.e., $K_U^{\mathrm{do}(V,\mathbb{Q})}(\omega,A)=\mathbb{P}^{\mathrm{do}(V,\mathbb{Q})}(A)$.

(ii) In the causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q})})$ *obtained after an intervention on* $\mathcal{H}_U$ *with any* $\mathbb{Q}$*, we have that* ω *continues to have no active* U*-causal effect on* A *after intervening on* $\mathcal{H}_V$*.*

Proof By the definition of active post-intervention causal effects, we have, for all $\omega_{V\setminus U} \in \Omega_{V\setminus U}$,

$$K_{U\cup V}(\omega_{U\cup V}, A) = K_V(\omega_V, A). \tag{*}$$

We will use this property in both parts of the proof.

(i) Since $U \cap V = \emptyset$, we have $V \setminus U = V$. Then see that, for all $\omega_U \in \Omega_U$,

$$\begin{aligned} K_U^{\mathrm{do}(V,\mathbb{Q})}(\omega_U, A) &= \int \mathbb{Q}(d\omega'_{V\setminus U}) K_{U\cup V}((\omega_U, \omega'_{V\setminus U}), A) && \text{Definition 2} \\ &= \int \mathbb{Q}(d\omega'_V) K_V(\omega'_V, A) && \text{by (*)} \\ &= \mathbb{P}^{\mathrm{do}(V,\mathbb{Q})}(A) && \text{Definition 2.} \end{aligned}$$

This is precisely the definition of no active causal effect in Definition 3.

(ii) Since $U \subseteq U \cup V$, we have $U \setminus (U \cup V) = \emptyset$. Hence,

$$K_{U\cup V}^{\mathrm{do}(U,\mathbb{Q})}(\omega, A) = \int \mathbb{Q}(d\omega'_\emptyset) K_{U\cup V}((\omega_{U\cup V}, \omega_\emptyset), A) = K_{U\cup V}(\omega, A).$$

But then by (*), we have

$$K_{U\cup V}^{\mathrm{do}(U,\mathbb{Q})}(\omega, A) = K_V(\omega, A).$$

On the other hand, we have

$$\begin{aligned} K_V^{\mathrm{do}(U,\mathbb{Q})}(\omega, A) &= \int \mathbb{Q}(d\omega'_{U\setminus V}) K_{U\cup V}((\omega_V, \omega'_{U\setminus V}), A) \\ &= \int \mathbb{Q}(d\omega'_{U\setminus V}) K_V(\omega, A) && \text{by (*)} \\ &= K_V(\omega, A). \end{aligned}$$

Putting the two together, we have

$$K_{U\cup V}^{\mathrm{do}(U,\mathbb{Q})}(\omega, A) = K_V(\omega, A).$$

This is precisely the definition of no active post-intervention causal effect in Definition 8.

■