Causal Inference Through the Structural Causal Marginal Problem

Luigi Gresele [*1] **Julius von Kügelgen** [*1 2] **Jonas M. Kübler** [*1] **Elke Kirschbaum** [3] **Bernhard Schölkopf** [1]
Dominik Janzing [3]

Abstract

We introduce an approach to counterfactual inference based on merging information from multiple datasets. We consider a causal reformulation of the statistical *marginal problem*: given a collection of *marginal* structural causal models (SCMs) over distinct but overlapping sets of variables, determine the set of *joint* SCMs that are counterfactually consistent with the marginal ones. We formalise this approach for categorical SCMs using the response function formulation and show that it reduces the space of allowed marginal and joint SCMs. Our work thus highlights a new mode of falsifiability through additional *variables*, in contrast to the statistical one via additional *data*.

1. Introduction

Counterfactual statements are ubiquitous in human judgement and reasoning. Consider the following example. A patient, Alice, is recommended a treatment X against her disease and agrees to take it. The effectiveness of the treatment has been rigorously established through a randomised control trial, which found a positive average causal effect (ACE). However, the ACE is an average of treatment efficacy over the whole population, including some individuals who respond better and others who respond worse. Alice might wonder what *her own* chances of recovery would have been, had she not taken X—a query called the effect of treatment on the treated (ETT) (Heckman, 1992; Shpitser & Pearl, 2009). This requires envisioning consequences of a hypothetical change (not taking the treatment), given that the opposite happened (in reality, she took it).

In a proposed hierarchy of causal reasoning termed the *ladder of causation* (Pearl & Mackenzie, 2018; Bareinboim et al., 2020), such counterfactual statements occupy the highest, third rung, whereas the second rung corresponds to interventions and experiments ("doing") and the lowest, first rung to passive observation ("seeing"). Counterfactual reasoning (e.g., answering personalised, individual-level questions such as Alice's) thus requires the most fine-grained causal modelling.[1] In the graphical approach to causal inference (Pearl, 2009b), counterfactuals are expressed using structural causal models (SCMs).

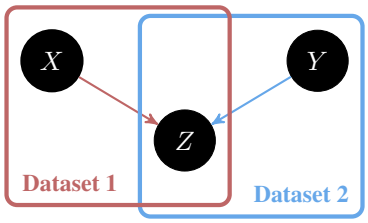

Figure 1: **Overview of the Causal Marginal Problem.** Given observations of subsets of variables in a causal graph, consistently merging the available *causal* marginal information imposes non-trivial constraints on the set of admissible joint and marginal causal models which can in turn be useful for counterfactual inference.

In practice, however, we typically do not have access to an SCM but only to observational or experimental data (rungs one and two) which may be insufficient to answer questions such as Alice's: we simply cannot perform an experiment where the same person is both given and not given a treatment, an issue also referred to as the *fundamental problem of causal inference* (Imbens & Rubin, 2015). Counterfactual queries thus need to be evaluated based on a partial state of knowledge and may be subject to an unresolvable degree of ambiguity, even in the absence of statistical uncertainty (Dawid, 2000). Pearl (2000) therefore postulates restrictions on the types of inference we can make given our data and modelling assumptions: counterfactual expressions should be evaluated subject to an *identifiability* requirement, specifying whether a given query can be estimated based on empirical observations, under conditions which can be phrased in the language of graphical models (Shpitser &

[*]Equal contribution [1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Cambridge, Cambridge, United Kingdom [3]Amazon Research, Tübingen, Germany. Correspondence to: Luigi Gresele <luigi.gresele@tue.mpg.de>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]the example concerns an individual causal effect; population-level counterfactuals can also be considered (Pearl, 2009a, § 3.4).

Pearl, 2007; 2008; Pearl, 2001; Correa et al., 2021).

When full identification is not achievable, partial identification sometimes still yields informative bounds based on empirically observable quantities (Robins, 1989; Manski, 1990; Balke & Pearl, 1997; Tian & Pearl, 2000). However, these methods typically rely on *joint information* over all variables, based on observational or experimental studies, or combinations thereof (Zhang et al., 2021).

What if we instead have studies involving distinct, but overlapping subsets of variables? Can we combine them to answer counterfactual questions? In Alice's case, knowing the effect of treatment X alone may be insufficient. Suppose, however, that a separate study characterises the interventional effect of a rare condition Y on her disease (cf. Fig. 1). Since the condition is rare, and testing for it is costly, there are no studies characterising the joint effect of X and Y on recovery. Could Alice nevertheless make use of the available information on the effect of Y and combine it with information on X to better answer her counterfactual question?

In order to answer these kinds of questions, in the present work we propose an approach to counterfactual causal inference which does not require joint observations of all variables: instead, our approach is based on *merging information from different datasets*, involving distinct but overlapping sets of variables. This can be seen as the *causal reformulation* of a classic problem in statistics called the *marginal problem* (Vorob'ev, 1962; Kellerer, 1964).

> **(Statistical) Marginal Problem:** *Given some distributions over non-identical but overlapping subsets of variables, determine existence and uniqueness of a consistent joint distribution over their union.*

For example, consider random variables X, Y, Z, and suppose that we are given the *"marginals"* $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$. Is there a joint $\mathbb{P}_{XYZ}$ that implies these marginals?[2]

In our proposed causal reformulation, we aim to merge *marginal causal models* such that they are consistent at various levels of the ladder of causation. In particular, we focus on the *counterfactual marginal problem*, in which counterfactual consistency across marginal and joint SCMs is enforced. We formalise this in the context of categorical SCMs by exploiting their *response function formulation* (Greenland & Robins, 1986; Balke & Pearl, 1994) and show that counterfactuals can acquire empirical content when considered in the broader context of a joint model, *even if only observations of the marginal models are available*.

Structure and contributions. Following a review of relevant notions of causal modelling (§ 2), we introduce the structural causal marginal problem (§ 3), describe how to treat it (§ 3.2) and illustrate its applications through examples (§ 3.3), theory (§ 3.4) and numerical simulations (§ 4). Finally, we describe extensions of the basic setting (§ 5) and discuss our findings in the context of existing literature (§ 6).

While focusing mostly on simple examples, the present work still makes a significant conceptual point: SCMs can sometimes be falsified as interventional models over additional *variables* become available. This provides causal models with an additional mode of falsifiability compared to statistical models, where the standard is to do this by means of additional *data*. The boundaries between the first two rungs of the ladder of causation and the third thus become more blurry as additional variables are observed.

2. Categorical Structural Causal Models

An SCM $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{F}, \mathbb{P}_{\mathbf{U}})$ consists of (Pearl, 2009b):

(i) a tuple $\mathbf{V} = (V_1, \dots, V_n)$ of observed, or *endogenous*, variables whose causal relations are modelled;

(ii) a tuple $\mathbf{U} = (U_1, \dots, U_n)$ of unobserved, or *exogenous*, variables which account for any stochasticity;

(iii) a tuple $\mathcal{F} = (f_1, \dots, f_n)$ of deterministic functions, or *mechanisms*, computing each V_i from its *causal parents*, or direct causes, $\mathbf{PA}_i \subseteq \mathbf{V} \setminus \{V_i\}$ and the corresponding U_i via the *structural equations*

$$\{V_i := f_i(\mathbf{PA}_i, U_i)\}_{i=1}^{n}; \tag{1}$$

(iv) a joint distribution $\mathbb{P}_{\mathbf{U}}$ over the exogenous $\mathbf{U}$.

Every SCM induces a directed *causal graph* $\mathcal{G}$ with nodes $\mathbf{V}$ and edges $V_j \to V_i$ $\forall i, \forall V_j \in \mathbf{PA}_i$ (see Fig. 1 for an example). We make the common assumption that $\mathcal{G}$ does not contain cycles,[3] which ensures that $\mathcal{M}$ induces a unique *observational distribution* $\mathbb{P}_{\mathbf{V}}$ over $\mathbf{V}$ (see below). In addition, we assume throughout that all V_i are categorical variables:

Assumption 1 (Finite domains)**.** The domains $\mathcal{V}_i$ of all endogenous variables V_i are finite, $\forall i : |\mathcal{V}_i| < \infty$.

Whereas for *general* SCMs the f_i are arbitrary unknown functions and the domains $\mathcal{U}_i$ of the exogenous U_i unspecified and potentially infinite, Asm. 1 permits an *equivalent representation* that makes such SCMs easier to study. The key observation is that, for categorical $\mathbf{V}$, there are only finitely many functions $\{\tilde{f}_{i,k}\}_k$ mapping $\mathbf{PA}_i$ to V_i. For each value u_i, the function $f_i(\cdot, u_i)$ corresponds to one such *response function* $\tilde{f}_{i,k}$, so U_i acts as a "random switch" that induces a distribution on $\{\tilde{f}_{i,k}\}_k$. We can thus partition the domain $\mathcal{U}_i$ into equivalence classes of values yielding the same $\tilde{f}_{i,k}$ and replace U_i with a *categorical response function variable* R_i defined over these equivalence classes:

[2] A trivial negative example is the case where $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ imply different $\mathbb{P}_Z$; in general, $\mathbb{P}_{XYZ}$ (if it exists) is not unique.

[3] For a treatment of cyclic SCMs, see Bongers et al. (2021).

$$\{V_i := \tilde{f}_{i,R_i}(\mathbf{PA}_i)\}_{i=1}^n, \qquad \mathbf{R} \sim \mathbb{P}_{\mathbf{R}} \tag{2}$$

with $\mathbf{R} = (R_1, ..., R_n)$ and each R_i taking values in

$$\mathcal{R}_i = \{0, ..., |\mathcal{V}_i|^{\prod_{V_j \in \mathbf{PA}_i} |\mathcal{V}_j|} - 1\} \tag{3}$$

if $\mathbf{PA}_i \neq \varnothing$, and $\mathcal{R}_i = \{0, ..., |\mathcal{V}_i| - 1\}$ otherwise.

This re-parametrisation of discrete SCMs is known as the *response function framework*[4] and we refer to Balke & Pearl (1994) for further details. Its main benefit is that the $\tilde{f}_{i,k}$ are easily enumerated, so that the categorical SCM (2) is entirely characterised by the unknown distribution $\mathbb{P}_{\mathbf{R}}$.

Using the shorthand $\mathbb{P}(\mathbf{x})$ for $\mathbb{P}_{\mathbf{X}}(\mathbf{X} = \mathbf{x})$, the *observational distribution* $\mathbb{P}_{\mathbf{V}}$ induced by (2) is given by:

$$\mathbb{P}(\mathbf{v}) = \sum_{\mathbf{r}} \mathbb{P}(\mathbf{r}) \prod_{i=1}^n \mathbb{I}\{v_i = \tilde{f}_{i,r_i}(\mathbf{pa}_i)\} \tag{4}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. If $\mathbb{P}_{\mathbf{V}}$ is known from empirical observation, (4) imposes a constraint on the space of allowed SCMs parametrised by different $\mathbb{P}_{\mathbf{R}}$.

Interventions in the form of external manipulations to subsets $\mathbf{V}_{\mathcal{I}} \subseteq \mathbf{V}$ of variables correspond to changes to the structural equations (1): e.g., setting $\mathbf{V}_{\mathcal{I}}$ to a constant $\mathbf{v}_{\mathcal{I}}$ is denoted using Pearl's *do-operator* by $do(\mathbf{V}_{\mathcal{I}} := \mathbf{v}_{\mathcal{I}})$, or $do(\mathbf{v}_{\mathcal{I}})$ for short. Interventional distributions are given by:

$$\mathbb{P}(\mathbf{v}_{\setminus\mathcal{I}}|do(\mathbf{v}_{\mathcal{I}})) = \sum_{\mathbf{r}} \mathbb{P}(\mathbf{r}) \prod_{i \notin \mathcal{I}} \mathbb{I}\{v_i = \tilde{f}_{i,r_i}(\mathbf{pa}_i)\} \tag{5}$$

Counterfactuals which condition on some observation $\mathbf{w}$ of a subset of variables $\mathbf{W} \subseteq \mathbf{V}$ when reasoning about a hypothetical intervention $do(\mathbf{v}_{\mathcal{I}})$ are modelled by using the posterior $\mathbb{P}_{\mathbf{R}|\mathbf{w}}$ computed via (4) in place of $\mathbb{P}_{\mathbf{R}}$ in (5). Note that the condition $\mathbf{W} = \mathbf{w}$ can contradict the assignment $\mathbf{V}_{\mathcal{I}} := \mathbf{v}_{\mathcal{I}}$, which renders the query counterfactual.

For now, we additionally make the following common assumption (which we will relax again in § 5).

Assumption 2 (Causal sufficiency)**.** The exogenous variables are mutually independent, i.e., $\mathbb{P}_{\mathbf{R}}$ factorises.

Asm. 2 means that there is no hidden confounding, i.e., no unobserved variable influences more than one V_i. It implies the following *Markov factorisation* (Spirtes et al., 2000):

$$\mathbb{P}(\mathbf{v}) = \prod_{i=1}^n \mathbb{P}(v_i|\mathbf{pa}_i)\,, \tag{6}$$

where each *causal Markov kernel* $\mathbb{P}(v_i|\mathbf{pa}_i)$ is given by

$$\mathbb{P}(v_i|\mathbf{pa}_i) = \sum_{r_i \in \mathcal{R}_i} \mathbb{P}(r_i)\, \mathbb{I}\{v_i = \tilde{f}_{i,r_i}(\mathbf{pa}_i)\}\,. \tag{7}$$

Asm. 2 has two important consequences: first, it suffices to consider the marginals of each R_i separately (rather than model their joint distribution $\mathbb{P}_{\mathbf{R}}$); second, interventional queries become identifiable from observational data via the g-formula (Robins, 1986), a.k.a. *truncated factorisation*

$$\mathbb{P}(\mathbf{v}_{\setminus\mathcal{I}}|do(\mathbf{v}_{\mathcal{I}})) = \delta(\mathbf{v}_{\mathcal{I}}) \prod_{i \notin \mathcal{I}} \mathbb{P}(v_i|\mathbf{pa}_i)\,. \tag{8}$$

Under Asm. 2, the boundary between interventional (rung 2) and observational (rung 1) quantities thus disappears once the causal graph is known. However, there is typically still a whole family of SCMs consistent with the available rung 1/2 information that imply different counterfactuals (rung 3), see, e.g., Peters et al. (2017, § 3.4) for an explicit description of this ambiguity. Next, we illustrate this point for Boolean SCMs which will be the main objects of study.

2.1. Causally-Sufficient Cause-Effect Models

Consider a bivariate, Boolean SCM $\mathcal{M}_X$ over $X \to Z$. Using response functions, this can be written *w.l.o.g.* as

$$X := R_X, \qquad Z := f_{R_Z}(X), \tag{9}$$

where R_Z indexes the four distinct functions f_k from $\{0,1\}$ to $\{0,1\}$: the two constant functions $f_0 \equiv 0$, and $f_1 \equiv 1$, as well as $f_2(X) = X$ ("ID"), and $f_3(X) = 1 - X$ ("NOT").

Here, $\mathbb{P}_{R_X}$ coincides with the (observed) marginal $\mathbb{P}_X$, and we assume that X is not constant, $0 < \mathbb{P}(X = 1) < 1$. Under Asm. 2, the SCM $\mathcal{M}_X$ from (9) is thus characterised entirely by the distribution $\mathbb{P}_{R_Z}$ over the four f_k. We represent this as a probability vector $\mathbf{a} \in \Delta^3$, where $\Delta^{K-1} = \{\mathbf{a} \in \mathbb{R}^K \,|\, a_k \geq 0, \sum_{k=0}^{K-1} a_k = 1\}$ denotes the probability simplex over K points. Due to the constraints imposed on $\mathbf{a}$ by the observed $\mathbb{P}_{Z|X}$ via (7), we can write it in terms of a *single free parameter* $\lambda_X \in [\lambda_X^{\min}, \lambda_X^{\max}]$ as:

$$\mathbf{a}(\lambda_X) = \begin{pmatrix} 0 \\ 1 - p_{00} - p_{01} \\ p_{00} \\ p_{01} \end{pmatrix} + \lambda_X \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \tag{10}$$

with $p_{ij} = \mathbb{P}(Z = i|X = j)$, $\lambda_X^{\min} = \max\{0, p_{00}+p_{01}-1\}$, and $\lambda_X^{\max} = \min\{p_{00}, p_{01}\}$, see Appx. A for details.

Different choices of $\lambda_X \in [\lambda_X^{\min}, \lambda_X^{\max}]$ thus define a *family of SCMs* that are observationally and interventionally equivalent[5] but imply *different counterfactual distributions*. In particular, for any given observation (x, z), the probability that "Z would have flipped had X been different" is given by $\gamma(\lambda_X) := a_2 + a_3 = p_{00} + p_{01} - 2\lambda_X$. For this reason, we call γ the *counterfactual influence* of X on Z. SCMs with larger λ_X thus exhibit a smaller counterfactual influence.

3. The Structural Causal Marginal Problem

We now formulate the *causal marginal problem*, which can be understood as a causal version of the (statistical) marginal

[4] also referred to as *principal stratification* (Frangakis & Rubin, 2002) or *canonical representation* (Peters et al., 2017)

[5] i.e., indistinguishable based on all *do*-interventions.

problem (Vorob'ev, 1962; Kellerer, 1964).

Causal Marginal Problem: *Can marginal causal models over subsets of variables with known causal graph be consistently merged? What constraints on marginal and joint causal models does this imply?*

We study this problem within the SCM framework, i.e., the *structural causal marginal problem*. To build intuition and gain a better understanding of the fundamental concepts, we first analyse the causally-sufficient, Boolean setting from § 2.1: we assume that in addition to $\mathbb{P}_{XZ}$, we observe $\mathbb{P}_{YZ}$ from another dataset where Y is a second independent Boolean cause of Z, as illustrated in Fig. 1.[6] Crucially, we do not have joint observations of all three variables, i.e., $\mathbb{P}_{XYZ}$ is unknown. While this case might appear rather simple, it already bears a number of nontrivial implications for counterfactual inference. We defer a more general definition and a discussion of extensions to § 5.

We denote the second marginal SCM over $Y \to Z$ by $\mathcal{M}_Y$,

$$Y := Q_Y, \qquad Z := f_{Q_Z}(Y), \tag{11}$$

using the same response functions f_k as in (9) for $\mathcal{M}_X$, and parametrise the family of SCMs consistent with the observed $\mathbb{P}_{YZ}$ with a probability vector $\mathbf{b} \in \Delta^3$ with a single free parameter $\lambda_Y \in [\lambda_Y^{\min}, \lambda_Y^{\max}]$, analogously to (10).

The space of marginal SCMs $(\mathcal{M}_X, \mathcal{M}_Y)$ parametrised by (λ_X, λ_Y) that are *separately* consistent with $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ (i.e., prior to considerations about consistently merging them into a joint model) is illustrated in Fig. 2 as the red rectangle. We will show that: (i) enforcing that the two marginal SCMs can be merged into a joint SCM (§ 3.1) reduces the space of admissible $(\mathcal{M}_X, \mathcal{M}_Y)$ (blue & green areas in Fig. 2; § 3.2 and § 4); (ii) knowing one of the marginal SCMs exactly (e.g., from prior knowledge or particular observations) further restricts the choices for the other marginal (horizontal green line in Fig. 2; § 3.3); and (iii) some marginal models are inherently easier to falsify than others (§ 3.4 and § 4).

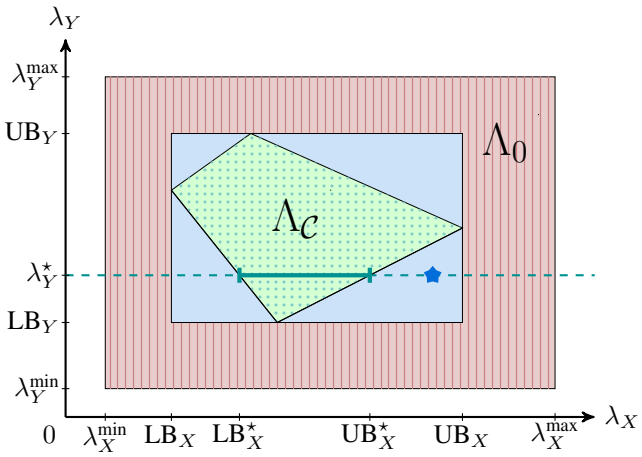

Figure 2: **2D Schematic of the Structural Causal Marginal Problem.** For the causal graph from Fig. 1 and model class from § 2.1, given $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$, the two marginal SCMs $\mathcal{M}_X$ and $\mathcal{M}_Y$ over $X \to Z$ and $Y \to Z$ are each parametrised by a single free parameter $\lambda_X \in [\lambda_X^{\min}, \lambda_X^{\max}]$ (x-axis) and $\lambda_Y \in [\lambda_Y^{\min}, \lambda_Y^{\max}]$ (y-axis), respectively. The outer dashed red area Λ_0 corresponds to combinations of counterfactual marginal models (λ_X, λ_Y) that are falsified in that they cannot be counterfactually consistent (Defn. 3); the inner dotted green polytope $\Lambda_{\mathcal{C}}$ corresponds to (λ_X, λ_Y) that are counterfactually consistent; and the solid blue area, defined as the surrounding rectangle of the latter, corresponds to (λ_X, λ_Y) that are not counterfactually consistent but cannot be falsified without additional assumptions or constraints. Enforcing consistency with the other marginal (interventional) model implies $\lambda_X \in [\mathrm{LB}_X, \mathrm{UB}_X]$ and $\lambda_Y \in [\mathrm{LB}_Y, \mathrm{UB}_Y]$, but without additional information about the other marginal this range cannot be reduced further. For a given $\mathcal{M}_Y$ corresponding to $\lambda_Y^\star$ (dashed horizontal green line), on the other hand, the interval of consistent λ_X shrinks further to $[\mathrm{LB}_X^\star, \mathrm{UB}_X^\star]$ (solid green line) so that the λ_X corresponding to the blue star marker can be ruled out.

3.1. Consistency Between Marginal and Joint SCMs

We now define the joint model and provide a systematic way of linking its representation to those of the marginal models. We write the joint SCM $\mathcal{M}$ over $\{X, Y\} \to Z$ as

$$X := R_X, \qquad Y := Q_Y, \qquad Z := h_S(X, Y), \tag{12}$$

where S indexes the 16 response functions $h_0, \ldots, h_{15}$ from $\{0,1\}^2$ to $\{0,1\}$ (listed in Tab. 1 in Appx. B). We denote the distribution $\mathbb{P}_S$ over the h_k by a probability vector $\mathbf{c} \in \Delta^{15}$. Note that unlike for the marginal models, we do not a priori have additional constraints reducing the number of free parameters of $\mathbf{c}$ since $\mathbb{P}_{XYZ}$ and thus $\mathbb{P}_{Z|XY}$ are unknown.

To relate the joint (12) and marginal SCMs (9) and (11), a key observation is that for any fixed value x of X (resp. y of Y), each two-variable function $h_k(X, Y)$ implicitly defines a single-variable function $f_j(Y)$ over the remaining variable Y (resp. $f_{j'}(X)$ over X). Formally, we define the following projection operators for $x, y \in \{0, 1\}$:

$$\begin{aligned} \mathcal{P}_x^X &: h_k \mapsto h_k(x, Y) = f_j(Y) \quad \text{for some} \quad j, \\ \mathcal{P}_y^Y &: h_k \mapsto h_k(X, y) = f_{j'}(X) \quad \text{for some} \quad j'. \end{aligned} \tag{13}$$

For example, we defined $h_0(X, Y) \equiv 0$ (see Tab. 1 in Appx. B), so $\mathcal{P}_0^X(h_0) = f_0(Y) \equiv 0$; and, similarly,

[6] Asm. 2 implies $X \perp\!\!\!\perp Y$, for otherwise Y (resp. X), which is unobserved in $\mathcal{M}_X$ (resp. $\mathcal{M}_Y$), would be a hidden confounder. This is, in principle, falsifiable through observation of $\mathbb{P}_{XY}$.

for $h_7(X,Y) = \neg(X \wedge Y)$ we have that $\mathcal{P}_1^Y(h_7) = f_3(X)$ since $h_7(X,1) = 1 - X$, i.e., the NOT function f_3.

Together with the marginal distributions of X and Y (obtained by marginalisation of Z in $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$), the distribution over the h_k in $\mathcal{M}$ parametrised by $\mathbf{c} \in \Delta^{15}$ thus induces distributions over the f_j in $\mathcal{M}_X$ and $\mathcal{M}_Y$ via (13). The latter are parametrised by $\mathbf{a}(\lambda_X)$ and $\mathbf{b}(\lambda_Y)$ (see (10)), and enforcing that they match the corresponding distributions induced by $\mathcal{M}$ yields the following *linear* constraints:

$$\begin{aligned} a_j(\lambda_X) &= \sum_{y=0}^{1} \mathbb{P}_Y(y) \sum_{k=0}^{15} \mathbb{I}\{\mathcal{P}_y^Y(h_k) = f_j(X)\}\, c_k\,, \\ b_j(\lambda_Y) &= \sum_{x=0}^{1} \mathbb{P}_X(x) \sum_{k=0}^{15} \mathbb{I}\{\mathcal{P}_x^X(h_k) = f_j(Y)\}\, c_k\,, \end{aligned} \tag{14}$$

for $j = 0,1,2,3$. Writing (14) in matrix form, we obtain:

$$\mathbf{a}(\lambda_X) = \mathbf{A}\mathbf{c}, \qquad \mathbf{b}(\lambda_Y) = \mathbf{B}\mathbf{c}\,, \tag{15}$$

where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{4\times 16}$ are constant matrices whose entries are given in terms of $\mathbb{P}_Y$ and $\mathbb{P}_X$, respectively.[7]

In general, (15) does not uniquely determine a joint SCM in terms of the marginal ones as it involves at most eight independent constraints. Nor does there always exist a joint model (parametrised by $\mathbf{c}$) that satisfies (15): take, e.g., any combination of $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ for which already the *statistical* marginal problem does not have a solution.

To discuss solutions to the structural causal marginal problem, we introduce the following notion of consistency.

Definition 3 (Counterfactual consistency)**.** An SCM $\mathcal{M}$ over observed variables $\mathbf{V}$ is *counterfactually consistent* with a (marginal) SCM $\mathcal{M}_1$ over a subset $\mathbf{W}_1 \subseteq \mathbf{V}$ if all counterfactual distributions of $\mathbf{W}_1$ in $\mathcal{M}_1$ coincide with those implied by $\mathcal{M}$ via marginalisation of $\mathbf{V} \setminus \mathbf{W}_1$, (see Bongers et al., 2021, Defn. 5.3 for marginalisation of SCMs). Two SCMs $\mathcal{M}_1, \mathcal{M}_2$ over subsets $\mathbf{W}_1, \mathbf{W}_2 \subseteq \mathbf{V}$ are counterfactually consistent if there is a joint SCM $\mathcal{M}$ over $\mathbf{V}$ which is counterfactually consistent with both $\mathcal{M}_1$ and $\mathcal{M}_2$.

Defn. 3 can be understood as a generalisation of counterfactual *equivalence* (see, e.g., Peters et al., 2017, Defn. 6.47) which also involves equality of counterfactual distributions, but applies to different SCMs over the *same* set of variables.

The counterfactual distributions implied by an SCM are fully determined by the structural equations and noise distribution (as parametrised by λ_X, λ_Y, and $\mathbf{c}$ here). In our case, a marginal SCM $\mathcal{M}_X$ (or $\mathcal{M}_Y$) is thus counterfactually consistent with a joint SCM $\mathcal{M}$ if the corresponding constraint in (15) holds. The two marginal SCMs $\mathcal{M}_X$ and $\mathcal{M}_Y$ are counterfactually consistent if both constraints in (15) hold simultaneously for some $\mathbf{c}$. In this case, we say that $\mathbf{c}$ (or $\mathcal{M}$) is a *solution* to the structural causal marginal problem.

3.2. Determining the Space of Solutions

As discussed, (15) and the simplex constraints $\mathbf{c} \in \Delta^{15}$, and $(\lambda_X, \lambda_Y) \in [\lambda_X^{\min}, \lambda_X^{\max}] \times [\lambda_Y^{\min}, \lambda_Y^{\max}] =: \Lambda_0$ define the solution space for the structural causal marginal problem. Specifically, they imply a set of *linear* equality and inequality constraints that, if satisfiable, yield a *convex polytope* $\mathcal{C}$ as the feasible set for $\mathbf{c}$ (Boyd & Vandenberghe, 2004):

$$\mathcal{C} := \{\mathbf{c} \in \Delta^{15} \mid \exists (\lambda_X, \lambda_Y) \in \Lambda_0 \text{ s.t. (15) holds}\} \tag{16}$$

see Appx. F for details. By (10) and (15), we have that $\lambda_X = [\mathbf{A}\mathbf{c}]_0$ and $\lambda_Y = [\mathbf{B}\mathbf{c}]_0$, so the set of jointly feasible (λ_X, λ_Y) is given by $\Lambda_{\mathcal{C}} := \{([\mathbf{A}\mathbf{c}]_0, [\mathbf{B}\mathbf{c}]_0)^\top \mid \mathbf{c} \in \mathcal{C}\}$. $\Lambda_{\mathcal{C}}$ is illustrated as the dotted green region in Fig. 2.

We could now minimise and maximise some (linear) causal query $\mathcal{Q}(\mathbf{c})$ over $\mathbf{c} \in \mathcal{C}$, to obtain bounds on counterfactuals of interest, e.g., the ETT for Alice mentioned in § 1. Since $\mathcal{C}$ is convex, this results in a linear program which can be solved easily and with global optimality guarantees (Dantzig, 1963; Karmarkar, 1984). Such an approach is closely related to partial identification (cf. § 1). Here, we focus instead on how the space of marginal and joint models is reduced when additional marginals are observed.

Does enforcing counterfactual consistency meaningfully restrict the space of admissible marginal SCMs? To check this, we can compare, e.g., the interval $[\lambda_X^{\min}, \lambda_X^{\max}]$ of allowed λ_X *prior* to enforcing (15) with the lower and upper bounds $[\mathrm{LB}_X, \mathrm{UB}_X]$ defined as $\min/\max_{\Lambda_{\mathcal{C}}} \lambda_X$, and similarly for $[\mathrm{LB}_Y, \mathrm{UB}_Y]$. The region $[\mathrm{LB}_X, \mathrm{UB}_X] \times [\mathrm{LB}_Y, \mathrm{UB}_Y]$ is illustrated as the solid blue area in Fig. 2. By definition, it is the rectangle delimiting the projection $\Lambda_{\mathcal{C}}$ of the polytope $\mathcal{C}$ of feasible solutions in the (λ_X, λ_Y)-plane. If the blue and dashed red rectangles coincide, neither of the marginal SCMs is further restricted by enforcing consistency. Otherwise, marginals that fall outside the blue area are falsified in that they cannot be counterfactually consistent.

We highlight a subtle point regarding the blue area in Fig. 2, counterfactual consistency, and falsifiability: If (λ_X, λ_Y) lies within the blue region but outside $\Lambda_{\mathcal{C}}$ (e.g., the blue star marker in Fig. 2), the corresponding marginal SCMs $\mathcal{M}_X$ and $\mathcal{M}_Y$ are not counterfactually consistent. However, neither of them is therefore falsified; it is only their combination that can be ruled out. Since we generally know neither of the marginal SCMs exactly (assuming we only observe $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$), for any $\lambda_X \in [\mathrm{LB}_X, \mathrm{UB}_X]$, by definition, there is a λ_Y' such that (λ_X, λ_Y') are counterfactually consistent. Hence, λ_X cannot be ruled out without additional knowledge about λ_Y. If, on the other hand, we know that $\lambda_Y = \lambda_Y^\star$ (illustrated as the horizontal dashed green

[7] Specifically, $(\mathbf{A})_{jk} = \sum_{y=0}^{1} \mathbb{P}_Y(y) \mathbb{I}\{\mathcal{P}_y^Y(h_k) = f_j(X)\}$, and $(\mathbf{B})_{jk} = \sum_{x=0}^{1} \mathbb{P}_X(x) \mathbb{I}\{\mathcal{P}_x^X(h_k) = f_j(Y)\}$.

line in Fig. 2), the red rectangle degenerates to the interval $[\lambda_X^{\min}, \lambda_X^{\max}] \times \{\lambda_Y^\star\}$, and the blue and green regions coincide and collapse to the sub-interval $[\mathrm{LB}_X^\star, \mathrm{UB}_X^\star] \times \{\lambda_Y^\star\}$ defined as $\min/\max_{(\lambda_X, \lambda_Y^\star) \in \Lambda_\mathcal{C}} \lambda_X$, shown as the solid green interval in Fig. 2. Next, we illustrate this with an example.

3.3. Worked-Out Example

Suppose that the observed marginal distributions $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ are such that they satisfy the following: (i) $X \perp\!\!\!\perp Z$, (ii) $\mathbb{P}(Y = 0, Z = 1) = 0$, (iii) $\mathbb{P}(Z = 1) = 0.5$, and (iv) $0 < \mathbb{P}(X = 0), \mathbb{P}(Y = 0) < 1$. Crucially, these assumptions only involve empirical quantities and do not require joint observations of (X, Y, Z). We focus on the main points here and refer to Appx. C for detailed derivations.

First, we consider the $Y \to Z$ marginal SCM $\mathcal{M}_Y$ in (11). Assumption (ii) implies zero probability for the constant one (f_1) and NOT (f_3) functions. Together with (iii), it turns out that this *uniquely* determines $\mathcal{M}_Y$: we must have $\lambda_Y = \lambda_Y^\star := \frac{2\theta - 1}{2\theta}$ where $\theta = \mathbb{P}(Y = 1)$, and the response function distribution is given by $\mathbf{b} = (\frac{2\theta-1}{2\theta}, 0, \frac{1}{2\theta}, 0)^\top$.[8] This can also be written more compactly as an AND model:

$$Z := Y \wedge N_Z, \qquad N_Z \sim \text{Bernoulli}(\tfrac{1}{2\theta}). \tag{17}$$

Next, we consider the $X \to Z$ marginal SCM $\mathcal{M}_X$ in (9). Intuitively, assumption (i) rules out SCMs that do not give equal weight to the constant zero (f_0) and one (f_1) functions, as well as to the ID (f_2) and NOT (f_3) functions, for otherwise X and Z could not be statistically independent. Substituting (i) and (iii) into (10), we indeed find the family of response function distributions implied by $\mathbb{P}_{XZ}$ to be $\mathbf{a}(\lambda_X) = (\lambda_X, \lambda_X, 0.5 - \lambda_X, 0.5 - \lambda_X)^\top$ with $0 \leq \lambda_X \leq 0.5$. For example, for $\lambda_X = 0.5$ this yields

$$Z := M_Z, \qquad M_Z \sim \text{Bernoulli}(0.5), \tag{18}$$

whereas for $\lambda_X = 0$ we obtain

$$Z := X \oplus M_Z, \qquad M_Z \sim \text{Bernoulli}(0.5). \tag{19}$$

As discussed, (18) and (19) are interventionally equivalent—Z is an unbiased coin toss regardless of X—but entail different counterfactuals: given some (x, z), the statement "Z would have been different, had X been $x' \neq x$" would be true only for the XOR model (19) but false for (18). This reflects that for (19) the counterfactual influence is $\gamma = 1$, while for (18) it is $\gamma = 0$. We also note that (19) violates faithfulness (Spirtes et al., 2000).[9]

Next, we analyse whether and how the problem is further constrained by enforcing counterfactual consistency. Recall that in $\mathcal{M}_Y$ we have $b_1 = b_3 = 0$. Together with assumption (iv), the second constraint in (14) for $j = 1, 3$ then implies that all but four of the c_k are zero. The first constraint in (15) then yields a system of four linear equations relating the non-zero components c_0, c_2, c_8, c_{10} of $\mathbf{c}$ to $\mathbf{a}(\lambda_X)$. By solving for c_0 and enforcing positivity, $c_0 \geq 0$, we finally obtain the consistency constraint: $1 - \theta \leq \lambda_X$.

In summary, if we *only* observe $\mathbb{P}_{XZ}$, any $\lambda_X \in [0, 0.5]$ is allowed; if we *additionally* know $\mathbb{P}_{YZ}$ and enforce counterfactual consistency, this interval shrinks to $\lambda_X \in [1-\theta, 0.5]$. The space of counterfactually consistent $\mathcal{M}_X$ can thus be arbitrarily small, depending on $0.5 \leq \theta = \mathbb{P}(Y = 1) < 1$. This is illustrated in Fig. 3 **(a)** for different values of θ, see § 4 for details. In particular, we note that the (unfaithful) XOR model (19) is falsified in that it can never be counterfactually consistent with $\mathcal{M}_Y$ from (17). Moreover, in the extreme case that $\theta = 0.5$, the interval collapses to a point and the only admissible $\mathcal{M}_X$ is (18) where X has no counterfactual influence on Z. This seems intuitive since $\mathcal{M}_Y$ puts all weight on the ID function ($Z := Y$) for $\theta = 0.5$, i.e., Y fully determines Z in that case.

3.4. Some Marginal SCMs Cannot Be Falsified

In the previous example, enforcing consistency with the interventional $Y \to Z$ marginal only affected the *lower* bound on λ_X. In fact, it can be shown that this holds more generally for both λ_X and λ_Y (proof in Appx. E):

Proposition 4. *Consider the structural causal marginal problem described in §§ 3.1 and 3.2, with $X \to Z \leftarrow Y$, causal sufficiency, and Boolean X, Y, Z. If a solution exists (i.e., $\mathcal{C}$ is non-empty), we have $(\lambda_X^{max}, \lambda_Y^{max})^\top \in \Lambda_\mathcal{C}$.*

In particular, this implies $\mathrm{UB}_X = \lambda_X^{\max}$ and $\mathrm{UB}_Y = \lambda_Y^{\max}$, as illustrated in Fig. 3 **(b)**.[10] As a result, the structural causal marginal problem cannot falsify models $\mathcal{M}_X$ or $\mathcal{M}_Y$ that assign the maximally allowed weight to the constant functions $Z := 0$ and $Z := 1$. Conversely, models corresponding to small values of λ_X, λ_Y can sometimes be falsified: note that these are the models where the cause X (or Y) has a stronger counterfactual influence on Z, as defined in § 2.1. We elaborate on the significance of this result in § 6.

4. Experiments

In Fig. 3 **(a)** we visualise the worked-out example from § 3.3. Recall that there the interventional $Y \to Z$ model uniquely determines $\mathcal{M}_Y$, and Λ_0 is therefore a segment with $\lambda_X \in [0, 0.5]$ and y-coordinate fixed by $\theta = \mathbb{P}(Y = 1)$. We take 20 linearly-spaced $\theta \in (0.5, 1)$ and plot both Λ_0 (thick red segments) and the reduced range $[\mathrm{LB}_X^\star, \mathrm{UB}_X^\star]$ (superimposed, thin blue lines). Decreasing θ from 1 (top line at

[8] Note that (ii) and (iii) together imply that $\theta \geq 0.5$ since $\mathbb{P}(Z = 1) = \mathbb{P}(Y = 1, Z = 1) = \theta\, \mathbb{P}(Z = 1 | Y = 1) = 0.5$

[9] Our point could, in principle, also be made for more generic causal models, but the math is less simple then.

[10] Fig. 2 should thus be understood as a conceptual visualisation rather than an exact representation; Fig. 3 **(b)** is a refinement.

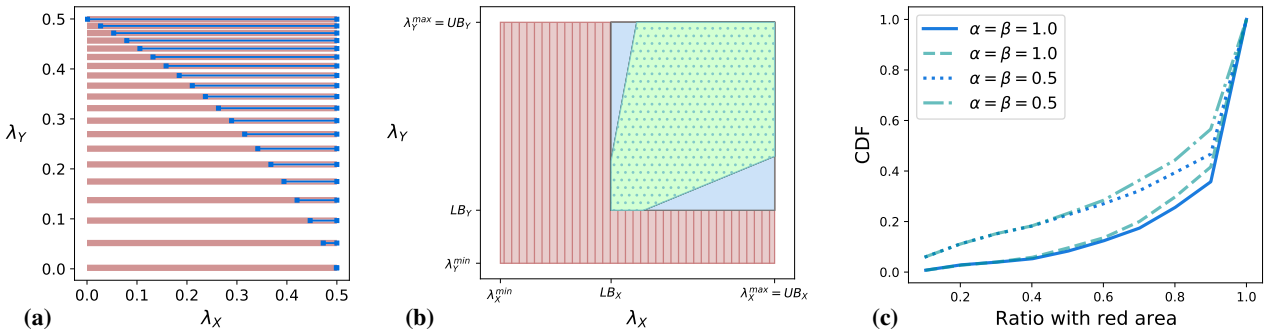

Figure 3: **(a)** For the example from § 3.3, both the unique allowed value of λ_Y and the range of consistent $X \to Z$ models $\mathcal{M}_X$ (thin blue lines) change as $\theta = \mathbb{P}(Y = 1)$ is varied. **(b)** An instance of the structural causal marginal problem that entails constraints for both λ_X and λ_Y. Note that $(\lambda_X^{\max}, \lambda_Y^{\max}) \in \Lambda_{\mathcal{C}}$, as implied by Prop. 4. **(c)** In solid/dotted blue (resp. dashed green), CDFs of the observed ratios between the blue (resp. green) and red area for different Beta priors over $\mathbb{P}_{Z|XY}$: often, counterfactual consistency induces meaningful constraints.

$\lambda_Y = 0.5$) to 0.5 (bottom line at $\lambda_Y = 0$) yields an increase in $\text{LB}_X^\star$, thereby restricting the range of allowed $\mathcal{M}_X$ models and fully specifying it for $\theta = 0.5$ when $Z := Y$.

Extending the analytical treatment of § 3.3 to more general settings is nontrivial. To characterise the entailed constraints in generic settings, we therefore resort to numerical simulations (see Appx. G for all technical details): we generate random instances of consistent $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$, compute the space of solutions $\Lambda_{\mathcal{C}}$, and compare it to Λ_0. A specific instance is shown in Fig. 3 **(b)**; see [GIF1] [GIF2] for additional visualisations, where we fix a conditional $\mathbb{P}_{Z|XY}$ and plot $\Lambda_{\mathcal{C}}$ and Λ_0 for different choices of $\mathbb{P}_X, \mathbb{P}_Y$.[11] The parameters used to generate Fig. 3 **(b)** violate some of the restrictive assumptions of § 3.3 (most notably $X \perp\!\!\!\perp Z$ and $P(Y = 0, Z = 1) = 0$), and show that the schematic visualisation in Fig. 2 captures some aspects of the general case: the structural causal marginal problem can yield constraints for both marginal SCMs $\mathcal{M}_X$ and $\mathcal{M}_Y$, and Λ_0 and $\Lambda_{\mathcal{C}}$ are different. Moreover, we see that $(\lambda_X^{\max}, \lambda_Y^{\max}) \in \Lambda_{\mathcal{C}}$, consistent with Prop. 4.

In Fig. 3 **(c)**, we plot the cumulative distribution functions (CDFs) of the ratios between the blue and red areas (i.e., $(\text{UB}_X - \text{LB}_X)(\text{UB}_Y - \text{LB}_Y)/|\Lambda_0|$) in blue, and the ratio between the green and red areas (i.e., $|\Lambda_{\mathcal{C}}|/|\Lambda_0|$) in green. The CDFs are estimated based on 1,000 independent samples of $\mathbb{P}(Z = 1|X = i, Y = j) \sim \text{Beta}(\alpha, \beta)$ for $i, j \in \{0, 1\}$ and $\mathbb{P}(X = 1), \mathbb{P}(Y = 1) \sim \text{U}[0, 1]$.[11] We compare two scenarios: $\alpha = \beta = 1$, i.e., a Uniform prior, shown as solid lines; and $\alpha = \beta = 0.5$, leading to more deterministic conditionals, shown as dashed lines. Across both scenarios, a reduction (i.e., ratios smaller than one) can be observed at least 30% of the time. Whereas many times there is no or only a small reduction, we also sometimes (with positive probability) observe quite substantial reductions of 50+%. Moreover, we find that $\alpha = \beta = 0.5$ leads to larger reductions, suggesting that more deterministic (joint) conditionals may impose stronger constraints. Finally, we remark that $(\lambda_X^{\max}, \lambda_Y^{\max}) \in \Lambda_{\mathcal{C}}$ indeed holds across all runs.

[11] We fix a $\mathbb{P}_{Z|XY}$ to ensure a solution to the statistical marginal problem exists; only the $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ derived from it are subsequently used to computate the solution spaces.

5. Limitations and Extensions

So far, we have focused on one of the simplest instances of the causal marginal problem involving (i) *only two marginals*, each consisting of a (ii) *Boolean* cause-effect model, under (iii) *causal sufficiency* (Asm. 2). This simplified setting allowed us to focus on the main points and to visualise the problem in 2D (see Figs. 2 and 3 **(b)**). We now discuss how each of these restrictions can be relaxed to allow for more general settings, see Fig. 4 for an illustration.

More than two marginals. Suppose that we have access to $m > 2$ marginals, e.g., by separately observing the effect of $(m - 1)$ additional causes $Y_1, \ldots, Y_{m-1}$ on Z for the setting from Fig. 1 and § 3. Enforcing counterfactual consistency for each marginal would yield m linear constraints in (15), subject to which we could, e.g., solve for $[\text{LB}_X, \text{UB}_X]$. While intuitively this should produce a tighter bound (if feasible), the number of response functions $h_k(X, Y_1, \ldots, Y_{m-1})$ in the joint model grows as 2^{2^m} (cf. (3)) which for large m may pose computational challenges. In this case, analogies to statistical learning theory (Vapnik, 1998) suggest restrictions on the capacity of response functions which may render SCMs with m variables falsifiable from marginal observations with $k \ll m$ variables (Janzing, 2018), see Appx. I for a more detailed discussion. Causal graphs different from Fig. 1 are, of course, also possible, leading to different parametrisations of the

marginal and joint models; the general procedure of deriving linear constraints by enforcing counterfactual consistency and finding the corresponding feasible set would still apply. Note that causal modularity implies that (under Asm. 2) the joint model can be specified by *separately* describing the relations between each variable and its causal parents. Hence, if each variable is observed jointly with *all* its parents in at least one marginal, this provides the same information as joint observation of all variables. This would for example be the case if we had the graph $X \leftarrow Z \leftarrow Y$ or $X \leftarrow Z \rightarrow Y$, but does not hold in the considered case of Fig. 1.

Dependent causes. In our example of Fig. 1, if additionally we have dependent causes, e.g., $X \rightarrow Y$, our approach still applies but needs to be modified: first, the *joint* model now also involves a distribution over the four response functions generating Y from X; second, the *marginal* $Y \rightarrow Z$ model is now confounded by X, which requires specifying a joint distribution over (Q_Y, Q_Z)—see also Appx. D.

Beyond Boolean variables. It is straightforward to extend our approach to arbitrary (non-binary) *categorical* variables. As described in generality in § 2, there would be more response functions, and the marginals may no longer be described by a single parameter but would still be constrained via (7). The projection operators (13) remain the same, and the constraints would be derived analogously to (14) with sums over the respective domains. For *continuous* variables, it is less clear how to proceed, as no simple parametrisation such as (2) exists in general. However, recent work suggests that assumptions on the allowed class of functions f_i in (1) such as Lipschitz-continuity have non-trivial implications for partial identification (Gunsilius, 2018; 2019), see also (Kilbertus et al., 2020; Zhang & Bareinboim, 2021) for recent progress. An alternative is to discretise continuous variables, e.g., by thresholding.

Unobserved confounding. When Asm. 2 is violated, the Markov factorisation (6) does not hold and we cannot consider the distributions $\mathbb{P}_{R_i}$ of each response function variable separately. Instead, we need to parametrise their joint distribution $\mathbb{P}_{\mathbf{R}}$ (which can drastically increase the number of parameters) and derive the constraints imposed by the observational distributions via (4) instead of (7).[12] With hidden confounders, a gap remains between the first and the second rung even when the causal graph is known: interventional distributions are no longer determined by the observed $\mathbb{P}_{\mathbf{V}}$ as (8) no longer holds. Knowing some of the do-probabilities from experimental data therefore provides additional information and imposes further constraints via (5). For a more detailed treatment of a confounded version of the setting from § 3, we refer to Appx. D. Finally, we note that in confounded settings it could also be interesting to consider an instantiation of the causal marginal problem based on interventional models such as causal Bayesian networks (CBNs; Spirtes et al., 2000) instead of SCMs.

[12] For known confounding structures, a partially factorised $\mathbb{P}_{\mathbf{R}}$ could be used, but this typically leads to nonlinear constraints.

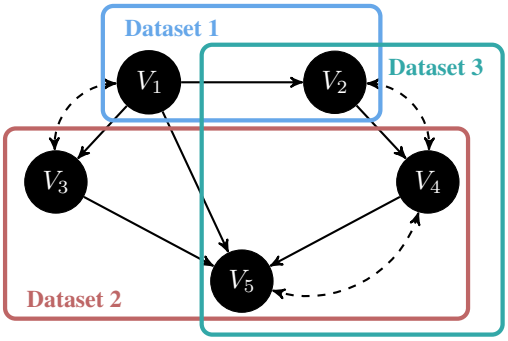

Figure 4: **Illustration of a More General Version of the Causal Marginal Problem.** Here, we have a causal graph over $n = 5$ causal variables and observe $m = 3$ marginals over the subsets $\mathbf{W}_1 = \{V_1, V_2\}$, $\mathbf{W}_2 = \{V_3, V_4, V_5\}$, and $\mathbf{W}_3 = \{V_2, V_4, V_5\}$. Dashed bi-directed arrows indicate unobserved confounding.

With these extensions in mind, we finally give a more general definition of the causal marginal problem (cf. Fig. 4).

Definition 5 (Causal marginal problem)**.** Consider m marginal causal (interventional or counterfactual) models $\mathcal{M}_1, \ldots, \mathcal{M}_m$ over distinct but overlapping sets of variables $\mathbf{W}_1, \ldots, \mathbf{W}_m \subseteq \{V_1, \ldots, V_n\}$, respectively. The *causal marginal problem* consists of determining the space of joint causal (interventional or counterfactual) models $\mathcal{M}$ over $\mathbf{W}_1 \cup \ldots \cup \mathbf{W}_m$ which are (interventionally or counterfactually) consistent with the marginal ones.

6. Related Work

The problem of merging causal models involving overlapping subsets of variables has also been considered by Janzing (2018); Mejia et al. (2022),[13] though focusing on interventional (rung 2) quantities. Marginalisation of SCMs, i.e., the inverse problem of merging, has been discussed by Bongers et al. (2016) and Rubenstein et al. (2017). The latter introduce a notion of *interventional* consistency between marginal and joint SCMs, which complements ours of *counterfactual* consistency (Defn. 3). A related type of consistency between different abstractions of the same underlying causal system has been studied by Chalupka et al. (2016; 2017); Beckers & Halpern (2019); Beckers et al. (2020).

[13] Federated learning (Kairouz et al., 2021) is loosely related: there, the aim is to learn from data from multiple sources (clients), and each client's data is accessible and processed *locally*, whereas in our setting all marginals are available and processed *globally*.

In the present work, we have explored the implications of merging for the space of allowed marginal and joint models, i.e., *partial identification of SCMs*. A parallel literature instead aims to identify *specific causal queries* from a given collection of observational and experimental datasets (involving subsets of variables), a task referred to as *transportability* or *data fusion* (Pearl & Bareinboim, 2014; Bareinboim & Pearl, 2016)—see also Chau et al. (2021) for uncertainty quantification in this context and Lee & Bareinboim (2021) for a combination with proxy-based approaches.

Both our and the aforementioned line of work assume that the causal graph is known a priori. For *causal structure learning* approaches for the setting of multiple datasets involving overlapping sets of variables, we refer to Triantafillou et al. (2010); Tillman & Eberhardt (2014); Triantafillou & Tsamardinos (2015); Huang et al. (2020).

7. Discussion

Empirical content of counterfactuals. The use of counterfactuals in causal inference has long been a subject of debate; as summarised by Shafer (1996): "*were counterfactuals to have objective meaning, we might take them as basic, and define probability and causality in terms of them*". Some prominent approaches to causal inference indeed regard counterfactuals (or potential outcomes) as foundational (Imbens & Rubin, 2015; Pearl, 2009b). Others question the legitimacy of models allowing for direct formulation of counterfactual queries (such as SCMs): Dawid (2000) terms them *"metaphysical"*, arguing that they either yield unscientific (i.e., empirically irrefutable) statements or are unnecessary in that the inferences for which they are used could also be rephrased in non-counterfactual terms.

Our work illustrates a possible mode of falsifiability for counterfactual models: some SCMs may be falsified when previously unobserved variables become observable together with subsets of the original ones (e.g., through a new experiment or study) and are consistently merged into a joint model. For example, in the setting of Fig. 1, the exogenous variable associated to Z could (partly) correspond to Y: together with $Y \to Z$, observing $\mathbb{P}_{YZ}$ would then provide (partial) information on what is otherwise unobserved. This reflects a view according to which counterfactuals do carry an empirical message and *"may earn predictive power"* when *"the uncertainty-producing variables offer the potential of being observed sometime in the future (before our next prediction or action)"* (Pearl, 2009b, § 7.2.2). We further illustrate this point with an example in Appx. H.

Another insight is that *interventional marginal models* (i.e., a causal graph and corresponding observational distribution) *can entail constraints for counterfactual ones* such as SCMs. In other words, questions regarding model consistency may also be meaningful when marginal and joint models do not refer to the same rung in the ladder of causation. This intertwines the two model classes (rungs two and three).

SCMs and falsifiability. Popper (2005) considers falsifiability a crucial property of *scientific* hypotheses: unfalsifiable ones belong to the realm of metaphysics, and falsifiable ones are increasingly corroborated as many attempts to falsify them fail. Prop. 4 suggests that some SCMs are intrinsically 'harder' to falsify in that the space of interventional models they can be consistently merged with is larger: those marginal models with the weakest counterfactual influence can always be consistently merged and are thus not falsifiable through additional variables. Conversely, if a marginal SCM with a strong counterfactual influence can (repeatedly) be merged consistently with new marginals, we obtain indirect evidence for it in the Popperian sense. *Which classes of causal models (beyond our Boolean setting) offer a larger space of possible falsifications?* This parallels the idea of capacity measures in supervised learning, where the generalisation gap is provably smaller if a class of allowed explanations has small capacity relative to the dataset size. The latter means that the space of datasets that would falsify it (in that they cannot be fitted by any explanation in the class) is large (Corfield et al., 2009). By analogy, this would suggest to prefer SCMs that are easier to falsify: the question may be investigated in future work as a first step towards a 'statistical learning theory of causal data fusion'.

Concluding remarks. We introduced the structural causal marginal problem as a framework for merging causal information from different datasets. While previous work focused on bounds on counterfactuals from *joint* observations, we have emphasised bounds and falsifiabiliy that come from *marginal* causal information involving different subsets of variables. This way, causal insights emerge from 'bringing puzzle pieces together' rather than from complete datasets.

Software and Data

Code is available at https://github.com/lgresele/structural-causal-marginal.

Acknowledgements

We thank Sergio Hernan Garrido Mejia, Claudia Shi, Shiva Kasiviswanathan, Filippo Camilloni, Krikamol Muandet, Sander Beckers, Armin Kekić, Atalanti Mastakouri and Kailash Budhathoki for valuable discussions; and the anonymous reviewers for helpful comments. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, 01IS18039B; and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pp. 46–54. Elsevier, 1994.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. On Pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2(3):4, 2020.

Beckers, S. and Halpern, J. Y. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2678–2685, 2019.

Beckers, S., Eberhardt, F., and Halpern, J. Y. Approximate causal abstractions. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pp. 606–615, 2020.

Bongers, S., Peters, J., Schölkopf, B., and Mooij, J. M. Structural causal models: Cycles, marginalizations, exogenous reparametrizations and reductions. *arXiv preprint arXiv*, 1611, 2016.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Caron, S. Polyhedron Manipulation in Python, 2018.

Chalupka, K., Eberhardt, F., and Perona, P. Multi-level cause-effect systems. In *Artificial Intelligence and Statistics*, pp. 361–369, 2016.

Chalupka, K., Eberhardt, F., and Perona, P. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

Chau, S. L., Ton, J.-F., González, J., Teh, Y. W., and Sejdinovic, D. Bayesimp: Uncertainty quantification for causal data fusion. In *Advances in Neural Information Processing Systems 34*, 2021.

Corfield, D., Schölkopf, B., and Vapnik, V. Falsificationism and statistical learning theory: Comparing the Popper and Vapnik-Chervonenkis dimensions. *Journal for General Philosophy of Science*, 40(1):51–58, July 2009.

Correa, J. D., Lee, S., and Bareinboim, E. Nested counterfactual identification from arbitrary surrogate experiments. In *Advances in Neural Information Processing Systems 34*, 2021.

Dantzig, G. *Linear programming and extensions*. Princeton University Press, 1963.

Dawid, A. P. Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450): 407–424, 2000.

Frangakis, C. E. and Rubin, D. B. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Greenland, S. and Robins, J. M. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419, 1986.

Gunsilius, F. Non-testability of instrument validity under continuous endogenous variables. *arXiv preprint arXiv:1806.09517*, 2018.

Gunsilius, F. A path-sampling method to partially identify causal effects in instrumental variable models. *arXiv preprint arXiv:1910.09502*, 2019.

Heckman, J. J. Randomization and social policy evaluation. *Evaluating welfare and training programs*, 1:201–30, 1992.

Huang, B., Zhang, K., Gong, M., and Glymour, C. Causal discovery from multiple data sets with non-identical variable sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10153–10161, 2020.

Imbens, G. W. and Angrist, J. D. Identification and estimation of local average treatment effects. *Econometrica*, 62 (2):467–475, 1994.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Janzing, D. Merging joint distributions via causal model classes with low VC dimension. preprint arXiv:1804.03206v2, 2018.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Karmarkar, N. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984.

Kellerer, H. Maßtheoretische Marginalprobleme. *Math. Ann.*, 153:168–198, 1964. In German.

Kilbertus, N., Kusner, M. J., and Silva, R. A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems 33*, pp. 20108–20119, 2020.

Lee, S. and Bareinboim, E. Causal identification with matrix equations. In *Advances in Neural Information Processing Systems 34*, 2021.

Manski, C. F. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Mejia, S. H. G., Kirschbaum, E., and Janzing, D. Obtaining causal information by merging datasets with maxent. In *International Conference on Artificial Intelligence and Statistics*, pp. 581–603. PMLR, 2022.

Pearl, J. The logic of counterfactuals in causal inference. *Journal of the American Statistical Association (Discussion of 'Causal Inference without Counterfactuals' by A.P. Dawid)*, 2000.

Pearl, J. Direct and indirect effects. In *7th Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, 2001.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009a.

Pearl, J. *Causality: Models, reasoning, and inference*. Cambridge university press, 2009b.

Pearl, J. and Bareinboim, E. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

Pearl, J. and Mackenzie, D. *The book of why: the new science of cause and effect*. Basic books, 2018.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Popper, K. R. *The logic of scientific discovery*. Routledge, 2005. Original in German: "Logik der Forschung" (1934).

Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

Robins, J. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pp. 113–159, 1989.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

Shafer, G. *The art of causal conjecture*. MIT press, 1996.

Shpitser, I. and Pearl, J. What counterfactuals can be tested. In *23rd Conference on Uncertainty in Artificial Intelligence*, pp. 352–359, 2007.

Shpitser, I. and Pearl, J. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Shpitser, I. and Pearl, J. Effects of treatment on the treated: Identification and generalization. In *25th Conference on Uncertainty in Artificial Intelligence*, 2009.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, 2000.

Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Tillman, R. E. and Eberhardt, F. Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika*, 41(1):41–64, 2014.

Triantafillou, S. and Tsamardinos, I. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, 16(1):2147–2205, 2015.

Triantafillou, S., Tsamardinos, I., and Tollis, I. Learning causal structure from overlapping variable sets. In *13th International Conference on Artificial Intelligence and Statistics*, pp. 860–867, 2010.

Vapnik, V. *Statistical learning theory*. John Wileys & Sons, New York, 1998.

Vapnik, V. N. and Chervonenkis, A. Y. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272, 2020.

Vorob'ev, N. N. Consistent families of measures and their extensions. *Theory of Probability & Its Applications*, 7 (2):147–163, 1962.

Zhang, J. and Bareinboim, E. Bounding causal effects on continuous outcome. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

Zhang, J., Tian, J., and Bareinboim, E. Partial counterfactual identification from observational and experimental data. *arXiv preprint arXiv:2110.05690*, 2021.

A. Parametrisation of Boolean Causally-Sufficient Cause-Effect Models

The definition of the response functions of the joint model is provided in Tab. 1. We now derive the family of SCMs that are (interventionally) consistent with $X \to Z$ and the observed $\mathbb{P}_{XZ}$ for the setting considered in § 2.1, i.e., assuming causal sufficiency and Boolean variables.

First, we note that the marginal distribution $\mathbb{P}_X$ completely determines the distribution of R_X in (9).

Next, we consider how $\mathbb{P}_{Z|X}$ constrains the distribution of R_Z, i.e., the probability vector $\mathbf{a} \in \Delta^3$.

From (7) and the definition of the response functions f_i, we obtain the following two independent constraints:

$$\mathbb{P}_{Z|X}(Z=0|X=0) =: p_{00} = a_0 + a_2 \tag{20}$$
$$\mathbb{P}_{Z|X}(Z=0|X=1) =: p_{01} = a_0 + a_3 \tag{21}$$

Additionally, we have the simplex constraint:

$$1 = a_0 + a_1 + a_2 + a_3 \tag{22}$$

We now solve the under-determined system of equations (20), (21), (22) by setting $a_0 =: \lambda_X$.

From (20) and (21), this yields

$$a_2 = p_{00} - \lambda_X \tag{23}$$
$$a_3 = p_{01} - \lambda_X \tag{24}$$

and finally, by substitution in (22),

$$a_1 = 1 - \lambda_X - (p_{00} - \lambda_X) - (p_{01} - \lambda_X) = 1 - p_{00} - p_{01} + \lambda_X \tag{25}$$

Writing this as a vector, we obtain the following form for $\mathbf{a}$:

$$\mathbf{a} = \begin{pmatrix} 0 \\ 1 - p_{00} - p_{01} \\ p_{00} \\ p_{01} \end{pmatrix} + \lambda_X \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \tag{26}$$

with a single free parameter $\lambda_X \in \mathbb{R}$.

Since we require $0 \leq a_i \leq 1, \forall i$ for $\mathbf{a}$ to be a valid probability vector, we find the admissible range of λ_X to be:

$$\max\{0, p_{00} + p_{01} - 1\} \leq \lambda_X \leq \min\{p_{00}, p_{01}\} \tag{27}$$

Similarly, we can characterise the other marginal SCM $\mathcal{M}^B$ over $Y \to Z$ in terms of its observational distribution. Denoting $p'_{ij} := \mathbb{P}(Z=i|Y=j)$, this yields analogously:

$$\mathbf{b} = \begin{pmatrix} 0 \\ 1 - p'_{00} - p'_{01} \\ p'_{00} \\ p'_{01} \end{pmatrix} + \lambda_Y \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \tag{28}$$

with

$$\max\{0, p'_{00} + p'_{01} - 1\} \leq \lambda_Y \leq \min\{p'_{00}, p'_{01}\} \tag{29}$$

B. Parametrisation of the Joint SCM with two Boolean causes

Table 1: Definition of the 16 response functions $h_k(X, Y)$ from (12) mapping two Boolean inputs to a Boolean output. Each row corresponds to one of the four different combinations (x, y) of the two inputs X and Y; columns correspond to different h_k; and cells indicate the corresponding output of $h_k(x, y)$.

X	Y	h_0	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}
0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
0	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

C. Details for the Example from § 3.3

We now provide a more detailed account of the worked-out example from § 3.3.

Recall that we make the following assumptions:

(i) $X \perp\!\!\!\perp Z$

(ii) $\mathbb{P}(Y = 0, Z = 1) = 0$,

(iii) $\mathbb{P}(Z = 1) = 0.5$,

(iv) $0 < \mathbb{P}(X = 0), \mathbb{P}(Y = 0) < 1$.

C.1. Derivation of the SCM $\mathcal{M}_Y$ over $Y \to Z$

Recall from § 2.1 that the SCM $\mathcal{M}_Y$ over $Y \to Z$ is characterised by the probability vector

$$\mathbf{b} = \begin{pmatrix} 0 \\ 1 - p'_{00} - p'_{01} \\ p'_{00} \\ p'_{01} \end{pmatrix} + \lambda_Y \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$$

where $p'_{ij} = \mathbb{P}(Z = i|Y = j)$, see Appx. A.

Now by assumption (ii) and (7), we have that

$$\mathbb{P}(Z = 1|Y = 0) = 0 = b_1 + b_3 \tag{30}$$

from which we conclude that

$$b_3 = p'_{01} - \lambda_Y = 0 \Leftrightarrow \lambda_Y = p'_{01} \tag{31}$$

and

$$b_1 = 1 - p'_{00} - p'_{01} + \lambda_Y = 0 \Leftrightarrow p'_{00} = 1. \tag{32}$$

This yields the following intermediate form of $\mathbf{b}$:

$$\mathbf{b} = \begin{pmatrix} p'_{01} \\ 0 \\ 1 - p'_{01} \\ 0 \end{pmatrix}$$

Next, by assumptions (ii) and (iii) we have that

$$\mathbb{P}(Z = 1) = 0.5 = \mathbb{P}(Y = 1)\mathbb{P}(Z = 1|Y = 1) = \mathbb{P}(Y = 1)(1 - p'_{01}) \tag{33}$$

Writing $\theta := \mathbb{P}(Y = 1)$ and solving for p'_{01} we find

$$p'_{01} = \frac{2\theta - 1}{2\theta} \tag{34}$$

which yields the final expression $\mathbf{b} = (\frac{2\theta-1}{2\theta}, 0, \frac{1}{2\theta}, 0)^\top$ as stated in the main paper.

In other words, $Z := Y$ with probability $\frac{1}{2\theta}$ and $Z := 0$ otherwise, which corresponds to the AND model from (17).

C.2. Derivation of the family of SCMs $\mathcal{M}_X$ over $X \to Z$

Next, we consider the family of SCMs $\mathcal{M}_X$ over $X \to Z$ which is characterised by the response function probability vector

$$\mathbf{a} = \begin{pmatrix} 0 \\ 1 - p_{00} - p_{01} \\ p_{00} \\ p_{01} \end{pmatrix} + \lambda_X \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$$

where $p_{ij} = \mathbb{P}(Z = i|X = j)$, see Appx. A.

By assumption (iii) and independence of X and Z (assumption (i)), we have that:

$$\mathbb{P}(Z = 0) = 0.5 = p_{00} = p_{01} \tag{35}$$

Substituting the above into the expression for $\mathbf{a}(\lambda_X)$, we obtain:

$$\mathbf{a}(\lambda_X) = \begin{pmatrix} \lambda_X \\ \lambda_X \\ 0.5 - \lambda_X \\ 0.5 - \lambda_X \end{pmatrix}. \tag{36}$$

as well as $\lambda_X \in [\lambda_X^{\min}, \lambda_X^{\max}] = [0, 0.5]$ as stated in § 3.3.

Moreover, for $\lambda_X = 0.5$, we have that $Z := 0$ or $Z := 1$, both with probability 0.5 corresponding to (18), whereas for $\lambda_X = 0$, we have that $Z := X$ or $Z := 1 - X$, both with probability 0.5 corresponding to (19).

C.3. Enforcing counterfactual consistency between the marginal SCMs

We now explore the implications of enforcing counterfactual consistency between the two marginal SCMs taking the forms derived in the previous two subsections. To this end, we consider what the valid choices for the joint model, i.e., for $\mathbf{c}$, are, and whether this imposes additional constraints on the marginal models.

First, note that for $\mathcal{M}_Y$ we have $b_1 = b_3 = 0$. According to (14), this implies:

$$b_1 = 0 = \mathbb{P}(X = 0)(c_3 + c_7 + c_{11} + c_{15}) + \mathbb{P}(X = 1)(c_{12} + c_{13} + c_{14} + c_{15}) \tag{37}$$

$$b_3 = 0 = \mathbb{P}(X = 0)(c_1 + c_5 + c_9 + c_{13}) + \mathbb{P}(X = 1)(c_4 + c_5 + c_6 + c_7) \tag{38}$$

Together with $0 < \mathbb{P}(X = 0) < 1$ from assumption (iv), and since $c_i \geq 0$, we must have that:

$$c_1 = c_3 = c_4 = c_5 = c_6 = c_7 = c_9 = c_{11} = c_{12} = c_{13} = c_{14} = c_{15} = 0. \tag{39}$$

This only leaves c_0, c_2, c_8, c_{10} as non-zero elements of $\mathbf{c}$.

We now consider counterfactual consistency with $\mathcal{M}_X$. Writing the constraint $\mathbf{a}(\lambda_X) = \mathbf{A}\mathbf{c}$ from (15) subject to (39), we obtain:

$$\begin{pmatrix} \lambda_X \\ \lambda_X \\ 0.5 - \lambda_X \\ 0.5 - \lambda_X \end{pmatrix} = \begin{pmatrix} 1 & \mathbb{P}(Y = 0) & \mathbb{P}(Y = 0) & \mathbb{P}(Y = 0) \\ 0 & 0 & 0 & \mathbb{P}(Y = 1) \\ 0 & 0 & \mathbb{P}(Y = 1) & 0 \\ 0 & \mathbb{P}(Y = 1) & 0 & 0 \end{pmatrix} \begin{pmatrix} c_0 \\ c_2 \\ c_8 \\ c_{10} \end{pmatrix} \tag{40}$$

Whereas $(c_2, c_8, c_{10}) = \frac{1}{\mathbb{P}(Y=1)}(\lambda_X, 0.5 - \lambda_X, 0.5 - \lambda_X)$ are always valid probabilities for $\lambda_X \in [0, 0.5]$ since we must have $\theta = \mathbb{P}(Y = 1) \geq 0.5$ by assumptions (ii) and (iii) (see footnote 7 in § 3.3), solving for c_0 yields:

$$c_0 = \lambda_X - \mathbb{P}(Y = 0)(c_2 + c_8 + c_{10}) = \lambda_X - \frac{\mathbb{P}(Y = 0)}{\mathbb{P}(Y = 1)}(1 - \lambda_X) \geq 0 \tag{41}$$

$$\Leftrightarrow \lambda_X \left(1 + \frac{1-\theta}{\theta}\right) \geq \frac{1-\theta}{\theta} \Leftrightarrow \lambda_X \geq 1 - \theta \tag{42}$$

That is, in order for there to be a valid solution $\mathbf{c}$, the additional constraint $\lambda_X \geq 1 - \theta$ must be satisfied.

D. Structural Causal Marginal Problem *With Unobserved Confounding*

We now provide a more detailed treatment of a version of the structural causal marginal problem from § 3 in which causal sufficiency (Asm. 2) is violated, i.e., allowing for arbitrary unobserved confounding. Specifically, we consider the setting in which both marginal SCMs $X \to Z$ and $Y \to Z$ are confounded (i.e., there exist unobserved variables influencing $\{X, Z\}$ and $\{Y, Z\}$, respectively). Likewise, the joint model $\{X, Y\} \to Z$ is also assumed to be potentially confounded.

On a technical level, unobserved confounding manifests in a potential *dependence of the exogenous noise terms* in the structural equations. In other words, the distribution over exogenous variables no longer factorises—in contrast to the unconfounded case. Within the response function framework, this means that the input or cause is no longer independent of the function or mechanism generating the effect from the cause(s). This means that we cannot parametrise the cause distribution and distribution over functions separately, but instead need to consider their joint distribution $\mathbb{P}_{\mathbf{R}}$.

D.1. Constraints imposed on the marginal SCMs by the observational marginal distributions

First we consider the marginal SCMs $\mathcal{M}_X$ and $\mathcal{M}_Y$ (defined as in the main paper) and investigate how they are constrained by the observed $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$.

For notational convenience, we denote the observational distributions by

$$\begin{aligned} \mathbb{P}(X = i, Z = j) &= \alpha_{ij}, \\ \mathbb{P}(Y = i, Z = j) &= \beta_{ij}, \end{aligned} \tag{43}$$

for $i, j \in \{0, 1\}$, and collect them in vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Delta^3 \subseteq [0, 1]^4$.

Similarly, we parametrise the joint distributions over the corresponding response function variables as follows:

$$\begin{aligned} \mathbb{P}(R_X = i, R_Z = j) &= \mathbb{P}(X = i, R_Z = j) = q^X_{ij}, \\ \mathbb{P}(Q_Y = i, Q_Z = j) &= \mathbb{P}(Y = i, Q_Z = j) = q^Y_{ij}, \end{aligned} \tag{44}$$

for $i \in \{0, 1\}, j \in \{0, 1, 2, 3\}$, and collect them in vectors $\mathbf{q}^X, \mathbf{q}^Y \in \Delta^7 \subseteq [0, 1]^8$.

Since the Markov factorisation (6) does not hold under hidden confounding, we need to derive the constraints imposed by the observational distributions $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ via (4) instead of (7). This yields:

$$\begin{aligned} \alpha_{ij} = \mathbb{P}(X = i, Z = j) &= \sum_{i'=0}^{1} \sum_{j'=0}^{3} \mathbb{P}(R_X = i', R_Z = j') \mathbb{I}\{i = i'\} \mathbb{I}\{j = f_{j'}(i')\} = \sum_{j'=0}^{3} q^X_{ij'} \mathbb{I}\{j = f_{j'}(i)\} \\ \beta_{ij} = \mathbb{P}(Y = i, Z = j) &= \sum_{i'=0}^{1} \sum_{j'=0}^{3} \mathbb{P}(Q_Y = i', Q_Z = j') \mathbb{I}\{i = i'\} \mathbb{I}\{j = f_{j'}(i')\} = \sum_{j'=0}^{3} q^Y_{ij'} \mathbb{I}\{j = f_{j'}(i)\} \end{aligned} \tag{45}$$

for $i, j \in \{0, 1\}$. Writing the above in matrix form, we thus obtain the constraints

$$\begin{aligned} \boldsymbol{\alpha} &= \mathbf{L}^X \mathbf{q}^X \\ \boldsymbol{\beta} &= \mathbf{L}^Y \mathbf{q}^Y \end{aligned} \tag{46}$$

where $\mathbf{L}^X, \mathbf{L}^Y \in \{0, 1\}^{4 \times 8}$ are binary constraint matrices.

We conclude that the space of all potentially confounded SCMs $\mathcal{M}_X$ over binary $X \to Z$ consistent with a given observational joint distribution $\boldsymbol{\alpha}$ is parametrised by all $\mathbf{q}^X \in \Delta^7$ which satisfy $\boldsymbol{\alpha} = \mathbf{L}^X \mathbf{q}^X$.

Unlike in the unconfounded case, this results in four free parameters: seven free parameters with three linearly independent constraints. (Note that matching the observational distribution only eliminates three instead of four free parameters since the fourth constraint is a linear combination of the other three: $\alpha_{11} = 1 - \alpha_{00} - \alpha_{01} - \alpha_{10}$.)

Analogously, any potentially confounded SCM $\mathcal{M}_Y$ over binary $Y \to Z$ consistent with a given observational joint distribution $\boldsymbol{\beta}$ is parametrised by all $\mathbf{q}^Y \in \Delta^7$ which satisfy $\boldsymbol{\beta} = \mathbf{L}^Y \mathbf{q}^Y$.

D.2. Additional constraints imposed via experimental data

Whereas in the unconfounded cases, the observational $\mathbb{P}_{Z|X}$ and interventional $\mathbb{P}_{Z|do(X)}$ conditionals are identical,

$$\forall (x, z): \qquad \mathbb{P}(Z = z | X = x) = \mathbb{P}(Z = z | do(X = x)),$$

this is not the case when unobserved confounding is allowed: there may exist some (x, z) such that

$$\mathbb{P}(Z = z | X = x) \neq \mathbb{P}(Z = z | do(X = x)).$$

Intuitively, with hidden confounding, the observational conditional captures two types of dependence: (i) the direct dependence between X and Z, and (ii) the (indirect) dependence due to their (unobserved) common cause. The interventional distribution, on the other hand, only comprises the first type (i). As a consequence, having access not only to the marginal observational distribution $\mathbb{P}_{XZ}$ but also to the do probabilities $\mathbb{P}(Z = z | do(X = x))$ may impose additional constraints.

Specifically, we have the following additional constraints:

$$\begin{aligned} \alpha_{ij}^{\text{IV}} &= \mathbb{P}(Z = j | do(X = i)) = \sum_{j'=0}^{3} (q_{0j'}^X + q_{1j'}^X) \mathbb{I}\{j = f_{j'}(i)\} \\ \beta_{ij}^{\text{IV}} &= \mathbb{P}(Z = j | do(Y = i)) = \sum_{j'=0}^{3} (q_{0j'}^Y + q_{1j'}^Y) \mathbb{I}\{j = f_{j'}(i)\} \end{aligned} \tag{47}$$

Note that in contrast to before, we are additionally summing over the first subscript of q leading to the $(q_{0j'}^A + q_{1j'}^A)$ terms. This is because—unlike in the observational case—the value of the exogenous variable R_X associated with X does not matter, since X is fixed by intervention, rather than taking on its natural value through the mechanism f.

The above can be written in matrix form as follows:

$$\begin{aligned} \boldsymbol{\alpha}^{\text{IV}} &= \mathbf{L}_{\text{IV}}^X \mathbf{q}^X \\ \boldsymbol{\beta}^{\text{IV}} &= \mathbf{L}_{\text{IV}}^Y \mathbf{q}^Y. \end{aligned} \tag{48}$$

If experimental data in the form of $\boldsymbol{\alpha}^{\text{IV}}, \boldsymbol{\beta}^{\text{IV}}$ (or parts thereof) are available, we can use it to additionally constrain $\mathbf{q}^X, \mathbf{q}^Y$. The number of free parameters for each of $\mathbf{q}^X, \mathbf{q}^Y$ can be reduced by at most two more this way, leaving a total of two free parameters each.

D.3. Additional constraints via assumptions such as monotonicity

Another way of reducing the number of free parameters is by means of additional assumptions such as the *monotonicity* assumption, which is common in epidemiology and economics, particularly in the context of instrumental variable (IV) models (Imbens & Angrist, 1994), and posits that there are no "defiers", i.e., the weight of the NOT function f_3 is zero:

$$\begin{aligned} \mathbb{P}(R_Z = 3) = 0 &= q_{03}^X + q_{13}^X \\ \mathbb{P}(Q_Z = 3) = 0 &= q_{03}^Y + q_{13}^Y \end{aligned} \tag{49}$$

D.4. Parametrisation of the joint SCM

Next, we parametrise the joint SCM $\mathcal{M}$ over $\{X, Y\} \to Z$. Whereas in the unconfounded case, we were able to conclude that $X \perp\!\!\!\perp Y$ (for otherwise one of the marginal SCMs would be confounded), this is not necessarily true in the more general

confounded case. Here, we will work under the assumption that we do not know the causal ordering between X and Y and therefore cannot specify a full SCM without additional assumptions or background knowledge. We therefore proceed with specifying a *partial* causal model, consisting of (i) a joint distribution $\mathbb{P}_{XY}$, and (ii) the structural equation generating Z

$$Z := h_S(X, Y) \tag{50}$$

Note that such a model will only allow us to reason interventionally and counterfactually about joint interventions of the form $do(X := x, Y := y)$ but not about single node interventions, $do(X := x)$ or $do(Y := y)$, since we are not modelling the causal relationship between X and Y. (And, since we never observe X and Y jointly, we may not be able to infer it.)

As in the unconfounded case, the response function variable S takes values in $\{0, 1, ..., 15\}$ indexing the 16 functions $h_k : \{0, 1\}^2 \to \{0, 1\}$ listed in Tab. 1 in Appx. B.

We parametrise this joint partial causal model over binary, potentially confounded $\{X, Y\} \to Z$ as follows:

$$\mathbb{P}(X = i, Y = j, S = k) = q_{ijk} \tag{51}$$

and collect these probabilities of the $2 \times 2 \times 16 = 64$ joint states in a vector $\mathbf{q} \in \Delta^{63} \subseteq [0, 1]^{64}$.

D.5. Enforcing consistency between the joint and marginal models

We now impose the additional constraint that the two marginal SCMs $\mathcal{M}_X$ and $\mathcal{M}_Y$ parametrised by $\mathbf{q}^X$ and $\mathbf{q}^Y$, respectively, must be counterfactually consistent (at the level of counterfactual involving Z under changes to X and Y) with the (partial) joint model parametrised by $\mathbf{q}$. To this end, we proceed as in the unconfounded case making use of the projection operators $\mathcal{P}_y^Y$ and $\mathcal{P}_x^X$ from (13) which given a particular value $Y = y$ or $X = x$ map the functions h_k to functions $f_j(X)$ or $f_{j'}(Y)$, respectively.

Specifically, for enforcing consistency of the joint model $\mathcal{M}$ with $\mathcal{M}_X$ over $X \to Z$ after marginalisation of Y, we obtain for all $i \in \{0, 1\}$ and $j \in \{0, 1, 2, 3\}$:

$$q_{ij}^X = \mathbb{P}(X = i, R_Z = j) = \sum_{y=0}^{1} \sum_{k:\mathcal{P}_y^Y(h_k)=f_j} \mathbb{P}(X = i, Y = y, S = k) = \sum_{y=0}^{1} \sum_{k:\mathcal{P}_y^Y(h_k)=f_j} q_{iyk} \tag{52}$$

Similarly, for consistency of $\mathcal{M}$ with $\mathcal{M}_Y$ after marginalisation of X we obtain for all $i \in \{0, 1\}$ and $j \in \{0, 1, 2, 3\}$:

$$q_{ij}^Y = \mathbb{P}(Y = i, Q_Z = j) = \sum_{x=0}^{1} \sum_{k:\mathcal{P}_x^X(h_k)=f_j} \mathbb{P}(X = x, Y = i, S = k) = \sum_{x=0}^{1} \sum_{k:\mathcal{P}_x^X(h_k)=f_j} q_{xik} \tag{53}$$

This can be written in matrix form as

$$\begin{aligned} \mathbf{q}^X &= \mathbf{K}^X \mathbf{q} \\ \mathbf{q}^Y &= \mathbf{K}^Y \mathbf{q} \end{aligned} \tag{54}$$

where $\mathbf{K}^X, \mathbf{K}^Y \in \{0, 1\}^{8 \times 64}$ are binary constraint matrices.

D.6. Linear program and polytope of solutions

We can now reason about different types of interventional and counterfactual queries that can be expressed in terms of $\mathbf{q}^X$, $\mathbf{q}^Y$, or $\mathbf{q}$ subject to the constraints imposed by enforcing consistency:

(i) between the two families of marginal SCMs $\mathcal{M}_X$ and $\mathcal{M}_Y$ parametrised by $\mathbf{q}^X$ and $\mathbf{q}^Y$ with their respective observational distributions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$;

(ii) between the joint (partial) SCM $\mathcal{M}$ parametrised by $\mathbf{q}$ with the two marginal SCMs $\mathcal{M}_X$ and $\mathcal{M}_Y$.

Denoting the query of interest by $\mathcal{Q}$, this leads to the following optimisation problem which is again a linear program:

$$\begin{aligned} \min/\max_{\mathbf{q}^X,\mathbf{q}^Y\in\Delta^7,\mathbf{q}\in\Delta^{63}} \quad & \mathcal{Q}(\mathbf{q}^X,\mathbf{q}^Y,\mathbf{q}) \\ \text{subject to:} \quad & \boldsymbol{\alpha} = \mathbf{L}^X\mathbf{q}^X \\ & \boldsymbol{\beta} = \mathbf{L}^Y\mathbf{q}^Y \\ & \boldsymbol{\alpha}^{\text{IV}} = \mathbf{L}^X_{\text{IV}}\mathbf{q}^X \\ & \boldsymbol{\beta}^{\text{IV}} = \mathbf{L}^Y_{\text{IV}}\mathbf{q}^Y \\ & \mathbf{q}^X = \mathbf{K}^X\mathbf{q} \\ & \mathbf{q}^Y = \mathbf{K}^Y\mathbf{q} \end{aligned} \tag{55}$$

where $\boldsymbol{\alpha},\boldsymbol{\beta}$ and $\boldsymbol{\alpha}^{\text{IV}},\boldsymbol{\beta}^{\text{IV}}$ (provided experimental data is available) are known constant probabilities, and $\mathbf{L}^X,\mathbf{L}^Y,\mathbf{L}^X_{\text{IV}},\mathbf{L}^Y_{\text{IV}},\mathbf{K}^X,\mathbf{K}^Y$ are known constant, binary constraint matrices.

Similarly to (16) (see Appx. F for more details) we can also first define the space of allowed joint SCMs as a polytope:

$$\mathcal{C}_{\text{conf}} := \left\{\mathbf{q}\in\Delta^{63} \mid \exists \mathbf{q}^X,\mathbf{q}^Y\in\Delta^7 :\ \boldsymbol{\alpha}=\mathbf{L}^X\mathbf{q}^X, \boldsymbol{\beta}=\mathbf{L}^Y\mathbf{q}^Y, \boldsymbol{\alpha}^{\text{IV}}=\mathbf{L}^X_{\text{IV}}\mathbf{q}^X, \boldsymbol{\beta}^{\text{IV}}=\mathbf{L}^Y_{\text{IV}}\mathbf{q}^Y, \mathbf{q}^X=\mathbf{K}^X\mathbf{q}, \mathbf{q}^Y=\mathbf{K}^Y\mathbf{q}\right\}.$$

The vertices of $\mathcal{C}_{\text{conf}}$ can be found using numerical solvers in analogy to how this is done in the unconfounded case, see Appx. F. We could then optimise queries over allowed $\mathbf{q}$ simply by optimising over $\mathcal{C}_{\text{conf}}$, an equivalent formulation to (55). Arguably, however, if one solely cares about a specific query, solving (55) is more direct. Furthermore, if one is interested in the consistent marginal models, we can compute the projection of the vertices of $\mathcal{C}_{\text{conf}}$ onto, say, $\mathbf{q}^X$ via $\mathbf{q}^X=\mathbf{K}^X\mathbf{q}$ and define the consistent SCMs from $X\to Z$ as the convex hull of the projected vertices.

D.7. What counterfactual queries can be addressed by which model?

We may wonder what types of counterfactual queries each of the marginal and joint SCMs may be able to answer, especially given that we only considered a partial specification of the joint SCM. We summarise this as follows:

$\mathcal{M}_X$: $\mathbb{P}(Z_{do(x)}|x',z')$
$\mathcal{M}_Y$: $\mathbb{P}(Z_{do(y)}|y',z')$
$\mathcal{M}$: $\mathbb{P}(Z_{do(x,y)}|x',y',z')$
None: $\mathbb{P}(Z_{do(x)}|x',y',z'), \mathbb{P}(Z_{do(y)}|x',y',z')$

Answering the last type of query would require either additional assumptions such as $X \perp\!\!\!\perp Y$, or knowledge of the qualitative causal relationship between X and Y, either whether we have $X\to Y$ or $Y\to X$.

E. Proof of Prop. 4

In order to prove Prop. 4, we will use the following Lemma, which we prove separately in Appx. E.1.

Lemma 6. *Consider the setting $X\to Z\leftarrow Y$ as in § 3.2. Assume the two statistical marginal models $\mathbb{P}_{XZ}$, $\mathbb{P}_{YZ}$ can successfully be merged, and*

$$\delta_X := \mathbb{P}(Z=0|X=1)-\mathbb{P}(Z=0|X=0)\geq 0 \quad \textit{and} \quad \delta_Y := \mathbb{P}(Z=0|Y=1)-\mathbb{P}(Z=0|Y=0)\geq 0.$$

Then there exist conditional probabilities $q_{i,j} := \mathbb{P}(Z=0|X=i,Y=j)$ such that

$$q_{00}\leq q_{01}\leq q_{11} \quad \textit{and} \quad q_{00}\leq q_{10}\leq q_{11}$$

and the distribution defined via $\mathbb{P}_{XYZ}(X=i,Y=j,Z=0)=q_{ij}\mathbb{P}(X=i)\mathbb{P}(Y=j)$, has marginals that coincide with $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$.

Analogous statements as in Lemma 6 hold by swapping the roles of $X=1$ and $X=0$, or $Y=1$ and $Y=0$, respectively. For convenience we now restate Prop. 4 and then provide its proof.

Proposition 4. Consider the Boolean setting $X \to Z \leftarrow Y$ with marginal and joint models as defined in § 3.2. If a solution to the structural causal marginal problem exists (i.e., $\mathcal{C}$ is non-empty), we have $(\lambda_X^{\max}, \lambda_Y^{\max})^\top \in \Lambda_{\mathcal{C}}$.

Proof of Prop. 4. As we do throughout we assume that $0 < \mathbb{P}(X = 1) < 1$, and $0 < \mathbb{P}(Y = 1) < 1$.

In the setting we consider and under the assumption that the statistical marginal models can be merged, there exists a conditional distribution $\mathbb{P}(Z|X, Y)$ such that we have $\mathbb{P}(X, Y, Z) = \mathbb{P}(Z|X, Y)\mathbb{P}(X)\mathbb{P}(Y)$ and all statistical constraints are satisfied. $\mathbb{P}(Z|X, Y)$ is completely characterised by four probabilities $0 \leq q_{i,j} \leq 1$ that are defined as

$$q_{i,j} := \mathbb{P}(Z = 0|X = i, Y = j)$$

for $i, j \in \{0, 1\}$. We will construct $\mathbf{c}$ for the case that

$$\textbf{(S1)} \quad \mathbb{P}(Z = 0|X = 0) \leq \mathbb{P}(Z = 0|X = 1) \quad \text{and} \quad \mathbb{P}(Z = 0|Y = 0) \leq \mathbb{P}(Z = 0|Y = 1).$$

For all the other cases, the construction of $\mathbf{c}$ follows in analogy.

By Lemma 6, there exists $q_{00}, q_{01}, q_{10}, q_{11}$ that are consistent with $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ and such that

$$q_{00} \leq q_{01} \leq q_{11} \quad \text{and} \quad q_{00} \leq q_{10} \leq q_{11}.$$

Given $\{q_{ij}\}$, the question is, can we find a corresponding vector $\mathbf{c} \in \Delta^{15}$ that implies $\lambda_X = \lambda_X^{\max}$ and $\lambda_Y = \lambda_Y^{\max}$? The connection between $\{q_{ij}\}$ and $\mathbf{c}$ is given in eq. (7) and can be derived from Tab. 1:

$$\begin{aligned} q_{00} &= c_0 + c_2 + c_4 + c_6 + c_8 + c_{10} + c_{12} + c_{14} \\ q_{01} &= c_0 + c_1 + c_4 + c_5 + c_8 + c_9 + c_{12} + c_{13} \\ q_{10} &= c_0 + c_1 + c_2 + c_3 + c_8 + c_9 + c_{10} + c_{11} \\ q_{11} &= c_0 + c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7 \end{aligned} \tag{56}$$

Recall that $(\mathbf{A})_{jk} = \sum_{y=0}^{1} \mathbb{P}_Y(y)\mathbb{I}\{\mathcal{P}_y^Y(h_k) = f_j(X)\}$, with $f_0 \equiv 0$, and $f_1 \equiv 1$, as well as $f_2(X) = X$ ("ID"), and $f_3(X) = 1 - X$ ("NOT"), see § 2.1 and § 3 and footnote 6.

By **(S1)** setting $\lambda_X = \lambda_X^{\max}$ results in $[\mathbf{a}(\lambda_X)]_2 = 0 = [\mathbf{Ac}]_2$. Writing out $[\mathbf{Ac}]_2$ we obtain

$$0 = c_4\mathbb{P}(Y = 0) + c_6\mathbb{P}(Y = 0) + c_8\mathbb{P}(Y = 1) + c_9\mathbb{P}(Y = 1) + c_{12} + c_{13}P(Y = 1) + c_{14}P(Y = 0).$$

Together with $0 < \mathbb{P}(Y = 0) < 1$ and since all entries of $\mathbf{c}$ are non-negative, this gives

$$0 = c_4 = c_6 = c_8 = c_9 = c_{12} = c_{13} = c_{14}.$$

Analogously, setting $\lambda_Y = \lambda_Y^{\max}$ results in

$$0 = c_2 = c_6 = c_8 = c_9 = c_{10} = c_{11} = c_{14}.$$

Thus, the only possible non-zero entries of $\mathbf{c}$ are $c_0, c_1, c_3, c_5, c_7, c_{15}$ and (56) reduces to

$$\begin{aligned} q_{00} &= c_0 \\ q_{01} &= c_0 + c_1 + c_5 \\ q_{10} &= c_0 + c_1 + c_3 \\ q_{11} &= c_0 + c_1 + c_3 + c_5 + c_7. \end{aligned} \tag{57}$$

Since we have $q_{00} \leq q_{01}, q_{10} \leq q_{11}$, we can make the following assignment:

if $q_{01} \geq q_{10}$**:**	**if** $q_{01} \leq q_{10}$**:**
$c_0 = q_{00}$	$c_0 = q_{00}$
$c_1 = q_{10} - q_{00}$	$c_1 = q_{01} - q_{00}$
$c_3 = 0$	$c_3 = q_{10} - q_{01}$
$c_5 = q_{01} - q_{10}$	$c_5 = 0$
$c_7 = q_{11} - q_{01}$	$c_7 = q_{11} - q_{10}$
$c_{15} = 1 - q_{11}$	$c_{15} = 1 - q_{11}$

Clearly the entries of $\mathbf{c}$ sum to 1 and since $q_{00} \leq q_{01} \leq q_{11}$ and $q_{00} \leq q_{10} \leq q_{11}$, all entries of $\mathbf{c}$ are valid probabilities. By construction we have now found a valid probability vector that is consistent with the marginals $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ and implies $\lambda_X = \lambda_X^{\max}, \lambda_Y = \lambda_Y^{\max}$.

Let us verify that the constructed $\mathbf{c}$ indeed fulfils (15) for $\lambda_X = \lambda_X^{\max} = \mathbb{P}(Z=0|X=0)$ (Recall that are working for the special case **(S1)**). Using the constructed $\mathbf{c}$, Tab. 1, the definition of $\mathbf{A}$ we obtain

$$
\begin{aligned}
[\mathbf{Ac}]_0 &= c_0 + c_1\mathbb{P}(Y=1) + c_5\mathbb{P}(Y=1) \\
&= q_{00} + \mathbb{P}(Y=1)(q_{01} - q_{00}) \\
&= \mathbb{P}(Y=0)q_{00} + \mathbb{P}(Y=1)q_{01} \\
&= \mathbb{P}(Z=0,Y=0|X=0) + \mathbb{P}(Z=0,Y=1|X=0) \\
&= \mathbb{P}(Z=0|X=0) = 0 + \lambda_X^{\max} \\
&= [\mathbf{a}(\lambda_X^{\max})]_0,
\end{aligned}
$$

$$
\begin{aligned}
[\mathbf{Ac}]_1 &= c_5\mathbb{P}(Y=0) + c_7\mathbb{P}(Y=0) + c_{15} \\
&= \mathbb{P}(Y=0)(c_5 + c_7 + c_{15}) + \mathbb{P}(Y=1)c_{15} \\
&= \mathbb{P}(Y=0)(q_{11} - q_{10} + 1 - q_{11}) + \mathbb{P}(Y=1)(1-q_{11}) \\
&= 1 - \mathbb{P}(Z=0,Y=0|X=1) - \mathbb{P}(Z=0,Y=1|X=1) \\
&= 1 - \mathbb{P}(Z=0|X=1) \\
&= 1 - \mathbb{P}(Z=0|X=1) - \mathbb{P}(Z=0|X=0) + \mathbb{P}(Z=0|X=0) \\
&= 1 - \mathbb{P}(Z=0|X=1) - \mathbb{P}(Z=0|X=0)\lambda_X^{\max} \\
&= [\mathbf{a}(\lambda_X^{\max})]_1,
\end{aligned}
$$

$$
\begin{aligned}
[\mathbf{Ac}]_2 &= 0 \\
&= \mathbb{P}(Z=0|X=0) - \lambda_X^{\max} \\
&= [\mathbf{a}(\lambda_X^{\max})]_2,
\end{aligned}
$$

$$
\begin{aligned}
[\mathbf{Ac}]_3 &= c_1\mathbb{P}(Y=0) + c_3 + c_7\mathbb{P}(Y=1) \\
&= \mathbb{P}(Y=0)(c_1 + c_3) + \mathbb{P}(Y=1)(c_3 + c_7) \\
&= \mathbb{P}(Y=0)(q_{10} - q_{00}) + \mathbb{P}(Y=1)(q_{11} - q_{01}) \\
&= \mathbb{P}(Z=0,Y=0|X=1) - \mathbb{P}(Z=0,Y=0|X=0) + \mathbb{P}(Z=0,Y=1|X=1) - \mathbb{P}(Z=0,Y=1|X=0) \\
&= \mathbb{P}(Z=0|X=1) - \mathbb{P}(Z=0|X=0) \\
&= \mathbb{P}(Z=0|X=1) - \lambda_X^{\max} \\
&= [\mathbf{a}(\lambda_X^{\max})]_3.
\end{aligned}
$$

Analogously, it follows that (15) is consistent for $\lambda_Y = \lambda_Y^{\max}$. □

E.1. Proof of Lemma 6

Proof of Lemma 6. As we do throughout we assume that $< \mathbb{P}(X=1) < 1$, and $0 < \mathbb{P}(Y=1) < 1$. A necessary condition for merging $\mathbb{P}_{XZ}$ and $\mathbb{P}_{YZ}$ is that they imply the same marginal distribution over Z

$$
\sum_{i\in\{0,1\}} \mathbb{P}(Z=0|X=i)\mathbb{P}(X=i) = \sum_{j\in\{0,1\}} \mathbb{P}(Z=0|Y=j)\mathbb{P}(Y=j). \tag{58}
$$

Using the definition of δ_X, δ_Y we obtain

$$
\begin{aligned}
& \mathbb{P}(Z=0|X=0)\mathbb{P}(X=0) + (\mathbb{P}(Z=0|X=0) + \delta_X)\mathbb{P}(X=1) && (59)\\
&\quad = (\mathbb{P}(Z=0|Y=1) - \delta_Y)\mathbb{P}(Y=0) + \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) && (60)\\
\Leftrightarrow\quad & \mathbb{P}(Z=0|X=0) + \delta_X\mathbb{P}(X=1) = \mathbb{P}(Z=0|Y=1) - \delta_Y\mathbb{P}(Y=0) && (61)\\
\Leftrightarrow\quad & \delta_{YX} := \mathbb{P}(Z=0|Y=1) - \mathbb{P}(Z=0|X=0) = \delta_Y\mathbb{P}(Y=0) + \delta_X\mathbb{P}(X=1). && (62)
\end{aligned}
$$

By Assumption $\delta_X \geq 0, \delta_Y \geq 0$, and thus $\delta_{YX} \geq 0$. Analogously we obtain

$$\delta_{XY} := \mathbb{P}(Z=0|X=1) - \mathbb{P}(Z=0|Y=0) = \delta_X \mathbb{P}(X=0) + \delta_Y \mathbb{P}(Y=1) \geq 0. \tag{63}$$

Since we assumed that a joint statistical model $\mathbb{P}_{XYZ}$ exists, which is consistent with the marginals $\mathbb{P}_{XZ}, \mathbb{P}_{YZ}$, there exists $q_{i,j} := \mathbb{P}(Z=0|X=i, Y=j)$ such that

$$\begin{aligned}
\mathbb{P}(Z=0|X=0) &= q_{00}\mathbb{P}(Y=0) + q_{01}\mathbb{P}(Y=1),\\
\mathbb{P}(Z=0|X=1) &= q_{10}\mathbb{P}(Y=0) + q_{11}\mathbb{P}(Y=1),\\
\mathbb{P}(Z=0|Y=0) &= q_{00}\mathbb{P}(X=0) + q_{10}\mathbb{P}(X=1),\\
\mathbb{P}(Z=0|Y=1) &= q_{01}\mathbb{P}(X=0) + q_{11}\mathbb{P}(X=1).
\end{aligned}$$

Thus under our assumption that none of the marginal probabilities $\mathbb{P}(X=0)$, $\mathbb{P}(Y=0)$ equals 0 or 1, choosing q_{00} uniquely determines all the other $\{q_{ij}\}$:

$$\begin{aligned}
(a)\quad & \mathbb{P}(Z=0|X=0) = q_{00}\mathbb{P}(Y=0) + q_{01}\mathbb{P}(Y=1)\\
& \Leftrightarrow q_{01} = \frac{\mathbb{P}(Z=0|X=0)}{\mathbb{P}(Y=1)} - \frac{q_{00}\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)},
\end{aligned} \tag{64}$$

$$\begin{aligned}
(b)\quad & \mathbb{P}(Z=0|Y=0) = q_{00}\mathbb{P}(X=0) + q_{10}\mathbb{P}(X=1)\\
& \Leftrightarrow q_{10} = \frac{\mathbb{P}(Z=0|Y=0)}{\mathbb{P}(X=1)} - \frac{q_{00}\mathbb{P}(X=0)}{\mathbb{P}(X=1)},
\end{aligned} \tag{65}$$

$$\begin{aligned}
(c)\quad & \mathbb{P}(Z=0|Y=1) = q_{01}\mathbb{P}(X=0) + q_{11}\mathbb{P}(X=1)\\
\Leftrightarrow\ & q_{11} = \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{q_{01}\mathbb{P}(X=0)}{\mathbb{P}(X=1)}\\
&= \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{\left(\frac{\mathbb{P}(Z=0|X=0)}{\mathbb{P}(Y=1)} - \frac{q_{00}\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)}\right)\mathbb{P}(X=0)}{\mathbb{P}(X=1)}\\
&= \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} + \frac{q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)}.
\end{aligned} \tag{66}$$

$\mathbb{P}(Z=0|X=1) = q_{10}\mathbb{P}(Y=0) + q_{11}\mathbb{P}(Y=1)$ is then ensured if the marginals can consistently be merged, which we assumed. Our goal is thus to check whether a q_{00} exists such that

$$0 \leq q_{00}, \quad q_{00} \leq q_{01}, \quad q_{00} \leq q_{10}, \quad q_{01} \leq q_{11}, \quad q_{10} \leq q_{11}, \quad q_{11} \leq 1. \tag{67}$$

Using the equalities defined above, we can express all these constraints in terms of q_{00}. Ensuring that a solution exists, will then complete the proof.

$$
\begin{aligned}
&(C1) \quad 0 \leq q_{00} \\
&(C2) \quad q_{00} \leq q_{01} \Leftrightarrow q_{00} \leq \frac{\mathbb{P}(Z=0|X=0)}{\mathbb{P}(Y=1)} - \frac{q_{00}\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \Leftrightarrow q_{00} \leq \mathbb{P}(Z=0|X=0), \\
&(C3) \quad q_{00} \leq q_{10} \Leftrightarrow q_{00} \leq \frac{\mathbb{P}(Z=0|Y=0)}{\mathbb{P}(X=1)} - \frac{q_{00}\mathbb{P}(X=0)}{\mathbb{P}(X=1)} \Leftrightarrow q_{00} \leq \mathbb{P}(Z=0|Y=0), \\
&(C4) \quad q_{01} \leq q_{11} \Leftrightarrow \frac{\mathbb{P}(Z=0|X=0)}{\mathbb{P}(Y=1)} - \frac{q_{00}\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \\
&\qquad\qquad \leq \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} + \frac{q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)}, \\
&(C5) \quad q_{10} \leq q_{11} \Leftrightarrow \frac{\mathbb{P}(Z=0|Y=0)}{\mathbb{P}(X=1)} - \frac{q_{00}\mathbb{P}(X=0)}{\mathbb{P}(X=1)} \\
&\qquad\qquad \leq \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} + \frac{q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} \\
&(C6) \quad q_{11} \leq 1 \Leftrightarrow \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} + \frac{q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} \leq 1.
\end{aligned}
$$

$(C1), (C2), (C3)$ are already in interpretable form, so next we rewrite $(C4)$

$$
\begin{aligned}
&\frac{\mathbb{P}(Z=0|X=0)}{\mathbb{P}(Y=1)} - \frac{q_{00}\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \leq \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} + \frac{q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} \\
\Leftrightarrow\ &\mathbb{P}(Z=0|X=0) - q_{00}\mathbb{P}(Y=0) \leq \frac{\mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(X=1)} \\
&\qquad\qquad + \frac{q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0)}{\mathbb{P}(X=1)} \\
\Leftrightarrow\ &q_{00}\mathbb{P}(Y=0)\left(1 + \frac{\mathbb{P}(X=0)}{\mathbb{P}(X=1)}\right) \geq \mathbb{P}(Z=0|X=0) + \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=1)} \\
\Leftrightarrow\ &q_{00}\mathbb{P}(Y=0) \geq \mathbb{P}(Z=0|X=0)\mathbb{P}(X=1) + \mathbb{P}(Z=0|X=0)\mathbb{P}(X=0) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) \\
\Leftrightarrow\ &q_{00}\mathbb{P}(Y=0) \geq \mathbb{P}(Z=0|X=0) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) \\
\Leftrightarrow\ &q_{00} \geq \frac{\mathbb{P}(Z=0|X=0) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} \\
\overset{(*)}{\Leftrightarrow}\ &q_{00} \geq \frac{\mathbb{P}(Z=0|X=0) - [\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0) + \mathbb{P}(Z=0|X=1)\mathbb{P}(X=1) - \mathbb{P}(Z=0|Y=0)\mathbb{P}(Y=0)]}{\mathbb{P}(Y=0)} \\
\Leftrightarrow\ &q_{00} \geq \mathbb{P}(Z=0|Y=0) - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0) + \mathbb{P}(Z=0|X=1)\mathbb{P}(X=1) - \mathbb{P}(Z=0|X=0)}{\mathbb{P}(Y=0)} \\
\Leftrightarrow\ &q_{00} \geq \mathbb{P}(Z=0|Y=0) - \frac{\mathbb{P}(X=1)}{\mathbb{P}(Y=0)}(\mathbb{P}(Z=0|X=1) - \mathbb{P}(Z=0|X=0)) \\
\Leftrightarrow\ &q_{00} \geq \mathbb{P}(Z=0|Y=0) - \frac{\mathbb{P}(X=1)}{\mathbb{P}(Y=0)}\delta_X, \qquad (C4)
\end{aligned}
$$

where at $(*)$ we used (58). We can also rewrite $(C4)$ once more in terms of $\mathbb{P}(Z=0|X=0)$

$$
\begin{aligned}
& q_{00} \geq \mathbb{P}(Z=0|Y=0) - \frac{\mathbb{P}(X=1)}{\mathbb{P}(Y=0)}\delta_X \\
\Leftrightarrow\, & q_{00} \geq \mathbb{P}(Z=0|Y=0) - \frac{\mathbb{P}(X=1)}{\mathbb{P}(Y=0)}(\mathbb{P}(Z=0|X=1) - \mathbb{P}(Z=0|X=0)) \\
\Leftrightarrow\, & q_{00}\mathbb{P}(Y=0) \geq \mathbb{P}(Z=0|Y=0)\mathbb{P}(Y=0) - \mathbb{P}(X=1)(\mathbb{P}(Z=0|X=1) - \mathbb{P}(Z=0|X=0)) \\
\overset{(*)}{\Leftrightarrow}\, & q_{00}\mathbb{P}(Y=0) \geq \mathbb{P}(Z=0|X=0)\mathbb{P}(X=0) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) + \mathbb{P}(Z=0|X=0)\mathbb{P}(X=1) \\
\Leftrightarrow\, & q_{00}\mathbb{P}(Y=0) \geq \mathbb{P}(Z=0|X=0) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) \\
\Leftrightarrow\, & q_{00}\mathbb{P}(Y=0) \geq \mathbb{P}(Z=0|X=0)(\mathbb{P}(Y=0) + \mathbb{P}(Y=1)) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) \\
& \qquad\qquad - \mathbb{P}(Z=0|X=0)\mathbb{P}(Y=1) + \mathbb{P}(Z=0|X=0)\mathbb{P}(Y=1) \\
\Leftrightarrow\, & q_{00}\mathbb{P}(Y=0) \geq \mathbb{P}(Z=0|X=0)\mathbb{P}(Y=0) + \mathbb{P}(Z=0|X=0)\mathbb{P}(Y=1) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) \\
\Leftrightarrow\, & q_{00} \geq \mathbb{P}(Z=0|X=0) - \delta_{YX}\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}. \qquad (C4)
\end{aligned}
$$

Analogously, we can work through $(C5)$ and obtain

$$
\begin{aligned}
& \frac{\mathbb{P}(Z=0|Y=0)}{\mathbb{P}(X=1)} - \frac{q_{00}\mathbb{P}(X=0)}{\mathbb{P}(X=1)} \leq \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} + \frac{q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} \\
\Leftrightarrow\quad & q_{00} \geq \mathbb{P}(Z=0|X=0) - \delta_Y\frac{\mathbb{P}(Y=1)}{\mathbb{P}(X=0)} \\
\Leftrightarrow\quad & q_{00} \geq \mathbb{P}(Z=0|Y=0) - \delta_{XY}\frac{\mathbb{P}(X=1)}{\mathbb{P}(X=0)}. \qquad (C5)
\end{aligned}
$$

Next we consider (C6)

$$
\begin{aligned}
& q_{11} \leq 1 \Leftrightarrow \frac{\mathbb{P}(Z=0|Y=1)}{\mathbb{P}(X=1)} - \frac{\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} + \frac{q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=1)\mathbb{P}(X=1)} \leq 1 \\
\Leftrightarrow\quad & \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) - \mathbb{P}(Z=0|X=0)\mathbb{P}(X=0) + q_{00}\mathbb{P}(Y=0)\mathbb{P}(X=0) \leq \mathbb{P}(X=1)\mathbb{P}(Y=1) \\
\Leftrightarrow\quad & q_{00} \leq \frac{\mathbb{P}(X=1)\mathbb{P}(Y=1) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) + \mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)}{\mathbb{P}(Y=0)\mathbb{P}(X=0)}.
\end{aligned}
$$

Using

$$
\begin{aligned}
& \mathbb{P}(X=1)\mathbb{P}(Y=1) - \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1) + \mathbb{P}(Z=0|X=0)\mathbb{P}(X=0) \\
&= \mathbb{P}(X=1,Y=1) - \mathbb{P}(Z=0,Y=1) + \mathbb{P}(Z=0,X=0) \\
&= [\mathbb{P}(Z=0,X=1,Y=1) + \mathbb{P}(Z=1,X=1,Y=1)] \\
&\quad - [\mathbb{P}(Z=0,X=0,Y=1) + \mathbb{P}(Z=0,X=1,Y=1)] \\
&\quad + [\mathbb{P}(Z=0,X=0,Y=0) + \mathbb{P}(Z=0,X=0,Y=1)] \\
&= \mathbb{P}(Z=1,X=1,Y=1) + \mathbb{P}(Z=0,X=0,Y=0) \geq 0,
\end{aligned}
$$

we obtain for $(C6)$

$$
q_{00} \leq \frac{\mathbb{P}(Z=1,X=1,Y=1) + \mathbb{P}(Z=0,X=0,Y=0)}{\mathbb{P}(X=0)\mathbb{P}(Y=0)}. \qquad (C6) \tag{68}
$$

Let us summarize all constraints once more:

$$
\begin{aligned}
&(C1) \quad q_{00} \geq 0 =: \eta_1 \\
&(C2) \quad q_{00} \leq \mathbb{P}(Z=0|X=0) =: \eta_2, \\
&(C3) \quad q_{00} \leq \mathbb{P}(Z=0|Y=0) =: \eta_3, \\
&(C4) \quad q_{00} \geq \mathbb{P}(Z=0|Y=0) - \frac{\mathbb{P}(X=1)}{\mathbb{P}(Y=0)}\delta_X = \mathbb{P}(Z=0|X=0) - \delta_{YX}\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} =: \eta_4, \\
&(C5) \quad q_{00} \geq \mathbb{P}(Z=0|X=0) - \delta_Y\frac{\mathbb{P}(Y=1)}{\mathbb{P}(X=0)} = \mathbb{P}(Z=0|Y=0) - \delta_{XY}\frac{\mathbb{P}(X=1)}{\mathbb{P}(X=0)} =: \eta_5, \\
&(C6) \quad q_{00} \leq \frac{\mathbb{P}(Z=1,X=1,Y=1) + \mathbb{P}(Z=0,X=0,Y=0)}{\mathbb{P}(X=0)\mathbb{P}(Y=0)} =: \eta_6.
\end{aligned}
$$

We now have to check whether all the lower bounds on q_{00} are smaller than all the upper bounds, in other words:

$$
\text{The proof is complete.} \quad \Leftrightarrow \quad \max(\eta_1,\eta_4,\eta_5) \leq \min(\eta_2,\eta_3,\eta_6). \tag{69}
$$

First consider the case where $\eta_2 = \min(\eta_2,\eta_3,\eta_6)$. Then all lower bounds are achievable, since $\delta_X, \delta_Y, \delta_{XY}, \delta_{YX}$ are per definition non-negative. The same holds if $\eta_3 = \min(\eta_2,\eta_3,\eta_6)$.

It is less apparent to see what holds in the case of $\eta_6 = \min(\eta_2,\eta_3,\eta_6)$. Since the numerator of η_6 is the sum of two probabilities we always have $\eta_6 \geq 0 = \eta_1$. We therefore need to show $\eta_6 \geq \eta_4, \eta_6 \geq \eta_5$.

$$
\begin{aligned}
& \eta_4 \leq \eta_6 \\
\Leftrightarrow \quad & \mathbb{P}(Z=0|Y=0) - \frac{\mathbb{P}(X=1)}{\mathbb{P}(Y=0)}\delta_X \leq \frac{\mathbb{P}(Z=1,X=1,Y=1) + \mathbb{P}(Z=0,X=0,Y=0)}{\mathbb{P}(X=0)\mathbb{P}(Y=0)} \\
\Leftrightarrow \quad & \mathbb{P}(Z=0,Y=0)\mathbb{P}(X=0) - \mathbb{P}(X=0)\mathbb{P}(X=1)\left[\mathbb{P}(Z=0|X=1) - \mathbb{P}(Z=0|X=0)\right] \\
& \qquad \leq \mathbb{P}(Z=1,X=1,Y=1) + \mathbb{P}(Z=0,X=0,Y=0) \\
\Leftrightarrow \quad & \left[\mathbb{P}(Z=0,X=0,Y=0) + \mathbb{P}(Z=0,X=1,Y=0)\right]\mathbb{P}(X=0) - \mathbb{P}(X=0)\mathbb{P}(Z=0,X=1) \\
& \qquad\qquad + \mathbb{P}(X=1)\mathbb{P}(Z=0,X=0) \\
& \qquad \leq \mathbb{P}(Z=1,X=1,Y=1) + \mathbb{P}(Z=0,X=0,Y=0) \\
\Leftrightarrow \quad & \left[\mathbb{P}(Z=0,X=0,Y=0) + \mathbb{P}(Z=0,X=1,Y=0)\right]\mathbb{P}(X=0) - \mathbb{P}(X=0)\mathbb{P}(Z=0,X=1) \\
& \qquad\qquad + \mathbb{P}(X=1)\left[\mathbb{P}(Z=0,X=0,Y=0) + \mathbb{P}(Z=0,X=0,Y=1)\right] \\
& \qquad \leq \mathbb{P}(Z=1,X=1,Y=1) + \mathbb{P}(Z=0,X=0,Y=0) \\
\Leftrightarrow \quad & \mathbb{P}(Z=0,X=1,Y=0)\mathbb{P}(X=0) - \mathbb{P}(X=0)\mathbb{P}(Z=0,X=1) + \mathbb{P}(X=1)\mathbb{P}(Z=0,X=0,Y=1) \\
& \qquad \leq \mathbb{P}(Z=1,X=1,Y=1) \\
\Leftrightarrow \quad & \mathbb{P}(X=0)\left[\mathbb{P}(Z=0,X=1,Y=0) - \mathbb{P}(Z=0,X=1)\right] + \mathbb{P}(X=1)\mathbb{P}(Z=0,X=0,Y=1) \\
& \qquad \leq \mathbb{P}(Z=1,X=1,Y=1) \\
\Leftrightarrow \quad & -\mathbb{P}(X=0)\mathbb{P}(Z=0,X=1,Y=1) + \mathbb{P}(X=1)\mathbb{P}(Z=0,X=0,Y=1) \\
& \qquad \leq \mathbb{P}(Z=1,X=1,Y=1) \\
\Leftrightarrow \quad & \mathbb{P}(X=1)\mathbb{P}(Z=0,X=0,Y=1) \leq \mathbb{P}(Z=1,X=1,Y=1) + \mathbb{P}(X=0)\mathbb{P}(Z=0,X=1,Y=1) \\
\Leftrightarrow \quad & \mathbb{P}(X=1)\mathbb{P}(Z=0,X=0,Y=1) \leq \mathbb{P}(Z=1,X=1,Y=1) + (1-\mathbb{P}(X=1))\mathbb{P}(Z=0,X=1,Y=1) \\
\Leftrightarrow \quad & \mathbb{P}(X=1)\mathbb{P}(Z=0,X=0,Y=1) \leq \mathbb{P}(X=1,Y=1) - \mathbb{P}(X=1)\mathbb{P}(Z=0,X=1,Y=1) \\
\Leftrightarrow \quad & \mathbb{P}(X=1)\mathbb{P}(Z=0,X=0,Y=1) \leq \mathbb{P}(X=1)\mathbb{P}(Y=1) - \mathbb{P}(X=1)\mathbb{P}(Z=0,X=1,Y=1) \\
\Leftrightarrow \quad & \mathbb{P}(X=1)\mathbb{P}(Z=0,Y=1) \leq \mathbb{P}(X=1)\left[\mathbb{P}(Z=0,Y=1) + \mathbb{P}(Z=1,Y=1)\right] \\
\Leftrightarrow \quad & 0 \leq \mathbb{P}(X=1)\mathbb{P}(Z=1,Y=1),
\end{aligned}
$$

which always is a true statement. Analogously we obtain

$$
\eta_5 \leq \eta_6 \Leftrightarrow 0 \leq \mathbb{P}(Y=1)\mathbb{P}(Z=1,X=1) \Leftrightarrow \textbf{TRUE}.
$$

Note that the validity of η_6 being larger than all lower bounds holds independently of our assumptions on $\delta_X \geq 0$ and $\delta_Y \geq 0$.

We have thus shown that under the assumptions stated in the Lemma, the interval $[\max(\eta_1, \eta_4, \eta_5), \min(\eta_2, \eta_3, \eta_6)]$ is non-empty. Hence, picking any q_{00} in this interval and computing q_{01}, q_{10}, q_{11} accordingly, leads to an example fulfilling the Lemma's statement.

□

F. Construction of the convex polytope $\mathcal{C}$

In § 3.2 we defined the set of feasible joint counterfactual models via eq. (16). To clarify that this actually is a polytope and to handle it with numerical solvers, we need to express the constraints on $\mathbf{c}$ as a set of equalities and inequalities. These should take the following form (see Section 2.2.4 of Boyd & Vandenberghe (2004)):

$$\tilde{\mathbf{A}}\mathbf{c} \preceq \tilde{\mathbf{b}} \tag{70}$$

$$\tilde{\mathbf{C}}\mathbf{c} = \tilde{\mathbf{d}}, \tag{71}$$

for some matrices $\tilde{\mathbf{A}}, \tilde{\mathbf{C}}$ that need to be determined. Here we take on the notation $\mathbf{a} \preceq \mathbf{b}$, to denote that $a_i \leq b_i$ for all entries.

Inequality constraints. Starting from (15) we have

$$\mathbf{A}\mathbf{c} = \mathbf{a}(\lambda^A) = \mathbf{a}_0 + \lambda^A \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \qquad \text{with} \qquad \mathbf{a}_0 = \begin{pmatrix} 0 \\ 1 - P(Z=0|X=0) - P(Z=0|X=1) \\ P(Z=0|X=0) \\ P(Z=0|X=1) \end{pmatrix}.$$

Therefore, using $\lambda_X^{\min} \leq \lambda^A \leq \lambda_X^{\max}$, we get the following inequalities:

$$\mathbf{A}\mathbf{c} \preceq \mathbf{a}_0 + \begin{pmatrix} \lambda_X^{\max} \\ \lambda_X^{\max} \\ -\lambda_X^{\min} \\ -\lambda_X^{\min} \end{pmatrix}. \tag{72}$$

And similarly for $\mathbf{b}(\lambda^B)$,

$$\mathbf{B}\mathbf{c} \preceq \mathbf{b}_0 + \begin{pmatrix} \lambda_Y^{\max} \\ \lambda_Y^{\max} \\ -\lambda_Y^{\min} \\ -\lambda_Y^{\min} \end{pmatrix}. \tag{73}$$

Additionally, from

$$-\mathbf{A}\mathbf{c} = -\mathbf{a}_0 + \lambda^A \begin{pmatrix} -1 \\ -1 \\ +1 \\ +1 \end{pmatrix} \tag{74}$$

and again using $\lambda_X^{\min} \leq \lambda^A \leq \lambda_X^{\max}$, we get

$$-\mathbf{A}\mathbf{c} \preceq -\mathbf{a}_0 + \begin{pmatrix} -\lambda_X^{\min} \\ -\lambda_X^{\min} \\ \lambda_X^{\max} \\ \lambda_X^{\max} \end{pmatrix}. \tag{75}$$

And similarly for $\mathbf{b}(\lambda^B)$,

$$-\mathbf{B}\mathbf{c} \preceq -\mathbf{b}_0 + \begin{pmatrix} -\lambda_Y^{\min} \\ -\lambda_Y^{\min} \\ \lambda_Y^{\max} \\ \lambda_Y^{\max} \end{pmatrix}. \tag{76}$$

Finally, from the positivity constraint $c_i \geq 0 \ \ \forall i$, we get

$$-\mathbb{I}\mathbf{c} \leq [0, \dots, 0]^\top \in \mathbb{R}^{16}$$

where $\mathbb{I}$ is the 16×16 identity matrix.

Overall, we can express the inequality constraints as

$$\tilde{\mathbf{A}}\mathbf{c} \preceq \tilde{\mathbf{b}} \tag{77}$$

by defining the 32×16 matrix $\tilde{\mathbf{A}}$ and the 32-dimensional vector $\tilde{\mathbf{b}}$ as

$$\tilde{\mathbf{A}} := \begin{pmatrix} \mathbf{A} \\ -\mathbf{A} \\ \mathbf{B} \\ -\mathbf{B} \\ -\mathbb{I} \end{pmatrix}, \qquad \tilde{\mathbf{b}} := \begin{pmatrix} \tilde{\mathbf{a}}_1 \\ \tilde{\mathbf{a}}_2 \\ \tilde{\mathbf{b}}_1 \\ \tilde{\mathbf{b}}_2 \\ \mathbf{0} \end{pmatrix}$$

Where $\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2$ respectively denote the RHS of eqs. (72), (75), (73), (76), and $\mathbf{0} = [0, \dots, 0]^\top \in \mathbb{R}^{16}$.

Equality constraints. Besides the inequalities above we also need to ensure that all 4 implied equations of $\mathbf{a}(\lambda^A) = \mathbf{A}\mathbf{c}$ are fulfilled simultaneously (i.e., they are fulfilled for the same λ^A). We can enforce this by ensuring that λ^A as computed from the first row equals the one computed from the second, third, and fourth row, respectively. Let us make it explicit for the equality of λ^A computed from the first two rows:

$$\begin{aligned} &\lambda^A = [\mathbf{A}\mathbf{c}]_0 - [\mathbf{a}_0]_0 = [\mathbf{A}\mathbf{c}]_1 - [\mathbf{a}_0]_1 = \lambda^A \\ &[\mathbf{A}\mathbf{c}]_0 - [\mathbf{A}\mathbf{c}]_1 = [\mathbf{a}_0]_0 - [\mathbf{a}_0]_1 \\ \Leftrightarrow &\begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} \mathbf{A}\mathbf{c} = \begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} \mathbf{a}_0 \end{aligned}$$

Doing this also for the third and fourth row, we obtain the constraints

$$\begin{aligned} \begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} \mathbf{A}\mathbf{c} &= \begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} \mathbf{a}_0, \\ \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix} \mathbf{A}\mathbf{c} &= \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix} \mathbf{a}_0, \\ \begin{pmatrix} 1 & 0 & 0 & 1 \end{pmatrix} \mathbf{A}\mathbf{c} &= \begin{pmatrix} 1 & 0 & 0 & 1 \end{pmatrix} \mathbf{a}_0, \end{aligned}$$

which we rewrite as one set of constraints

$$\tilde{\mathbf{C}}_\mathbf{A}\mathbf{c} = \tilde{\mathbf{d}}_{\mathbf{a}_0}, \quad \text{with } \tilde{\mathbf{C}}_\mathbf{A} := \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \mathbf{A}, \quad \text{and } \tilde{\mathbf{d}}_{\mathbf{a}_0} := \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \mathbf{a}_0. \tag{78}$$

Proceeding similarly, we obtain

$$\tilde{\mathbf{C}}_\mathbf{B}\mathbf{c} = \tilde{\mathbf{d}}_{\mathbf{b}_0}, \quad \text{with } \tilde{\mathbf{C}}_\mathbf{B} := \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \mathbf{B}, \quad \text{and } \tilde{\mathbf{d}}_{\mathbf{b}_0} := \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \mathbf{b}_0. \tag{79}$$

Additionally we obtain one equality constraint ensuring that the probabilities of $\mathbf{c}$ sum to one $\sum_i c_i = 1$, i.e.,

$$\tilde{\mathbf{C}}_\mathbf{1}\mathbf{c} = 1, \quad \text{with } \tilde{\mathbf{C}}_\mathbf{1} := [1.0, \dots, 1.0], \mathbb{R}^{1\times 16}.$$

Overall we can thus collect all equality constraints as $\tilde{\mathbf{C}}\mathbf{c} = \tilde{\mathbf{d}}$ with

$$\tilde{\mathbf{C}} := \begin{pmatrix} \tilde{\mathbf{C}}_\mathbf{A} \\ \tilde{\mathbf{C}}_\mathbf{B} \\ \tilde{\mathbf{C}}_\mathbf{1} \end{pmatrix}, \qquad \text{and } \tilde{\mathbf{d}} := \begin{pmatrix} \tilde{\mathbf{d}}_{\mathbf{a}_0} \\ \tilde{\mathbf{d}}_{\mathbf{b}_0} \\ 1 \end{pmatrix}$$

Characterization of $\mathcal{C}$. To summarise we can characterise the polytope of feasible joint models as

$$\boxed{\mathcal{C} := \{\mathbf{c} \in \Delta^{15} \mid \exists(\lambda_X, \lambda_Y) \in \Lambda_0 \text{ s.t. (15) holds}\} = \{\mathbf{c} \in \mathbb{R}^{16} \mid \tilde{\mathbf{C}}\mathbf{c} = \tilde{\mathbf{d}}, \tilde{\mathbf{A}}\mathbf{c} = \tilde{\mathbf{b}}\}}.$$

The convex polyhedron $\mathcal{C}$ can alternatively be represented as the convex hull of its vertices $\mathcal{V}_\mathcal{C} := \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_m\} \subset \mathbb{R}^{16}$, where $m \in \mathbb{N}$ depends on the number of constraints. We use the `pypoman` package (Caron, 2018), to compute those vertices of the polyhedron. After that we can project each vertex into the (λ_X, λ_Y)-plane. Since $[\mathbf{Ac}]_0 = \lambda_X$ and $[\mathbf{Bc}]_0 = \lambda_Y$, each $\mathbf{c}$ corresponds to

$$\begin{pmatrix} \lambda_X \\ \lambda_Y \end{pmatrix} = \tilde{\mathbf{E}}\mathbf{c}, \tag{80}$$

with

$$\tilde{\mathbf{E}} := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix}. \tag{81}$$

Thus the region of admissible pairs (λ_X, λ_Y) (green region in Fig. 2) is given as the convex hull of the projected vertices of $\mathcal{C}$

$$\Lambda_\mathcal{C} = \text{Conv}\left(\{\tilde{\mathbf{E}}\boldsymbol{v} \mid \boldsymbol{v} \in \mathcal{V}_\mathcal{C}\}\right). \tag{82}$$

G. Technical details on the Experiments

All the code used for the experiments and to generate the plots in Fig. 3 can be found in the supplementary material.

Parametrising $\mathbb{P}_{XYZ}$. In order to ensure that a solution to the marginal problem exists, we start from parameters of the joint distribution $\mathbb{P}_{XYZ}$ (i.e. a joint distribution exists by construction), and compute the parameters of the marginal distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ by marginalisation. We factorise $\mathbb{P}_{XYZ}$ according to the DAG in Fig. 1 as

$$\theta_X = P(X = 1) \tag{83}$$
$$\theta_Y = P(Y = 1) \tag{84}$$
$$\theta_{Z|X=x,Y=y} = P(Z = 1|X = x, Y = y), \quad x, y \in \{0, 1\}, \tag{85}$$

which overall requires 6 parameters.

Finding solutions: projections of the 16-dimensional polytope. The polygon $\Lambda_\mathcal{C}$ can be determined by projecting the vertices of the high-dimensional polytope $\mathcal{C}$ in 2-dimensions and computing their convex hull. For the polytope vertices projection, the equality and inequality constraints and affine projection are described in Appx. F. For polyhedra manipulation in Python, we use the `pypoman` package (Caron, 2018), which allows to compute the 2-d projection of the vertices of our $\mathcal{C}$ polytope. Once the projected vertices are computed, $\Lambda_\mathcal{C}$ can be found as their convex hull, which we compute using `scipy` (Virtanen et al., 2020).

Alternatively, if we are only interested in $\text{LB}^\star_X, \text{UB}^\star_X$, the computation could be formulated as a linear program (Boyd & Vandenberghe, 2004),

$$\begin{aligned} \min/\max_{\lambda_X, \lambda_Y \in \mathbb{R}, \mathbf{c} \in \Delta^{15}} \quad & \lambda_X \\ \text{subject to} \quad & \mathbf{a}(\lambda_X) = \mathbf{Ac} \\ & \mathbf{a}(\lambda_Y) = \mathbf{Bc} \\ & \lambda_X^{\min} \leq \lambda_X \leq \lambda_X^{\max} \\ & \lambda_Y^{\min} \leq \lambda_Y \leq \lambda_Y^{\max} \end{aligned} \tag{86}$$

and similarly for $\text{LB}^\star_Y, \text{UB}^\star_Y$. This can be solved using e.g. the `linprog` module in `scipy` (Virtanen et al., 2020).

Sampling problem instances. The parameters in (83) and (84) are sampled from a Uniform distribution on $[0, 1]$, while those in (85) are sampled from a Beta distribution, whose parameters α and β are set either to 1 (corresponding to a

Uniform distribution) or to 0.5 (which puts more mass towards the extremes, thus resulting in more deterministic conditional distributions $\mathbb{P}_{Z|XY}$).

Parameter sweeps and GIF visualisations. To generate [GIF1] we use generic (i.e., not inducing a unique SCM) conditionals

$$\mathbb{P}(Z=1|X=0,Y=0)=0.3, \tag{87}$$
$$\mathbb{P}(Z=1|X=0,Y=1)=0.8, \tag{88}$$
$$\mathbb{P}(Z=1|X=1,Y=0)=0.3, \tag{89}$$
$$\mathbb{P}(Z=1|X=1,Y=1)=0.7. \tag{90}$$

Whereas [GIF2] is generated through a joint SCM $Z := X \oplus Y$:

$$\mathbb{P}(Z=1|X=0,Y=0)=0, \tag{91}$$
$$\mathbb{P}(Z=1|X=0,Y=1)=1, \tag{92}$$
$$\mathbb{P}(Z=1|X=1,Y=0)=1, \tag{93}$$
$$\mathbb{P}(Z=1|X=1,Y=1)=0. \tag{94}$$

In both cases we then sweep over different probabilities $\mathbb{P}(X=1), \mathbb{P}(Y=1)$ as shown in the respective plots on the right.

The green points in the plots represent the projected vertices of the high-dimensional polytope $\mathcal{C}$.

H. An illustrative example

We now provide an example which—despite arguably being slightly contrived—is meant to illustrate the potential usefulness of our approach and the structural causal marginal problem in a real-world context.

Suppose that we are interested in investigating a disease Z for which $Z=1$ indicates that a person recovers completely after ten days (fast recovery), while $Z=0$ indicates that the disease went on for more than ten days (long symptoms). We assume that there exists some medication X against disease Z, such that $X=1$ denotes that a person took the medication, while $X=0$ denotes that a person did not take the medication.

Clearly, the disease does not cause the medication, but potentially vice versa, so we can take the causal graph to be $X \to Z$. For sake of simplicity, suppose further that X and Z are unconfounded (see Appx. D for a detailed treatment of confounding).

We have access to an observational study in the form of a distribution $\mathbb{P}_{XZ}$ which indicates that *without* medication the chances of having long symptoms are 50%, i.e., $\mathbb{P}(Z=0|X=0)=1/2$, whereas *with* medication the chances of long symptoms reduce to 40%, i.e., $\mathbb{P}(Z=0|X=1)=0.4$. Thus, overall, the medication has a positive ACE.

The family of marginal SCMs $\mathcal{M}_X$ over $X \to Z$ that can explain these findings can be found via (10) and are given by:

$$\mathbf{a}(\lambda_X) = \begin{pmatrix} 0 \\ 0.1 \\ 0.5 \\ 0.4 \end{pmatrix} + \lambda_X \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \tag{95}$$

with $\lambda_X \in [0, 0.4]$.

Since $\lambda_X = 0$ is allowed, we cannot exclude that what happens is the following:

- For 10% of the people the medication has no effect and they always recover fast ($a_1 = 0.1$, $Z := f_1(X) \equiv 1$).
- For 50% of the people the medication causes the fast recovery, while without medication, they have long symptoms ($a_2 = 0.5$, $Z := f_2(X) = X$).
- For 40% of the people the medication causes active harm: If they take it, they experience long symptoms, while without, they recover fast ($a_3 = 0.4$, $Z := f_3(X) = 1 - X$).

Now it is plausible that this scenario would be quite frightening and could cause some people to refuse to take the medication because they are afraid that it harms. (Although from a purely statistical perspective it is still advisable to take it and that's why we assume the medication was approved.)

But now imagine that another study is conducted that investigates the (unconfounded) effect of the presence of some specific genotype $Y = 1$ ($Y = 0$ denotes that a person has a different genotype than the one under investigation) on the chances of fast recovery $Z = 1$. For, say, privacy reasons, however, *this study does not document whether or not subjects undertook the medication* X, so it only provides data from $\mathbb{P}_{YZ}$ and we have no joint observations of $\mathbb{P}_{XYZ}$.

Suppose the second study finds that 40% of people have the genotype, $\mathbb{P}(Y = 1) = 0.4$, and *all* of those experience long symptoms $\mathbb{P}(Z = 0|Y = 1) = 1$.

We fix the remaining probabilities to $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = 1/2$ and $\mathbb{P}(Z = 0|Y = 0) = 1/12$, although other choices can also lead to the same conclusion.

With the methods proposed in this paper, we can then show that enforcing consistency of both datasets constrains the possible SCMs over $X \to Z$ to a *unique* $\lambda_X = 0.4$, see Appx. H.1 below for details. Now this SCM has a totally different interpretation to the one (previously still possible) given above:

- For 40% of people the medication has no effect and they always experience long symptoms ($a_0 = 0.4$, $Z := f_0(X) \equiv 0$).
- For 50% of people the medication has no effect and they always recover fast ($a_1 = 0.5$, $Z := f_1(X) \equiv 1$).
- For 10% of people the medication causes the fast recovery, while without medication, they have long symptoms ($a_2 = 0.1$, $Z := f_2(X) = X$).
- For 0% of people the medication causes active harm ($a_3 = 0$, $Z := f_3(X) = 1 - X$).

It seems plausible that people would be much more willing to take the medication now that they know '*it cannot harm*'—even if the ACE remains unchanged. However, note that we now also know that the medication only helps in 10% of the cases.

H.1. Explicit Calculation

For conciseness when we presented the example we simply stated that $\mathbb{P}_{YZ}$ forces $\lambda_X = 0.4$. For completeness we now provide the explicit calculation. We assumed $\mathbb{P}(Z = 0|Y = 1) = 1$, thus from (10) we obtain

$$\mathbf{b}(\lambda_Y) = \begin{pmatrix} 0 \\ -\mathbb{P}(Z = 0|Y = 0) \\ \mathbb{P}(Z = 0|Y = 0) \\ 1 \end{pmatrix} + \lambda_Y \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}. \tag{96}$$

The only value of λ_Y that ensures this is a valid probability vector is $\lambda_Y = \mathbb{P}(Z = 0|Y = 0)$. This results in

$$\mathbf{b} = \begin{pmatrix} \mathbb{P}(Z = 0|Y = 0) \\ 0 \\ 0 \\ 1 - \mathbb{P}(Z = 0|Y = 0) \end{pmatrix}. \tag{97}$$

By enforcing $\mathbf{Bc} = \mathbf{b}$, for the two zero entries we obtain (by similar considerations as in the proof of Prop. 4):

$$\begin{aligned} 0 = [\mathbf{Bc}]_1 &= c_3\mathbb{P}(X = 0) + c_7\mathbb{P}(X = 0) + c_{11}\mathbb{P}(X = 0) + c_{12}\mathbb{P}(X = 1) + c_{13}\mathbb{P}(X = 1) + c_{14}\mathbb{P}(X = 1) + c_{15} \\ &\Leftrightarrow 0 = c_3 = c_7 = c_{11} = c_{12} = c_{13} = c_{14} = c_{15}. \end{aligned}$$

Similarly

$$\begin{aligned} 0 = [\mathbf{Bc}]_2 &= c_2\mathbb{P}(X = 0) + c_6\mathbb{P}(X = 0) + c_{10} + c_{14}\mathbb{P}(Y = 0) + c_8\mathbb{P}(X = 1) + c_9\mathbb{P}(X = 1) + c_{11}\mathbb{P}(X = 1) \\ &\Leftrightarrow 0 = c_2 = c_6 = c_{10} = c_{14} = c_8 = c_9 = c_{11}. \end{aligned}$$

So overall the non-zero entries can only be c_0, c_1, c_4, c_5

Furthermore, we have (considering only non-zero entries of $\mathbf{c}$)

$$[\mathbf{Ac}]_0 = c_0 + c_1\mathbb{P}(Y=1) + c_4\mathbb{P}(Y=1) + c_5\mathbb{P}(Y=1) \tag{98}$$
$$= \mathbb{P}(Y=0)c_0 + \mathbb{P}(Y=1)(c_0 + c_1 + c_4 + c_5) \tag{99}$$
$$= 0.6c_0 + 0.4 \geq 0.4 \tag{100}$$

where we used $\mathbb{P}(Y=1) = 0.4$ and $c_0 + c_1 + c_4 + c_5 = 1$ as they are the only non-zero entries. On the other hand from (95) we have $[\mathbf{a}(\lambda_X)]_0 \leq 0.4$. Hence the only way $[\mathbf{Ac}]_0 = [\mathbf{a}(\lambda_X)]_0 = \lambda_X$, happens if $c_0 = 0$ and $\lambda_X = 0.4$.

To conclude, we show that setting $c_1 = 0$, $c_4 = \frac{1}{6}$, $c_5 = \frac{5}{6}$ leads to the correct marginals.

We have

$$\mathbb{P}(Z=0|X=0) = c_4 + \mathbb{P}(Y=1)c_5 = \frac{1}{6} + \frac{5}{6}\cdot 0.4 = 1/2 \quad \checkmark \tag{101}$$
$$\mathbb{P}(Z=0|X=1) = c_4\mathbb{P}(Y=1) + c_5\mathbb{P}(Y=1) = (c_4 + c_5)\cdot\mathbb{P}(Y=1) = 0.4 \quad \checkmark \tag{102}$$
$$\mathbb{P}(Z=0|Y=0) = c_4\mathbb{P}(X=0) = \frac{1}{6}\cdot\frac{1}{2} = \frac{1}{12} \quad \checkmark \tag{103}$$
$$\mathbb{P}(Z=0|Y=1) = c_4 + c_5 = 1 \qquad \checkmark. \tag{104}$$

Furthermore we have

$$\mathbb{P}(Z=0|X=0)\mathbb{P}(X=0) + \mathbb{P}(Z=0|X=1)\mathbb{P}(X=1) = \frac{1}{2}(0.5 + 0.4)$$
$$= 0.45$$
$$= \frac{1}{12}\cdot\frac{3}{5} + 1\cdot\frac{4}{10} = \mathbb{P}(Z=0|Y=0)\mathbb{P}(Y=0) + \mathbb{P}(Z=0|Y=1)\mathbb{P}(Y=1),$$

so also the marginal distribution over Z is consistent.

I. More on the connection to statistical learning theory and capacity measures

The following remarks are meant to illustrate to what extent the ambiguity in the space of allowed SCMs may be reduced by only considering function classes with low VC dimension (Vapnik & Chervonenkis, 1971). Intuitively, if we allow all (arbitrarily complex) response functions, the space of consistent (joint) models can be quite large. If, on the other hand, we constrain their allowed capacity and only allow for 'simple' functions, this couples their behaviour across different input values and consequently can reduce the model space substantially.

Since the space of possible SCMs compatible with all observed probabilities is a convex polytope, a simple measure for its size is the entropy of its unique maximum entropy distribution. Let us first compute this entropy for the case where all response functions are allowed (i.e., without restrictions on their VC dimension).

To generate any conditional $\mathbb{P}_{Y|X}$ with cause X and effect Y attaining values in finite sets $\mathcal{X}$ and $\mathcal{Y}$ with $|\mathcal{X}| = n$ by an SCM, following Peters et al. (2017, § 3.4) we represent each function $f : \mathcal{X} \to \mathcal{Y}$ as an element in the n-fold Cartesian product $\mathcal{Y}^n := \mathcal{Y} \times \cdots \times \mathcal{Y}$ such that the j-th component indicates $f(x_j)$. Then, each distribution $\mathbb{P}(Y|X = x_j)$ determines only the marginal distribution of the j^{th} component of $\mathcal{Y}^n$. Thus, the MaxEnt joint distribution on $\mathcal{Y}^n$ having these n marginal distributions is simply given by their product. In other words, we obtain a distribution of functions in which observing what f does with the input x_j tells us nothing on what f does with a different input $x_i \neq x_j$. The observed probabilities are always compatible with such a 'decoupling of inputs' since we don't observe one draw of the function applied to different inputs. The SCM obtained this way has the entropy

$$\sum_{j=1}^{n} H(Y|X = x_j), \tag{105}$$

where the sum runs over all n possible values x_j of X (without weighting factor $p(x_j)$) and thus grows as $O(n)$.

Restricted function classes, however, couple different inputs: If Y is binary, and we consider a function class C with VC dimension h, the size $|C|$ (which here coincides with the shattering coefficient) is bounded from above by $\log|C| \in$

$O(h \log n)$ (Vapnik, 1998). Hence, for fixed h, the MaxEnt distribution on C grows at most logarithmically in n as opposed to the linear growth in (105).

In summary, this means that the space of allowed models (as measured by MaxEnt here), is reduced when restrictions on the function class are enforced.